(I) Some Home Truths About Hypothesis and Significance Testing, and (II) The Jaynes Information Criterion (JIC) and the Role of Parsimony in Bayes Factors

David Draper

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

`draper@ucsc.edu`

CRiSM Workshop: Contemporary Issues in Hypothesis Testing (Warwick)

15 Sep 2016

The Big Picture

- Problems addressed by the discipline of **statistics** typically have the following structure.

The Big Picture

- Problems addressed by the discipline of **statistics** typically have the following structure.
- You (Good 1950) [note the capital Y]: a generic person wishing to reason sensibly in the presence of uncertainty) are given a **problem** $\mathbb{P} = (\mathbb{Q}, \mathbb{C})$ involving **uncertainty** about $\boxed{\theta}$, the unknown aspect of $\mathbb{P}$ of principal interest.

The Big Picture

- Problems addressed by the discipline of **statistics** typically have the following structure.
- You (Good 1950) [note the capital Y]: a generic person wishing to reason sensibly in the presence of uncertainty) are given a **problem** $\mathbb{P} = (\mathbb{Q}, \mathbb{C})$ involving **uncertainty** about $\boxed{\theta}$, the unknown aspect of $\mathbb{P}$ of principal interest.
- Here $\mathbb{Q}$ identifies the main **questions** to be answered, and $\mathbb{C}$ represents the (real-world) **context** in which the questions are raised, instantiated through a finite set $\mathcal{B}$ of **(true/false) propositions**, all rendered true by problem context.

The Big Picture

- Problems addressed by the discipline of **statistics** typically have the following structure.
- You (Good 1950) [note the capital Y]: a generic person wishing to reason sensibly in the presence of uncertainty) are given a **problem** $\mathbb{P} = (\mathbb{Q}, \mathbb{C})$ involving **uncertainty** about $\boxed{\theta}$, the unknown aspect of $\mathbb{P}$ of principal interest.
- Here $\mathbb{Q}$ identifies the main **questions** to be answered, and $\mathbb{C}$ represents the (real-world) **context** in which the questions are raised, instantiated through a finite set $\mathcal{B}$ of **(true/false) propositions**, all rendered true by problem context.
- You examine Your resources and find that it's possible to obtain a new **data set** D to decrease Your uncertainty about θ.

The Big Picture

- Problems addressed by the discipline of **statistics** typically have the following structure.
- You (Good 1950) [note the capital Y]: a generic person wishing to reason sensibly in the presence of uncertainty) are given a **problem** $\mathbb{P} = (\mathbb{Q}, \mathbb{C})$ involving **uncertainty** about $\boxed{\theta}$, the unknown aspect of $\mathbb{P}$ of principal interest.
- Here $\mathbb{Q}$ identifies the main **questions** to be answered, and $\mathbb{C}$ represents the (real-world) **context** in which the questions are raised, instantiated through a finite set $\mathcal{B}$ of **(true/false) propositions**, all rendered true by problem context.
- You examine Your resources and find that it's possible to obtain a new **data set** D to decrease Your uncertainty about θ.
- In this setting, a **Theorem** due to Cox (1946) and Jaynes (2002) — recently rigorized and extended by Terenin and Draper (2016) — says that

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \,|\, \mathcal{B})$, encoding Your information about θ **external** to D, and $p(D \,|\, \theta \, \mathcal{B}) \propto \ell(\theta \,|\, D \, \mathcal{B})$, capturing Your information about θ **internal** to D — then **optimal inference** about θ is based (**Bayes's Theorem**) on the distribution $p(\theta \,|\, D \, \mathcal{B}) \propto p(\theta \,|\, \mathcal{B}) \, \ell(\theta \,|\, D \, \mathcal{B})$,*

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \mid \mathcal{B})$, encoding Your information about θ **external** to D, and $p(D \mid \theta \, \mathcal{B}) \propto \ell(\theta \mid D \, \mathcal{B})$, capturing Your information about θ **internal** to D — then **optimal inference** about θ is based (**Bayes's Theorem**) on the distribution $p(\theta \mid D \, \mathcal{B}) \propto p(\theta \mid \mathcal{B}) \, \ell(\theta \mid D \, \mathcal{B})$, and **optimal prediction** of new data D^* is based on the distribution $p(D^* \mid D \, \mathcal{B}) = \int_\Theta p(D^* \mid \theta \, D \, \mathcal{B}) \, p(\theta \mid D \, \mathcal{B}) \, d\theta$,*

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \mid \mathcal{B})$, encoding Your information about θ **external** to D, and $p(D \mid \theta \, \mathcal{B}) \propto \ell(\theta \mid D \, \mathcal{B})$, capturing Your information about θ **internal** to D — then **optimal inference** about θ is based (**Bayes's Theorem**) on the distribution $p(\theta \mid D \, \mathcal{B}) \propto p(\theta \mid \mathcal{B}) \, \ell(\theta \mid D \, \mathcal{B})$, and **optimal prediction** of new data D^* is based on the distribution $p(D^* \mid D \, \mathcal{B}) = \int_\Theta p(D^* \mid \theta D \, \mathcal{B}) \, p(\theta \mid D \, \mathcal{B}) \, d\theta$, where Θ is the set of possible values of θ; and*

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \mid \mathcal{B})$, encoding Your information about θ* ***external*** *to D, and $p(D \mid \theta \, \mathcal{B}) \propto \ell(\theta \mid D \, \mathcal{B})$, capturing Your information about θ* ***internal*** *to D — then* ***optimal inference*** *about θ is based (**Bayes's Theorem**) on the distribution $p(\theta \mid D \, \mathcal{B}) \propto p(\theta \mid \mathcal{B}) \, \ell(\theta \mid D \, \mathcal{B})$, and* ***optimal prediction*** *of new data D^* is based on the distribution $p(D^* \mid D \, \mathcal{B}) = \int_\Theta p(D^* \mid \theta D \, \mathcal{B}) \, p(\theta \mid D \, \mathcal{B}) \, d\theta$, where Θ is the set of possible values of θ; and*
- *If You're prepared to specify two additional ingredients —*

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \mid \mathcal{B})$, encoding Your information about θ* ***external*** *to D, and $p(D \mid \theta\, \mathcal{B}) \propto \ell(\theta \mid D\, \mathcal{B})$, capturing Your information about θ* ***internal*** *to D — then* ***optimal inference*** *about θ is based (**Bayes's Theorem**) on the distribution $p(\theta \mid D\, \mathcal{B}) \propto p(\theta \mid \mathcal{B})\, \ell(\theta \mid D\, \mathcal{B})$, and* ***optimal prediction*** *of new data D^* is based on the distribution $p(D^* \mid D\, \mathcal{B}) = \int_\Theta p(D^* \mid \theta D\, \mathcal{B})\, p(\theta \mid D\, \mathcal{B})\, d\theta$, where Θ is the set of possible values of θ; and*
- *If You're prepared to specify two additional ingredients — Your* ***action space*** *$\{a \in (\mathcal{A} \mid \mathcal{B})\}$, an exhaustive set of possible actions, and*

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \,|\, \mathcal{B})$, encoding Your information about θ* ***external*** *to D, and $p(D \,|\, \theta \, \mathcal{B}) \propto \ell(\theta \,|\, D \, \mathcal{B})$, capturing Your information about θ* ***internal*** *to D — then* ***optimal inference*** *about θ is based (****Bayes's Theorem****) on the distribution $p(\theta \,|\, D \, \mathcal{B}) \propto p(\theta \,|\, \mathcal{B}) \, \ell(\theta \,|\, D \, \mathcal{B})$, and* ***optimal prediction*** *of new data D^* is based on the distribution $p(D^* \,|\, D \, \mathcal{B}) = \int_\Theta p(D^* \,|\, \theta \, D \, \mathcal{B}) \, p(\theta \,|\, D \, \mathcal{B}) \, d\theta$, where Θ is the set of possible values of θ; and*

- *If You're prepared to specify two additional ingredients — Your* ***action space*** *$\{a \in (\mathcal{A} \,|\, \mathcal{B})\}$, an exhaustive set of possible actions, and Your real-valued* ***utility function*** *$U(a, \theta^* \,|\, \mathcal{B})$, quantifying the costs and benefits that would result if You took action a and the unknown θ actually had the value θ^* —*

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \,|\, \mathcal{B})$, encoding Your information about θ **external** to D, and $p(D \,|\, \theta \, \mathcal{B}) \propto \ell(\theta \,|\, D \, \mathcal{B})$, capturing Your information about θ **internal** to D — then **optimal inference** about θ is based (**Bayes's Theorem**) on the distribution $p(\theta \,|\, D \, \mathcal{B}) \propto p(\theta \,|\, \mathcal{B}) \, \ell(\theta \,|\, D \, \mathcal{B})$, and **optimal prediction** of new data D^* is based on the distribution $p(D^* \,|\, D \, \mathcal{B}) = \int_\Theta p(D^* \,|\, \theta D \, \mathcal{B}) \, p(\theta \,|\, D \, \mathcal{B}) \, d\theta$, where Θ is the set of possible values of θ; and*

- *If You're prepared to specify two additional ingredients — Your **action space** $\{a \in (\mathcal{A} \,|\, \mathcal{B})\}$, an exhaustive set of possible actions, and Your real-valued **utility function** $U(a, \theta^* \,|\, \mathcal{B})$, quantifying the costs and benefits that would result if You took action a and the unknown θ actually had the value θ^* — then **optimal decision-making** is attained by finding the action a^* that maximizes the expected utility $E_{(\theta \,|\, D \, \mathcal{B})} \, U(a^*, \theta \,|\, \mathcal{B})$.*

The Big Picture (continued)

- *If You're prepared to specify two probability distributions — $p(\theta \mid \mathcal{B})$, encoding Your information about θ* ***external*** *to D, and $p(D \mid \theta\, \mathcal{B}) \propto \ell(\theta \mid D\, \mathcal{B})$, capturing Your information about θ* ***internal*** *to D — then* ***optimal inference*** *about θ is based (**Bayes's Theorem**) on the distribution $p(\theta \mid D\, \mathcal{B}) \propto p(\theta \mid \mathcal{B})\, \ell(\theta \mid D\, \mathcal{B})$, and* ***optimal prediction*** *of new data D^* is based on the distribution $p(D^* \mid D\, \mathcal{B}) = \int_\Theta p(D^* \mid \theta D\, \mathcal{B})\, p(\theta \mid D\, \mathcal{B})\, d\theta$, where Θ is the set of possible values of θ; and*

- *If You're prepared to specify two additional ingredients — Your* ***action space*** *$\{a \in (\mathcal{A} \mid \mathcal{B})\}$, an exhaustive set of possible actions, and Your real-valued* ***utility function*** *$U(a, \theta^* \mid \mathcal{B})$, quantifying the costs and benefits that would result if You took action a and the unknown θ actually had the value θ^* — then* ***optimal decision-making*** *is attained by finding the action a^* that maximizes the expected utility $E_{(\theta \mid D\, \mathcal{B})}\, U(a^*, \theta \mid \mathcal{B})$.*

(**Bayesian game theory** is more general than **Bayesian decision theory** ...)

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$,

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$, let's agree to call $M = \{p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B})\}$ Your **model** for Your uncertainty about θ and D^*; and

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$, let's agree to call $M = \{p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B})\}$ Your **model** for Your uncertainty about θ and D^*; and
- If instead **decision-making** is the goal defined by $\mathbb{Q}$,

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$, let's agree to call $M = \{p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B})\}$ Your **model** for Your uncertainty about θ and D^*; and
- If instead **decision-making** is the goal defined by $\mathbb{Q}$, let's agree to call $M_d = \{p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B}), (\mathcal{A} \,|\, \mathcal{B}), U(a, \theta \,|\, \mathcal{B})\}$ Your **model** for Your uncertainty about a^*.

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$, let's agree to call $M = \{p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B})\}$ Your **model** for Your uncertainty about θ and D^*; and
- If instead **decision-making** is the goal defined by $\mathbb{Q}$, let's agree to call $M_d = \{p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta \, \mathcal{B}), (\mathcal{A} \,|\, \mathcal{B}), U(a, \theta \,|\, \mathcal{B})\}$ Your **model** for Your uncertainty about a^*.
- The two main **practical challenges** in using Cox's Theorem are

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$, let's agree to call $M = \{p(\theta \mid \mathcal{B}), p(D \mid \theta\, \mathcal{B})\}$ Your **model** for Your uncertainty about θ and D^*; and

- If instead **decision-making** is the goal defined by $\mathbb{Q}$, let's agree to call $M_d = \{p(\theta \mid \mathcal{B}), p(D \mid \theta\, \mathcal{B}), (\mathcal{A} \mid \mathcal{B}), U(a, \theta \mid \mathcal{B})\}$ Your **model** for Your uncertainty about a^*.

- The two main **practical challenges** in using Cox's Theorem are

  - **(technical)** **Integrals** arising in **computing** the inferential and predictive distributions may be difficult to approximate accurately,

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$, let's agree to call $M = \{p(\theta \mid \mathcal{B}), p(D \mid \theta\, \mathcal{B})\}$ Your **model** for Your uncertainty about θ and D^*; and
- If instead **decision-making** is the goal defined by $\mathbb{Q}$, let's agree to call $M_d = \{p(\theta \mid \mathcal{B}), p(D \mid \theta\, \mathcal{B}), (\mathcal{A} \mid \mathcal{B}), U(a, \theta \mid \mathcal{B})\}$ Your **model** for Your uncertainty about a^*.
- The two main **practical challenges** in using Cox's Theorem are
  - **(technical)** **Integrals** arising in **computing** the inferential and predictive distributions may be difficult to approximate accurately, and the **optimization** over the **action space** may be difficult to perform; and

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$, let's agree to call $M = \{p(\theta \mid \mathcal{B}), p(D \mid \theta\, \mathcal{B})\}$ Your **model** for Your uncertainty about θ and D^*; and
- If instead **decision-making** is the goal defined by $\mathbb{Q}$, let's agree to call $M_d = \{p(\theta \mid \mathcal{B}), p(D \mid \theta\, \mathcal{B}), (\mathcal{A} \mid \mathcal{B}), U(a, \theta \mid \mathcal{B})\}$ Your **model** for Your uncertainty about a^*.
- The two main **practical challenges** in using Cox's Theorem are
  - **(technical)** **Integrals** arising in **computing** the inferential and predictive distributions may be difficult to approximate accurately, and the **optimization** over the **action space** may be difficult to perform; and
  - **(substantive)** The mapping from $\mathbb{P}$ to M or M_d is rarely unique, giving rise to **model uncertainty**.

The Big Picture (continued)

- If **inference** and/or **prediction** are the goals defined by $\mathbb{Q}$, let's agree to call $M = \{p(\theta \mid \mathcal{B}), p(D \mid \theta\, \mathcal{B})\}$ Your **model** for Your uncertainty about θ and D^*; and
- If instead **decision-making** is the goal defined by $\mathbb{Q}$, let's agree to call $M_d = \{p(\theta \mid \mathcal{B}), p(D \mid \theta\, \mathcal{B}), (\mathcal{A} \mid \mathcal{B}), U(a, \theta \mid \mathcal{B})\}$ Your **model** for Your uncertainty about a^*.
- The two main **practical challenges** in using Cox's Theorem are
  - **(technical)** **Integrals** arising in **computing** the inferential and predictive distributions may be difficult to approximate accurately, and the **optimization** over the **action space** may be difficult to perform; and
  - **(substantive)** The mapping from $\mathbb{P}$ to M or M_d is rarely unique, giving rise to **model uncertainty**.
- How do **hypothesis** and **significance testing** fit into this framework?

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Your sampling distribution $p(D \,|\, \theta \, \mathcal{B})$ is assumed by You to be a member of a family of densities with **known mathematical form** but indexed by an **unknown parameter vector** $\theta \in \Theta = \mathbb{R}^k$, for some positive integer k.

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Your sampling distribution $p(D \mid \theta \, \mathcal{B})$ is assumed by You to be a member of a family of densities with **known mathematical form** but indexed by an **unknown parameter vector** $\theta \in \Theta = \mathbb{R}^k$, for some positive integer k.

A subset Θ_1 of Θ is singled out in some way; for example, $(\theta \in \Theta_1)$ corresponds to a **scientific theory** being true or false.

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Your sampling distribution $p(D \mid \theta \, \mathcal{B})$ is assumed by You to be a member of a family of densities with **known mathematical form** but indexed by an **unknown parameter vector** $\theta \in \Theta = \mathbb{R}^k$, for some positive integer k.

A subset Θ_1 of Θ is singled out in some way; for example, $(\theta \in \Theta_1)$ corresponds to a **scientific theory** being true or false.

The **frequentist** testing story now has a **bifurcation**:

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Your sampling distribution $p(D \mid \theta \, \mathcal{B})$ is assumed by You to be a member of a family of densities with **known mathematical form** but indexed by an **unknown parameter vector** $\theta \in \Theta = \mathbb{R}^k$, for some positive integer k.

A subset Θ_1 of Θ is singled out in some way; for example, $(\theta \in \Theta_1)$ corresponds to a **scientific theory** being true or false.

The **frequentist** testing story now has a **bifurcation**:

(Fisher significance testing) "Every experiment may be said to exist only in order to give the [data] a chance of disproving [the truth of the (true/false) proposition $(\theta \in \Theta_1)$]":

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Your sampling distribution $p(D \mid \theta \, \mathcal{B})$ is assumed by You to be a member of a family of densities with **known mathematical form** but indexed by an **unknown parameter vector** $\theta \in \Theta = \mathbb{R}^k$, for some positive integer k.

A subset Θ_1 of Θ is singled out in some way; for example, $(\theta \in \Theta_1)$ corresponds to a **scientific theory** being true or false.

The **frequentist** testing story now has a **bifurcation**:

(Fisher significance testing) "Every experiment may be said to exist only in order to give the [data] a chance of disproving [the truth of the (true/false) proposition $(\theta \in \Theta_1)$]": use D either to reject $(\theta \in \Theta_1)$ or to fail to reject $(\theta \in \Theta_1)$,

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Your sampling distribution $p(D \mid \theta \, \mathcal{B})$ is assumed by You to be a member of a family of densities with **known mathematical form** but indexed by an **unknown parameter vector** $\theta \in \Theta = \mathbb{R}^k$, for some positive integer k.

A subset Θ_1 of Θ is singled out in some way; for example, $(\theta \in \Theta_1)$ corresponds to a **scientific theory** being true or false.

The **frequentist** testing story now has a **bifurcation**:

(Fisher significance testing) "Every experiment may be said to exist only in order to give the [data] a chance of disproving [the truth of the (true/false) proposition $(\theta \in \Theta_1)$]": use D either to reject $(\theta \in \Theta_1)$ or to fail to reject $(\theta \in \Theta_1)$, but **WITHOUT** regard for the plausibility of D under the opposite proposition $(\theta \notin \Theta_1)$; versus

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Your sampling distribution $p(D \mid \theta \, \mathcal{B})$ is assumed by You to be a member of a family of densities with **known mathematical form** but indexed by an **unknown parameter vector** $\theta \in \Theta = \mathbb{R}^k$, for some positive integer k.

A subset Θ_1 of Θ is singled out in some way; for example, $(\theta \in \Theta_1)$ corresponds to a **scientific theory** being true or false.

The **frequentist** testing story now has a **bifurcation**:

(Fisher significance testing) "Every experiment may be said to exist only in order to give the [data] a chance of disproving [the truth of the (true/false) proposition $(\theta \in \Theta_1)$]": use D either to reject $(\theta \in \Theta_1)$ or to fail to reject $(\theta \in \Theta_1)$, but **WITHOUT** regard for the plausibility of D under the opposite proposition $(\theta \notin \Theta_1)$; versus

(Neyman-Pearson hypothesis testing) Use D either to reject $(\theta \in \Theta_1)$ or to fail to reject $(\theta \in \Theta_1)$,

Significance and Hypothesis Testing

In the context of **parametric statistical modeling**, testing typically looks like this:

Your sampling distribution $p(D \mid \theta \, \mathcal{B})$ is assumed by You to be a member of a family of densities with **known mathematical form** but indexed by an **unknown parameter vector** $\theta \in \Theta = \mathbb{R}^k$, for some positive integer k.

A subset Θ_1 of Θ is singled out in some way; for example, $(\theta \in \Theta_1)$ corresponds to a **scientific theory** being true or false.

The **frequentist** testing story now has a **bifurcation**:

(Fisher significance testing) "Every experiment may be said to exist only in order to give the [data] a chance of disproving [the truth of the (true/false) proposition $(\theta \in \Theta_1)$]": use D either to reject $(\theta \in \Theta_1)$ or to fail to reject $(\theta \in \Theta_1)$, but **WITHOUT** regard for the plausibility of D under the opposite proposition $(\theta \notin \Theta_1)$; versus

(Neyman-Pearson hypothesis testing) Use D either to reject $(\theta \in \Theta_1)$ or to fail to reject $(\theta \in \Theta_1)$, but **WITH** regard for the plausibility of D under the opposite proposition $(\theta \notin \Theta_1)$.

Bayesian Testing

Bayesian testing would seem to be completely straightforward:

Bayesian Testing

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \mid \theta \, \mathcal{B})$ with a **prior distribution** $p(\theta \mid \mathcal{B})$ specified by problem context $\mathbb{C}$,

Bayesian Testing

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \mid \theta \, \mathcal{B})$ with a **prior distribution** $p(\theta \mid \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \mid D \, \mathcal{B})$,

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \,|\, \theta \, \mathcal{B})$ with a **prior distribution** $p(\theta \,|\, \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \,|\, D \, \mathcal{B})$, and compute

$$p(\theta \in \Theta_1 \,|\, D \, \mathcal{B}) = \int_{\Theta_1} p(\theta \,|\, D \, \mathcal{B}) \, d\theta \, . \qquad (1)$$

Bayesian Testing

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \,|\, \theta\, \mathcal{B})$ with a **prior distribution** $p(\theta \,|\, \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \,|\, D\, \mathcal{B})$, and compute

$$p(\theta \in \Theta_1 \,|\, D\, \mathcal{B}) = \int_{\Theta_1} p(\theta \,|\, D\, \mathcal{B})\, d\theta\,. \qquad (1)$$

However, **not so fast:**

Bayesian Testing

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \mid \theta \, \mathcal{B})$ with a **prior distribution** $p(\theta \mid \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \mid D \, \mathcal{B})$, and compute

$$p(\theta \in \Theta_1 \mid D \, \mathcal{B}) = \int_{\Theta_1} p(\theta \mid D \, \mathcal{B}) \, d\theta \, . \tag{1}$$

However, **not so fast:**

- If Θ_1 defines a **subspace** of $\mathbb{R}^k$ of dimension less than k,

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \mid \theta \, \mathcal{B})$ with a **prior distribution** $p(\theta \mid \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \mid D \, \mathcal{B})$, and compute

$$p(\theta \in \Theta_1 \mid D \, \mathcal{B}) = \int_{\Theta_1} p(\theta \mid D \, \mathcal{B}) \, d\theta \, . \tag{1}$$

However, **not so fast:**

- If Θ_1 defines a **subspace** of $\mathbb{R}^k$ of dimension less than k, the integral in (1) will be **0** unless Your prior $p(\theta \mid \mathcal{B})$ places **non-zero probability** on the lower-dimensional subspace,

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \mid \theta \, \mathcal{B})$ with a **prior distribution** $p(\theta \mid \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \mid D \, \mathcal{B})$, and compute

$$p(\theta \in \Theta_1 \mid D \, \mathcal{B}) = \int_{\Theta_1} p(\theta \mid D \, \mathcal{B}) \, d\theta \, . \qquad (1)$$

However, **not so fast:**

- If Θ_1 defines a **subspace** of $\mathbb{R}^k$ of dimension less than k, the integral in (1) will be **0** unless Your prior $p(\theta \mid \mathcal{B})$ places **non-zero probability** on the lower-dimensional subspace, which in many settings is **inappropriate** (more about this later);

Bayesian Testing

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \mid \theta\, \mathcal{B})$ with a **prior distribution** $p(\theta \mid \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \mid D\, \mathcal{B})$, and compute

$$p(\theta \in \Theta_1 \mid D\, \mathcal{B}) = \int_{\Theta_1} p(\theta \mid D\, \mathcal{B})\, d\theta \, . \qquad (1)$$

However, **not so fast:**

- If Θ_1 defines a **subspace** of $\mathbb{R}^k$ of dimension less than k, the integral in (1) will be **0** unless Your prior $p(\theta \mid \mathcal{B})$ places **non-zero probability** on the lower-dimensional subspace, which in many settings is **inappropriate** (more about this later);
- You may well have **model uncertainty** about either or both of $p(D \mid \theta\, \mathcal{B})$ and $p(\theta \mid \mathcal{B})$,

Bayesian Testing

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \mid \theta\, \mathcal{B})$ with a **prior distribution** $p(\theta \mid \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \mid D\, \mathcal{B})$, and compute

$$p(\theta \in \Theta_1 \mid D\, \mathcal{B}) = \int_{\Theta_1} p(\theta \mid D\, \mathcal{B})\, d\theta\,. \tag{1}$$

However, **not so fast:**

- If Θ_1 defines a **subspace** of $\mathbb{R}^k$ of dimension less than k, the integral in (1) will be **0** unless Your prior $p(\theta \mid \mathcal{B})$ places **non-zero probability** on the lower-dimensional subspace, which in many settings is **inappropriate** (more about this later);
- You may well have **model uncertainty** about either or both of $p(D \mid \theta\, \mathcal{B})$ and $p(\theta \mid \mathcal{B})$, which may not be **uniquely specified** by problem context,

Bayesian Testing

Bayesian testing would seem to be completely straightforward:

Augment the previously specified **sampling distribution** $p(D \mid \theta \, \mathcal{B})$ with a **prior distribution** $p(\theta \mid \mathcal{B})$ specified by problem context $\mathbb{C}$, update to Your **posterior distribution** $p(\theta \mid D \, \mathcal{B})$, and compute

$$p(\theta \in \Theta_1 \mid D \, \mathcal{B}) = \int_{\Theta_1} p(\theta \mid D \, \mathcal{B}) \, d\theta \, . \qquad (1)$$

However, **not so fast:**

- If Θ_1 defines a **subspace** of $\mathbb{R}^k$ of dimension less than k, the integral in (1) will be **0** unless Your prior $p(\theta \mid \mathcal{B})$ places **non-zero probability** on the lower-dimensional subspace, which in many settings is **inappropriate** (more about this later);
- You may well have **model uncertainty** about either or both of $p(D \mid \theta \, \mathcal{B})$ and $p(\theta \mid \mathcal{B})$, which may not be **uniquely specified** by problem context, so that $p(\theta \in \Theta_1 \mid D \, \mathcal{B})$ may not even be **approximately uniquely specified**; and

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

  > *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake),*

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

  > *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake), or a **decision** problem (using that knowledge to choose an action),*

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

  > *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake), or a **decision** problem (using that knowledge to choose an action), or **both**?*

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

> *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake), or a **decision** problem (using that knowledge to choose an action), or **both**?*

- It's arguable that **testing virtually always involves both inference and decision**,

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

  > *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake), or a **decision** problem (using that knowledge to choose an action), or **both**?*

- It's arguable that **testing virtually always involves both inference and decision**, even when inference appears to be the only goal.

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

  > *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake), or a **decision** problem (using that knowledge to choose an action), or **both**?*

- It's arguable that **testing virtually always involves both inference and decision**, even when inference appears to be the only goal.
  - **Example: Finding the Higgs boson.** **(Louis Lyon)**

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

  > *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake), or a **decision** problem (using that knowledge to choose an action), or **both**?*

- It's arguable that **testing virtually always involves both inference and decision**, even when inference appears to be the only goal.
  - **Example: Finding the Higgs boson.** **(Louis Lyon)** On 4 Jul 2012 researchers at the **Large Hadron Collider** (LHC) in Geneva, Switzerland made this announcement:

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

  > *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake), or a **decision** problem (using that knowledge to choose an action), or **both**?*

- It's arguable that **testing virtually always involves both inference and decision**, even when inference appears to be the only goal.
  - **Example: Finding the Higgs boson.** **(Louis Lyon)** On 4 Jul 2012 researchers at the **Large Hadron Collider** (LHC) in Geneva, Switzerland made this announcement:

    > *"CMS observes an excess of events at a mass of approximately 125 GeV with a statistical significance of **five standard deviations** (5 sigma) above background expectations.*

Testing As Decision

- The simplicity of equation (1) sidesteps an **important issue**, equally crucial for frequentists and Bayesians alike:

  *Is this an **inferential** problem (the scientific acquisition of knowledge for its own sake), or a **decision** problem (using that knowledge to choose an action), or **both**?*

- It's arguable that **testing virtually always involves both inference and decision**, even when inference appears to be the only goal.
  - **Example: Finding the Higgs boson.** **(Louis Lyon)** On 4 Jul 2012 researchers at the **Large Hadron Collider** (LHC) in Geneva, Switzerland made this announcement:

    *"CMS observes an excess of events at a mass of approximately 125 GeV with a statistical significance of **five standard deviations** (5 sigma) above background expectations. The probability of the background alone fluctuating up by this amount or more is about **one in three million**."*

This Sure Looks Like Inference

The 1 in 3 million figure is a **frequentist P–value**

This Sure Looks Like Inference

The 1 in 3 million figure is a **frequentist *P*–value** (and would actually be $\Phi(-5) \doteq$ **1 in about 3.5 million** if a Gaussian approximation had been used):

This Sure Looks Like Inference

The 1 in 3 million figure is a **frequentist P–value** (and would actually be $\Phi(-5) \doteq$ **1 in about 3.5 million** if a Gaussian approximation had been used):

Let $\theta \in \Theta = \mathbb{R}$ be the **underlying excess fluctuation above background** at about 125 GeV (a value predicted by Higg's theory),

This Sure Looks Like Inference

The 1 in 3 million figure is a **frequentist *P*–value** (and would actually be $\Phi(-5) \doteq$ **1 in about 3.5 million** if a Gaussian approximation had been used):

Let $\theta \in \Theta = \mathbb{R}$ be the **underlying excess fluctuation above background** at about 125 GeV (a value predicted by Higg's theory), so that in this problem $\Theta_1 = \{0\}$,

This Sure Looks Like Inference

The 1 in 3 million figure is a **frequentist P–value** (and would actually be $\Phi(-5) \doteq$ **1 in about 3.5 million** if a Gaussian approximation had been used):

Let $\theta \in \Theta = \mathbb{R}$ be the **underlying excess fluctuation above background** at about 125 GeV (a value predicted by Higg's theory), so that in this problem $\Theta_1 = \{0\}$, and let $t(D)$ be a **one-dimensional summary** of the data set D that (after standardization) has — by assumption (i.e., no bias in the measuring process) and the Central Limit Theorem — an approximately $N(\theta, 1)$ **sampling distribution**;

This Sure Looks Like Inference

The 1 in 3 million figure is a **frequentist *P*–value** (and would actually be $\Phi(-5) \doteq$ **1 in about 3.5 million** if a Gaussian approximation had been used):

Let $\theta \in \Theta = \mathbb{R}$ be the **underlying excess fluctuation above background** at about 125 GeV (a value predicted by Higg's theory), so that in this problem $\Theta_1 = \{0\}$, and let $t(D)$ be a **one-dimensional summary** of the data set D that (after standardization) has — by assumption (i.e., no bias in the measuring process) and the Central Limit Theorem — an approximately $N(\theta, 1)$ **sampling distribution**; then the LHC researchers computed $P_{RS,\theta=0}[t(D) > 5] \doteq \Phi(-5)$, where RS stands for **repeated-sampling**.

This was a standard **Fisherian significance test**:

This Sure Looks Like Inference

The 1 in 3 million figure is a **frequentist** ***P*–value** (and would actually be $\Phi(-5) \doteq$ **1 in about 3.5 million** if a Gaussian approximation had been used):

Let $\theta \in \Theta = \mathbb{R}$ be the **underlying excess fluctuation above background** at about 125 GeV (a value predicted by Higg's theory), so that in this problem $\Theta_1 = \{0\}$, and let $t(D)$ be a **one-dimensional summary** of the data set D that (after standardization) has — by assumption (i.e., no bias in the measuring process) and the Central Limit Theorem — an approximately $N(\theta, 1)$ **sampling distribution**; then the LHC researchers computed $P_{RS,\theta=0}[t(D) > 5] \doteq \Phi(-5)$, where RS stands for **repeated-sampling**.

This was a standard **Fisherian significance test**: the researchers were interested in rejecting the hypothesis that

$$\theta = 0 \longleftrightarrow \textbf{(the Higgs boson doesn't exist)}$$

This Sure Looks Like Inference

The 1 in 3 million figure is a **frequentist *P*–value** (and would actually be $\Phi(-5) \doteq$ **1 in about 3.5 million** if a Gaussian approximation had been used):

Let $\theta \in \Theta = \mathbb{R}$ be the **underlying excess fluctuation above background** at about 125 GeV (a value predicted by Higg's theory), so that in this problem $\Theta_1 = \{0\}$, and let $t(D)$ be a **one-dimensional summary** of the data set D that (after standardization) has — by assumption (i.e., no bias in the measuring process) and the Central Limit Theorem — an approximately $N(\theta, 1)$ **sampling distribution**; then the LHC researchers computed $P_{RS, \theta=0}[t(D) > 5] \doteq \Phi(-5)$, where RS stands for **repeated-sampling**.

This was a standard **Fisherian significance test**: the researchers were interested in rejecting the hypothesis that

$$\theta = 0 \longleftrightarrow \textbf{(the Higgs boson doesn't exist)}$$

and they gathered data (400 "Higgs-like events" out of **6 trillion particle-particle collisions**) until they achieved a **5–sigma** *P*–value.

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

But It Was Actually Both Inference and Decision

We now go through the usual ***inferential* stochastic proof by contradiction**:

(a) assume the Higgs **doesn’t** exist;

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**:

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**: $\boxed{Q:}$ **Why 5 sigma?**

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**: *Q*: **Why 5 sigma?**

A: The LHC people were worried about the **consequences of a false positive**,

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**: $\boxed{Q:}$ **Why 5 sigma?**

$\boxed{A:}$ The LHC people were worried about the **consequences of a false positive**, for their careers and for the scientific reputation of the LHC;

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**: *Q*: **Why 5 sigma?**

A: The LHC people were worried about the **consequences of a false positive**, for their careers and for the scientific reputation of the LHC; over time the physics community has arrived at 5 sigma as a **convention**,

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**: Q: **Why 5 sigma?**

A: The LHC people were worried about the **consequences of a false positive**, for their careers and for the scientific reputation of the LHC; over time the physics community has arrived at 5 sigma as a **convention**, not as the result of careful calculation (why 1 in 3–3.5 million?).

Thus the LHC **significance test** represented both **inference** (the particle exists)

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**: *Q*: **Why 5 sigma?**

A: The LHC people were worried about the **consequences of a false positive**, for their careers and for the scientific reputation of the LHC; over time the physics community has arrived at 5 sigma as a **convention**, not as the result of careful calculation (why 1 in 3–3.5 million?).

Thus the LHC **significance test** represented both **inference** (the particle exists) and **decision** (whether to announce their findings earlier,

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**: *Q*: **Why 5 sigma?**

A: The LHC people were worried about the **consequences of a false positive**, for their careers and for the scientific reputation of the LHC; over time the physics community has arrived at 5 sigma as a **convention**, not as the result of careful calculation (why 1 in 3–3.5 million?).

Thus the LHC **significance test** represented both **inference** (the particle exists) and **decision** (whether to announce their findings earlier, now (5 sigma)

But It Was Actually Both Inference and Decision

We now go through the usual **inferential stochastic proof by contradiction**:

(a) assume the Higgs **doesn't** exist;

(b) the data are **exceedingly unlikely** under supposition (a); therefore

(c) (a) must be **wrong** and the Higgs **exists** after all.

Peter Higgs and another theoretician got the **Nobel Prize** in physics for this discovery, only one year later.

Just one nagging **question**: *Q*: **Why 5 sigma?**

A: The LHC people were worried about the **consequences of a false positive**, for their careers and for the scientific reputation of the LHC; over time the physics community has arrived at 5 sigma as a **convention**, not as the result of careful calculation (why 1 in 3–3.5 million?).

Thus the LHC **significance test** represented both **inference** (the particle exists) and **decision** (whether to announce their findings earlier, now (5 sigma) or later).

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**,

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well,

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Example: Evaluating a hypertension drug.

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Example: Evaluating a hypertension drug. Consider **assessing** the **performance** of a **drug**, for **lowering systolic blood pressure** (SBP) in **hypertensive** patients, in a **phase–II clinical trial**,

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Example: Evaluating a hypertension drug. 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**).

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Example: Evaluating a hypertension drug. 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**

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Example: Evaluating a hypertension drug. 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**.

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Example: Evaluating a hypertension drug. 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,

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Example: Evaluating a hypertension drug. 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**,

Home Truth #1

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Example: Evaluating a hypertension drug. 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** (MacGregor et al., 1979: *BMJ*: **Captopril**).

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**,

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

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**;

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**),

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**;

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 **(Roche)**.

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 **(Roche)**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**:

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 **(Roche)**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

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 **(Roche)**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

The **action space** here is $(\mathcal{A} \mid \mathcal{B}) = (a_1, a_2) =$ (**don't take the drug forward** to **phase III**, **do take it forward**),

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 **(Roche)**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

The **action space** here is $(\mathcal{A} \mid \mathcal{B}) = (a_1, a_2) =$ (**don't take the drug forward** to **phase III**, **do take it forward**), and a **sensible utility function** $U(a_j, \theta \mid \mathcal{B})$ should be **continuous** and **monotonically increasing** in θ over a **broad range** of **positive** θ values

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

However, to facilitate a comparison between **Neyman-Pearson hypothesis testing** and **Bayesian methods**,

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

However, to facilitate a comparison between **Neyman-Pearson hypothesis testing** and **Bayesian methods**, here I'll **compare two models** M_1 and M_2 that **dichotomize** the θ range, **but not at 0**:

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

However, to facilitate a comparison between **Neyman-Pearson hypothesis testing** and **Bayesian methods**, 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**,

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

However, to facilitate a comparison between **Neyman-Pearson hypothesis testing** and **Bayesian methods**, 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**

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

However, to facilitate a comparison between **Neyman-Pearson hypothesis testing** and **Bayesian methods**, 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**).

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

However, to facilitate a comparison between **Neyman-Pearson hypothesis testing** and **Bayesian methods**, 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 $\theta > \Delta$,

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

However, to facilitate a comparison between **Neyman-Pearson hypothesis testing** and **Bayesian methods**, 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**

Nothing Special About $\theta = 0$

(the **bigger** the **SBP decline** for **hypertensive patients** who **start** at (say) **160 mmHg**, the **better**, up to a **drop** of about **60 mmHg**, **beyond** which the **drug** starts inducing **fainting spells**).

However, to facilitate a comparison between **Neyman-Pearson hypothesis testing** and **Bayesian methods**, 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** (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**).

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous**

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Suppose (as above) that the **parameter space** is $\Theta = \mathbb{R}^k$ for k a positive integer.

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Suppose (as above) that the **parameter space** is $\Theta = \mathbb{R}^k$ for k a positive integer.

Definition: A **structural subspace**

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Suppose (as above) that the **parameter space** is $\Theta = \mathbb{R}^k$ for k a positive integer.

Definition: A **structural subspace** is any $\Theta_1 \subset \Theta$ of dimension less than k for which the **conclusion** that $\theta \in \Theta_1$ would have **different scientific and behavioral consequences** than those arising from the less restrictive statement that $\theta \in \Theta$.

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Suppose (as above) that the **parameter space** is $\Theta = \mathbb{R}^k$ for k a positive integer.

Definition: A **structural subspace** is any $\Theta_1 \subset \Theta$ of dimension less than k for which the **conclusion** that $\theta \in \Theta_1$ would have **different scientific and behavioral consequences** than those arising from the less restrictive statement that $\theta \in \Theta$. If Θ_1 consists of a **single point** $\{\theta_1\}$, such a point is a **structural singleton**.

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Suppose (as above) that the **parameter space** is $\Theta = \mathbb{R}^k$ for k a positive integer.

Definition: A **structural subspace** is any $\Theta_1 \subset \Theta$ of dimension less than k for which the **conclusion** that $\theta \in \Theta_1$ would have **different scientific and behavioral consequences** than those arising from the less restrictive statement that $\theta \in \Theta$. If Θ_1 consists of a **single point** $\{\theta_1\}$, such a point is a **structural singleton**.

Home Truth #2(b): **Sharp-null** $(\theta = \theta_1)$ **hypothesis testing** is only appropriate when θ_1 is a **structural singleton**.

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Suppose (as above) that the **parameter space** is $\Theta = \mathbb{R}^k$ for k a positive integer.

Definition: A **structural subspace** is any $\Theta_1 \subset \Theta$ of dimension less than k for which the **conclusion** that $\theta \in \Theta_1$ would have **different scientific and behavioral consequences** than those arising from the less restrictive statement that $\theta \in \Theta$. If Θ_1 consists of a **single point** $\{\theta_1\}$, such a point is a **structural singleton**.

Home Truth #2(b): **Sharp-null** ($\theta = \theta_1$) **hypothesis testing** is only appropriate when θ_1 is a **structural singleton**. This **rules out** a great deal of testing performed in **routine practice** **(Andrew Gelman)**;

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Suppose (as above) that the **parameter space** is $\Theta = \mathbb{R}^k$ for k a positive integer.

Definition: A **structural subspace** is any $\Theta_1 \subset \Theta$ of dimension less than k for which the **conclusion** that $\theta \in \Theta_1$ would have **different scientific and behavioral consequences** than those arising from the less restrictive statement that $\theta \in \Theta$. If Θ_1 consists of a **single point** $\{\theta_1\}$, such a point is a **structural singleton**.

Home Truth #2(b): **Sharp-null** $(\theta = \theta_1)$ **hypothesis testing** is only appropriate when θ_1 is a **structural singleton**. This **rules out** a great deal of testing performed in **routine practice** **(Andrew Gelman)**; in the **absence** of a structural subspace,

When *Not* To Test

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Suppose (as above) that the **parameter space** is $\Theta = \mathbb{R}^k$ for k a positive integer.

Definition: A **structural subspace** is any $\Theta_1 \subset \Theta$ of dimension less than k for which the **conclusion** that $\theta \in \Theta_1$ would have **different scientific and behavioral consequences** than those arising from the less restrictive statement that $\theta \in \Theta$. If Θ_1 consists of a **single point** $\{\theta_1\}$, such a point is a **structural singleton**.

Home Truth #2(b): **Sharp-null** $(\theta = \theta_1)$ **hypothesis testing** is only appropriate when θ_1 is a **structural singleton**. This **rules out** a great deal of testing performed in **routine practice** **(Andrew Gelman)**; in the **absence** of a structural subspace, the most scienfically useful approach to **inference** is **estimation** via appropriate summaries of the **posterior distribution** $p(\theta \mid D\,\mathcal{B})$.

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued).

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued). I argued above that, if **dichotomization** of $\Theta = \mathbb{R}$ is to be pursued at all,

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued). I argued above that, if **dichotomization** of $\Theta = \mathbb{R}$ is to be pursued at all, the **right dichotomization** is

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued). I argued above that, if **dichotomization** of $\Theta = \mathbb{R}$ is to be pursued at all, the **right dichotomization** is

$$\begin{aligned} \theta \leq \Delta \quad &\longleftrightarrow \quad \textbf{don't take} \text{ drug to Phase III} \\ \theta > \Delta \quad &\longleftrightarrow \quad \textbf{take} \text{ drug to Phase III} \,. \end{aligned} \tag{2}$$

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued). I argued above that, if **dichotomization** of $\Theta = \mathbb{R}$ is to be pursued at all, the **right dichotomization** is

$$\begin{aligned} \theta \leq \Delta \quad &\longleftrightarrow \quad \textbf{don't take} \text{ drug to Phase III} \\ \theta > \Delta \quad &\longleftrightarrow \quad \textbf{take} \text{ drug to Phase III} \, . \end{aligned} \tag{2}$$

Suppose that **little information** about θ **external** to the **experimental data set** You're about to collect is available.

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued). I argued above that, if **dichotomization** of $\Theta = \mathbb{R}$ is to be pursued at all, the **right dichotomization** is

$$\begin{aligned} \theta \leq \Delta \quad &\longleftrightarrow \quad \textbf{don't take} \text{ drug to Phase III} \\ \theta > \Delta \quad &\longleftrightarrow \quad \textbf{take} \text{ drug to Phase III} \,. \end{aligned} \tag{2}$$

Suppose that **little information** about θ **external** to the **experimental data set** You're about to collect is available.

Then, from a **Bayesian** point of view, **hypothesis testing** amounts to **comparing the two models**

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued). I argued above that, if **dichotomization** of $\Theta = \mathbb{R}$ is to be pursued at all, the **right dichotomization** is

$$\begin{aligned} \theta \le \Delta & \quad \longleftrightarrow \quad \textbf{don't take} \text{ drug to Phase III} \\ \theta > \Delta & \quad \longleftrightarrow \quad \textbf{take} \text{ drug to Phase III} \, . \end{aligned} \qquad (2)$$

Suppose that **little information** about θ **external** to the **experimental data set** You're about to collect is available.

Then, from a **Bayesian** point of view, **hypothesis testing** amounts to **comparing the two models**

$$M_1 \colon \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } \theta \le \Delta \\ (y_i | \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \qquad (3)$$

$$M_2 \colon \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } \theta > \Delta \\ (y_i | \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} , \qquad (4)$$

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued). I argued above that, if **dichotomization** of $\Theta = \mathbb{R}$ is to be pursued at all, the **right dichotomization** is

$$\begin{array}{lcl} \theta \leq \Delta & \longleftrightarrow & \textbf{don't take} \text{ drug to Phase III} \\ \theta > \Delta & \longleftrightarrow & \textbf{take} \text{ drug to Phase III}\,. \end{array} \quad (2)$$

Suppose that **little information** about θ **external** to the **experimental data set** You're about to collect is available.

Then, from a **Bayesian** point of view, **hypothesis testing** amounts to **comparing the two models**

$$M_1\text{:} \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } \theta \leq \Delta \\ (y_i | \theta\, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \quad (3)$$

$$M_2\text{:} \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } \theta > \Delta \\ (y_i | \theta\, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \quad (4)$$

in which **for simplicity** I'll take σ to be **known** (the **results** presented below are **similar** with σ **learned** from the **data**).

The Bigger Picture: Bayesian Model Specification

Here's a **rather general algorithm** for finding **good Bayesian models**:

The Bigger Picture: Bayesian Model Specification

Here's a **rather general algorithm** for finding **good Bayesian models**:

(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).

The Bigger Picture: Bayesian Model Specification

Here's a **rather general algorithm** for finding **good Bayesian models**:

(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).

The question in **step (a) — Where to start? —** is often easy to answer;

The Bigger Picture: Bayesian Model Specification

Here's a **rather general algorithm** for finding **good Bayesian models**:

(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).

The question in **step (a)** — **Where to start?** — is often easy to answer; by contrast, the question in **step (c)** is **so hard to answer** that we currently don't have any **reliable Bayesian modeling robots/AIs**.

The Bigger Picture: Bayesian Model Specification

Here's a **rather general algorithm** for finding **good Bayesian models**:

(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).

The question in **step (a) — Where to start? —** is often easy to answer; by contrast, the question in **step (c)** is **so hard to answer** that we currently don't have any **reliable Bayesian modeling robots/AIs**.

Implementing the algorithm above involves facing **two additional important questions**, in **steps (d) and (b)** (respectively):

The Bigger Picture: Bayesian Model Specification

Here's a **rather general algorithm** for finding **good Bayesian models**:

(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).

The question in **step (a) — Where to start? —** is often easy to answer; by contrast, the question in **step (c)** is **so hard to answer** that we currently don't have any **reliable Bayesian modeling robots/AIs**.

Implementing the algorithm above involves facing **two additional important questions**, in **steps (d) and (b)** (respectively):

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

The Bigger Picture: Bayesian Model Specification

Here's a **rather general algorithm** for finding **good Bayesian models**:

(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).

The question in **step (a)** — **Where to start?** — is often easy to answer; by contrast, the question in **step (c)** is **so hard to answer** that we currently don't have any **reliable Bayesian modeling robots/AIs**.

Implementing the algorithm above involves facing **two additional important questions**, in **steps (d) and (b)** (respectively):

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

A Hypothesis Is Just a Prior Specification

Example: Evaluating a hypertension drug (continued). I argued above that, if **dichotomization** of $\Theta = \mathbb{R}$ is to be pursued at all, the **right dichotomization** is

$$\begin{aligned} \theta \leq \Delta & \quad\longleftrightarrow\quad \textbf{don't take} \text{ drug to Phase III} \\ \theta > \Delta & \quad\longleftrightarrow\quad \textbf{take} \text{ drug to Phase III}\,. \end{aligned} \tag{5}$$

Suppose that **little information** about θ **external** to the **experimental data set** You're about to collect is available.

Then, from a **Bayesian** point of view, **hypothesis testing** amounts to **comparing the two models**

$$M_1\colon \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } \theta \leq \Delta \\ (y_i | \theta\, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{6}$$

$$M_2\colon \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } \theta > \Delta \\ (y_i | \theta\, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{7}$$

in which **for simplicity** I'll take σ to be **known** (the **results** presented below are **similar** with σ **learned** from the **data**).

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed:

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose?**

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose? [utility]**

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose? [utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**,

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose?** **[utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**, and again highlights the importance of **decision** in **Bayesian hypothesis testing**.

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose? [utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**, and again highlights the importance of **decision** in **Bayesian hypothesis testing**.

Strictly speaking, **better for what purpose?** can only be answered on a **problem-by-problem basis**,

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose? [utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**, and again highlights the importance of **decision** in **Bayesian hypothesis testing**.

Strictly speaking, **better for what purpose?** can only be answered on a **problem-by-problem basis**, with a **utility function** tailored to the problem at hand;

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose? [utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**, and again highlights the importance of **decision** in **Bayesian hypothesis testing**.

Strictly speaking, **better for what purpose?** can only be answered on a **problem-by-problem basis**, with a **utility function** tailored to the problem at hand; but people have a powerful need for **general-purpose tools** whose implied utility structure may be a **decent approximation** in the problem they're working on.

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose? [utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**, and again highlights the importance of **decision** in **Bayesian hypothesis testing**.

Strictly speaking, **better for what purpose?** can only be answered on a **problem-by-problem basis**, with a **utility function** tailored to the problem at hand; but people have a powerful need for **general-purpose tools** whose implied utility structure may be a **decent approximation** in the problem they're working on.

Three such tools are **Bayes factors**,

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose?** **[utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**, and again highlights the importance of **decision** in **Bayesian hypothesis testing**.

Strictly speaking, **better for what purpose?** can only be answered on a **problem-by-problem basis**, with a **utility function** tailored to the problem at hand; but people have a powerful need for **general-purpose tools** whose implied utility structure may be a **decent approximation** in the problem they're working on.

Three such tools are **Bayes factors**, **log scores**, and

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose?** **[utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**, and again highlights the importance of **decision** in **Bayesian hypothesis testing**.

Strictly speaking, **better for what purpose?** can only be answered on a **problem-by-problem basis**, with a **utility function** tailored to the problem at hand; but people have a powerful need for **general-purpose tools** whose implied utility structure may be a **decent approximation** in the problem they're working on.

Three such tools are **Bayes factors**, **log scores**, and **posterior probabilities** (more on this later);

Hypothesis Testing = Model Comparison

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**:

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

This question cannot be answered until a **more fundamental question** is addressed: **better for what purpose? [utility]**

This means that **Bayesian model specification** is **fundamentally decision-theoretic**, and again highlights the importance of **decision** in **Bayesian hypothesis testing**.

Strictly speaking, **better for what purpose?** can only be answered on a **problem-by-problem basis**, with a **utility function** tailored to the problem at hand; but people have a powerful need for **general-purpose tools** whose implied utility structure may be a **decent approximation** in the problem they're working on.

Three such tools are **Bayes factors**, **log scores**, and **posterior probabilities** (more on this later); any such method appropriate to model comparison is **equally appropriate to hypothesis testing**.

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

also **cannot be answered** using general-purpose methodology,

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

also **cannot be answered** using general-purpose methodology, because answering it also raises a **more fundamental question**:

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

also **cannot be answered** using general-purpose methodology, because answering it also raises a **more fundamental question**: **good enough for what purpose?**

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

also **cannot be answered** using general-purpose methodology, because answering it also raises a **more fundamental question**: **good enough for what purpose?**

This again fundamentally requires **special-purpose decision-theory**,

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

also **cannot be answered** using general-purpose methodology, because answering it also raises a **more fundamental question**: **good enough for what purpose?**

This again fundamentally requires **special-purpose decision-theory**, but a **related question** CAN be answered rather generally:

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

also **cannot be answered** using general-purpose methodology, because answering it also raises a **more fundamental question**: **good enough for what purpose?**

This again fundamentally requires **special-purpose decision-theory**, but a **related question** CAN be answered rather generally:

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:**

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

also **cannot be answered** using general-purpose methodology, because answering it also raises a **more fundamental question**: **good enough for what purpose?**

This again fundamentally requires **special-purpose decision-theory**, but a **related question** CAN be answered rather generally:

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Bayesian Significance Testing Can Be Meaningful Too

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

The **other model specification question**

Q_2: Is M_1 **good enough** (to stop looking for a better model)?

also **cannot be answered** using general-purpose methodology, because answering it also raises a **more fundamental question**: **good enough for what purpose?**

This again fundamentally requires **special-purpose decision-theory**, but a **related question** CAN be answered rather generally:

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Q_2': **Could** the data set D have **arisen** from M_1?

PPPs Are Often Badly Calibrated

$\boxed{Q_2':}$ **Could** the data set D have **arisen** from M_1?

PPPs Are Often Badly Calibrated

$\boxed{Q_2'}$: **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do,

PPPs Are Often Badly Calibrated

$\boxed{Q_2':}$ **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

PPPs Are Often Badly Calibrated

$\boxed{Q_2':}$ **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

if Gelman gives You a P–value of **0.04**, that's bad for M_1;

PPPs Are Often Badly Calibrated

Q_2': **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

if Gelman gives You a P–value of **0.04**, that's bad for M_1; but if You get 0.4 from Gelman,

PPPs Are Often Badly Calibrated

Q_2': **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

if Gelman gives You a P–value of **0.04**, that's bad for M_1; but if You get 0.4 from Gelman, a **well-calibrated version** of that "P–value" could easily be more like **0.04**

PPPs Are Often Badly Calibrated

Q_2': **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

if Gelman gives You a P–value of **0.04**, that's bad for M_1; but if You get 0.4 from Gelman, a **well-calibrated version** of that "P–value" could easily be more like **0.04** (Draper and Krnjajić, 2015, document this and show how to **fix it**).

PPPs Are Often Badly Calibrated

Q_2': **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

if Gelman gives You a P–value of **0.04**, that's bad for M_1; but if You get 0.4 from Gelman, a **well-calibrated version** of that "P–value" could easily be more like **0.04** (Draper and Krnjajić, 2015, document this and show how to **fix it**).

Example: Evaluating a hypertension drug (continued).

PPPs Are Often Badly Calibrated

Q_2': **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

if Gelman gives You a P–value of **0.04**, that's bad for M_1; but if You get 0.4 from Gelman, a **well-calibrated version** of that "P–value" could easily be more like **0.04** (Draper and Krnjajić, 2015, document this and show how to **fix it**).

Example: Evaluating a hypertension drug (continued). An enlightened version of the **frequentist Neyman–Pearson approach** would test H_1: $\theta \leq \Delta$ against H_2: $\theta > \Delta$,

PPPs Are Often Badly Calibrated

Q_2': **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

if Gelman gives You a P–value of **0.04**, that's bad for M_1; but if You get 0.4 from Gelman, a **well-calibrated version** of that "P–value" could easily be more like **0.04** (Draper and Krnjajić, 2015, document this and show how to **fix it**).

Example: Evaluating a hypertension drug (continued). An enlightened version of the **frequentist Neyman–Pearson approach** would test H_1: $\theta \leq \Delta$ against H_2: $\theta > \Delta$, using the following **implied utility structure** with $(\alpha, \beta) =$ (type **I** error rate, type **II** error rate):

PPPs Are Often Badly Calibrated

Q_2': **Could** the data set D have **arisen** from M_1?

This is what methods such as **posterior predictive** P**–values** (**PPP**; Gelman et al., 1996) try to do, but PPP is typically **badly calibrated** (Bayarri and Berger, 2000; Robins et al., 2000):

if Gelman gives You a P–value of **0.04**, that's bad for M_1; but if You get 0.4 from Gelman, a **well-calibrated version** of that "P–value" could easily be more like **0.04** (Draper and Krnjajić, 2015, document this and show how to **fix it**).

Example: Evaluating a hypertension drug (continued). An enlightened version of the **frequentist Neyman–Pearson approach** would test H_1: $\theta \leq \Delta$ against H_2: $\theta > \Delta$, using the following **implied utility structure** with $(\alpha, \beta) =$ (type **I** error rate, type **II** error rate):

N–P	**Truth**	
Action	$\theta \leq \Delta$	$\theta > \Delta$
a_1 (**stop**)	0	$-\alpha$
a_2 (**phase III**)	$-\beta$	0

Neyman–Pearson Utility Structure Is Wrong

N–P Action	Truth: $\theta \leq \Delta$	Truth: $\theta > \Delta$
a_1 **(stop)**	0	$-\alpha$
a_2 **(phase III)**	$-\beta$	0

Neyman–Pearson Utility Structure Is Wrong

N–P Action	**Truth** $\theta \le \Delta$	$\theta > \Delta$
a_1 (**stop**)	0	$-\alpha$
a_2 (**phase III**)	$-\beta$	0

But this **utility structure is wrong** in all 4 cells: with $\{u_{ij}\} \ge 0$,

Neyman–Pearson Utility Structure Is Wrong

N–P Action	Truth $\theta \leq \Delta$	$\theta > \Delta$
a_1 (**stop**)	0	$-\alpha$
a_2 (**phase III**)	$-\beta$	0

But this **utility structure is wrong** in all 4 cells: with $\{u_{ij}\} \geq 0$,

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

Neyman–Pearson Utility Structure Is Wrong

N–P Action	Truth $\theta \le \Delta$	$\theta > \Delta$
a_1 **(stop)**	0	$-\alpha$
a_2 **(phase III)**	$-\beta$	0

But this **utility structure is wrong** in all 4 cells: with $\{u_{ij}\} \ge 0$,

Bayes Action	Truth $\theta \le \Delta$	$\theta > \Delta$
a_1 **(stop)**	u_{11}	$-u_{12}$
a_2 **(phase III)**	$-u_{21}$	u_{22}

- $u_{11} > 0$ is the **gain** from **correctly not going forward** to **phase III**;

Neyman–Pearson Utility Structure Is Wrong

N–P Action	Truth $\theta \leq \Delta$	$\theta > \Delta$
a_1 (**stop**)	0	$-\alpha$
a_2 (**phase III**)	$-\beta$	0

But this **utility structure is wrong** in all 4 cells: with $\{u_{ij}\} \geq 0$,

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

- $u_{11} > 0$ is the **gain** from **correctly not going forward** to **phase III**;
- $-u_{12} < 0$ is the **loss** from **incorrectly failing to go forward**;

Neyman–Pearson Utility Structure Is Wrong

N–P	**Truth**	
Action	$\theta \leq \Delta$	$\theta > \Delta$
a_1 (**stop**)	0	$-\alpha$
a_2 (**phase III**)	$-\beta$	0

But this **utility structure is wrong** in all 4 cells: with $\{u_{ij}\} \geq 0$,

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

- $u_{11} > 0$ is the **gain** from **correctly not going forward** to **phase III**;
- $-u_{12} < 0$ is the **loss** from **incorrectly failing to go forward**;
- $-u_{21} < 0$ is the **loss** from **incorrectly going forward**; and

Neyman–Pearson Utility Structure Is Wrong

N–P Action	Truth $\theta \le \Delta$	$\theta > \Delta$
a_1 (**stop**)	0	$-\alpha$
a_2 (**phase III**)	$-\beta$	0

But this **utility structure is wrong** in all 4 cells: with $\{u_{ij}\} \ge 0$,

Bayes Action	Truth $\theta \le \Delta$	$\theta > \Delta$
a_1 (**stop**)	u_{11}	$-u_{12}$
a_2 (**phase III**)	$-u_{21}$	u_{22}

- $u_{11} > 0$ is the **gain** from **correctly not going forward** to **phase III**;
- $-u_{12} < 0$ is the **loss** from **incorrectly failing to go forward**;
- $-u_{21} < 0$ is the **loss** from **incorrectly going forward**; and
- $u_{22} > 0$ is the **gain** from **correctly going forward**.

Neyman–Pearson Utility Structure Is Wrong

N–P Action	Truth $\theta \le \Delta$	$\theta > \Delta$
a_1 (**stop**)	0	$-\alpha$
a_2 (**phase III**)	$-\beta$	0

But this **utility structure is wrong** in all 4 cells: with $\{u_{ij}\} \ge 0$,

Bayes Action	Truth $\theta \le \Delta$	$\theta > \Delta$
a_1 (**stop**)	u_{11}	$-u_{12}$
a_2 (**phase III**)	$-u_{21}$	u_{22}

- $u_{11} > 0$ is the **gain** from **correctly not going forward** to **phase III**;
- $-u_{12} < 0$ is the **loss** from **incorrectly failing to go forward**;
- $-u_{21} < 0$ is the **loss** from **incorrectly going forward**; and
- $u_{22} > 0$ is the **gain** from **correctly going forward**.

The $\{u_{ij}\}$ need to be in **money, or QALYs, or ...**;

Neyman–Pearson Utility Structure Is Wrong

N–P Action	Truth $\theta \le \Delta$	$\theta > \Delta$
a_1 (**stop**)	0	$-\alpha$
a_2 (**phase III**)	$-\beta$	0

But this **utility structure is wrong** in all 4 cells: with $\{u_{ij}\} \geq 0$,

Bayes Action	Truth $\theta \le \Delta$	$\theta > \Delta$
a_1 (**stop**)	u_{11}	$-u_{12}$
a_2 (**phase III**)	$-u_{21}$	u_{22}

- $u_{11} > 0$ is the **gain** from **correctly not going forward** to **phase III**;
- $-u_{12} < 0$ is the **loss** from **incorrectly failing to go forward**;
- $-u_{21} < 0$ is the **loss** from **incorrectly going forward**; and
- $u_{22} > 0$ is the **gain** from **correctly going forward**.

The $\{u_{ij}\}$ need to be in **money, or QALYs, or ...**;
α and β are **incorrectly on the probability scale**.

Don't Use Inferential Tools To Make Decisions

The **optimal Bayesian decision** turns out to be:

Don't Use Inferential Tools To Make Decisions

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

$$P(\theta > \Delta | y\, \mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \,. \qquad (8)$$

Don't Use Inferential Tools To Make Decisions

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

$$P(\theta > \Delta | y\, \mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \,. \qquad (8)$$

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

Don't Use Inferential Tools To Make Decisions

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

$$P(\theta > \Delta | y \, \mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \, . \quad (8)$$

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

$$\alpha = (1 - u^*) = \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}} \, . \quad (9)$$

Don't Use Inferential Tools To Make Decisions

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

$$P(\theta > \Delta | y\,\mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \,. \tag{8}$$

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

$$\alpha = (1 - u^*) = \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}} \,. \tag{9}$$

The **built-in trade-off** between **false positives and false negatives** in **level–α hypothesis-testing** for any given α

Don't Use Inferential Tools To Make Decisions

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

$$P(\theta > \Delta | y\, \mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \,. \qquad (8)$$

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

$$\alpha = (1 - u^*) = \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}} \,. \qquad (9)$$

The **built-in trade-off** between **false positives and false negatives** in **level–α hypothesis-testing** for any given α may be **close to optimal** or **not**, according to the **real-world values** of $\{u_{11}, u_{12}, u_{21}, u_{22}\}$.

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

$$P(\theta > \Delta | y\, \mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \,. \tag{8}$$

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

$$\alpha = (1 - u^*) = \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}} \,. \tag{9}$$

The **built-in trade-off** between **false positives and false negatives** in **level–α hypothesis-testing** for any given α may be **close to optimal** or **not**, according to the **real-world values** of $\{u_{11}, u_{12}, u_{21}, u_{22}\}$.

In **phase-II clinical trials** or **micro-array experiments**,

Don't Use Inferential Tools To Make Decisions

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

$$P(\theta > \Delta | y\, \mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \,. \qquad (8)$$

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

$$\alpha = (1 - u^*) = \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}} \,. \qquad (9)$$

The **built-in trade-off** between **false positives and false negatives** in **level–α hypothesis-testing** for any given α may be **close to optimal** or **not**, according to the **real-world values** of $\{u_{11}, u_{12}, u_{21}, u_{22}\}$.

In **phase-II clinical trials** or **micro-array experiments**, when You're **screening many drugs** or **genes** for those that **may lead** to an **effective treatment**

Don't Use Inferential Tools To Make Decisions

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

$$P(\theta > \Delta | y\, \mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \,. \tag{8}$$

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

$$\alpha = (1 - u^*) = \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}} \,. \tag{9}$$

The **built-in trade-off** between **false positives and false negatives** in **level–α hypothesis-testing** for any given α may be **close to optimal** or **not**, according to the **real-world values** of $\{u_{11}, u_{12}, u_{21}, u_{22}\}$.

In **phase-II clinical trials** or **micro-array experiments**, when You're **screening many drugs** or **genes** for those that **may lead** to an **effective treatment** and — from the **drug company's point of view** — a **false-negative error** (of **failing to move forward** with a **drug** or **gene** that's actually **worth further investigation**) can be **much more costly** than a **false-positive mistake**,

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

$$P(\theta > \Delta | y\,\mathcal{B}) \geq \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^* \,. \qquad (8)$$

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

$$\alpha = (1 - u^*) = \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}} \,. \qquad (9)$$

The **built-in trade-off** between **false positives and false negatives** in **level–α hypothesis-testing** for any given α may be **close to optimal** or **not**, according to the **real-world values** of $\{u_{11}, u_{12}, u_{21}, u_{22}\}$.

In **phase-II clinical trials** or **micro-array experiments**, when You're **screening many drugs** or **genes** for those that **may lead** to an **effective treatment** and — from the **drug company's point of view** — a **false-negative error** (of **failing to move forward** with a **drug** or **gene** that's actually **worth further investigation**) can be **much more costly** than a **false-positive mistake**, this **corresponds** to $u_{12} \gg u_{21}$

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**,

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**,

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger α values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**, this **approach** led to α **values** on the order of

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**, this **approach** led to α **values** on the order of **0.45**.

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**, this **approach** led to α **values** on the order of **0.45**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**:

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**, this **approach** led to α **values** on the order of **0.45**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**, this **approach** led to α **values** on the order of **0.45**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

- If the problem had instead been **inferential**,

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**, this **approach** led to α **values** on the order of **0.45**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

- If the problem had instead been **inferential**, the **optimal conclusion** would simply be based on the posterior for θ:

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**, this **approach** led to α **values** on the order of **0.45**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

- If the problem had instead been **inferential**, the **optimal conclusion** would simply be based on the posterior for θ: let

$$M^* = \{(\theta|\mathcal{B}) \sim \textbf{diffuse} \text{ on } \mathbb{R}, (y_i|\theta\,\mathcal{B}) \stackrel{\text{IID}}{\sim} N(\theta, \sigma^2)\}$$

Home Truth #1(b) Revisited

and **leads** in the **hypothesis-testing approach** in **phase-II trials** to a **willingness** to use **(much) larger** α **values** than the **conventional 0.01** or **0.05**, something that **good frequentist biostatisticians** have **long known intuitively**.

In **work** I've done with the **Swiss pharmaceutical company Roche**, this **approach** led to α **values** on the order of **0.45**.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

- If the problem had instead been **inferential**, the **optimal conclusion** would simply be based on the posterior for θ: let

$$M^* = \{(\theta|\mathcal{B}) \sim \textbf{diffuse} \text{ on } \mathbb{R}, (y_i|\theta\,\mathcal{B}) \stackrel{\text{IID}}{\sim} N(\theta, \sigma^2)\}$$

and **choose** M_2 if $p(\theta > \Delta | y\, M^*\, \mathcal{B}) > 0.5$.

Establishing Bio-Equivalence

- **(establishing bio-equivalence)** In this case there's a **previous hypertension drug** B (call the **new drug** A),

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**.

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,

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**:

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,

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

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**).

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**

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**

$$[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})] \qquad (10)$$

in the **population** of **patients** to which it's **appropriate** to **generalize** from the **patients in Your trial**,

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**

$$[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})] \qquad (10)$$

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

- **(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**

$$[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})] \qquad (10)$$

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

Again in this **setting** there's **nothing special** about $\theta = 0$,

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**

$$[(SBP_{before,A} - SBP_{after,A}) - (SBP_{before,B} - SBP_{after,B})] \qquad (10)$$

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 $|\theta| \leq \lambda$,

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 **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

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

$$M_3\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \qquad (11)$$

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

$$M_3\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{11}$$

$$M_4\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{12}$$

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

$$M_3\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{11}$$

$$M_4\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{12}$$

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

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

$$M_3\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{11}$$

$$M_4\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{12}$$

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

Bayesian decision theory (as in the drug evaluation above) again leads to the **optimal action**;

Bio-Equivalence Modeling

what **matters** here is whether $|\theta| \le \lambda$, 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

$$M_3\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \le \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{11}$$

$$M_4\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} , \tag{12}$$

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

Bayesian decision theory (as in the drug evaluation above) again leads to the **optimal action**; if **inference** were instead the goal, again just look at the **posterior** for θ:

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

$$M_3\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{11}$$

$$M_4\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{12}$$

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

Bayesian decision theory (as in the drug evaluation above) again leads to the **optimal action**; if **inference** were instead the goal, again just look at the **posterior** for θ: as before, let

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

$$M_3\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{11}$$

$$M_4\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{12}$$

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

Bayesian decision theory (as in the drug evaluation above) again leads to the **optimal action**; if **inference** were instead the goal, again just look at the **posterior** for θ: as before, let

$$M^* = \{(\theta|\mathcal{B}) \sim \text{diffuse on } \mathbb{R}, (y_i|\theta\,\mathcal{B}) \stackrel{\text{IID}}{\sim} N(\theta, \sigma^2)\},$$

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

$$M_3\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| \leq \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\} \quad \text{and} \tag{11}$$

$$M_4\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } |\theta| > \lambda \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & N(\theta, \sigma^2) \end{array} \right\}, \tag{12}$$

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

Bayesian decision theory (as in the drug evaluation above) again leads to the **optimal action**; if **inference** were instead the goal, again just look at the **posterior** for θ: as before, let

$$M^* = \{(\theta|\mathcal{B}) \sim \text{diffuse on } \mathbb{R}, (y_i|\theta\,\mathcal{B}) \stackrel{\text{IID}}{\sim} N(\theta, \sigma^2)\},$$

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

Case Study: Mendel's Peas

- Between 1856 and 1863 the Augustinian monk Gregor Mendel cultivated about 28,000 plants, most of them garden peas (*Pisum sativum*), to study the **nature of inheritance**, publishing his results in Mendel (1866).

Case Study: Mendel's Peas

- Between 1856 and 1863 the Augustinian monk Gregor Mendel cultivated about 28,000 plants, most of them garden peas (*Pisum sativum*), to study the **nature of inheritance**, publishing his results in Mendel (1866).
- He examined seven observable (**phenotypic**) characteristics of his pea plants, including whether the seeds were round or wrinkled.

Case Study: Mendel's Peas

- Between 1856 and 1863 the Augustinian monk Gregor Mendel cultivated about 28,000 plants, most of them garden peas (*Pisum sativum*), to study the **nature of inheritance**, publishing his results in Mendel (1866).
- He examined seven observable (**phenotypic**) characteristics of his pea plants, including whether the seeds were round or wrinkled.
- He grew multiple generations of many lines of peas for two years, to ensure that they **bred true**, meaning that — in the case of seed shape — every new generation always had round seeds in some of the lines and always wrinkled seeds in other lines.

Case Study: Mendel's Peas

- Between 1856 and 1863 the Augustinian monk Gregor Mendel cultivated about 28,000 plants, most of them garden peas (*Pisum sativum*), to study the **nature of inheritance**, publishing his results in Mendel (1866).
- He examined seven observable (**phenotypic**) characteristics of his pea plants, including whether the seeds were round or wrinkled.
- He grew multiple generations of many lines of peas for two years, to ensure that they **bred true**, meaning that — in the case of seed shape — every new generation always had round seeds in some of the lines and always wrinkled seeds in other lines.
- He then crossed pure-round and pure-wrinkled plants; all of the (first-generation) offspring came out round, demonstrating in his nascent genetic theory that round is the **dominant** phenotype and wrinkled the **recessive**.

Case Study: Mendel's Peas

- Between 1856 and 1863 the Augustinian monk Gregor Mendel cultivated about 28,000 plants, most of them garden peas (*Pisum sativum*), to study the **nature of inheritance**, publishing his results in Mendel (1866).
- He examined seven observable (**phenotypic**) characteristics of his pea plants, including whether the seeds were round or wrinkled.
- He grew multiple generations of many lines of peas for two years, to ensure that they **bred true**, meaning that — in the case of seed shape — every new generation always had round seeds in some of the lines and always wrinkled seeds in other lines.
- He then crossed pure-round and pure-wrinkled plants; all of the (first-generation) offspring came out round, demonstrating in his nascent genetic theory that round is the **dominant** phenotype and wrinkled the **recessive**.
- But when he crossed the **first-generation** offspring with each other, only about $\boxed{\theta_1 = \frac{3}{4}}$ had **second-generation** offspring with round seeds.

Second-Generation Hybrid Results

- Precisely the same thing happened with the **other six phenotype characters**.

Second-Generation Hybrid Results

- Precisely the same thing happened with the **other six phenotype characters**.
- The table below presents **Mendel's raw data** (Griffiths et al. (2000)) for all seven phenotypes;

Second-Generation Hybrid Results

- Precisely the same thing happened with the **other six phenotype characters**.
- The table below presents **Mendel's raw data** (Griffiths et al. (2000)) for all seven phenotypes; here s is the number of **dominants** he observed out of n plants, and $\bar{y} = \frac{s}{n}$:

Second-Generation Hybrid Results

- Precisely the same thing happened with the **other six phenotype characters**.
- The table below presents **Mendel's raw data** (Griffiths et al. (2000)) for all seven phenotypes; here s is the number of **dominants** he observed out of n plants, and $\bar{y} = \frac{s}{n}$:

```
         dataset                s    n  y.bar

round x wrinkled seeds   5474 7324 0.7474
yellow x green seeds     6022 8023 0.7506     A   a
purple x white petals     705  929 0.7589   +---+---+
inflated x pinched pods   882 1181 0.7468 A | A | A |
green x yellow pods       428  580 0.7379   +---+---+
axial x terminal flowers  651  858 0.7587 a | A | a |
long x short stems        787 1064 0.7397   +---+---+
```

Second-Generation Hybrid Results

- Precisely the same thing happened with the **other six phenotype characters**.
- The table below presents **Mendel's raw data** (Griffiths et al. (2000)) for all seven phenotypes; here s is the number of **dominants** he observed out of n plants, and $\bar{y} = \frac{s}{n}$:

dataset	s	n	y.bar
round x wrinkled seeds	5474	7324	0.7474
yellow x green seeds	6022	8023	0.7506
purple x white petals	705	929	0.7589
inflated x pinched pods	882	1181	0.7468
green x yellow pods	428	580	0.7379
axial x terminal flowers	651	858	0.7587
long x short stems	787	1064	0.7397

```
    A   a
  +---+---+
A | A | A |
  +---+---+
a | A | a |
  +---+---+
```

From this data, Mendel formulated his now-familiar **theory of inheritance** with dominant-recessive characteristics:

Second-Generation Hybrid Results

- Precisely the same thing happened with the **other six phenotype characters**.
- The table below presents **Mendel's raw data** (Griffiths et al. (2000)) for all seven phenotypes; here s is the number of **dominants** he observed out of n plants, and $\bar{y} = \frac{s}{n}$:

dataset	s	n	y.bar
round x wrinkled seeds	5474	7324	0.7474
yellow x green seeds	6022	8023	0.7506
purple x white petals	705	929	0.7589
inflated x pinched pods	882	1181	0.7468
green x yellow pods	428	580	0.7379
axial x terminal flowers	651	858	0.7587
long x short stems	787	1064	0.7397

```
    A   a
  +---+---+
A | A | A |
  +---+---+
a | A | a |
  +---+---+
```

From this data, Mendel formulated his now-familiar **theory of inheritance** with dominant-recessive characteristics: each parent contributes one of the two **alleles**, A (dominant) or a (recessive) they got from their parents,

Second-Generation Hybrid Results

- Precisely the same thing happened with the **other six phenotype characters**.
- The table below presents **Mendel's raw data** (Griffiths et al. (2000)) for all seven phenotypes; here s is the number of **dominants** he observed out of n plants, and $\bar{y} = \frac{s}{n}$:

dataset	s	n	y.bar
round x wrinkled seeds	5474	7324	0.7474
yellow x green seeds	6022	8023	0.7506
purple x white petals	705	929	0.7589
inflated x pinched pods	882	1181	0.7468
green x yellow pods	428	580	0.7379
axial x terminal flowers	651	858	0.7587
long x short stems	787	1064	0.7397

```
    A   a
  +---+---+
A | A | A |
  +---+---+
a | A | a |
  +---+---+
```

From this data, Mendel formulated his now-familiar **theory of inheritance** with dominant-recessive characteristics: each parent contributes one of the two **alleles**, A (dominant) or a (recessive) they got from their parents, as in the 2×2 **Punnett square** above.

Mendel's Model Comparison

Roll the clock back mentally to **1865**,

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**;

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time —

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say —

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**,

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**:

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\colon \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i | \theta\, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \qquad (13)$$

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\text{:} \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i | \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \qquad (13)$$

$$M_2\text{:} \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } 0 < \theta < 1 \\ (y_i | \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} , \qquad (14)$$

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\text{:} \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i | \theta\, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{13}$$

$$M_2\text{:} \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } 0 < \theta < 1 \\ (y_i | \theta\, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} , \tag{14}$$

in which —

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{13}$$

$$M_2\text{:} \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } 0 < \theta < 1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\}, \tag{14}$$

in which — without loss of much generality —

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{13}$$

$$M_2\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } 0 < \theta < 1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\}, \tag{14}$$

in which — without loss of much generality — the **prior** in M_2 can be **instantiated** with a Beta(α, β) distribution with **small positive** (α, β).

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{13}$$

$$M_2\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{diffuse for } 0 < \theta < 1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\}, \tag{14}$$

in which — without loss of much generality — the **prior** in M_2 can be **instantiated** with a Beta(α, β) distribution with **small positive** (α, β).

How to **compare** these two models?

Mendel's Model Comparison

Roll the clock back mentally to **1865**, and imagine Mendel proposing a **theory** involving a **structural singleton** at $\theta_1 = \frac{3}{4}$ in the context of a **Bernoulli sampling model**; how strongly do these data **support or refute** such a theory?

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\colon \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i | \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{13}$$

$$M_2\colon \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{diffuse for } 0 < \theta < 1 \\ (y_i | \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} , \tag{14}$$

in which — without loss of much generality — the **prior** in M_2 can be **instantiated** with a Beta(α, β) distribution with **small positive** (α, β).

How to **compare** these two models? One approach: **Bayes factors**.

Bayes Factors

Suppose that the number m of models in Your **ensemble** $\mathcal{M} = \{M_1, \ldots, M_m\}$ of models under comparison is **finite**.

Bayes Factors

Suppose that the number m of models in Your **ensemble** $\mathcal{M} = \{M_1, \ldots, M_m\}$ of models under comparison is **finite**.

In such cases it suffices to make **pairwise comparisons** of the M_j;

Bayes Factors

Suppose that the number m of models in Your **ensemble** $\mathcal{M} = \{M_1, \ldots, M_m\}$ of models under comparison is **finite**.

In such cases it suffices to make **pairwise comparisons** of the M_j; so specialize to the case $m = 2$ and $\mathcal{M} = \{M_1, M_2\}$.

Bayes Factors

Suppose that the number m of models in Your **ensemble** $\mathcal{M} = \{M_1, \ldots, M_m\}$ of models under comparison is **finite**.

In such cases it suffices to make **pairwise comparisons** of the M_j; so specialize to the case $m = 2$ and $\mathcal{M} = \{M_1, M_2\}$.

Bayes factors arise as the **data-driven component** of a decision-theoretic approach to model comparison that selects the model with the **highest posterior probability**:

$$\left[\frac{p(M_2 \mid D \, \mathcal{B})}{p(M_1 \mid D \, \mathcal{B})} \right] = \left[\frac{p(M_2 \mid \mathcal{B})}{p(M_1 \mid \mathcal{B})} \right] \cdot \left[\frac{p(D \mid M_2 \, \mathcal{B})}{p(D \mid M_1 \, \mathcal{B})} \right]$$
$$\left[\begin{array}{c} \text{posterior odds} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] = \left[\begin{array}{c} \text{prior odds} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] \cdot \left[\begin{array}{c} \text{Bayes factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] . \tag{15}$$

Bayes Factors

Suppose that the number m of models in Your **ensemble** $\mathcal{M} = \{M_1, \ldots, M_m\}$ of models under comparison is **finite**.

In such cases it suffices to make **pairwise comparisons** of the M_j; so specialize to the case $m = 2$ and $\mathcal{M} = \{M_1, M_2\}$.

Bayes factors arise as the **data-driven component** of a decision-theoretic approach to model comparison that selects the model with the **highest posterior probability**:

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

$$\left[\begin{array}{c}\text{posterior odds}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right] = \left[\begin{array}{c}\text{prior odds}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right] \cdot \left[\begin{array}{c}\text{Bayes factor}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right]. \qquad (15)$$

Specifying the **prior odds ratio** in applied settings seems to me to be a **more difficult problem** than acknowledged by such writers as Jeffreys (1939)

Bayes Factors

Suppose that the number m of models in Your **ensemble** $\mathcal{M} = \{M_1, \dots, M_m\}$ of models under comparison is **finite**.

In such cases it suffices to make **pairwise comparisons** of the M_j; so specialize to the case $m = 2$ and $\mathcal{M} = \{M_1, M_2\}$.

Bayes factors arise as the **data-driven component** of a decision-theoretic approach to model comparison that selects the model with the **highest posterior probability**:

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

$$\left[\begin{array}{c}\text{posterior odds}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right] = \left[\begin{array}{c}\text{prior odds}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right] \cdot \left[\begin{array}{c}\text{Bayes factor}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right]. \quad (15)$$

Specifying the **prior odds ratio** in applied settings seems to me to be a **more difficult problem** than acknowledged by such writers as Jeffreys (1939) — e.g., I see **nothing remotely "objective"** about taking this ratio to be **1**;

Bayes Factors

Suppose that the number m of models in Your **ensemble** $\mathcal{M} = \{M_1, \ldots, M_m\}$ of models under comparison is **finite**.

In such cases it suffices to make **pairwise comparisons** of the M_j; so specialize to the case $m = 2$ and $\mathcal{M} = \{M_1, M_2\}$.

Bayes factors arise as the **data-driven component** of a decision-theoretic approach to model comparison that selects the model with the **highest posterior probability**:

$$\left[\frac{p(M_2 \mid D\,\mathcal{B})}{p(M_1 \mid D\,\mathcal{B})}\right] = \left[\frac{p(M_2 \mid \mathcal{B})}{p(M_1 \mid \mathcal{B})}\right] \cdot \left[\frac{p(D \mid M_2\,\mathcal{B})}{p(D \mid M_1\,\mathcal{B})}\right]$$
$$\left[\begin{array}{c}\text{posterior odds}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right] = \left[\begin{array}{c}\text{prior odds}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right] \cdot \left[\begin{array}{c}\text{Bayes factor}\\ \text{in favor of}\\ M_2 \text{ over } M_1\end{array}\right]. \tag{15}$$

Specifying the **prior odds ratio** in applied settings seems to me to be a **more difficult problem** than acknowledged by such writers as Jeffreys (1939) — e.g., I see **nothing remotely "objective"** about taking this ratio to be **1**; in my view this should be approached with **sensitivity analysis**.

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

$$\begin{aligned} (\theta_j \mid M_j \, \mathcal{B}) & \sim p(\theta_j \mid M_j \, \mathcal{B}) \\ (y_i \mid M_j \, \theta_j \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} p(y_i \mid M_j \, \theta_j \, \mathcal{B}) \, , \end{aligned} \tag{16}$$

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

$$\begin{aligned} (\theta_j \mid M_j \, \mathcal{B}) &\sim p(\theta_j \mid M_j \, \mathcal{B}) \\ (y_i \mid M_j \, \theta_j \, \mathcal{B}) &\stackrel{\text{IID}}{\sim} p(y_i \mid M_j \, \theta_j \, \mathcal{B}) \, , \end{aligned} \tag{16}$$

in which $(i = 1, \ldots, n)$; $(j = 1, 2)$;

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

$$\begin{array}{rcl} (\theta_j \mid M_j \, \mathcal{B}) & \sim & p(\theta_j \mid M_j \, \mathcal{B}) \\ (y_i \mid M_j \, \theta_j \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & p(y_i \mid M_j \, \theta_j \, \mathcal{B}) \, , \end{array} \qquad (16)$$

in which $(i = 1, \ldots, n)$; $(j = 1, 2)$; the y_i are $(d \times 1)$ vectors of **outcome values** that live in $\mathbb{R}^d$ (often in what follows $d = 1$);

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

$$\begin{aligned} (\theta_j \,|\, M_j \, \mathcal{B}) & \sim & p(\theta_j \,|\, M_j \, \mathcal{B}) \\ (y_i \,|\, M_j \, \theta_j \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & p(y_i \,|\, M_j \, \theta_j \, \mathcal{B}) \, , \end{aligned} \quad (16)$$

in which $(i = 1, \ldots, n)$; $(j = 1, 2)$; the y_i are $(d \times 1)$ vectors of **outcome values** that live in $\mathbb{R}^d$ (often in what follows $d = 1$); and the **functional forms** of the prior $p(\theta_j \,|\, M_j \, \mathcal{B})$ and sampling distribution $p(y_i \,|\, M_j \, \theta_j \, \mathcal{B})$ are assumed **known**.

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

$$\begin{aligned} (\theta_j \mid M_j \, \mathcal{B}) & \sim & p(\theta_j \mid M_j \, \mathcal{B}) \\ (y_i \mid M_j \, \theta_j \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & p(y_i \mid M_j \, \theta_j \, \mathcal{B}) \, , \end{aligned} \tag{16}$$

in which $(i = 1, \ldots, n)$; $(j = 1, 2)$; the y_i are $(d \times 1)$ vectors of **outcome values** that live in $\mathbb{R}^d$ (often in what follows $d = 1$); and the **functional forms** of the prior $p(\theta_j \mid M_j \, \mathcal{B})$ and sampling distribution $p(y_i \mid M_j \, \theta_j \, \mathcal{B})$ are assumed **known**.

In this context, with $D = y \triangleq (y_1, \ldots, y_n)$, the **Bayes factor** in favor of M_2 over M_1 may be written

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

$$\begin{aligned} (\theta_j \mid M_j \, \mathcal{B}) & \sim & p(\theta_j \mid M_j \, \mathcal{B}) \\ (y_i \mid M_j \, \theta_j \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & p(y_i \mid M_j \, \theta_j \, \mathcal{B}) \, , \end{aligned} \tag{16}$$

in which $(i = 1, \ldots, n)$; $(j = 1, 2)$; the y_i are $(d \times 1)$ vectors of **outcome values** that live in $\mathbb{R}^d$ (often in what follows $d = 1$); and the **functional forms** of the prior $p(\theta_j \mid M_j \, \mathcal{B})$ and sampling distribution $p(y_i \mid M_j \, \theta_j \, \mathcal{B})$ are assumed **known**.

In this context, with $D = y \triangleq (y_1, \ldots, y_n)$, the **Bayes factor** in favor of M_2 over M_1 may be written

$$BF_{21} \triangleq BF \left[(M_2 \,||\, M_1) \mid y \, \mathcal{B} \right] \triangleq \left[\frac{IL(M_2 \mid y \, \mathcal{B})}{IL(M_1 \mid y \, \mathcal{B})} \right] . \tag{17}$$

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

$$\begin{aligned} (\theta_j \mid M_j \, \mathcal{B}) &\sim p(\theta_j \mid M_j \, \mathcal{B}) \\ (y_i \mid M_j \, \theta_j \, \mathcal{B}) &\stackrel{\text{IID}}{\sim} p(y_i \mid M_j \, \theta_j \, \mathcal{B}) \, , \end{aligned} \tag{16}$$

in which $(i = 1, \ldots, n)$; $(j = 1, 2)$; the y_i are $(d \times 1)$ vectors of **outcome values** that live in $\mathbb{R}^d$ (often in what follows $d = 1$); and the **functional forms** of the prior $p(\theta_j \mid M_j \, \mathcal{B})$ and sampling distribution $p(y_i \mid M_j \, \theta_j \, \mathcal{B})$ are assumed **known**.

In this context, with $D = y \triangleq (y_1, \ldots, y_n)$, the **Bayes factor** in favor of M_2 over M_1 may be written

$$BF_{21} \triangleq BF \left[(M_2 \,||\, M_1) \mid y \, \mathcal{B} \right] \triangleq \left[\frac{IL(M_2 \mid y \, \mathcal{B})}{IL(M_1 \mid y \, \mathcal{B})} \right] . \tag{17}$$

Here $IL(M_j \mid y \, \mathcal{B})$ is the **integrated likelihood** for model j:

Bayes Factors (continued)

For now let's focus only on the **Bayes factor** and concentrate on **parametric models** of the form

$$\begin{aligned} (\theta_j \,|\, M_j \, \mathcal{B}) &\sim p(\theta_j \,|\, M_j \, \mathcal{B}) \\ (y_i \,|\, M_j \, \theta_j \, \mathcal{B}) &\stackrel{\text{IID}}{\sim} p(y_i \,|\, M_j \, \theta_j \, \mathcal{B}) \, , \end{aligned} \qquad (16)$$

in which $(i = 1, \ldots, n)$; $(j = 1, 2)$; the y_i are $(d \times 1)$ vectors of **outcome values** that live in $\mathbb{R}^d$ (often in what follows $d = 1$); and the **functional forms** of the prior $p(\theta_j \,|\, M_j \, \mathcal{B})$ and sampling distribution $p(y_i \,|\, M_j \, \theta_j \, \mathcal{B})$ are assumed **known**.

In this context, with $D = y \triangleq (y_1, \ldots, y_n)$, the **Bayes factor** in favor of M_2 over M_1 may be written

$$BF_{21} \triangleq BF\left[(M_2 \,||\, M_1) \,|\, y \, \mathcal{B} \right] \triangleq \left[\frac{IL(M_2 \,|\, y \, \mathcal{B})}{IL(M_1 \,|\, y \, \mathcal{B})} \right] . \qquad (17)$$

Here $IL(M_j \,|\, y \, \mathcal{B})$ is the **integrated likelihood** for model j:

$$IL(M_j \,|\, y \, \mathcal{B}) = p(y \,|\, M_j \, \mathcal{B}) = \int_{\Theta_j} p(y \,|\, M_j \, \theta_j \, \mathcal{B}) \, p(\theta_j \,|\, M_j \, \mathcal{B}) \, d\theta_j \quad (18)$$

Bayes Factors (continued)

$$
\begin{aligned}
IL(M_j \mid y\, \mathcal{B}) &= p(y \mid M_j\, \mathcal{B}) = \int_{\Theta_j} p(y \mid M_j\, \theta_j\, \mathcal{B})\, p(\theta_j \mid M_j\, \mathcal{B})\, d\theta_j \\
&= \int_{\Theta_j} \ell(\theta_j \mid M_j\, y\, \mathcal{B})\, p(\theta_j \mid M_j\, \mathcal{B})\, d\theta_j\,, \qquad (19)
\end{aligned}
$$

Bayes Factors (continued)

$$\begin{aligned} \ell L(M_j \mid y\,\mathcal{B}) &= p(y \mid M_j\,\mathcal{B}) = \int_{\Theta_j} p(y \mid M_j\,\theta_j\,\mathcal{B})\, p(\theta_j \mid M_j\,\mathcal{B})\, d\theta_j \\ &= \int_{\Theta_j} \ell(\theta_j \mid M_j\,y\,\mathcal{B})\, p(\theta_j \mid M_j\,\mathcal{B})\, d\theta_j\,, \qquad (19) \end{aligned}$$

in which Θ_j is the **parameter space** for model j, of dimension k_j (in all of my examples $\Theta_j = \mathbb{R}^{k_j}$),

Bayes Factors (continued)

$$\begin{aligned} IL(M_j \mid y \, \mathcal{B}) &= p(y \mid M_j \, \mathcal{B}) = \int_{\Theta_j} p(y \mid M_j \, \theta_j \, \mathcal{B}) \, p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \\ &= \int_{\Theta_j} \ell(\theta_j \mid M_j \, y \, \mathcal{B}) \, p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, , \end{aligned} \tag{19}$$

in which Θ_j is the **parameter space** for model j, of dimension k_j (in all of my examples $\Theta_j = \mathbb{R}^{k_j}$), and in which

$$\ell(\theta_j \mid M_j \, y \, \mathcal{B}) = \prod_{i=1}^{n} p(y_i \mid M_j \, \theta_j \, \mathcal{B}) \tag{20}$$

is the **likelihood function** for model M_j.

Bayes Factors (continued)

$$
\begin{aligned}
IL(M_j \mid y\,\mathcal{B}) &= p(y \mid M_j\,\mathcal{B}) = \int_{\Theta_j} p(y \mid M_j\,\theta_j\,\mathcal{B})\, p(\theta_j \mid M_j\,\mathcal{B})\, d\theta_j \\
&= \int_{\Theta_j} \ell(\theta_j \mid M_j\,y\,\mathcal{B})\, p(\theta_j \mid M_j\,\mathcal{B})\, d\theta_j\,, \qquad (19)
\end{aligned}
$$

in which Θ_j is the **parameter space** for model j, of dimension k_j (in all of my examples $\Theta_j = \mathbb{R}^{k_j}$), and in which

$$
\ell(\theta_j \mid M_j\,y\,\mathcal{B}) = \prod_{i=1}^{n} p(y_i \mid M_j\,\theta_j\,\mathcal{B}) \qquad (20)
$$

is the **likelihood function** for model M_j.

I suppose here that $k_1 < k_2$, so that M_1 is the **simpler** of the two models.

An interesting **approximate special case** of Bayes factors was developed by Schwarz (1978), who

Bayes Factors (continued)

$$
\begin{aligned}
IL(M_j \mid y\,\mathcal{B}) &= p(y \mid M_j\,\mathcal{B}) = \int_{\Theta_j} p(y \mid M_j\,\theta_j\,\mathcal{B})\, p(\theta_j \mid M_j\,\mathcal{B})\, d\theta_j \\
&= \int_{\Theta_j} \ell(\theta_j \mid M_j\, y\,\mathcal{B})\, p(\theta_j \mid M_j\,\mathcal{B})\, d\theta_j \,, \qquad (19)
\end{aligned}
$$

in which Θ_j is the **parameter space** for model j, of dimension k_j (in all of my examples $\Theta_j = \mathbb{R}^{k_j}$), and in which

$$
\ell(\theta_j \mid M_j\, y\,\mathcal{B}) = \prod_{i=1}^{n} p(y_i \mid M_j\,\theta_j\,\mathcal{B}) \qquad (20)
$$

is the **likelihood function** for model M_j.

I suppose here that $k_1 < k_2$, so that M_1 is the **simpler** of the two models.

An interesting **approximate special case** of Bayes factors was developed by Schwarz (1978), who — in the context of parametric models belonging to the class of **regular exponential families** —

Bayes Factors (continued)

$$
\begin{aligned}
IL(M_j \mid y\, \mathcal{B}) &= p(y \mid M_j\, \mathcal{B}) = \int_{\Theta_j} p(y \mid M_j\, \theta_j\, \mathcal{B})\, p(\theta_j \mid M_j\, \mathcal{B})\, d\theta_j \\
&= \int_{\Theta_j} \ell(\theta_j \mid M_j\, y\, \mathcal{B})\, p(\theta_j \mid M_j\, \mathcal{B})\, d\theta_j \, , \qquad (19)
\end{aligned}
$$

in which Θ_j is the **parameter space** for model j, of dimension k_j (in all of my examples $\Theta_j = \mathbb{R}^{k_j}$), and in which

$$
\ell(\theta_j \mid M_j\, y\, \mathcal{B}) = \prod_{i=1}^{n} p(y_i \mid M_j\, \theta_j\, \mathcal{B}) \qquad (20)
$$

is the **likelihood function** for model M_j.

I suppose here that $k_1 < k_2$, so that M_1 is the **simpler** of the two models.

An interesting **approximate special case** of Bayes factors was developed by Schwarz (1978), who — in the context of parametric models belonging to the class of **regular exponential families** — developed an $O_p(1)$ Taylor series approximation to $\log\left[IL(M_j \mid y\, \mathcal{B})\right]$, namely

Bayes Factors (continued)

$$IL(M_j \mid y\,\mathcal{B}) = p(y \mid M_j\,\mathcal{B}) = \int_{\Theta_j} p(y \mid M_j\,\theta_j\,\mathcal{B})\, p(\theta_j \mid M_j\,\mathcal{B})\, d\theta_j$$
$$= \int_{\Theta_j} \ell(\theta_j \mid M_j\,y\,\mathcal{B})\, p(\theta_j \mid M_j\,\mathcal{B})\, d\theta_j\,, \qquad (19)$$

in which Θ_j is the **parameter space** for model j, of dimension k_j (in all of my examples $\Theta_j = \mathbb{R}^{k_j}$), and in which

$$\ell(\theta_j \mid M_j\,y\,\mathcal{B}) = \prod_{i=1}^{n} p(y_i \mid M_j\,\theta_j\,\mathcal{B}) \qquad (20)$$

is the **likelihood function** for model M_j.

I suppose here that $k_1 < k_2$, so that M_1 is the **simpler** of the two models.

An interesting **approximate special case** of Bayes factors was developed by Schwarz (1978), who — in the context of parametric models belonging to the class of **regular exponential families** — developed an $O_p(1)$ Taylor series approximation to $\log\left[IL(M_j \mid y\,\mathcal{B})\right]$, namely

$$\log\left[IL(M_j \mid y\,\mathcal{B})\right] = \log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] - \frac{k_j}{2}\log(n) + O_p(1)\,; \qquad (21)$$

BIC

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

BIC

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log\left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})\right] - \frac{k_j}{2}\log(n)$; this is equivalent to minimizing

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log\left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})\right] - \frac{k_j}{2}\log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y \, \mathcal{B}) \triangleq -2 \log\left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})\right] + k_j \log(n)\,, \qquad (22)$$

in which BIC is the **Bayesian information criterion**

BIC

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] - \frac{k_j}{2}\log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y\, \mathcal{B}) \triangleq -2 \log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] + k_j \log(n)\,, \qquad (22)$$

in which BIC is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of BIC or multiplication by -2;

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] - \frac{k_j}{2}\log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y\, \mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] + k_j\,\log(n)\,, \qquad (22)$$

in which BIC is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of BIC or multiplication by -2; this **rescaling**, which was intended to put the log-likelihood contribution to BIC on the **deviance scale**,

BIC

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log \left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B}) \right] - \frac{k_j}{2} \log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y \, \mathcal{B}) \triangleq -2 \log \left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B}) \right] + k_j \log(n) \,, \qquad (22)$$

in which BIC is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of BIC or multiplication by -2; this **rescaling**, which was intended to put the log-likelihood contribution to BIC on the **deviance scale**, was first suggested by Akaike (1980), who does not cite Schwarz).

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] - \frac{k_j}{2}\log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y\, \mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] + k_j\,\log(n)\,, \tag{22}$$

in which BIC is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of BIC or multiplication by -2; this **rescaling**, which was intended to put the log-likelihood contribution to BIC on the **deviance scale**, was first suggested by Akaike (1980), who does not cite Schwarz).

The attractive feature of BIC is that it neatly **decomposes model comparison**

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log \left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B}) \right] - \frac{k_j}{2} \log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y \, \mathcal{B}) \triangleq -2 \log \left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B}) \right] + k_j \log(n) \, , \qquad (22)$$

in which BIC is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of BIC or multiplication by -2; this **rescaling**, which was intended to put the log-likelihood contribution to BIC on the **deviance scale**, was first suggested by Akaike (1980), who does not cite Schwarz).

The attractive feature of BIC is that it neatly **decomposes model comparison** into an **additive balance** between **model fit** (the log-likelihood term)

BIC

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log\left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})\right] - \frac{k_j}{2}\log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y \, \mathcal{B}) \triangleq -2 \, \log\left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})\right] + k_j \, \log(n) \,, \qquad (22)$$

in which *BIC* is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of *BIC* or multiplication by -2; this **rescaling**, which was intended to put the log-likelihood contribution to *BIC* on the **deviance scale**, was first suggested by Akaike (1980), who does not cite Schwarz).

The attractive feature of *BIC* is that it neatly **decomposes model comparison** into an **additive balance** between **model fit** (the log-likelihood term) and **parsimony** (the $k_j \, \log(n)$ term).

BIC

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] - \frac{k_j}{2}\log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y\, \mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] + k_j\,\log(n)\,, \qquad (22)$$

in which BIC is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of BIC or multiplication by -2; this **rescaling**, which was intended to put the log-likelihood contribution to BIC on the **deviance scale**, was first suggested by Akaike (1980), who does not cite Schwarz).

The attractive feature of BIC is that it neatly **decomposes model comparison** into an **additive balance** between **model fit** (the log-likelihood term) and **parsimony** (the $k_j\,\log(n)$ term).

Two centuries earlier,

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log \left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B}) \right] - \frac{k_j}{2} \log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y \, \mathcal{B}) \triangleq -2 \log \left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B}) \right] + k_j \log(n) \,, \qquad (22)$$

in which BIC is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of BIC or multiplication by -2; this **rescaling**, which was intended to put the log-likelihood contribution to BIC on the **deviance scale**, was first suggested by Akaike (1980), who does not cite Schwarz).

The attractive feature of BIC is that it neatly **decomposes model comparison** into an **additive balance** between **model fit** (the log-likelihood term) and **parsimony** (the $k_j \log(n)$ term).

Two centuries earlier, **Laplace** (1774) developed a **more accurate** $O_p\left(\frac{1}{n}\right)$ **approximation** to the log integrated likelihood,

here $\hat{\theta}_j$ is the **maximum-likelihood estimate** (MLE) of θ_j under model M_j, assumed to exist and to be unique.

Schwarz advocated a preference for the model in $\mathcal{M}$ that **maximizes** $\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] - \frac{k_j}{2}\log(n)$; this is equivalent to minimizing

$$BIC(M_j \mid y\, \mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})\right] + k_j\,\log(n)\,, \qquad (22)$$

in which BIC is the **Bayesian information criterion** (interestingly, Schwarz (1978) makes no mention of BIC or multiplication by -2; this **rescaling**, which was intended to put the log-likelihood contribution to BIC on the **deviance scale**, was first suggested by Akaike (1980), who does not cite Schwarz).

The attractive feature of BIC is that it neatly **decomposes model comparison** into an **additive balance** between **model fit** (the log-likelihood term) and **parsimony** (the $k_j\,\log(n)$ term).

Two centuries earlier, **Laplace** (1774) developed a **more accurate** $O_p\left(\frac{1}{n}\right)$ **approximation** to the log integrated likelihood, of which Schwarz was apparently unaware:

BIC and Laplace

$$\log\left[IL(M_j \mid y\,\mathcal{B})\right] = \log\left[\ell(\hat{\theta}_j \mid M_j\, y\,\mathcal{B})\right] + \log\left[p(\hat{\theta}_j \mid M_j\,\mathcal{B})\right] + \frac{k_j}{2}\log(2\,\pi) - \frac{1}{2}\log|\hat{I}_j| + O_p\left(\frac{1}{n}\right), \quad (23)$$

$$\begin{aligned}\log\left[IL(M_j \mid y\,\mathcal{B})\right] &= \log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] + \log\left[p(\hat{\theta}_j \mid M_j\,\mathcal{B})\right] \\ &\quad + \frac{k_j}{2}\log(2\,\pi) - \frac{1}{2}\log|\hat{I}_j| + O_p\left(\frac{1}{n}\right)\,, \quad (23)\end{aligned}$$

in which $\log|\hat{I}_j|$ is the determinant of the **observed information matrix** for model M_j.

BIC and Laplace

$$\log\left[IL(M_j \mid y\,\mathcal{B})\right] = \log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] + \log\left[p(\hat{\theta}_j \mid M_j\,\mathcal{B})\right] + \frac{k_j}{2}\log(2\,\pi) - \frac{1}{2}\log|\hat{I}_j| + O_p\left(\frac{1}{n}\right), \quad (23)$$

in which $\log|\hat{I}_j|$ is the determinant of the **observed information matrix** for model M_j.

A comparison of expressions (21) and (23) immediately begs the following **question**:

BIC and Laplace

$$\log\left[IL(M_j \mid y\,\mathcal{B})\right] = \log\left[\ell(\hat{\theta}_j \mid M_j\, y\,\mathcal{B})\right] + \log\left[p(\hat{\theta}_j \mid M_j\,\mathcal{B})\right] + \frac{k_j}{2}\log(2\,\pi) - \frac{1}{2}\log|\hat{I}_j| + O_p\left(\frac{1}{n}\right), \quad (23)$$

in which $\log|\hat{I}_j|$ is the determinant of the **observed information matrix** for model M_j.

A comparison of expressions (21) and (23) immediately begs the following **question**: is there a **prior distribution** $p(\theta_j \mid M_j\,\mathcal{B})$ for which the approximations of Laplace and Schwarz **coincide**?

BIC and Laplace

$$\log\left[IL(M_j \mid y\,\mathcal{B})\right] = \log\left[\ell(\hat{\theta}_j \mid M_j\, y\,\mathcal{B})\right] + \log\left[p(\hat{\theta}_j \mid M_j\,\mathcal{B})\right] + \frac{k_j}{2}\log(2\,\pi) - \frac{1}{2}\log|\hat{I}_j| + O_p\left(\frac{1}{n}\right), \quad (23)$$

in which $\log|\hat{I}_j|$ is the determinant of the **observed information matrix** for model M_j.

A comparison of expressions (21) and (23) immediately begs the following **question**: is there a **prior distribution** $p(\theta_j \mid M_j\,\mathcal{B})$ for which the approximations of Laplace and Schwarz **coincide**?

Suppose that all of the components of θ_j have been transformed to live on $\mathbb{R}$, so that it becomes reasonable to try a **multivariate normal prior**;

BIC and Laplace

$$\log \left[IL(M_j \mid y \, \mathcal{B}) \right] = \log \left[\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B}) \right] + \log \left[p(\hat{\theta}_j \mid M_j \, \mathcal{B}) \right] + \frac{k_j}{2} \log(2 \, \pi) - \frac{1}{2} \log |\hat{I}_j| + O_p \left(\frac{1}{n} \right) \, , \quad (23)$$

in which $\log |\hat{I}_j|$ is the determinant of the **observed information matrix** for model M_j.

A comparison of expressions (21) and (23) immediately begs the following **question**: is there a **prior distribution** $p(\theta_j \mid M_j \, \mathcal{B})$ for which the approximations of Laplace and Schwarz **coincide**?

Suppose that all of the components of θ_j have been transformed to live on $\mathbb{R}$, so that it becomes reasonable to try a **multivariate normal prior**; the result that succeeds in making Laplace and Schwarz agree is

$$\begin{aligned} \log\left[IL(M_j \mid y\,\mathcal{B})\right] &= \log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] + \log\left[p(\hat{\theta}_j \mid M_j\,\mathcal{B})\right] \\ &\quad + \frac{k_j}{2}\log(2\,\pi) - \frac{1}{2}\log|\hat{I}_j| + O_p\left(\frac{1}{n}\right)\,, \quad (23) \end{aligned}$$

in which $\log|\hat{I}_j|$ is the determinant of the **observed information matrix** for model M_j.

A comparison of expressions (21) and (23) immediately begs the following **question**: is there a **prior distribution** $p(\theta_j \mid M_j\,\mathcal{B})$ for which the approximations of Laplace and Schwarz **coincide**?

Suppose that all of the components of θ_j have been transformed to live on $\mathbb{R}$, so that it becomes reasonable to try a **multivariate normal prior**; the result that succeeds in making Laplace and Schwarz agree is

$$(\theta_j \mid M_j\,\mathcal{B}) \sim N_{k_j}\left(\hat{\theta}_j, n\,\hat{I}_j^{-1}\right)\,. \quad (24)$$

BIC and Laplace

$$\log\left[IL(M_j \mid y\,\mathcal{B})\right] = \log\left[\ell(\hat{\theta}_j \mid M_j\, y\,\mathcal{B})\right] + \log\left[p(\hat{\theta}_j \mid M_j\,\mathcal{B})\right] + \frac{k_j}{2}\log(2\,\pi) - \frac{1}{2}\log|\hat{I}_j| + O_p\left(\frac{1}{n}\right), \quad (23)$$

in which $\log|\hat{I}_j|$ is the determinant of the **observed information matrix** for model M_j.

A comparison of expressions (21) and (23) immediately begs the following **question**: is there a **prior distribution** $p(\theta_j \mid M_j\,\mathcal{B})$ for which the approximations of Laplace and Schwarz **coincide**?

Suppose that all of the components of θ_j have been transformed to live on $\mathbb{R}$, so that it becomes reasonable to try a **multivariate normal prior**; the result that succeeds in making Laplace and Schwarz agree is

$$(\theta_j \mid M_j\,\mathcal{B}) \sim N_{k_j}\left(\hat{\theta}_j, n\,\hat{I}_j^{-1}\right). \quad (24)$$

This has been referred to as a **unit-information** prior (Kass and Wasserman(1995)),

BIC and Laplace

$$\log\left[IL(M_j \mid y\,\mathcal{B})\right] = \log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] + \log\left[p(\hat{\theta}_j \mid M_j\,\mathcal{B})\right] + \frac{k_j}{2}\log(2\,\pi) - \frac{1}{2}\log|\hat{I}_j| + O_p\left(\frac{1}{n}\right), \quad (23)$$

in which $\log|\hat{I}_j|$ is the determinant of the **observed information matrix** for model M_j.

A comparison of expressions (21) and (23) immediately begs the following **question**: is there a **prior distribution** $p(\theta_j \mid M_j\,\mathcal{B})$ for which the approximations of Laplace and Schwarz **coincide**?

Suppose that all of the components of θ_j have been transformed to live on $\mathbb{R}$, so that it becomes reasonable to try a **multivariate normal prior**; the result that succeeds in making Laplace and Schwarz agree is

$$(\theta_j \mid M_j\,\mathcal{B}) \sim N_{k_j}\left(\hat{\theta}_j, n\,\hat{I}_j^{-1}\right). \quad (24)$$

This has been referred to as a **unit-information** prior (Kass and Wasserman(1995)), because it adds information to the posterior for θ_j equivalent to **1 observation** that's consistent with a **maximum-likelihood analysis**.

Enter Ed Jaynes

(This prior is **gently data-determined**,

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior

—

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, (This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n + 1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

(a) it's implicitly based on a **reasonable diffuse prior**, and

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

(a) it's implicitly based on a **reasonable diffuse prior**, and

(b) it explicitly **trades off model fit against model complexity**.

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

(a) it's implicitly based on a **reasonable diffuse prior**, and

(b) it explicitly **trades off model fit against model complexity**.

All of this so far is **routine**,

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

(a) it's implicitly based on a **reasonable diffuse prior**, and

(b) it explicitly **trades off model fit against model complexity**.

All of this so far is **routine**, but at the point in the story summarized by equation (18), **Jaynes** (2003) did something interesting:

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n + 1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

(a) it's implicitly based on a **reasonable diffuse prior**, and

(b) it explicitly **trades off model fit against model complexity**.

All of this so far is **routine**, but at the point in the story summarized by equation (18), **Jaynes** (2003) did something interesting: assuming (as above) that the MLE $\hat{\theta}_j$ for θ_j exists and is unique, (This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

(a) it's implicitly based on a **reasonable diffuse prior**, and

(b) it explicitly **trades off model fit against model complexity**.

All of this so far is **routine**, but at the point in the story summarized by equation (18), **Jaynes** (2003) did something interesting: assuming (as above) that the MLE $\hat{\theta}_j$ for θ_j exists and is unique, and that the maximum value $\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})$ attained by the likelihood function for model M_j is **strictly positive**,

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

(a) it's implicitly based on a **reasonable diffuse prior**, and

(b) it explicitly **trades off model fit against model complexity**.

All of this so far is **routine**, but at the point in the story summarized by equation (18), **Jaynes** (2003) did something interesting: assuming (as above) that the MLE $\hat{\theta}_j$ for θ_j exists and is unique, and that the maximum value $\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})$ attained by the likelihood function for model M_j is **strictly positive**, Jaynes can write

Enter Ed Jaynes

(This prior is **gently data-determined**, but — with 1 prior "observation" and n data observations in the resulting $(n+1)$–"observation" posterior — the data dependence in the prior is clearly **minimal**, even for modest n).

BIC thus has **two salient properties**:

(a) it's implicitly based on a **reasonable diffuse prior**, and

(b) it explicitly **trades off model fit against model complexity**.

All of this so far is **routine**, but at the point in the story summarized by equation (18), **Jaynes** (2003) did something interesting: assuming (as above) that the MLE $\hat{\theta}_j$ for θ_j exists and is unique, and that the maximum value $\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})$ attained by the likelihood function for model M_j is **strictly positive**, Jaynes can write

$$IL(M_j \mid y \, \mathcal{B}) = \ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B}) \left\{ \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \right\} .$$

Jaynes therefore defines

The Ockham Factor

Jaynes therefore defines

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, , \tag{25}$$

The Ockham Factor

Jaynes therefore defines

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \,, \tag{25}$$

and — although Jaynes doesn't use this name —

The Ockham Factor

Jaynes therefore defines

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \,, \qquad (25)$$

and — although Jaynes doesn't use this name — the **Bayes factor** in favor of M_2 over M_1 becomes

Jaynes therefore defines

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, , \qquad (25)$$

and — although Jaynes doesn't use this name — the **Bayes factor** in favor of M_2 over M_1 becomes

$$\begin{array}{ccccc} BF\left[(M_2 \,||\, M_1) \mid y \, \mathcal{B} \right] & = & \left[\frac{\ell(\hat{\theta}_2 \mid M_2 \, y \, \mathcal{B})}{\ell(\hat{\theta}_1 \mid M_1 \, y \, \mathcal{B})} \right] & \cdot & \left[\frac{W_2}{W_1} \right] \\ \left[\begin{array}{c} \text{Bayes factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] & = & \left[\begin{array}{c} \text{likelihood ratio} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] & \cdot & \left[\begin{array}{c} \text{Ockham factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] \\ BF_{21} & = & LR_{21} & \cdot & OF_{21} \, . \end{array} \qquad (26)$$

The Ockham Factor

Jaynes therefore defines

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \,, \tag{25}$$

and — although Jaynes doesn't use this name — the **Bayes factor** in favor of M_2 over M_1 becomes

$$\begin{array}{ccccc} BF\left[(M_2 \,||\, M_1) \mid y \, \mathcal{B}\right] & = & \left[\frac{\ell(\hat{\theta}_2 \mid M_2 \, y \, \mathcal{B})}{\ell(\hat{\theta}_1 \mid M_1 \, y \, \mathcal{B})}\right] & \cdot & \left[\frac{W_2}{W_1}\right] \\ \left[\begin{array}{c} \text{Bayes factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array}\right] & = & \left[\begin{array}{c} \text{likelihood ratio} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array}\right] & \cdot & \left[\begin{array}{c} \text{Ockham factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array}\right] \\ BF_{21} & = & LR_{21} & \cdot & OF_{21} \,. \end{array} \tag{26}$$

In this manner Jaynes has **decomposed** the Bayes factor BF_{21} into the product of two quantities that play **completely different roles** in its calculation:

The Ockham Factor

Jaynes therefore defines

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j\, y\, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B})} \right] p(\theta_j \mid M_j\, \mathcal{B})\, d\theta_j \,, \qquad (25)$$

and — although Jaynes doesn't use this name — the **Bayes factor** in favor of M_2 over M_1 becomes

$$\begin{array}{ccccc} BF\,[(M_2 \,||\, M_1) \mid y\, \mathcal{B}] & = & \left[\frac{\ell(\hat{\theta}_2 \mid M_2\, y\, \mathcal{B})}{\ell(\hat{\theta}_1 \mid M_1\, y\, \mathcal{B})} \right] & \cdot & \left[\frac{W_2}{W_1} \right] \\ \left[\begin{array}{c} \text{Bayes factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] & = & \left[\begin{array}{c} \text{likelihood ratio} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] & \cdot & \left[\begin{array}{c} \text{Ockham factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] \\ BF_{21} & = & LR_{21} & \cdot & OF_{21} \,. \end{array} \qquad (26)$$

In this manner Jaynes has **decomposed** the Bayes factor BF_{21} into the product of two quantities that play **completely different roles** in its calculation: the **likelihood ratio** LR_{21} in favor of M_2 over M_1,

The Ockham Factor

Jaynes therefore defines

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, , \tag{25}$$

and — although Jaynes doesn't use this name — the **Bayes factor** in favor of M_2 over M_1 becomes

$$\begin{array}{ccccc} BF\left[(M_2 \,||\, M_1) \mid y \, \mathcal{B}\right] & = & \left[\frac{\ell(\hat{\theta}_2 \mid M_2 \, y \, \mathcal{B})}{\ell(\hat{\theta}_1 \mid M_1 \, y \, \mathcal{B})} \right] & \cdot & \left[\frac{W_2}{W_1} \right] \\ \left[\begin{array}{c} \text{Bayes factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] & = & \left[\begin{array}{c} \text{likelihood ratio} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] & \cdot & \left[\begin{array}{c} \text{Ockham factor} \\ \text{in favor of} \\ M_2 \text{ over } M_1 \end{array} \right] \\ BF_{21} & = & LR_{21} & \cdot & OF_{21} \, . \end{array} \tag{26}$$

In this manner Jaynes has **decomposed** the Bayes factor BF_{21} into the product of two quantities that play **completely different roles** in its calculation: the **likelihood ratio** LR_{21} in favor of M_2 over M_1, and what Jaynes referred to as the **Ockham factor** OF_{21} in favor of M_2 over M_1.

The Jaynes Information Criterion (JIC)

The name **Ockham factor** is an allusion to Ockham's Razor,

The Jaynes Information Criterion (JIC)

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**,

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new);

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new); I prefer to motivate OF_{21} with a **simple Gaussian example** (below).

The Jaynes Information Criterion (JIC)

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new); I prefer to motivate OF_{21} with a **simple Gaussian example** (below).

First, however, to **faciliate comparison** with BIC,

The Jaynes Information Criterion (JIC)

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new); I prefer to motivate OF_{21} with a **simple Gaussian example** (below).

First, however, to **faciliate comparison** with BIC, let's transform BF_{21} affinely to the **log scale**:

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new); I prefer to motivate OF_{21} with a **simple Gaussian example** (below).

First, however, to **faciliate comparison** with BIC, let's transform BF_{21} affinely to the **log scale**:

$$\begin{array}{ccccc} -2\,\log(BF_{21}) & = & -2\,\log(LR_{21}) & + & [-2\,\log(OF_{21})] \\ \left\{ \begin{array}{c} -2\,IL(M_2 \mid y\,\mathcal{B}) \\ -\,[-2\,IL(M_1 \mid y\,\mathcal{B})] \end{array} \right\} & = & \left\{ \begin{array}{c} -2\,\log \ell(\hat{\theta}_2 \mid M_2\,y\,\mathcal{B}) \\ -\left[-2\,\log \ell(\hat{\theta}_1 \mid M_1\,y\,\mathcal{B})\right] \end{array} \right\} & + & \left\{ \begin{array}{c} -2\,\log(W_2) \\ -\,[-2\,\log(W_1)] \end{array} \right\} . \end{array} \tag{27}$$

The Jaynes Information Criterion (JIC)

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new); I prefer to motivate OF_{21} with a **simple Gaussian example** (below).

First, however, to **faciliate comparison** with BIC, let's transform BF_{21} affinely to the **log scale**:

$$\begin{array}{ccccc} -2\,\log(BF_{21}) & = & -2\,\log(LR_{21}) & + & [-2\,\log(OF_{21})] \\ \left\{ \begin{array}{c} -2\,IL(M_2 \mid y\,\mathcal{B}) \\ -\left[-2\,IL(M_1 \mid y\,\mathcal{B})\right] \end{array} \right\} & = & \left\{ \begin{array}{c} -2\,\log \ell(\hat{\theta}_2 \mid M_2\,y\,\mathcal{B}) \\ -\left[-2\,\log \ell(\hat{\theta}_1 \mid M_1\,y\,\mathcal{B})\right] \end{array} \right\} & + & \left\{ \begin{array}{c} -2\,\log(W_2) \\ -\left[-2\,\log(W_1)\right] \end{array} \right\} . \end{array} \tag{27}$$

Therefore I define, in Jaynes's honor, the **Jaynes Information Criterion**

The Jaynes Information Criterion (JIC)

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new); I prefer to motivate OF_{21} with a **simple Gaussian example** (below).

First, however, to **faciliate comparison** with BIC, let's transform BF_{21} affinely to the **log scale**:

$$\begin{array}{ccccc} -2\,\log(BF_{21}) & = & -2\,\log(LR_{21}) & + & [-2\,\log(OF_{21})] \\ \left\{ \begin{array}{c} -2\,IL(M_2 \mid y\,\mathcal{B}) \\ -\left[-2\,IL(M_1 \mid y\,\mathcal{B})\right] \end{array} \right\} & = & \left\{ \begin{array}{c} -2\,\log \ell(\hat{\theta}_2 \mid M_2\,y\,\mathcal{B}) \\ -\left[-2\,\log \ell(\hat{\theta}_1 \mid M_1\,y\,\mathcal{B})\right] \end{array} \right\} & + & \left\{ \begin{array}{c} -2\,\log(W_2) \\ -\left[-2\,\log(W_1)\right] \end{array} \right\} . \end{array} \tag{27}$$

Therefore I define, in Jaynes's honor, the **Jaynes Information Criterion**

$$JIC(M_j \mid D\,\mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] - 2\,\log(W_j)\,; \tag{28}$$

The Jaynes Information Criterion (JIC)

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new); I prefer to motivate OF_{21} with a **simple Gaussian example** (below).

First, however, to **faciliate comparison** with BIC, let's transform BF_{21} affinely to the **log scale**:

$$\begin{array}{ccccc} -2\,\log(BF_{21}) & = & -2\,\log(LR_{21}) & + & [-2\,\log(OF_{21})] \\ \left\{ \begin{array}{c} -2\,IL(M_2 \mid y\,\mathcal{B}) \\ -[-2\,IL(M_1 \mid y\,\mathcal{B})] \end{array} \right\} & = & \left\{ \begin{array}{c} -2\,\log \ell(\hat{\theta}_2 \mid M_2\,y\,\mathcal{B}) \\ -\left[-2\,\log \ell(\hat{\theta}_1 \mid M_1\,y\,\mathcal{B})\right] \end{array} \right\} & + & \left\{ \begin{array}{c} -2\,\log(W_2) \\ -[-2\,\log(W_1)] \end{array} \right\} . \end{array} \tag{27}$$

Therefore I define, in Jaynes's honor, the **Jaynes Information Criterion**

$$JIC(M_j \mid D\,\mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] - 2\,\log(W_j)\,; \tag{28}$$

this implies that, in the data-driven part of equation (15),

The Jaynes Information Criterion (JIC)

The name **Ockham factor** is an allusion to Ockham's Razor, and suggests that this term in the product in (26) has something to do with **parsimony**.

I don't find Jaynes's motivation for OF_{21} **compelling**, and in fact at this point I **part company** with him (everything below is new); I prefer to motivate OF_{21} with a **simple Gaussian example** (below).

First, however, to **faciliate comparison** with BIC, let's transform BF_{21} affinely to the **log scale**:

$$\begin{array}{ccccc} -2\,\log(BF_{21}) & = & -2\,\log(LR_{21}) & + & [-2\,\log(OF_{21})] \\ \left\{ \begin{array}{c} -2\,IL(M_2 \mid y\,\mathcal{B}) \\ -\,[-2\,IL(M_1 \mid y\,\mathcal{B})] \end{array} \right\} & = & \left\{ \begin{array}{c} -2\,\log \ell(\hat{\theta}_2 \mid M_2\,y\,\mathcal{B}) \\ -\left[-2\,\log \ell(\hat{\theta}_1 \mid M_1\,y\,\mathcal{B})\right] \end{array} \right\} & + & \left\{ \begin{array}{c} -2\,\log(W_2) \\ -\,[-2\,\log(W_1)] \end{array} \right\} . \end{array} \tag{27}$$

Therefore I define, in Jaynes's honor, the **Jaynes Information Criterion**

$$JIC(M_j \mid D\,\mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] - 2\,\log(W_j)\,; \tag{28}$$

this implies that, in the data-driven part of equation (15), models with **lower** JIC **values** are to be preferred.

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition,

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**,

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**:

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

$$(y_i \mid M_1 \, \mathcal{B}) \stackrel{\text{IID}}{\sim} N\left(\theta_1, \sigma^2\right) \,, \tag{29}$$

in which the standard deviation (SD) $\sigma > 0$ is assumed **known**

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

$$(y_i \mid M_1 \, \mathcal{B}) \stackrel{\text{IID}}{\sim} N\left(\theta_1, \sigma^2\right) \,, \tag{29}$$

in which the standard deviation (SD) $\sigma > 0$ is assumed **known** and where θ_1 is a **known structural singleton** arising from a scientific theory.

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

$$(y_i \mid M_1 \, \mathcal{B}) \stackrel{\text{IID}}{\sim} N\left(\theta_1, \sigma^2\right) \, , \tag{29}$$

in which the standard deviation (SD) $\sigma > 0$ is assumed **known** and where θ_1 is a **known structural singleton** arising from a scientific theory.

M_2 has the same form but with **unknown mean** θ (here, and throughout, I use **conjugate priors** when they exist):

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

$$(y_i \mid M_1 \, \mathcal{B}) \stackrel{\text{IID}}{\sim} N\left(\theta_1, \sigma^2\right) \, , \tag{29}$$

in which the standard deviation (SD) $\sigma > 0$ is assumed **known** and where θ_1 is a **known structural singleton** arising from a scientific theory.

M_2 has the same form but with **unknown mean** θ (here, and throughout, I use **conjugate priors** when they exist):

$$\begin{aligned} (\theta \mid M_2 \, \mathcal{B}) & \sim & N(\theta_0, \sigma_0^2) \\ (y_i \mid M_2 \, \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & N\left(\theta, \sigma^2\right) \, , \end{aligned} \tag{30}$$

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

$$(y_i \mid M_1 \, \mathcal{B}) \stackrel{\text{IID}}{\sim} N\left(\theta_1, \sigma^2\right) , \tag{29}$$

in which the standard deviation (SD) $\sigma > 0$ is assumed **known** and where θ_1 is a **known structural singleton** arising from a scientific theory.

M_2 has the same form but with **unknown mean** θ (here, and throughout, I use **conjugate priors** when they exist):

$$\begin{aligned} (\theta \mid M_2 \, \mathcal{B}) & \sim N(\theta_0, \sigma_0^2) \\ (y_i \mid M_2 \, \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} N\left(\theta, \sigma^2\right) , \end{aligned} \tag{30}$$

with known $(\theta_0, \sigma_0, \sigma)$;

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

$$(y_i \mid M_1 \, \mathcal{B}) \stackrel{\text{IID}}{\sim} N\left(\theta_1, \sigma^2\right) \, , \tag{29}$$

in which the standard deviation (SD) $\sigma > 0$ is assumed **known** and where θ_1 is a **known structural singleton** arising from a scientific theory.

M_2 has the same form but with **unknown mean** θ (here, and throughout, I use **conjugate priors** when they exist):

$$\begin{aligned} (\theta \mid M_2 \, \mathcal{B}) & \sim N(\theta_0, \sigma_0^2) \\ (y_i \mid M_2 \, \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} N\left(\theta, \sigma^2\right) \, , \end{aligned} \tag{30}$$

with known $(\theta_0, \sigma_0, \sigma)$; note that the **dimensions** of Θ_1 and Θ_2 in this setup are 0 and 1, respectively.

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

$$(y_i \mid M_1 \, \mathcal{B}) \stackrel{\text{IID}}{\sim} N\left(\theta_1, \sigma^2\right) \, , \tag{29}$$

in which the standard deviation (SD) $\sigma > 0$ is assumed **known** and where θ_1 is a **known structural singleton** arising from a scientific theory.

M_2 has the same form but with **unknown mean** θ (here, and throughout, I use **conjugate priors** when they exist):

$$\begin{aligned} (\theta \mid M_2 \, \mathcal{B}) & \sim & N(\theta_0, \sigma_0^2) \\ (y_i \mid M_2 \, \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & N\left(\theta, \sigma^2\right) \, , \end{aligned} \tag{30}$$

with known $(\theta_0, \sigma_0, \sigma)$; note that the **dimensions** of Θ_1 and Θ_2 in this setup are 0 and 1, respectively.

The **joint sampling distribution** for y under M_1 is

Gaussian Mean, Known Variance

To **motivate** the **Ockham factor** in the *JIC* definition, consider the following **model comparison**, which is identical to the large-sample approximate **Higgs boson setup**: M_1 is defined by

$$(y_i \mid M_1 \, \mathcal{B}) \stackrel{\text{IID}}{\sim} N\left(\theta_1, \sigma^2\right) \, , \tag{29}$$

in which the standard deviation (SD) $\sigma > 0$ is assumed **known** and where θ_1 is a **known structural singleton** arising from a scientific theory.

M_2 has the same form but with **unknown mean** θ (here, and throughout, I use **conjugate priors** when they exist):

$$\begin{aligned} (\theta \mid M_2 \, \mathcal{B}) & \sim N(\theta_0, \sigma_0^2) \\ (y_i \mid M_2 \, \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} N\left(\theta, \sigma^2\right) \, , \end{aligned} \tag{30}$$

with known $(\theta_0, \sigma_0, \sigma)$; note that the **dimensions** of Θ_1 and Θ_2 in this setup are 0 and 1, respectively.

The **joint sampling distribution** for y under M_1 is

$$p(y \mid M_1 \, \mathcal{B}) = \sigma^{-n} \, (2 \, \pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2 \, \sigma^2} \sum_{i=1}^n (y_i - \theta_1)^2 \right] \, ; \tag{31}$$

Gaussian Example (continued)

here the **sum** in the last term may be rewritten as

Gaussian Example (continued)

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n}(y_i - \theta_1)^2 = \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] + n(\bar{y} - \theta_1)^2\,, \tag{32}$$

Gaussian Example (continued)

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n}(y_i - \theta_1)^2 = \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] + n(\bar{y} - \theta_1)^2\,, \tag{32}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Gaussian Example (continued)

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n}(y_i - \theta_1)^2 = \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] + n(\bar{y} - \theta_1)^2 \,, \tag{32}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

There are **no unknown parameters** in M_1,

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n}(y_i - \theta_1)^2 = \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] + n(\bar{y} - \theta_1)^2 \,, \qquad (32)$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

There are **no unknown parameters** in M_1, so the log $\left[\ell(\hat{\theta}_1 \mid M_1\, y\, \mathcal{B})\right]$ term in $JIC(M_1 \mid D\, \mathcal{B})$ is to be interpreted as simply

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n} (y_i - \theta_1)^2 = \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \right] + n(\bar{y} - \theta_1)^2 \, , \qquad (32)$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

There are **no unknown parameters** in M_1, so the log $\left[\ell(\hat{\theta}_1 \mid M_1 \, y \, \mathcal{B}) \right]$ term in $JIC(M_1 \mid D \, \mathcal{B})$ is to be interpreted as simply

$$\log \left[p(y \mid M_1 \, \mathcal{B}) \right] = -n \, \log(\sigma) - \frac{n}{2} \log(2 \, \pi) - \frac{n \left[s^2 + (\bar{y} - \theta_1)^2 \right]}{2 \, \sigma^2} \, , \qquad (33)$$

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n}(y_i - \theta_1)^2 = \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] + n(\bar{y} - \theta_1)^2 \,, \qquad (32)$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

There are **no unknown parameters** in M_1, so the log $\left[\ell(\hat{\theta}_1 \mid M_1\, y\, \mathcal{B})\right]$ term in $JIC(M_1 \mid D\, \mathcal{B})$ is to be interpreted as simply

$$\log\left[p(y \mid M_1\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{n\left[s^2 + (\bar{y} - \theta_1)^2\right]}{2\,\sigma^2} \,, \qquad (33)$$

in which $s^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Gaussian Example (continued)

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n}(y_i - \theta_1)^2 = \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] + n(\bar{y} - \theta_1)^2 \, , \qquad (32)$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

There are **no unknown parameters** in M_1, so the log $\left[\ell(\hat{\theta}_1 \mid M_1 \, y \, \mathcal{B})\right]$ term in $JIC(M_1 \mid D \, \mathcal{B})$ is to be interpreted as simply

$$\log\left[p(y \mid M_1 \, \mathcal{B})\right] = -n \log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{n\left[s^2 + (\bar{y} - \theta_1)^2\right]}{2\,\sigma^2} \, , \qquad (33)$$

in which $s^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Similarly, there's **nothing to maximize** as a function of unknowns in (33) and M_1 has no prior distribution;

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n} (y_i - \theta_1)^2 = \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \right] + n (\bar{y} - \theta_1)^2 \, , \qquad (32)$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

There are **no unknown parameters** in M_1, so the log $\left[\ell(\hat{\theta}_1 \mid M_1 \, y \, \mathcal{B}) \right]$ term in $JIC(M_1 \mid D \, \mathcal{B})$ is to be interpreted as simply

$$\log \left[p(y \mid M_1 \, \mathcal{B}) \right] = -n \log(\sigma) - \frac{n}{2} \log(2 \, \pi) - \frac{n \left[s^2 + (\bar{y} - \theta_1)^2 \right]}{2 \, \sigma^2} \, , \qquad (33)$$

in which $s^2 \triangleq \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$.

Similarly, there's **nothing to maximize** as a function of unknowns in (33) and M_1 has no prior distribution; in situations like this (i.e., whenever $k_1 = 0$) I adopt the **convention** $W_1 \triangleq 1$. **Thus**

Gaussian Example (continued)

here the **sum** in the last term may be rewritten as

$$\sum_{i=1}^{n}(y_i - \theta_1)^2 = \left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] + n(\bar{y} - \theta_1)^2\,, \qquad (32)$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

There are **no unknown parameters** in M_1, so the log $\left[\ell(\hat{\theta}_1 \mid M_1\, y\, \mathcal{B})\right]$ term in $JIC(M_1 \mid D\,\mathcal{B})$ is to be interpreted as simply

$$\log\left[p(y \mid M_1\,\mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{n\left[s^2 + (\bar{y} - \theta_1)^2\right]}{2\,\sigma^2}\,, \quad (33)$$

in which $s^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Similarly, there's **nothing to maximize** as a function of unknowns in (33) and M_1 has no prior distribution; in situations like this (i.e., whenever $k_1 = 0$) I adopt the **convention** $W_1 \triangleq 1$. **Thus**

$$\begin{aligned} JIC(M_1 \mid D\,\mathcal{B}) &= -2\log\left[p(y \mid M_1\,\mathcal{B})\right] \\ &= 2\,n\log(\sigma) + 2\,n\log(2\,\pi) + \frac{n\left[s^2 + (\bar{y} - \theta_1)^2\right]}{\sigma^2}\,; \quad (34) \end{aligned}$$

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \mid D\,\mathcal{B}) = O_p(n)$.

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \mid D\,\mathcal{B}) = O_p(n)$.

As for M_2, its **log likelihood function** is

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \mid D\,\mathcal{B}) = O_p(n)$.

As for M_2, its **log likelihood function** is

$$\log\left[\ell(\theta \mid M_2\, y\, \mathcal{B})\right] = -n\, \log(\sigma) - \frac{n}{2} \log(2\,\pi) - \frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2\,, \quad (35)$$

which is **maximized** at $\hat{\theta} = \bar{y}$;

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \mid D\,\mathcal{B}) = O_p(n)$.

As for M_2, its **log likelihood function** is

$$\log\left[\ell(\theta \mid M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\,, \quad (35)$$

which is **maximized** at $\hat{\theta} = \bar{y}$; using an expression similar to (32),

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \mid D\,\mathcal{B}) = O_p(n)$.

As for M_2, its **log likelihood function** is

$$\log\left[\ell(\theta \mid M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\,, \quad (35)$$

which is **maximized** at $\hat{\theta} = \bar{y}$; using an expression similar to (32), the **maximum log likelihood contribution** to $JIC(M_2 \mid D\,\mathcal{B})$ simplifies to

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \,|\, D\,\mathcal{B}) = O_p(n)$.

As for M_2, its **log likelihood function** is

$$\log\left[\ell(\theta \,|\, M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\,, \quad (35)$$

which is **maximized** at $\hat{\theta} = \bar{y}$; using an expression similar to (32), the **maximum log likelihood contribution** to $JIC(M_2 \,|\, D\,\mathcal{B})$ simplifies to

$$\log\left[\ell(\hat{\theta} \,|\, M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{n\,s^2}{2\,\sigma^2}\,, \qquad (36)$$

which is **also** $O_p(n)$ as n increases.

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \mid D\,\mathcal{B}) = O_p(n)$.

As for M_2, its **log likelihood function** is

$$\log\left[\ell(\theta \mid M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\,, \quad (35)$$

which is **maximized** at $\hat{\theta} = \bar{y}$; using an expression similar to (32), the **maximum log likelihood contribution** to $JIC(M_2 \mid D\,\mathcal{B})$ simplifies to

$$\log\left[\ell(\hat{\theta} \mid M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{n\,s^2}{2\,\sigma^2}\,, \qquad (36)$$

which is **also** $O_p(n)$ as n increases.

This leads to a **difference** between the **log likelihood components** of $JIC(M_1 \mid D\,\mathcal{B})$ and $JIC(M_2 \mid D\,\mathcal{B})$ of the form

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \mid D\,\mathcal{B}) = O_p(n)$.

As for M_2, its **log likelihood function** is

$$\log\left[\ell(\theta \mid M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\,, \quad (35)$$

which is **maximized** at $\hat{\theta} = \bar{y}$; using an expression similar to (32), the **maximum log likelihood contribution** to $JIC(M_2 \mid D\,\mathcal{B})$ simplifies to

$$\log\left[\ell(\hat{\theta} \mid M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{n\,s^2}{2\,\sigma^2}\,, \quad (36)$$

which is **also** $O_p(n)$ as n increases.

This leads to a **difference** between the **log likelihood components** of $JIC(M_1 \mid D\,\mathcal{B})$ and $JIC(M_2 \mid D\,\mathcal{B})$ of the form

$$-\,2\,\log\left[\ell(\hat{\theta} \mid M_2\, y\, \mathcal{B})\right] - \{-2\,\log\left[p(y \mid M_1\,\mathcal{B})\right]\} \quad = \quad -n\left(\frac{\bar{y} - \theta_1}{\sigma}\right)^2 .$$

Gaussian Example (continued)

note that as the **sample size increases** $JIC(M_1 \mid D\,\mathcal{B}) = O_p(n)$.

As for M_2, its **log likelihood function** is

$$\log\left[\ell(\theta \mid M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\,, \quad (35)$$

which is **maximized** at $\hat{\theta} = \bar{y}$; using an expression similar to (32), the **maximum log likelihood contribution** to $JIC(M_2 \mid D\,\mathcal{B})$ simplifies to

$$\log\left[\ell(\hat{\theta} \mid M_2\, y\, \mathcal{B})\right] = -n\,\log(\sigma) - \frac{n}{2}\log(2\,\pi) - \frac{n\,s^2}{2\,\sigma^2}\,, \qquad (36)$$

which is **also** $O_p(n)$ as n increases.

This leads to a **difference** between the **log likelihood components** of $JIC(M_1 \mid D\,\mathcal{B})$ and $JIC(M_2 \mid D\,\mathcal{B})$ of the form

$$-2\,\log\left[\ell(\hat{\theta} \mid M_2\, y\, \mathcal{B})\right] - \{-2\,\log[p(y \mid M_1\,\mathcal{B})]\} \;=\; -n\left(\frac{\bar{y} - \theta_1}{\sigma}\right)^2.$$

Both the **minus sign** and the **structure** of this expression make good sense:

Gaussian Example (continued)

$$-2\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right]-\{-2\log[p(y\,|\,M_1\,\mathcal{B})]\} \quad = \quad -n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

Gaussian Example (continued)

$$-2\,\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right] - \{-2\,\log\left[p(y\,|\,M_1\,\mathcal{B})\right]\} \quad = \quad -n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ),

Gaussian Example (continued)

$$-2\,\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right] - \{-2\,\log\left[p(y\,|\,M_1\,\mathcal{B})\right]\} \quad = \quad -n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes,

Gaussian Example (continued)

$$-2\,\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right] - \{-2\,\log\left[p(y\,|\,M_1\,\mathcal{B})\right]\} \;=\; -n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

Gaussian Example (continued)

$$-2\,\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right] - \{-2\,\log\left[p(y\,|\,M_1\,\mathcal{B})\right]\} \quad = \quad -n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

The calculation of W_2 requires an **integration,**

Gaussian Example (continued)

$$- 2 \, \log \left[\ell(\hat{\theta} \mid M_2 \, y \, \mathcal{B}) \right] - \{ -2 \, \log \left[p(y \mid M_1 \, \mathcal{B}) \right] \} \quad = \quad -n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

The calculation of W_2 requires an **integration,** which in this problem (and many other parametric settings) produces an answer in closed form:

Gaussian Example (continued)

$$-2\,\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right]-\{-2\,\log\left[p(y\,|\,M_1\,\mathcal{B})\right]\} \quad = \quad -n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

The calculation of W_2 requires an **integration,** which in this problem (and many other parametric settings) produces an answer in closed form:

$$W_2 = \int_{-\infty}^{\infty} \frac{\sigma^{-n}(2\,\pi)^{-\frac{n}{2}}\,\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]}{\sigma^{-n}(2\,\pi)^{-\frac{n}{2}}\,\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}\left\{\frac{1}{\sigma\sqrt{2\,\pi}}\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(\theta-\theta_0)^2\right]\right\}d\theta\,. \tag{37}$$

Gaussian Example (continued)

$$-2\,\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right]-\{-2\,\log\left[p(y\,|\,M_1\,\mathcal{B})\right]\}\;\;=\;\;-n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

The calculation of W_2 requires an **integration,** which in this problem (and many other parametric settings) produces an answer in closed form:

$$W_2=\int_{-\infty}^{\infty}\frac{\sigma^{-n}(2\,\pi)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]}{\sigma^{-n}(2\,\pi)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}\left\{\frac{1}{\sigma\sqrt{2\,\pi}}\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(\theta-\theta_0)^2\right]\right\}d\theta\,. \tag{37}$$

After **simplification** and affine transformation to the log scale,

Gaussian Example (continued)

$$-2\,\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right] - \{-2\,\log\left[p(y\,|\,M_1\,\mathcal{B})\right]\} \;=\; -n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

The calculation of W_2 requires an **integration,** which in this problem (and many other parametric settings) produces an answer in closed form:

$$W_2 = \int_{-\infty}^{\infty} \frac{\sigma^{-n}(2\,\pi)^{-\frac{n}{2}}\,\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]}{\sigma^{-n}(2\,\pi)^{-\frac{n}{2}}\,\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}\left\{\frac{1}{\sigma\sqrt{2\,\pi}}\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(\theta-\theta_0)^2\right]\right\}d\theta\,. \tag{37}$$

After **simplification** and affine transformation to the log scale, You get

Gaussian Example (continued)

$$-2\,\log\left[\ell(\hat{\theta}\,|\,M_2\,y\,\mathcal{B})\right] - \{-2\,\log\left[p(y\,|\,M_1\,\mathcal{B})\right]\} \;=\; -n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

The calculation of W_2 requires an **integration,** which in this problem (and many other parametric settings) produces an answer in closed form:

$$W_2 = \int_{-\infty}^{\infty} \frac{\sigma^{-n}(2\,\pi)^{-\frac{n}{2}}\,\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]}{\sigma^{-n}(2\,\pi)^{-\frac{n}{2}}\,\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}\left\{\frac{1}{\sigma\sqrt{2\,\pi}}\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(\theta-\theta_0)^2\right]\right\}d\theta\,. \tag{37}$$

After **simplification** and affine transformation to the log scale, You get

$$-2\,\log(W_2) = \log(n) + \log\left(\frac{\sigma_0^2}{\sigma^2}+\frac{1}{n}\right) + \frac{(\bar{y}-\theta_0)^2}{\sigma_0^2+\frac{\sigma^2}{n}}\,, \tag{38}$$

Gaussian Example (continued)

$$- 2 \log \left[\ell(\hat{\theta} \,|\, M_2 \, y \, \mathcal{B}) \right] - \{ -2 \log \left[p(y \,|\, M_1 \, \mathcal{B}) \right] \} \quad = \quad -n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

The calculation of W_2 requires an **integration,** which in this problem (and many other parametric settings) produces an answer in closed form:

$$W_2 = \int_{-\infty}^{\infty} \frac{\sigma^{-n} (2\,\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\,\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 \right]}{\sigma^{-n} (2\,\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\,\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2 \right]} \left\{ \frac{1}{\sigma \sqrt{2\,\pi}} \exp\left[-\frac{1}{2\,\sigma^2} \sum_{i=1}^n (\theta - \theta_0)^2 \right] \right\} d\theta \, . \tag{37}$$

After **simplification** and affine transformation to the log scale, You get

$$- 2 \log(W_2) = \log(n) + \log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \, , \tag{38}$$

so that

Gaussian Example (continued)

$$- 2 \log \left[\ell(\hat{\theta} \mid M_2 \, y \, \mathcal{B}) \right] - \{ -2 \log [p(y \mid M_1 \, \mathcal{B})] \} \quad = \quad -n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2$$

the **farther** $\bar{y}$ is from θ_1 (in units of the SD σ), the **stronger** the evidence for M_2 becomes, increasing at an $O_p(n)$ rate on the log likelihood scale.

The calculation of W_2 requires an **integration,** which in this problem (and many other parametric settings) produces an answer in closed form:

$$W_2 = \int_{-\infty}^{\infty} \frac{\sigma^{-n} (2\,\pi)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\,\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 \right]}{\sigma^{-n} (2\,\pi)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\,\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2 \right]} \left\{ \frac{1}{\sigma \sqrt{2\,\pi}} \exp \left[-\frac{1}{2\,\sigma^2} \sum_{i=1}^n (\theta - \theta_0)^2 \right] \right\} d\theta \, . \tag{37}$$

After **simplification** and affine transformation to the log scale, You get

$$- 2 \log(W_2) = \log(n) + \log \left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \, , \tag{38}$$

so that

$$JIC(M_2 \mid D \, \mathcal{B}) \quad = \quad \left[2\,n \log(\sigma) + 2\,n \log(2\,\pi) + \frac{n s^2}{\sigma^2} \right] + \left[\log(n) + \log \left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right] \tag{39}$$

Now the Role of the Ockham Factor Is Clear

and **finally**

Now the Role of the Ockham Factor Is Clear

and **finally**

$$JIC(M_2 \mid D\,\mathcal{B}) - JIC(M_1 \mid D\,\mathcal{B}) = \left[-n\left(\frac{\bar{y}-\theta_1}{\sigma}\right)^2\right] + \left[\underline{\log(n)} + \log\left(\frac{\sigma_0^2}{\sigma^2}+\frac{1}{n}\right) + \frac{(\bar{y}-\theta_0)^2}{\sigma_0^2+\frac{\sigma^2}{n}}\right].$$

Now the Role of the Ockham Factor Is Clear

and **finally**

$$JIC(M_2 \mid D\,\mathcal{B}) - JIC(M_1 \mid D\,\mathcal{B}) = \left[-n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2 \right] + \left[\underline{\log(n)} + \log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right] .$$

Thus **in this problem**

Now the Role of the Ockham Factor Is Clear

and **finally**

$$JIC(M_2 \mid D\,\mathcal{B}) - JIC(M_1 \mid D\,\mathcal{B}) = \left[-n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2 \right] + \left[\underline{\log(n)} + \log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right].$$

Thus **in this problem**

$$JIC(M_1 \mid D\,\mathcal{B}) = BIC(M_1 \mid D\,\mathcal{B}) \tag{40}$$

Now the Role of the Ockham Factor Is Clear

and **finally**

$$JIC(M_2 \mid D\,\mathcal{B}) - JIC(M_1 \mid D\,\mathcal{B}) = \left[-n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2 \right] + \left[\underline{\log(n)} + \log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right] .$$

Thus **in this problem**

$$JIC(M_1 \mid D\,\mathcal{B}) = BIC(M_1 \mid D\,\mathcal{B}) \tag{40}$$

and

$$JIC(M_2 \mid D\,\mathcal{B}) = BIC(M_2 \mid D\,\mathcal{B}) + \left[\log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right] . \tag{41}$$

Now the Role of the Ockham Factor Is Clear

and **finally**

$$JIC(M_2 \mid D\,\mathcal{B}) - JIC(M_1 \mid D\,\mathcal{B}) = \left[-n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2 \right] + \left[\underline{\log(n)} + \log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right].$$

Thus **in this problem**

$$JIC(M_1 \mid D\,\mathcal{B}) = BIC(M_1 \mid D\,\mathcal{B}) \tag{40}$$

and

$$JIC(M_2 \mid D\,\mathcal{B}) = BIC(M_2 \mid D\,\mathcal{B}) + \left[\log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right]. \tag{41}$$

Now the nature of the **Ockham factor** $\frac{W_2}{W_1}$ becomes clear:

Now the Role of the Ockham Factor Is Clear

and **finally**

$$JIC(M_2 \,|\, D\,\mathcal{B}) - JIC(M_1 \,|\, D\,\mathcal{B}) = \left[-n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2 \right] + \left[\underline{\log(n)} + \log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right] .$$

Thus **in this problem**

$$JIC(M_1 \,|\, D\,\mathcal{B}) = BIC(M_1 \,|\, D\,\mathcal{B}) \tag{40}$$

and

$$JIC(M_2 \,|\, D\,\mathcal{B}) = BIC(M_2 \,|\, D\,\mathcal{B}) + \left[\log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right] . \tag{41}$$

Now the nature of the **Ockham factor** $\frac{W_2}{W_1}$ becomes clear: on the $-2\,\log\left(\frac{W_2}{W_1}\right)$ scale the Ockham factor **reproduces** BIC's $O_p[(k_2 - k_1)\log(n)]$ **approximate parsimony penalty**,

Now the Role of the Ockham Factor Is Clear

and **finally**

$$JIC(M_2 \,|\, D\,\mathcal{B}) - JIC(M_1 \,|\, D\,\mathcal{B}) = \left[-n \left(\frac{\bar{y} - \theta_1}{\sigma} \right)^2 \right] + \left[\underline{\log(n)} + \log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right] .$$

Thus **in this problem**

$$JIC(M_1 \,|\, D\,\mathcal{B}) = BIC(M_1 \,|\, D\,\mathcal{B}) \tag{40}$$

and

$$JIC(M_2 \,|\, D\,\mathcal{B}) = BIC(M_2 \,|\, D\,\mathcal{B}) + \left[\log\left(\frac{\sigma_0^2}{\sigma^2} + \frac{1}{n} \right) + \frac{(\bar{y} - \theta_0)^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \right] . \tag{41}$$

Now the nature of the **Ockham factor** $\frac{W_2}{W_1}$ becomes clear: on the $-2 \log\left(\frac{W_2}{W_1}\right)$ scale the Ockham factor **reproduces** BIC's $O_p[(k_2 - k_1)\log(n)]$ **approximate parsimony penalty**, but JIC is based on an **exact** Bayes factor that in addition includes $O_p(1)$ **correction terms** that arise from the priors in the two models.

Interpreting the Prior Correction Terms

As a result,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, JIC will do a **better job of model comparison** than BIC, while retaining BIC's appealing **fit-parsimony decomposition**.

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \tag{42}$$

which makes **good intuitive sense**:

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \qquad (42)$$

which makes **good intuitive sense**:

- as σ_0 **increases**,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \qquad (42)$$

which makes **good intuitive sense**:

- as σ_0 **increases**, the evidence for M_2 **weakens**,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \tag{42}$$

which makes **good intuitive sense**:

- as σ_0 **increases**, the evidence for M_2 **weakens**, because You then have **more uncertainty** about the underlying data-generating θ_{DG} in M_2;

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \qquad (42)$$

which makes **good intuitive sense**:

- as σ_0 **increases**, the evidence for M_2 **weakens**, because You then have **more uncertainty** about the underlying data-generating θ_{DG} in M_2;
- as σ **increases**,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \qquad (42)$$

which makes **good intuitive sense**:

- as σ_0 **increases**, the evidence for M_2 **weakens**, because You then have **more uncertainty** about the underlying data-generating θ_{DG} in M_2;
- as σ **increases**, the evidence for M_2 **strengthens**,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \qquad (42)$$

which makes **good intuitive sense**:

- as σ_0 **increases**, the evidence for M_2 **weakens**, because You then have **more uncertainty** about the underlying data-generating θ_{DG} in M_2;
- as σ **increases**, the evidence for M_2 **strengthens**, because it becomes **harder** to demonstrate that $\theta_{DG} = \theta_1$; and

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \tag{42}$$

which makes **good intuitive sense**:

- as σ_0 **increases**, the evidence for M_2 **weakens**, because You then have **more uncertainty** about the underlying data-generating θ_{DG} in M_2;
- as σ **increases**, the evidence for M_2 **strengthens**, because it becomes **harder** to demonstrate that $\theta_{DG} = \theta_1$; and
- as $\bar{y}$ **moves away** from its prior expectation θ_0 under M_2,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \qquad (42)$$

which makes **good intuitive sense**:

- as σ_0 **increases**, the evidence for M_2 **weakens**, because You then have **more uncertainty** about the underlying data-generating θ_{DG} in M_2;
- as σ **increases**, the evidence for M_2 **strengthens**, because it becomes **harder** to demonstrate that $\theta_{DG} = \theta_1$; and
- as $\bar{y}$ **moves away** from its prior expectation θ_0 under M_2, this **undermines** the evidence in favor of M_2,

Interpreting the Prior Correction Terms

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.

Neglecting $O\left(\frac{1}{n}\right)$ terms, the *JIC* **prior correction** in this problem is of the approximate form

$$\left[\log\left(\sigma_0^2\right) - \log\left(\sigma^2\right)\right] + \left(\frac{\bar{y} - \theta_0}{\sigma_0}\right)^2 , \qquad (42)$$

which makes **good intuitive sense**:

- as σ_0 **increases**, the evidence for M_2 **weakens**, because You then have **more uncertainty** about the underlying data-generating θ_{DG} in M_2;
- as σ **increases**, the evidence for M_2 **strengthens**, because it becomes **harder** to demonstrate that $\theta_{DG} = \theta_1$; and
- as $\bar{y}$ **moves away** from its prior expectation θ_0 under M_2, this **undermines** the evidence in favor of M_2, because of **conflict** between the prior and the data.

Back To Mendel

Taking **one phenotype** at a time —

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say —

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**,

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**:

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \qquad (43)$$

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{43}$$

$$M_2\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{Beta}(\alpha, \beta) \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} . \tag{44}$$

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{43}$$

$$M_2\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{Beta}(\alpha, \beta) \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} . \tag{44}$$

Calculation reveals that $\hat{\theta}_{MLE} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1: \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i | \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{43}$$

$$M_2: \left\{ \begin{array}{ccc} (\theta | \mathcal{B}) & \sim & \text{Beta}(\alpha, \beta) \\ (y_i | \theta \, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} . \tag{44}$$

Calculation reveals that $\hat{\theta}_{MLE} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and

$$\begin{aligned} -2 \log \left[\ell(\hat{\theta} \,|\, M_1 \, y \, \mathcal{B}) \right] &= -2 \, n \, [\bar{y} \log(\theta_1) + (1 - \bar{y}) \log(1 - \theta_1)] \\ &= -O_p(n) \, ; \end{aligned} \tag{45}$$

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{43}$$

$$M_2\colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{Beta}(\alpha, \beta) \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} . \tag{44}$$

Calculation reveals that $\hat{\theta}_{MLE} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and

$$\begin{aligned} -2 \log \left[\ell(\hat{\theta} \,|\, M_1\, y\, \mathcal{B}) \right] &= -2\, n \left[\bar{y} \log(\theta_1) + (1 - \bar{y}) \log(1 - \theta_1) \right] \\ &= -O_p(n)\,; \end{aligned} \tag{45}$$

$$-2 \log(W_1) = 0\,; \tag{46}$$

Back To Mendel

Taking **one phenotype** at a time — green (dominant) versus yellow pods, say — and letting $y_i = 1$ if second-generation pea plant i is **green** and 0 if **yellow**, Mendel's experimental setup leads without ambiguity to the **comparison of two models**: for $(i = 1, \ldots, n)$,

$$M_1 \colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{point mass at } \theta = \theta_1 \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} \quad \text{and} \tag{43}$$

$$M_2 \colon \left\{ \begin{array}{ccc} (\theta|\mathcal{B}) & \sim & \text{Beta}(\alpha, \beta) \\ (y_i|\theta\,\mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) \end{array} \right\} . \tag{44}$$

Calculation reveals that $\hat{\theta}_{MLE} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and

$$\begin{aligned} -2 \log \left[\ell(\hat{\theta} \,|\, M_1\, y\, \mathcal{B}) \right] &= -2\, n \left[\bar{y} \log(\theta_1) + (1 - \bar{y}) \log(1 - \theta_1) \right] \\ &= -O_p(n)\,; \end{aligned} \tag{45}$$

$$-2 \log(W_1) = 0\,; \tag{46}$$

$$\begin{aligned} -2 \log \left[\ell(\hat{\theta} \,|\, M_2\, y\, \mathcal{B}) \right] &= -2\, n \left[\bar{y} \log(\bar{y}) + (1 - \bar{y}) \log(1 - \bar{y}) \right] \\ &= -O_p(n)\,; \end{aligned} \tag{47}$$

Mendel (continued)

$$
\begin{aligned}
-2\,\log(W_2) &= -2\,\log\Gamma(\alpha+\beta) - 2\,\log\Gamma(\alpha+n\,\bar{y}) \\
&\quad -2\,\log\left[\beta+n(1-\bar{y})\right] + 2\,n\,\bar{y}\log(\bar{y}) \\
&\quad +2\,n(1-\bar{y})\log(1-\bar{y}) + 2\,\log\Gamma(\alpha) \\
&\quad +2\log\Gamma(\beta) + 2\log\Gamma(\alpha+\beta+n) \\
&= +O_p\left[\log(n)\right]\,. \qquad (48)
\end{aligned}
$$

Mendel (continued)

$$
\begin{aligned}
-2\,\log(W_2) &= -2\,\log\Gamma(\alpha+\beta) - 2\,\log\Gamma(\alpha + n\,\bar{y}) \\
&\quad -2\,\log\left[\beta + n(1-\bar{y})\right] + 2\,n\,\bar{y}\,\log(\bar{y}) \\
&\quad +2\,n(1-\bar{y})\log(1-\bar{y}) + 2\,\log\Gamma(\alpha) \\
&\quad +2\log\Gamma(\beta) + 2\log\Gamma(\alpha+\beta+n) \\
&= +O_p\left[\log(n)\right]\,. \qquad (48)
\end{aligned}
$$

With $\alpha = \beta = 1$ in *JIC* for **illustration**,

				--- model 1 ---			---- model 2 -----			
dataset	s	n	y.bar	-2 LL	-2 lOF	jic	-2 LL	-2 lOF	jic	jic-m2 minus jic-m1
round x wrinkled seeds	5474	7324	0.7474	8278.8	0	8278.8	8278.6	8.728	8287.3	8.466
bic				8278.8	0	8278.8	8278.6	8.899	8287.5	8.637
yellow x green seeds	6022	8023	0.7506	9012.8	0	9012.8	9012.8	8.828	9021.6	8.813
bic				9012.8	0	9012.8	9012.8	8.990	9021.8	8.975

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**,

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well,

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**:

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous**

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Home Truth #2(b): **Sharp-null** ($\theta = \theta_1$) **hypothesis testing** is only appropriate when θ_1 is a **structural singleton**.

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Home Truth #2(b): **Sharp-null** ($\theta = \theta_1$) **hypothesis testing** is only appropriate when θ_1 is a **structural singleton**. This **rules out** a great deal of testing performed in **routine practice**;

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Home Truth #2(b): **Sharp-null** ($\theta = \theta_1$) **hypothesis testing** is only appropriate when θ_1 is a **structural singleton**. This **rules out** a great deal of testing performed in **routine practice**; in the **absence** of a structural subspace,

Conclusions

Home Truth #1(a): **Hypothesis** and **significance testing** may look purely **inferential**, but there's almost always a **decision-theoretic** component as well, and it's worthwhile to be as explicit as possible about the real-world **consequences** of **false-positive** and **false-negative** mistakes.

Home Truth #1(b): It's **good** to **get out of the habit** of **using inferential methods** to **make decisions**: their **implicit utility structure** is often **far from optimal**.

Home Truth #2(a): It's both **silly** and **inappropriate** to test a **sharp hypothesis** of the form $\theta = \theta_1$ in problems in which (a) Your **uncertainty** about θ is **continuous** and (b) other values near θ_1 would have the **same real-world consequences**.

Home Truth #2(b): **Sharp-null** $(\theta = \theta_1)$ **hypothesis testing** is only appropriate when θ_1 is a **structural singleton**. This **rules out** a great deal of testing performed in **routine practice**; in the **absence** of a structural subspace, the most useful approach to **inference** is **estimation** via summarization of the **posterior distribution** $p(\theta \,|\, D\,\mathcal{B})$.

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:**

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Q_2': **Could** the data set D have **arisen** from M_1?

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Q_2': **Could** the data set D have **arisen** from M_1?

Posterior predictive P**–values** are in general an **uncalibrated approach** to answering Q_2':,

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Q_2': **Could** the data set D have **arisen** from M_1?

Posterior predictive P–**values** are in general an **uncalibrated approach** to answering Q_2':, but (Draper and Krnjajić, 2015) this can be **fixed**.

The **Jaynes Information Criterion** is

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Q_2': **Could** the data set D have **arisen** from M_1?

Posterior predictive P**–values** are in general an **uncalibrated approach** to answering Q_2':, but (Draper and Krnjajić, 2015) this can be **fixed**.

The **Jaynes Information Criterion** is

$$JIC(M_j \mid D\,\mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] - 2\,\log(W_j)\,; \qquad (49)$$

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Q_2': **Could** the data set D have **arisen** from M_1?

Posterior predictive P**–values** are in general an **uncalibrated approach** to answering Q_2':, but (Draper and Krnjajić, 2015) this can be **fixed**.

The **Jaynes Information Criterion** is

$$JIC(M_j \mid D\,\mathcal{B}) \triangleq -2 \log \left[\ell(\hat{\theta}_j \mid M_j\, y\, \mathcal{B}) \right] - 2 \log(W_j) \, ; \qquad (49)$$

models with **lower** JIC **values** are to be preferred;

Conclusions (continued)

Home Truth #3(a): **Bayesian hypothesis testing** is nothing less, and nothing more, than **Bayesian model comparison**.

Home Truth #3(b): The **model comparison** in 3(a) nearly always involves nothing less, and nothing more, than the **comparison of two prior distributions**, holding the sampling distribution constant.

Home Truth #3(c): **Bayesian significance testing** typically involves another important task in **Bayesian model specification:** answering the question

Q_2': **Could** the data set D have **arisen** from M_1?

Posterior predictive P**–values** are in general an **uncalibrated approach** to answering Q_2':, but (Draper and Krnjajić, 2015) this can be **fixed**.

The **Jaynes Information Criterion** is

$$JIC(M_j \mid D\,\mathcal{B}) \triangleq -2\,\log\left[\ell(\hat{\theta}_j \mid M_j\,y\,\mathcal{B})\right] - 2\,\log(W_j)\,; \qquad (49)$$

models with **lower** JIC **values** are to be preferred; here

Conclusions (concluded)

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, . \tag{50}$$

Conclusions (concluded)

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, . \tag{50}$$

JIC is based on an **exact** Bayes factor that

Conclusions (concluded)

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, . \qquad (50)$$

JIC is based on an **exact** Bayes factor that — when compared with *BIC* —

Conclusions (concluded)

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, . \tag{50}$$

JIC is based on an **exact** Bayes factor that — when compared with *BIC* — includes $O_p(1)$ **correction terms** arising from the priors in the models under comparison.

Conclusions (concluded)

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, . \tag{50}$$

JIC is based on an **exact** Bayes factor that — when compared with *BIC* — includes $O_p(1)$ **correction terms** arising from the priors in the models under comparison.

As a result,

Conclusions (concluded)

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, . \qquad (50)$$

JIC is based on an **exact** Bayes factor that — when compared with *BIC* — includes $O_p(1)$ **correction terms** arising from the priors in the models under comparison.

As a result, if You have **non-trivial** and **well-calibrated prior information**,

Conclusions (concluded)

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, . \tag{50}$$

JIC is based on an **exact** Bayes factor that — when compared with *BIC* — includes $O_p(1)$ **correction terms** arising from the priors in the models under comparison.

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*,

Conclusions (concluded)

$$W_j \triangleq \int_{\Theta_j} \left[\frac{\ell(\theta_j \mid M_j \, y \, \mathcal{B})}{\ell(\hat{\theta}_j \mid M_j \, y \, \mathcal{B})} \right] p(\theta_j \mid M_j \, \mathcal{B}) \, d\theta_j \, . \tag{50}$$

JIC is based on an **exact** Bayes factor that — when compared with *BIC* — includes $O_p(1)$ **correction terms** arising from the priors in the models under comparison.

As a result, if You have **non-trivial** and **well-calibrated prior information**, *JIC* will do a **better job of model comparison** than *BIC*, while retaining *BIC*'s appealing **fit-parsimony decomposition**.