Theory of Probability revisited: A reassessment of a Bayesian classic

Christian P. Robert

Université Paris Dauphine and CREST-INSEE
http://www.ceremade.dauphine.fr/~xian

March 8, 2006

Outline

1 Fundamental notions
2 Direct Probabilities
3 Estimation problems
4 Asymptotics & DT& ...
5 Significance tests: one new parameter
6 Significance tests: various complications
7 Frequency definitions and direct methods
8 General questions

First chapter: Fundamental notions

Who's Jeffreys?

Wikipedia article

Sir Harold Jeffreys (1891–1989)
Mathematician, statistician, geophysicist, and astronomer. He went to St John's College, Cambridge and became a fellow in 1914, where he taught mathematics then geophysics and astronomy. He was knighted in 1953 and received the Gold Medal of the Royal Astronomical Society in 1937.
Jeffreys and Science

Jeffreys married another mathematician and physicist, Bertha Swirles (1903-1999) and together they wrote *Methods of Mathematical Physics*.

Jeffreys is the founder of modern British geophysics. Many of his contributions are summarised in his book *The Earth*. One of his main discoveries is that the core of the Earth is liquid.

Jeffreys and Inference

Jeffreys first used probability to deal with problems in the Philosophy of Science. K. Pearson’s *Grammar of Science* made a great impression on him, with its emphasis on the probabilistic basic of scientific inference. Jeffreys treated probability as a degree of reasonable belief, an epistemic conception common to several Cambridge philosophers, including J.M. Keynes. He used probability to explicate induction and investigate the reasonableness of scientific theories.

Jeffreys and Statistics

Jeffreys wrote more than 400 papers, mostly on his own, on subjects ranging across celestial mechanics, fluid dynamics, meteorology, geophysics and probability.


The statistics papers are in volume 6, *Mathematics, Probability & Miscellaneous Other Sciences*. The coverage is not comprehensive for Jeffreys omitted papers that had been superseded by his books *Scientific Inference* and *Theory of Probability*.

For appraising scientific theories, Venn’s probability as a limiting frequency was useless but Jeffreys considered it mathematically unsound as well.

Around 1930 Jeffreys began devising methods for analysing geophysical data based on epistemic probability. He was extending the methods used by physical scientists and did not know much about, or greatly esteem, the efforts of statisticians.
Jeffreys and Fisher

Ronald Fisher

Meanwhile Ronald Fisher (1890–1962), had rejected the Bayesian approach (1922–1924) and based his work, including maximum likelihood, on frequentist foundations (\(\ast\)).

Fisher and Jeffreys first took serious notice of each another in 1933. About all they knew of each other’s work was that it was founded on a flawed notion of probability.

The Fisher–Jeffreys controversy

The Biometrika papers

Jeffreys (1933a) criticised Fisher (1932) and Fisher (1933) criticised Jeffreys (1932) with a rejoinder by Jeffreys (1933b).

Biometrika called a halt to the controversy by getting the parties to coordinate their last words, in Fisher (1934) and Jeffreys (1934).

Jeffreys' Theory of Probability

While Jeffreys conceded nothing to Fisher, the encounter affected the course of his work. He reacted to the dose of Statistics Fisher administered by reconstructing Fisher’s subject on his own foundations.

Theory of Probability (1939) was the outcome, as a theory of inductive inference founded on the principle of inverse probability, not a branch of pure mathematics, not a description of natural phenomena as with Kolmogorov and von Mises.

Theory of Probability begins with probability, refining the treatment in Scientific Inference, and proceeds to cover a range of applications comparable to that in Fisher’s book.

Jeffreys was very impressed by the solutions Fisher had found for many statistical problems—the trouble was that they had no real foundations! He also tried to place Fisher’s creations like sufficiency in his own system.
First chapter **Fundamental Notions** sets goals for a theory of induction rather than the mathematical bases of probability.

- **Objection to Kolmogorov’s axiomatic definition**
  
  *The above principles (...) rule out any definition of probability that attempts to define probability in terms of infinite sets of possible observations (I, §1.1, 8).*

- **No measure theoretic basis**, e.g.
  
  *If the law concerns a measure capable of any value in a continuous set we could reduce to a finite or an enumerable set (I, §1.62).*

**Conditional probabilities**

Probabilities of events defined as degrees of belief and conditional on past (prior) experience

*Our fundamental idea will not be simply the probability of a proposition \( p \) but the probability of \( p \) on data \( q \) (I, §1.2).*

Subjective flavour of probabilities due to different data, \( P(p|q) \), with same classical definition, e.g.

\[
P(RS|p) = P(R|p)P(S|Rp)
\]

proved for uniform distributions on finite sets (equiprobable events)

**Bayes Theorem**

**Bayes theorem = Inversion of probabilities**

If \( A \) and \( E \) are events such that \( P(E) \neq 0 \), \( P(A|E) \) and \( P(E|A) \) are related by

\[
P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)}
\]

\[
= \frac{P(E|A)P(A)}{P(E)}
\]

[Thomas Bayes (?)]
Who’s Bayes?

**Reverend Thomas Bayes (ca. 1702–1761)**

Presbyterian minister in Tunbridge Wells (Kent) from 1731, son of Joshua Bayes, nonconformist minister. Election to the Royal Society based on a tract of 1736 where he defended the views and philosophy of Newton.

His sole probability paper, *“Essay Towards Solving a Problem in the Doctrine of Chances”*, published posthumously in 1763 by Pierce and containing the seeds of Bayes’ Theorem.

(Modern) parametric model

Observations $x_1, \ldots, x_n$ generated from a probability distribution

$$x = (x_1, \ldots, x_n) \sim f(x|\theta), \quad \theta = (\theta_1, \ldots, \theta_n)$$

Fisher’s associated likelihood

$$\ell(\theta|x) = f(x|\theta)$$

[Inverse density]

Bayesian perspective

**Jeffreys’ premises**

- *Prior beliefs* on the parameters $\theta$ of a model modeled through a probability distribution $\pi$ on $\Theta$, called *prior distribution*
- *Inference* based on the distribution of $\theta$ conditional on $x$, $\pi(\theta|x)$, called *posterior distribution*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta) \, d\theta}.$$

The posterior probabilities of the hypotheses are proportional to the products of the prior probabilities and the likelihoods (I, §.1.22).

Modern Bayesian representation

**Definition (Bayesian model)**

A Bayesian statistical model is made of a parametric statistical model,

$$(\mathcal{X}, f(x|\theta)),$$

and a prior distribution on the parameters,

$$(\Theta, \pi(\theta)).$$
Jeffreys’ Justifications

- All probability statements are conditional
- Actualization of the information on $\theta$ by extracting the information on $\theta$ contained in the observation $x$
  
  *The principle of inverse probability does correspond to ordinary processes of learning (I, §1.5)*

- Allows incorporation of imperfect information in the decision process
  
  *A probability number [sic!] can be regarded as a generalization of the assertion sign (I, §1.51).*

- Unique mathematical way to condition upon the observations (conditional perspective) [Jeffreys?]

Bayes’ 1763 paper:

Billiard ball $W$ rolled on a line of length one, with a uniform probability of stopping anywhere: $W$ stops at $p$.
Second ball $O$ then rolled $n$ times under the same assumptions. $X$ denotes the number of times the ball $O$ stopped on the left of $W$.

**Bayes’ question:**

*Given $X$, what inference can we make on $p$?*

Modern translation:

Derive the posterior distribution of $p$ given $X$, when

$p \sim \mathcal{U}([0, 1])$ and $X|p \sim B(n, p)$

Resolution

Since

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$P(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp$$

and

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp,$$
Resolution (2)

then

\[
P(a < p < b | X = x) = \frac{\int_a^b p^x (1-p)^{n-x} \, dp}{\int_0^1 p^x (1-p)^{n-x} \, dp} = \frac{\int_a^b p^x (1-p)^{n-x} \, dp}{B(x+1, n-x+1)},
\]

i.e.

\[p|x \sim Be(x+1, n-x+1)\]

Beta distribution

Prior and posterior distributions

Given \( f(x|\theta) \) and \( \pi(\theta) \), several distributions of interest:

(a) the joint distribution of \((\theta, x)\),

\[\varphi(\theta, x) = f(x|\theta)\pi(\theta);\]

(b) the marginal distribution of \(x\),

\[m(x) = \int \varphi(\theta, x) \, d\theta = \int f(x|\theta)\pi(\theta) \, d\theta;\]

(c) the posterior distribution of \(\theta\),

\[\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)};\]

(d) the predictive distribution of \(y\), when \(y \sim g(y|\theta, x)\),

\[g(y|x) = \int g(y|\theta, x)\pi(\theta|x) \, d\theta.
\]

Prior Selection

First chapter of ToP quite obscure about choice of \(\pi\). Seems to advocate use of uniform priors:

- If there is originally no ground to believe one of a set of alternatives rather than another, the prior probabilities are equal (I, §1.22).
- To take the prior probabilities different in the absence of observational reason would be an expression of sheer prejudice (I, §1.4).
Prior Selection (2)

Still perceives a potential problem:

...possible to derive theorems by equating probabilities found in different ways (...) We must not expect too much in the nature of a general proof of consistency (I, §1.5).

but evacuates the difficulty:

...the choice in practice, within the range permitted, makes very little difference to the results (I, §1.5).

Posterior distribution

- Operates conditional upon the observations
- Incorporates the requirement of the Likelihood Principle
  ...the whole of the information contained in the observations that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give the likelihood (II, §2.0).
- Avoids averaging over the unobserved values of $x$
- Coherent updating of the information available on $\theta$, independent of the order in which i.i.d. observations are collected
  ...can be used as the prior probability in taking account of a further set of data, and the theory can therefore always take account of new information (I, §1.5).

Additional themes in ToP Chapter 1

- General remarks on model choice and the pervasive Occam’s razor rule
- Bayes factor for testing purposes
- Utility theory that evaluates decisions
- Fairly obscure disgressions on Logic and Gödel’s Theorem.

Who’s Occam?

Pluralitas non est ponenda sine neccesitate

William d’Occam (ca. 1290–ca. 1349)

William d’Occam or d’Ockham was a English theologian (and a Franciscan monk) from Oxford who worked on the bases of empirical induction, nominalism and logic and, in particular, posed the above principle later called Occam’s razor. Also tried for heresy in Avignon and excommunicated by John XXII.
Second chapter: Direct Probabilities

1 Fundamental notions

2 Direct Probabilities
   - Contents
   - Subjective determination
   - Conjugate priors

3 Estimation problems

4 Asymptotics & DT& ...

5 Significance tests: one new parameter

6 Significance tests: various complications

Comments

Physicist’s approach (approximations, intuition, series expansion)
Strange mix of Math (more measure theory than in Chapter I) and pseudo-common sense

*The normal law of error cannot therefore be theoretically proved* (II, §2.68).

Use of χ² test in a frequentist sense!
Advocates normal distributions on the Fourier coefficients

Prior remarks on prior Distributions

The most critical and most criticized point of Bayesian analysis!

**Because...**

the prior distribution is the key to Bayesian inference
But...  

In practice, it seldom occurs that the available prior information is precise enough to lead to an exact determination of the prior distribution.  

There is no such thing as the prior distribution!

Rather...  

The prior is a tool summarizing available information as well as uncertainty related with this information,  

And...  

Ungrounded prior distributions produce unjustified posterior inference

Subjective priors  

In situations with prior information, choice of prior mostly subjective.

Example (Capture probabilities)  

Capture-recapture experiment on migrations between zones  

Prior information on capture and survival probabilities, \( p_t \) and \( q_{it} \)

<table>
<thead>
<tr>
<th>Time</th>
<th>( p_t ) Mean</th>
<th>( p_t ) 95% cred. int.</th>
<th>( q_{it} ) Mean</th>
<th>( q_{it} ) 95% cred. int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.3</td>
<td>[0.1, 0.5]</td>
<td>0.7</td>
<td>[0.4, 0.95]</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>[0.2, 0.6]</td>
<td>0.65</td>
<td>[0.35, 0.9]</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>[0.3, 0.7]</td>
<td>0.7</td>
<td>[0.4, 0.95]</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>[0.05, 0.4]</td>
<td>0.7</td>
<td>[0.4, 0.95]</td>
</tr>
<tr>
<td>6</td>
<td>0.2</td>
<td>[0.05, 0.4]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example (Capture probabilities (2))  

Corresponding prior modeling

<table>
<thead>
<tr>
<th>Time</th>
<th>( \text{Dist.} )</th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( \text{Be}(6, 14) )</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
</tr>
<tr>
<td>3</td>
<td>( \text{Be}(8, 12) )</td>
<td>[0, 3.5]</td>
<td>[0, 3.5]</td>
<td>[0, 3.5]</td>
<td>[0, 3.5]</td>
<td>[0, 3.5]</td>
<td>[0, 3.5]</td>
</tr>
<tr>
<td>4</td>
<td>( \text{Be}(12, 12) )</td>
<td>[0, 5]</td>
<td>[0, 5]</td>
<td>[0, 5]</td>
<td>[0, 5]</td>
<td>[0, 5]</td>
<td>[0, 5]</td>
</tr>
<tr>
<td>5</td>
<td>( \text{Be}(15, 14) )</td>
<td>[0, 7]</td>
<td>[0, 7]</td>
<td>[0, 7]</td>
<td>[0, 7]</td>
<td>[0, 7]</td>
<td>[0, 7]</td>
</tr>
<tr>
<td>6</td>
<td>( \text{Be}(3, 14) )</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site</th>
<th>Time</th>
<th>( t=1.35 )</th>
<th>A</th>
<th>( t=2.4 )</th>
<th>B</th>
<th>( t=1.35 )</th>
<th>B</th>
<th>( t=2.4 )</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_t )</td>
<td>Mean</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>( q_{it} )</td>
<td>Mean</td>
<td>0.7</td>
<td>0.65</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>95% cred. int.</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td>[0, 2.5]</td>
<td></td>
</tr>
</tbody>
</table>
Strategies for prior determination

- Use a partition of $\Theta$ in sets (e.g., intervals), determine the probability of each set, and approach $\pi$ by an histogram.
- Select significant elements of $\Theta$, evaluate their respective likelihoods and deduce a likelihood curve proportional to $\pi$.
- Use the marginal distribution of $x$,
$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)\,d\theta$$

- Empirical and hierarchical Bayes techniques.

- Select a maximum entropy prior when prior characteristics are known:
$$\mathbb{E}^\pi[g_k(\theta)] = \omega_k \quad (k = 1, \ldots, K)$$
with solution, in the discrete case
$$\pi^*(\theta_i) = \frac{\exp\left\{ \sum_{k=1}^K \lambda_k g_k(\theta_i) \right\}}{\sum_j \exp\left\{ \sum_{k=1}^K \lambda_k g_k(\theta_j) \right\}},$$
and, in the continuous case,
$$\pi^*(\theta) = \frac{\exp\left\{ \sum_{k=1}^K \lambda_k g_k(\theta) \right\} \pi(\theta)}{\int \exp\left\{ \sum_{k=1}^K \lambda_k g_k(\eta) \right\} \pi(\eta)\,d\eta},$$
the $\lambda_k$'s being Lagrange multipliers and $\pi_0$ a reference measure. [Caveat]

- Parametric approximations
  - Restrict choice of $\pi$ to a parameterised density $\pi(\theta|\lambda)$
  - and determine the corresponding (hyper-)parameters $\lambda$
  - through the moments or quantiles of $\pi$.

Example
For the normal model $x \sim \mathcal{N}(\theta, 1)$, ranges of the posterior moments for fixed prior moments $\mu_1 = 0$ and $\mu_2$.

<table>
<thead>
<tr>
<th>$\mu_2$</th>
<th>$x$</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean</td>
<td>mean</td>
<td>variance</td>
</tr>
<tr>
<td>3 0</td>
<td>-1.05</td>
<td>1.05</td>
<td>3.00</td>
<td></td>
</tr>
<tr>
<td>3 1</td>
<td>-0.70</td>
<td>1.69</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td>3 2</td>
<td>-0.50</td>
<td>2.85</td>
<td>5.78</td>
<td></td>
</tr>
<tr>
<td>1.5 0</td>
<td>-0.59</td>
<td>0.59</td>
<td>1.50</td>
<td></td>
</tr>
<tr>
<td>1.5 1</td>
<td>-0.37</td>
<td>1.05</td>
<td>1.97</td>
<td></td>
</tr>
<tr>
<td>1.5 2</td>
<td>-0.27</td>
<td>2.08</td>
<td>3.80</td>
<td></td>
</tr>
</tbody>
</table>

[Goutis, 1990]
Conjugate priors

Specific parametric family with analytical properties

Definition

A family \( \mathcal{F} \) of probability distributions on \( \Theta \) is conjugate for a likelihood function \( f(x|\theta) \) if, for every \( \pi \in \mathcal{F} \), the posterior distribution \( \pi(\theta|x) \) also belongs to \( \mathcal{F} \).

[Raiffa & Schlaifer, 1961]

Only of interest when \( \mathcal{F} \) is parameterised: switching from prior to posterior distribution is reduced to an updating of the corresponding parameters.

Justifications

- Limited/finite information conveyed by \( x \)
- Preservation of the structure of \( \pi(\theta) \)
- Exchangeability motivations
- Device of virtual past observations
- Linearity of some estimators
- Tractability and simplicity
- First approximations to adequate priors, backed up by robustness analysis

Exponential families

Definition

The family of distributions

\[
f(x|\theta) = C(\theta)h(x) \exp\{R(\theta) \cdot T(x)\}
\]

is called an exponential family of dimension \( k \). When \( \Theta \subset \mathbb{R}^k \), \( \mathcal{F} \subset \mathbb{R}^k \) and

\[
f(x|\theta) = C(\theta)h(x) \exp\{\theta \cdot x\},
\]

the family is said to be natural.

Interesting analytical properties:

- Sufficient statistics (Pitman–Koopman Lemma)
- Common enough structure (normal, binomial, Poisson, Wishart, &tc...)
- Analyticity (\( \mathbb{E}_\theta[x] = \nabla \psi(\theta) \), ...)
- Allow for conjugate priors

\[
\pi(\theta|\mu, \lambda) = K(\mu, \lambda) e^{\theta \cdot \mu - \lambda \psi(\theta)}
\]
**Linearity of the posterior mean**

If

\[ \theta \sim \pi_{x_0}(\theta) \propto e^{\theta - x_0 - \lambda \psi(\theta)} \]

with \( x_0 \in X \), then

\[ \mathbb{E}^{\pi}[\nabla \psi(\theta)] = \frac{x_0}{\lambda}. \]

Therefore, if \( x_1, \ldots, x_n \) are i.i.d. \( f(x|\theta) \),

\[ \mathbb{E}^{\pi}[\nabla \psi(\theta)|x_1, \ldots, x_n] = \frac{x_0 + n\bar{x}}{\lambda + n}. \]

**But...**

**Example**

When \( x \sim Be(\alpha, \theta) \) with known \( \alpha \),

\[ f(x|\theta) \propto \frac{\Gamma(\alpha + \theta)(1 - x)^\theta}{\Gamma(\theta)}, \]

conjugate distribution not so easily manageable

\[ \pi(\theta|x_0, \lambda) \propto \left( \frac{\Gamma(\alpha + \theta)}{\Gamma(\theta)} \right)^\lambda (1 - x_0)^\theta. \]
Example
Coin spun on its edge, proportion \( \theta \) of heads
When spinning \( n \) times a given coin, number of heads
\[ x \sim B(n, \theta) \]
Flat prior, or mixture prior
\[ \frac{1}{2} [B(10, 20) + B(20, 10)] \]
or
\[ 0.5 B(10, 20) + 0.2 B(15, 15) + 0.3 B(20, 10). \]

Mixtures of natural conjugate distributions also make conjugate families

Three prior distributions for a spinning-coin experiment

Chapter 3: Estimation Problems

1 Fundamental notions

2 Direct Probabilities

3 Estimation problems
   - Improper prior distributions
   - Noninformative prior distributions
   - Bayesian inference
   - Sampling models
   - Normal models and linear regression
   - More sufficiency
   - More noninformative priors
   - The Jeffreys prior

4 Asymptotics & DT&
Improper prior distributions

Necessary extension from a prior distribution to a prior $\sigma$-finite measure $\pi$ such that

$$\int_{\Theta} \pi(\theta) \, d\theta = +\infty$$

...the fact that $\int_0^\infty \frac{dv}{v}$ diverges at both limits is a satisfactory feature (III, §3.1).

Modern justifications

Often automatic/noninformative prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More robust answer against possible misspecifications of the prior

5. Generally more acceptable to non-Bayesians, with frequentist justifications, such as:
   - (i) minimaxity
   - (ii) admissibility
   - (iii) invariance

6. Improper priors preferred to vague proper priors such as a $\mathcal{N}(0, 100^2)$ distribution

7. Penalization factor in

$$\min_d \int \mathcal{L}(\theta, d) \pi(\theta) f(x|\theta) \, dx \, d\theta$$

Validation

Extension of the posterior distribution $\pi(\theta|x)$ associated with an improper prior $\pi$ as given by Bayes’s formula

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta) \, d\theta},$$

when

$$\int_{\Theta} f(x|\theta)\pi(\theta) \, d\theta < \infty$$
If the parameter may have any value in a finite range, or from \(-\infty\) to \(+\infty\), its prior probability should be taken as uniformly distributed (III, §3.1).

\[
\text{Example (Flat prior)}
\]

If \(x \sim \mathcal{N}(\theta, 1)\) and \(\pi(\theta) = \omega\), constant, the pseudo marginal distribution is

\[
m(x) = \omega \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \theta)^2}{2} \right\} d\theta = \omega
\]

and the posterior distribution of \(\theta\) is

\[
\pi(\theta | x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \theta)^2}{2} \right\}
\]

i.e., corresponds to a \(\mathcal{N}(x, 1)\) distribution.

[Independent of \(\omega\)]

**Over-interpretation**

If we take

\[
P(d\sigma | H) \propto d\sigma
\]

as a statement that \(\sigma\) may have any value between 0 and \(\infty\) (\(\ldots\)), we must use \(\infty\) instead of 1 to denote certainty on data \(H\). (\(\ldots\)) But (\(\ldots\)) the number for the probability that \(\sigma < \alpha\) will be finite, and the number for \(\sigma > \alpha\) will be infinite. Thus (\(\ldots\)) the probability that \(\sigma < \alpha\) is 0. This is inconsistent with the statement that we know nothing about \(\sigma\) (III, §3.1)

---


*The mistake is to think of them [non-informative priors] as representing ignorance*

[Lindley, 1990]
Over-interpretation (2)

Example (Flat prior (2))

Consider a $\theta \sim \mathcal{N}(0, \tau^2)$ prior. Then, for any $(a, b)$

$$\lim_{\tau \to \infty} P^\pi(\theta \in [a, b]) = 0$$

...we usually have some vague knowledge initially that fixes upper and lower bounds [but] the truncation of the distribution makes a negligible change in the results (III, §3.1)

Noninformative setting

What if all we know is that we know “nothing” ?!

...how can we assign the prior probability when we know nothing about the value of the parameter except the very vague knowledge just indicated? (III, §3.1)

Example (Haldane prior)

For a binomial observation, $x \sim \mathcal{B}(n, p)$, and prior $\pi^*(p) \propto [p(1-p)]^{-1}$, the marginal distribution,

$$m(x) = \int_0^1 [p(1-p)]^{-1} \left(\frac{n}{x}\right) p^x(1-p)^{n-x} dp$$

$$= B(x, n-x),$$

is only defined for $x \neq 0, n$.

Missed by Jeffreys:

If a sample is of one type with respect to some property there is probability 1 that the population is of that type (III, §3.1)

Noninformative distributions

...provide a formal way of expressing ignorance of the value of the parameter over the range permitted (III, §3.1).

In the absence of prior information, prior distributions solely derived from the sample distribution $f(x|\theta)$

It says nothing about the value of the parameter, except the bare fact that it may possibly by its very nature be restricted to lie within certain definite limits (III, §3.1)
Re-Warning

Noninformative priors cannot be expected to represent exactly total ignorance about the problem at hand, but should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing.

[Kass and Wasserman, 1996]

Difficulties

Lack of reparameterization invariance/coherence

\[ \psi = e^\theta \quad \pi_1(\psi) = \frac{1}{\psi} \neq \pi_2(\psi) = 1 \]

There are cases of estimation where a law can be equally well expressed in terms of several different sets of parameters, and it is desirable to have a rule that will lead to the same results whichever set we choose. Otherwise we shall again be in danger of using different rules arbitrarily to suit our taste (III, §3.1)

Difficulties (2)

Problems of properness

If \( x \sim N(\theta, \sigma^2) \), \( \pi(\theta, \sigma) = 1 \)

\[ \pi(\theta, \sigma|x) \propto e^{-(x-\theta)^2/2\sigma^2} \sigma^{-1} \]

\[ \Rightarrow \pi(\sigma|x) \propto 1 \quad (!!!) \]
Difficulties (3)

Inappropriate for testing point null hypotheses:

*The fatal objection to the universal application of the uniform distribution is that it would make any significance test impossible. If a new parameter is being considered, the uniform distribution of prior probability for it would practically always lead to the result that the most probable value is different from zero (III, §3.1)*

but so would any continuous prior!

A strange conclusion

“The way out is in fact very easy”:

If \( v \) is capable of any value from 0 to \( \infty \), and we take its prior probability distribution as proportional to \( dv/v \), then \( \rho = 1/v \) is also capable of any value from 0 to \( \infty \), and if we take its prior probability as proportional to \( d\rho/\rho \) we have two perfectly consistent statements of the same form (III, §3.1)

Seems to consider that the objection of 0 probability result only applies to parameters with \((0, \infty)\) support.

ToP difficulties (§3.1)

End of §3.1 tries to justify the prior \( \pi(v) \propto 1/v \) as “correct” prior. E.g., usual argument that this corresponds to flat prior on log \( v \), although Jeffreys rejects Haldane’s prior which is based on flat prior on the logistic transform \( v/(1 - v) \)

...not regard the above as showing that \( dx/x(1-x) \) is right for their problem. Other transformations would have the same properties and would be mutually inconsistent if the same rule was taken for all. ...[even though] there is something to be said for the rule (III, §3.1)

\[
P(dx|H) = \frac{1}{\pi} \frac{dx}{\sqrt{x(1-x)}}.
\]

Very shaky from a mathematical point of view:

*the ratio of the probabilities that \( v \) is less or greater than \( a \) is*

\[
\int_0^a v^n \, dv / \int_a^\infty v^n \, dv.
\]

(...) If \( n < -1 \), the numerator is infinite and the denominator finite and the rule would say that the probability that \( v \) is greater than any finite value is 0. (...) But if \( n = -1 \) both integrals diverge and the ratio is indeterminate. (...) Thus we attach no value to the probability that \( v \) is greater or less than \( a \), which is a statement that we know nothing about \( v \) except that it is between 0 and \( \infty \) (III, §3.1)

See also the footnote † in §3.4!
Posterior distribution

\[ \pi(\theta|x) \propto f(x|\theta) \pi(\theta) \]

- extensive summary of the information available on \( \theta \)
- integrate simultaneously prior information and information brought by \( x \)
- unique motor of inference

Bayesian Decision Theory

For a loss \( L(\theta, \delta) \) and a prior \( \pi \), the Bayes rule is

\[ \delta^\pi(x) = \arg \min_d \mathbb{E}_\pi[L(\theta, d)|x]. \]

**Note:** Practical computation not always possible analytically.

Conjugate priors

For conjugate distributions, the posterior expectations of the natural parameters can be expressed analytically, for one or several observations.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Conjugate prior</th>
<th>Posterior mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{N}(\theta, \sigma^2) )</td>
<td>Normal</td>
<td>( \frac{\mu \sigma^2 + \tau^2 x}{\sigma^2 + \tau^2} )</td>
</tr>
<tr>
<td>Poisson ( \mathcal{P}(\theta) )</td>
<td>( \mathcal{N}(\mu, \tau^2) ) Gamma</td>
<td>( \frac{\alpha + x}{\beta + 1} )</td>
</tr>
<tr>
<td>( \mathcal{G}(\alpha, \beta) )</td>
<td>( \mathcal{G}(\alpha, \beta) )</td>
<td>( \frac{\alpha + \nu}{\beta + x} )</td>
</tr>
<tr>
<td>( \mathcal{B}(n, \theta) )</td>
<td>Beta</td>
<td>( \frac{\alpha + x}{\alpha + \beta + n} )</td>
</tr>
<tr>
<td>( \mathcal{N}(n, \theta) )</td>
<td>Beta</td>
<td>( \frac{\alpha + n}{\alpha + \beta + x + n} )</td>
</tr>
<tr>
<td>( \mathcal{D}(\alpha_1, \ldots, \alpha_k) ) Dirichlet</td>
<td>( \mathcal{D}(\alpha_1, \ldots, \alpha_k) )</td>
<td>( \frac{\alpha_i + x_i}{\left( \sum_j \alpha_j \right) + n} )</td>
</tr>
<tr>
<td>( \mathcal{G}(\alpha/2, \beta/2) ) Gamma</td>
<td>( \mathcal{G}(\alpha/2, \beta/2) )</td>
<td>( \frac{\alpha + 1}{\beta + (\mu - x)^2} )</td>
</tr>
</tbody>
</table>
Example

Consider

\[ x_1, \ldots, x_n \sim \mathcal{U}([0, \theta]) \]

and \( \theta \sim \mathcal{P}(\theta_0, \alpha) \). Then

\[ \theta | x_1, \ldots, x_n \sim \mathcal{P}(\max(\theta_0, x_1, \ldots, x_n), \alpha + n) \]

and

\[ \delta_\pi(x_1, \ldots, x_n) = \frac{\alpha + n}{\alpha + n - 1} \max(\theta_0, x_1, \ldots, x_n). \]

Hypergeometric and binomial inference

Case of an \( \mathcal{H}(N, n, r) \) distribution under uniform prior \( \pi(r) = 1/(N + 1) \)

Posterior

\[ P(r|N, l, H) = \binom{r}{l} \binom{N - r}{n - l} / \binom{N + 1}{n + 1} \]

Prediction

If \( x \sim f(x|\theta) \) and \( z \sim g(z|x, \theta) \), the predictive of \( z \) is

\[ g^\pi(z|x) = \int_\Theta g(z|x, \theta)\pi(\theta|x) \, d\theta. \]
Darroch model for capture–recapture

Alternative formulation:

\[ n_{11} \sim \mathcal{H}(N, n_2, n_1/N) \]

Classical (MLE) estimator of \( N \)

\[ \hat{N} = \frac{n_1}{(n_{11}/n_2)} \]

It cannot be used when \( n_{11} = 0 \)

Example (Deers)

Herd of deer on an island of Newfoundland (Canada) w/o any predator, thus culling necessary for ecological equilibrium.

Annual census too time-consuming, but birth and death patterns for the deer imply that the number of deer varies between 36 and 50. Prior:

\[ N \sim \mathcal{U}\{36, \ldots, 50\} \]

Table: Posterior distribution of the deer population size, \( \pi(N|n_{11}) \).

| \( N \) \( | n_{11} \) | 0   | 1   | 2   | 3   | 4   | 5   |
|----------------|-----|-----|-----|-----|-----|-----|
| 36             | 0.058 | 0.072 | 0.089 | 0.106 | 0.125 | 0.144 |
| 37             | 0.059 | 0.072 | 0.085 | 0.098 | 0.111 | 0.124 |
| 38             | 0.061 | 0.071 | 0.081 | 0.090 | 0.100 | 0.108 |
| 39             | 0.062 | 0.070 | 0.077 | 0.084 | 0.089 | 0.094 |
| 40             | 0.063 | 0.069 | 0.074 | 0.076 | 0.081 | 0.082 |
| 41             | 0.065 | 0.068 | 0.071 | 0.072 | 0.073 | 0.072 |
| 42             | 0.066 | 0.068 | 0.067 | 0.067 | 0.066 | 0.064 |
| 43             | 0.067 | 0.067 | 0.065 | 0.063 | 0.060 | 0.056 |
| 44             | 0.068 | 0.066 | 0.062 | 0.059 | 0.054 | 0.050 |
| 45             | 0.069 | 0.065 | 0.060 | 0.055 | 0.050 | 0.044 |
| 46             | 0.070 | 0.064 | 0.058 | 0.051 | 0.045 | 0.040 |
| 47             | 0.071 | 0.063 | 0.056 | 0.048 | 0.041 | 0.035 |
| 48             | 0.072 | 0.063 | 0.054 | 0.045 | 0.038 | 0.032 |
| 49             | 0.073 | 0.062 | 0.052 | 0.043 | 0.035 | 0.028 |
| 50             | 0.074 | 0.061 | 0.050 | 0.040 | 0.032 | 0.026 |
**Table:** Posterior mean of the deer population size, $N$.

<table>
<thead>
<tr>
<th>$n_{11}$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(N</td>
<td>n_{11})$</td>
<td>43.32</td>
<td>42.77</td>
<td>42.23</td>
<td>41.71</td>
<td>41.23</td>
</tr>
</tbody>
</table>

**Different loss function**

$$L(N, \delta) = \begin{cases} 10(\delta - N) & \text{if } \delta > N, \\ N - \delta & \text{otherwise,} \end{cases}$$

in order to avoid overestimation

Bayes estimator is $(1/11)$-quantile of $\pi(N | n_{11})$.

**Table:** Estimated deer population

<table>
<thead>
<tr>
<th>$n_{11}$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta^p(n_{11})$</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>36</td>
<td>36</td>
<td>36</td>
</tr>
</tbody>
</table>

**Laplace succession rule**

Example of a predictive distribution

...considering the probability that the next specimen will be of the first type. The population being of number $N$, of which $n$ have already been removed, and the members of the first type being $r$ in number, of which $l$ have been removed, the probability that the next would be of the type, given $r, N$ and the sample is (III, §3.2)

$$P(p | l, m, N, H) = \frac{r - l}{N - m} \frac{r}{l} \frac{N - r}{n - l} \left/ \frac{N + 1}{n + 1} \right.$$ 

the marginal posterior of $p$ is

$$P(p | l, m, N, H) = \frac{l + 1}{N - m} \frac{(N + 1)}{(n + 1)} \frac{N + 1}{n + 2}$$

which is independent of $N$. (...) Neither Bayes nor Laplace, however, seem to have considered the case of finite $N$ (III, §3.2)
New criticism of uniform prior

The fundamental trouble is that the prior probabilities $1/(N + 1)$ attached by the theory to the extreme values are utterly so small that they amount to saying, without any evidence at all, that it is practically certain that the population is not homogeneous in respect to the property to be investigated. (...) Now I say that for this reason the uniform assessment must be abandoned for ranges including the extreme values. (III, §3.21)

**Explanation:** This is a preparatory step for the introduction of specific priors fitted to point null hypotheses (using Dirac masses).

Local resolution

Different weight on the boundaries

$$P(r = 0 | NH) = P(r = N | NH) = k$$

...we are therefore restricted to values of $k$ between $\frac{1}{3}$ and $\frac{1}{2}$. A possible alternative form would be to take

$$k = \frac{1}{4} + \frac{1}{2(N + 1)}$$

which puts half the prior probability into the extremes and leaves the other half distributed over all values (III, §3.21)

Another contradiction

For the multinomial model

$$\mathcal{M}(n; p_1, \ldots, p_r),$$

under the uniform prior

$$(p_1, \ldots, p_r) \sim D(1, \ldots, 1),$$

the marginal on $p_1$ is not uniform:

$$p_1 \sim B(1, r - 1).$$

This expresses the fact that the average value of all the $p$’s is not $1/r$ instead of $1/2$ (III, §3.23)

The Poisson model

For $m \sim P(r)$, Jeffreys justifies the prior $P(dr | H) \propto dr/r$ by

*This parameter is not a chance but a chance per unit time, and therefore is dimensional (III, §3.3)*

Posterior distribution conditional on observations $m_1, \ldots, m_n$

$$P(dr | m_1, \ldots, m_n, H) \propto \frac{n^{Sm}}{(S_m - 1)!} r^{S_m - 1} e^{nr} dr$$

given by the incomplete $\Gamma$ function. We notice that the only function of the observations that appears in the posterior probability is $S_m$, therefore a sufficient statistic for $r$ (III, §3.3)
The normal model

Importance of the normal model in many fields

\[ N_p(\theta, \Sigma) \]

with known \( \Sigma \), normal conjugate distribution, \( N_p(\mu, A) \).

Under quadratic loss, the Bayes estimator is

\[
\delta^\pi(x) = x - \Sigma(\Sigma + A)^{-1}(x - \mu)
\]

\[
= (\Sigma^{-1} + A^{-1})^{-1}(\Sigma^{-1}x + A^{-1}\mu);
\]

Laplace approximation

\textbf{ToP} presents the normal distribution as a second order approximation of slowly varying densities,

\[ P(dx|x_1, \ldots, x_n, H) \propto f(x) \exp \left\{ -\frac{n}{2\sigma^2}(x - \bar{x})^2 \right\} \]

(with the weird convention that \( \bar{x} \) is the empirical mean of the \( x_i \)'s and \( x \) is the true mean, i.e. \( \theta \)...)

Estimation of variance

If

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

the likelihood is

\[
\ell(\theta, \sigma | \bar{x}, s^2) \propto \sigma^{-n} \exp \left\{ -\frac{n}{2\sigma^2} \left( s^2 + (\bar{x} - \theta)^2 \right) \right\}
\]

Jeffreys then argues in favour of

\[
\pi^*(\theta, \sigma) = 1/\sigma
\]

assuming independence between \( \theta \) and \( \sigma \) \[Warnin!\]

Reminder

- Not defined for \( n = 1, 2 \)
- \( \theta \) and \( \sigma^2 \) are not a posteriori independent.
- Conjugate posterior distributions have the same form
- but require a careful determination of the hyperparameters
More of the weird stuff!

Jeffreys also considers degenerate cases:

If \( n = 1, \bar{x} = x_1, \) and \( s = 0 \) [!!!], then

\[
P(dx|\bar{x}, H) \propto \sigma^{-2} \exp \left\{ \frac{(x - \bar{x})^2}{\sigma^2} \right\} \, dx\sigma
\]

Integrating with respect to \( \sigma \) we get

\[
P(dx|x_1, H) \propto \frac{dx}{|x - x_1|}
\]

that is, the most probable value of \( x \) is \( x_1 \) but we have no information about the accuracy of the determination (III, §3.41).

...even though \( P(dx|x_1, H) \) is not integrable...

(Modern) basics

The least-squares estimator \( \hat{\beta} \) has a normal distribution

\[
\hat{\beta} \sim \mathcal{N}_m(\beta, \sigma^2(X^T X)^{-1})
\]

Corresponding (Zellner’s) conjugate distributions on \((\beta, \sigma^2)\)

\[
\beta|\sigma^2 \sim \mathcal{N}_m(\mu, \frac{\sigma^2}{n_0}(X^T X)^{-1})
\]

\[
\sigma^2 \sim \mathcal{IG}(\nu/2, s_0^2/2)
\]

since, if \( s^2 = ||y - X\hat{\beta}||^2 \),

\[
\beta|\hat{\beta}, s^2, \sigma^2 \sim \mathcal{N}_{p}\left(\frac{n_0\mu + \hat{\beta}}{n_0 + 1}, \frac{\sigma^2}{n_0 + 1}(X^T X)^{-1}\right),
\]

\[
\sigma^2|\hat{\beta}, s^2 \sim \mathcal{IG}\left(\frac{k - p + \nu}{2}, \frac{s^2 + s_0^2 + \frac{n_0}{n_0 + 1}\mu - \hat{\beta})^T X^T X (\mu - \beta)}{2}\right)
\]

More general conjugate distributions of the type

\[
\beta \sim \mathcal{N}_m(A\theta, C),
\]

where \( \theta \in \mathbb{R}^q \) (\( q \leq m \)).
Prior modelling

In ToP

\[ P(dx_1, \ldots, dx_m, d\sigma|H) \propto dx_1 \cdots dx_m d\sigma/\sigma \]

i.e. \( \pi(\beta, \sigma) = 1/\sigma \)

...the posterior probability of \( \zeta_m \) is distributed as for \( t \) with \( n - m \) degrees of freedom (III, §3.5)

Explanation

- the \( \zeta_i \)'s are the transforms of the \( \beta_i \)'s in the eigenbasis of \( X^TX \)
- \( \zeta_i \) is also distributed as a \( t \) with \( n - m \) degrees of freedom

Sigma unknown

In this general case, the (apocryphal) Jeffreys prior is

\[ \pi^J(\beta, \Sigma) = \frac{1}{|\Sigma|^{(k+1)/2}}. \]

with likelihood

\[ \ell(\beta, \Sigma|y) \propto |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} \sum_{i=1}^{n} (y_i - X_i\beta)(y_i - X_i\beta)^T \right] \right\} \]

Consequences

- suggests (inverse) Wishart distribution on \( \Sigma \)
- posterior marginal distribution on \( \beta \) only defined for sample size large enough
- no closed form expression for posterior marginal

Plusses of the Bayesian approach

Example (Truncated normal)

When \( a_1, \ldots, a_n \) are \( \mathcal{N}(\alpha, \sigma_r^2) \), with \( \sigma_r^2 \) known, and \( \alpha > 0 \),

the posterior probability of \( \alpha \) is therefore a normal one about the weighted mean by the \( a_r \), but it is truncated at \( \alpha = 0 \) (III, §3.55).

Separation of likelihood (observations) from prior (\( \alpha > 0 \))
Preliminary example

Case of a quasi-exponential setting:

\[ x_1, \ldots, x_n \sim \mathcal{U}(\alpha - \sigma, \alpha + \sigma) \]

Under prior

\[ P(d\alpha, d\sigma|H) \propto d\alpha \, d\sigma / \sigma \]

the two extreme observations are sufficient statistics for \( \alpha \) and \( \sigma \). Then

\[ P(d\alpha|x_1, \ldots, x_n, H) \propto \begin{cases} (\alpha - x(1))^{-n}d\alpha & (\alpha > (x(1) + x(2))/2, \\ (x(2) - \alpha)^{-n}d\alpha & (\alpha < (x(1) + x(2))/2, \end{cases} \]

[with] a sharp peak at the mean of the extreme values

(III, §3.6)

Reassessment of sufficiency and Pitman–Koopman lemma

- in ToP, sufficiency is defined via a poor man’s factorisation theorem, rather than through Fisher’s conditional property (§3.7)
- Pitman–Koopman lemma is re-demonstrated while no mention is made of the support being independent of the parameter(s)...
- ...but Jeffreys concludes with an example where the support is \((\alpha, +\infty)\) (!)

Partial sufficiency

Example (Normal correlation)

Case of a \( \mathcal{N}_2(\theta, \Sigma) \) sample under the prior

\[ \pi(\theta, \Sigma) = 1/\sigma_{11}\sigma_{22} \quad \Sigma = \begin{bmatrix} \sigma_{11}^2 & \rho \sigma_{11}\sigma_{22} \\ \rho \sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{bmatrix} \]

Then (III, §3.9) \( \pi(\sigma_{11}, \sigma_{22}, \rho|\text{Data}) \) is proportional to

\[ \frac{1}{(\sigma_{11}\sigma_{22})^n (1 - \rho^2)^{(n-1)/2}} \exp \left\{ \frac{-n}{2(1 - \rho^2)} \left( \frac{s^2}{\sigma_{11}^2} + \frac{t^2}{\sigma_{22}^2} - \frac{2\rho \sigma_{11}\sigma_{22}}{\sigma_{11}\sigma_{22}} \right)^2 \right\} \]

Example (Normal correlation (2))

and

\[ \pi(\rho|\text{Data}) \propto \frac{(1 - \rho^2)^{(n-1)/2}}{(1 - \rho \sigma_{11}\sigma_{22})} S_{n-1}(\rho \sigma_{11}\sigma_{22}) \]

only depends on \( r \), which amounts to an additional proof that \( r \) is a sufficient statistic for \( \rho \) (III, §3.9)

...Jeffreys unaware of marginalisation paradoxes...
Marginalisation paradoxes

In the case of correlation, posterior on $\rho$ could have been derived from prior $\pi(\rho) = 1/2$ and distribution of $r$. This is not always the case:

**Marginalisation paradox**

- $\pi(\theta_1|x_1, x_2)$ only depends on $x_1$
- $f(x_1|\theta_1, \theta_2)$ only depends on $\theta_1$
- ...but $\pi(\theta_1|x_1, x_2)$ is not the same as

$$\pi(\theta_1|x_1) \propto \pi(\theta_1)f(x_1|\theta_1) \quad (!)$$

[Dawid, Stone & Zidek, 1973]

Example (Normal MP)

Case when

$$u_1 \sim \mathcal{N}(\mu_1, \sigma^2), \quad u_2 \sim \mathcal{N}(\mu_2, \sigma^2), \quad s^2 \sim \sigma^2 \chi^2_\nu / \nu,$$

and when $\zeta = (\mu_1 - \mu_2) / (\sigma \sqrt{2})$ parameter of interest, under prior $\pi(\mu_1, \mu_2, \sigma) = 1/\sigma$

Then

- $\pi(\zeta|x)$ only depends on $z = u_1 - u_2 / s \sqrt{2}$
- and $z$ only depends on $\zeta$
- ...but impossible to derive $\pi(\zeta|x)$ from $f(z|\zeta)$
- ...and no paradox when $\pi(\mu_1, \mu_2, \sigma) = 1/\sigma^2 \quad [!!]$  

(Modern, not ToP) invariant priors

**Principle:** Agree with the natural symmetries of the problem

- Identify invariance structures as group action

$$\mathcal{G} : x \to g(x) \sim f(g(x)|\bar{g}(\theta))$$
$$\bar{\mathcal{G}} : \theta \to \bar{g}(\theta)$$
$$\mathcal{G}^* : L(d, \theta) = L(g^*(d), \bar{g}(\theta))$$

- Determine an invariant prior

$$\pi(\bar{g}(A)) = \pi(A)$$

**Right Haar measure**

**But...**

- Requires invariance to be part of the decision problem
- Missing in most discrete setups (Poisson)
- Invariance must somehow belong to prior setting

Opening towards left- and right-Haar measures at the end of $\S 3.10$. 

[ToP = Theory of Probability]
Invariant divergences

Interesting point made by Jeffreys that both
\[ L_m = \int \left| (dP)^{1/m} - (dP')^{1/m} \right|, \quad L^c = \int \log \frac{dP'}{dP} d(P' - P) \]

...are invariant for all non-singular transformations of \( x \) and of the parameters in the laws (III, §3.10)

Examples:

1. the entropy distance (or Kullback–Leibler divergence)
\[ L_e(\theta, \delta) = \mathbb{E}_\theta \left[ \log \left( \frac{f(x|\theta)}{f(x|\delta)} \right) \right], \]

2. the Hellinger distance
\[ L_H(\theta, \delta) = \frac{1}{2} \mathbb{E}_\theta \left[ \left( \sqrt{\frac{f(x|\delta)}{f(x|\theta)}} - 1 \right)^2 \right]. \]

Intrinsic losses

Noninformative settings w/o natural parameterisation: the estimators should be invariant under reparameterisation

[Ultimate invariance!]

Principle

Corresponding parameterisation-free loss functions:
\[ L(\theta, \delta) = d(f(\cdot|\theta), f(\cdot|\delta)), \]

Example (Normal mean)

Consider \( x \sim \mathcal{N}(\theta, 1) \). Then
\[ L_e(\theta, \delta) = \frac{1}{2} \mathbb{E}_\theta \left[ -(x - \theta)^2 + (x - \delta)^2 \right] = \frac{1}{2}(\delta - \theta)^2, \]
\[ L_H(\theta, \delta) = 1 - \exp\{- (\delta - \theta)^2/8 \}. \]

When \( \pi(\theta|x) \) is \( \mathcal{N}(\mu(x), \sigma^2) \), Bayes estimator of \( \theta \)
\[ \delta^\pi(x) = \mu(x) \]
in both cases.
Example (Normal everything)
Consider \( x \sim \mathcal{N}(\lambda, \sigma^2) \) then

\[
L_2((\lambda, \sigma), (\lambda', \sigma')) = 2 \sinh^2 \zeta + \cosh \zeta \frac{(\lambda - \lambda')^2}{\sigma_0^2} \\
L^\varepsilon((\lambda, \sigma), (\lambda', \sigma')) = 2 \left[ 1 - \text{sech}^{1/2} \zeta \exp \left\{ -\frac{(\lambda - \lambda')^2}{8 \sigma_0^2 \cosh \zeta} \right\} \right]
\]

if \( \sigma = \sigma_0 e^{-\zeta/2} \) and \( \sigma = \sigma_0 e^{+\zeta/2} \) (III, §3.10, (14) & (15))

The Jeffreys prior

Based on Fisher information

\[
I(\theta) = \mathbb{E}_\theta \left[ \frac{\partial \ell}{\partial \theta^T} \frac{\partial \ell}{\partial \theta} \right]
\]

The Jeffreys prior distribution is

\[
\pi^*(\theta) \propto |I(\theta)|^{1/2}
\]

Note

This general presentation is not to be found in ToP! And not all priors of Jeffreys’ are Jeffreys priors!

Where did Jeffreys hid his prior?!

Starts with second order approximation to both \( L_2 \) and \( L^\varepsilon \):

\[
4L_2(\theta, \theta') \approx (\theta - \theta')^T I(\theta)(\theta - \theta') \approx L^\varepsilon(\theta, \theta')
\]

This expression is therefore invariant for all non-singular transformations of the parameters. It is not known whether any analogous forms can be derived from \( [L_m] \) if \( m \neq 2 \). (III, §3.10)

The fundamental prior

...if we took the prior probability density for the parameters to be proportional to \( ||g_{ik}||^{1/2} \) \( = |I(\theta)|^{1/2} \), it could stated for any law that is differentiable with respect to all parameters that the total probability in any region of the \( \alpha_i \) would be equal to the total probability in the corresponding region of the \( \alpha'_i \); in other words, it satisfies the rule that equivalent propositions have the same probability (III, §3.10)

Jeffreys never mentions Fisher information in connection with \( (g_{ik}) \)

Main point

Fisher information equivariant under reparameterisation:

\[
\frac{\partial \ell}{\partial \eta^T} \frac{\partial \ell}{\partial \eta} = \frac{\partial \ell}{\partial \theta^T} \frac{\partial \ell}{\partial \theta} \times \frac{\partial \eta^T}{\partial \theta^T} \frac{\partial \eta}{\partial \theta}
\]
Jeffreys' objections

**Example (Normal everything)**

In the case of a normal \( \mathcal{N}(\lambda, \sigma^2) \), \(|I(\theta)|^{1/2} = 1/\sigma^2\) instead of the prior \(\pi(\theta) = 1/\sigma\) advocated earlier:

> If the same method was applied to a joint distribution for several variables about independent true values, an extra factor \(1/\sigma\) would appear for each. This is unacceptable: (...) \(\lambda\) and \(\sigma\) are each capable of any value over a considerable range and neither gives any appreciable information about the other (III, §3.10)

**Example (Variable support)**

If the support of \( f(\cdot|\theta) \) depends on \( \theta \), e.g. \( U([\theta_1, \theta_2]) \) usually no Fisher information [e.g. exception \( \propto (x - \theta_1)^a (\theta_2 - x)^b \) with \( a, b > 1 \)]. Jeffreys suggests to condition on non-differentiable parameters to derive a prior on the other parameters, and to use a flat prior on the bounds of the support.

**Example (Mixture model)**

For \( f(x|\theta) = \sum_{i=1}^{k} \omega_i f_i(x|\alpha_i) \), Jeffreys suggests to separate the \( \omega_i \)'s from the \( \alpha_i \)'s:

\[
\pi_J(\omega, \alpha) \propto \prod_{i=1}^{k} |I(\alpha_i)|^{1/2} / \sqrt{\omega_i}
\]

(36)

**Poisson distribution**

The Poisson parameter, however, is in rather a special position. It is usually the product of a scale factor with an arbitrary sample size, which is not chosen until we have already have some information about the probable range of values for the scale parameter. It does however point a warning for all designed experiments. The whole point of general rules for the prior probability is to give a starting-point, which we take to represent ignorance. They will not be correct if previous knowledge is being used (...) In the case of the Poisson law the sample size is chosen so that \( \lambda \) will be a moderate number, usually 1 to 10. The \( d\lambda/\lambda \) rule, in fact, may express complete ignorance of the scale parameter; but \( d\lambda/\sqrt{\lambda} \) may express just enough information to suggest that the experiment is worth making.
Exponential families

Jeffreys makes yet another exception for *Huzurbazar* distributions

\[ f(x) = \phi(\alpha)\psi(x)\exp\{u(\alpha)v(x)\} \]

namely exponential families.

Using the reparameterisation \( \beta = u(\alpha) \), he considers three cases

1. \( \beta \in (-\infty, +\infty) \), then \( \pi^*(\beta) \propto 1 \)
2. \( \beta \in (0, +\infty) \), then \( \pi^*(\beta) \propto 1/\beta \)
3. \( \beta \in (0, 1) \), then \( \pi^*(\beta) = 1/\beta(1 - \beta) \)

Pros & Cons

- Parameterization invariant
- Relates to information theory
- Agrees with most invariant priors (e.g., location/scale)
- Suffers from dimensionality curse (e.g., Jeffreys' correction)
- Not coherent for Likelihood Principle (e.g., Binomial versus Negative binomial)

Example (Normal norm)

\( x \sim \mathcal{N}_p(\theta, I_p) \), \( \eta = \|\theta\|^2 \), \( \pi(\eta) = \eta^{p/2-1} \)

\[ \mathbb{E}^\theta[\eta|x] = \|x\|^2 + p \quad \text{Bias} 2p \]

Example (Likelihood paradox)

If \( x \sim \mathcal{B}(n, \theta) \), Jeffreys’ prior is

\[ \mathcal{B}(1/2, 1/2) \]

and, if \( n \sim \mathcal{N}eg(x, \theta) \), Jeffreys’ prior is

\[
\pi_2(\theta) = -\mathbb{E}_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right] \\
= \mathbb{E}_\theta \left[ \frac{x}{\theta^2} + \frac{n - x}{(1 - \theta)^2} \right] = \frac{x}{\theta^2(1 - \theta)}, \\
\propto \theta^{-1}(1 - \theta)^{-1/2}
\]
Bernardo’s reference priors

Generalizes Jeffreys priors by distinguishing between nuisance and interest parameters

**Principle**: maximize the information brought by the data

\[
E^n \left[ \int \pi(\theta|x_n) \log(\pi(\theta|x_n)/\pi(\theta))d\theta \right]
\]

and consider the limit of the \( \pi_n \)

**Outcome**: most usually, Jeffreys prior

---

Example (Neyman–Scott problem)

Observation of \( x_{ij} \) iid \( \mathcal{N}(\mu_i, \sigma^2) \), \( i = 1, \ldots, n, \) \( j = 1,2 \).

The usual Jeffreys prior for this model is

\[
\pi(\mu_1, \ldots, \mu_n, \sigma) = \sigma^{-n-1}
\]

which is inconsistent because

\[
E[\sigma^2|x_{11}, \ldots, x_{n2}] = s^2/(2n - 2),
\]

where

\[
s^2 = \sum_{i=1}^{n} \frac{(x_{i1} - x_{i2})^2}{2},
\]

---

Example (Neyman–Scott problem (2))

Associated reference prior with \( \theta_1 = \sigma \) and \( \theta_2 = (\mu_1, \ldots, \mu_n) \) gives

\[
\pi(\theta_2|\theta_1) \propto 1,
\]

\[
\pi(\sigma) \propto 1/\sigma
\]

Therefore,

\[
E[\sigma^2|x_{11}, \ldots, x_{n2}] = s^2/(n - 2)
\]
Matching priors

**Frequency-validated priors:**
Some posterior probabilities
\[\pi(g(\theta) \in C_x|x) = 1 - \alpha\]
must coincide with the corresponding frequentist coverage
\[P_\theta(C_x \ni g(\theta)) = \int I_{C_x}(g(\theta)) f(x|\theta) \, dx,\]
...asymptotically

For instance, Welch and Peers’ identity
\[P_\theta(\theta \leq k_\alpha(x)) = 1 - \alpha + O(n^{-1/2})\]
and for Jeffreys’ prior,
\[P_\theta(\theta \leq k_\alpha(x)) = 1 - \alpha + O(n^{-1})\]

In general, choice of a matching prior dictated by the cancelation of a first order term in an Edgeworth expansion, like
\[\left[I''(\theta)\right]^{-1/2} I'(\theta) \nabla \log \pi(\theta) + \nabla^T \{I'(\theta)[I''(\theta)]^{-1/2}\} = 0.\]

Example (Linear calibration)
\[(i = 1, \ldots, n, j = 1, \ldots, k)\]
\[y_i = \alpha + \beta x_i + \varepsilon_i, \quad y_{0j} = \alpha + \beta x_0 + \varepsilon_{0j},\]
with \(\theta = (x_0, \alpha, \beta, \sigma^2)\) and \(x_0\) quantity of interest
Example (Linear calibration (2))

One-sided differential equation:

\[
|\beta|^{-1}s^{-1/2} \frac{\partial}{\partial x_0} \left\{ e(x_0)\pi(\theta) \right\} - e^{-1/2}(x_0)\text{sgn}(\beta)n^{-1}s^{1/2} \frac{\partial}{\partial x_0} \pi(\theta)
\]

\[
- e^{-1/2}(x_0)(x_0 - \bar{x})s^{-1/2} \frac{\partial}{\partial \beta} \{\text{sgn}(\beta)\pi(\theta)\} = 0
\]

with

\[
s = \sum(x_i - \bar{x})^2, \quad e(x_0) = \frac{[(n + k)s + nk(x_0 - \bar{x})^2]}{nk}
\]

Example (Linear calibration (3))

Solutions

\[
\pi(x_0, \alpha, \beta, \sigma^2) \propto e(x_0)^{d-1/2}|\beta|^d g(\sigma^2),
\]

where \( g \) arbitrary.

Other approaches

Reference priors

<table>
<thead>
<tr>
<th>Partition</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>((x_0, \alpha, \beta, \sigma^2))</td>
<td>(</td>
</tr>
<tr>
<td>(x_0, \alpha, \beta, \sigma^2)</td>
<td>(e(x_0)^{-1/2}(\sigma^2)^{-1})</td>
</tr>
<tr>
<td>(x_0, \alpha, (\sigma^2, \beta))</td>
<td>(e(x_0)^{-1/2}(\sigma^2)^{-3/2})</td>
</tr>
<tr>
<td>(x_0, (\alpha, \beta), \sigma^2)</td>
<td>(e(x_0)^{-1/2}(\sigma^2)^{-1})</td>
</tr>
<tr>
<td>(x_0, (\alpha, \beta, \sigma^2))</td>
<td>(e(x_0)^{-1/2}(\sigma^2)^{-2})</td>
</tr>
</tbody>
</table>

- Rissanen's transmission information theory and minimum
  length priors
- Testing priors
- Stochastic complexity
Fourth chapter: Approximate methods and simplifications

1. Fundamental notions
2. Direct Probabilities
3. Estimation problems
4. Asymptotics & DT& ...
   - Some asymptotics
   - Evaluation of estimators
   - Loss functions
   - Admissibility
   - Usual loss functions
4.1. Chapter summary

MAP

Equivalence of MAP and ML estimators:

...the differences between the values that make the likelihood and the posterior density maxima are only of order \(1/n\) (IV, §4.0)

...extrapolated into

...in the great bulk of cases the results of [the method of maximum likelihood] are undistinguishable from those given by the principle of inverse probability (IV, §4.0)

The tramcar comparison

A man travelling in a foreign country has to change trains at a junction, and goes into the town, of the existence of which he has just heard. The first thing that he sees is a tramcar numbered \(m = 100\). What can he infer about the number \(N\) of tramcars in the town? (IV, §4.8)

Famous opposition: Bayes posterior expectation vs. MLE

- Exclusion of flat prior on \(N\)
- Choice of the scale prior \(\pi(N) \propto 1/N\)
- MLE is \(\hat{N} = m\)

Under \(\pi(N) \propto 1/N + O(n^{-2})\), posterior is

\[\pi(N|m) \propto 1/N^2 + O(n^{-3})\]

and

\[P(N > n_0|m,H) = \sum_{n_0+1}^{\infty} n^{-2} / \sum_{m}^{\infty} n^{-2} = \frac{m}{n_0}\]

Therefore posterior median is \(2m\)

No mention made of either MLE or unbiasedness
Laplace analytic approximation

When integrating a regular function
\[
\mathbb{E}^\pi[g(\theta)|x] = \frac{\int_{\Theta} g(\theta) f(x|\theta)\pi(\theta) d\theta}{\int_{\Theta} f(x|\theta)\pi(\theta) d\theta} = \frac{\int_{\Theta} b_N(\theta) \exp\{-nh_N(\theta)\} d\theta}{\int_{\Theta} b_D(\theta) \exp\{-nh_D(\theta)\} d\theta},
\]
Laplace’s approximation given by
\[
\frac{\int_{\Theta} b_N(\theta) \exp\{-nh_N(\theta)\} d\theta}{\int_{\Theta} b_D(\theta) \exp\{-nh_D(\theta)\} d\theta} = \frac{\sigma_N}{\sigma_D} e^{-n(h_N-h_D)} \left[ \hat{b}_N \frac{\sigma_D^2}{2nb_D^2} \left( \hat{b}_D \hat{b}_D'' - \hat{b}_N \hat{b}_D'' \right) \right] + O(n^{-2}).
\]

Consequence

\[
\mathbb{E}^\pi[g(\theta)|x] = \hat{g} + \frac{\sigma_D^2 b'_D g'}{n b_D} + \frac{\sigma_D^2 g''}{2n} - \frac{\sigma_D^4 h'' g'}{2n} + O(n^{-2}).
\]

Example (Binomial model)

\[
\pi(\theta|x) \text{ density of } \mathcal{B}(\alpha, \beta) \text{ distribution and posterior expectation of } \theta
\]
\[
\delta^\pi(x) = \frac{\alpha}{\alpha + \beta},
\]
compared with
\[
\delta^\pi(x) = \frac{\alpha^2 + \alpha \beta + 2 - 4\alpha}{(\alpha + \beta - 2)^2} + O((\alpha + \beta - 2)^{-2}),
\]

Fighting un-sufficiency

When maximum likelihood estimators not easily computed (e.g., outside exponential families), Jeffreys suggests use of Pearson’s minimum \(\chi^2\) estimation, which is a form of MLE for multinomial settings.

Asymptotic difficulties of

In practice, (...) it is enough to group [observations] so that there are no empty groups, \(m_r\) for a terminal group being calculated for a range extending to infinity (IV, §4.1)

bypassed in ToP

Unbiasedness

Searching for unbiased estimators presented in §4.3 as a way of fighting un-sufficiency and attributed to Neyman and Pearson.

Introduction of Decision Theory via a multidimensional loss function:

There are apparently an infinite number of unbiased statistics associated with any law (...) The estimates of \(\alpha, \beta, \ldots\) obtained will therefore be \(a, b, \ldots\) which differ little from \(\alpha, \beta, \ldots\) The choice is then made so that all of \(E(\alpha - a)^2, E(\beta - b)^2, \ldots\) will be as small as possible

Note the first sentence above: meaningless!
Besides, unbiasedness is a property almost never shared by Bayesian estimators!
Evaluating estimators

**Purpose of most inferential studies**

To provide the statistician/client with a decision \( d \in D \)

Requires an evaluation criterion for decisions and estimators

\[
L(\theta, d)
\]

[a.k.a. loss function]

Bayesian Decision Theory

**Three spaces/factors:**

1. On \( X \), distribution for the observation, \( f(x|\theta) \);
2. On \( \Theta \), prior distribution for the parameter, \( \pi(\theta) \);
3. On \( \Theta \times D \), loss function associated with the decisions, \( L(\theta, \delta) \);

Foundations

**Theorem (Existence)**

There exists an axiomatic derivation of the existence of a loss function.

[DeGroot, 1970]

Estimators

Decision procedure \( \delta \) usually called estimator (while its value \( \delta(x) \) called estimate of \( \theta \))

**Fact**

Impossible to uniformly minimize (in \( d \)) the loss function

\[
L(\theta, d)
\]

when \( \theta \) is unknown
Frequentist Principle

Average loss (or frequentist risk)

\[ R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(x))] = \int_X L(\theta, \delta(x)) f(x|\theta) \, dx \]

Principle

Select the best estimator based on the risk function

Difficulties with frequentist paradigm

(1) Error averaged over the different values of \( x \) proportionally to the density \( f(x|\theta) \): not so appealing for a client, who wants optimal results for her data \( x \)!

(2) Assumption of repeatability of experiments not always grounded.

(3) \( R(\theta, \delta) \) is a function of \( \theta \): there is no total ordering on the set of procedures.

Bayesian principle

Principle

Integrate over the space \( \Theta \) to get the posterior expected loss

\[ \rho(\pi, d|x) = \mathbb{E}^\pi[L(\theta, d)|x] = \int_\Theta L(\theta, d)\pi(\theta|x) \, d\theta, \]

Bayesian principle (2)

Alternative

Integrate over the space \( \Theta \) and compute integrated risk

\[ r(\pi, \delta) = \mathbb{E}^\pi[R(\theta, \delta)] = \int_\Theta \int_X L(\theta, \delta(x)) f(x|\theta) \, dx \, \pi(\theta) \, d\theta \]

which induces a total ordering on estimators.

Existence of an optimal decision
Bayes estimator

**Theorem (Construction of Bayes estimators)**

An estimator minimizing \( r(\pi, \delta) \)

can be obtained by selecting, for every \( x \in \mathcal{X} \), the value \( \delta(x) \) which minimizes

\[
\rho(\pi, \delta|x)
\]

since

\[
r(\pi, \delta) = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x) m(x) \, dx.
\]

Both approaches give the same estimator

Infinite Bayes risk

Above result valid for both proper and improper priors when

\[
r(\pi) < \infty
\]

Otherwise, generalized Bayes estimator that must be defined pointwise:

\[
\delta^\pi(x) = \arg \min_d \rho(\pi, d|x)
\]

if \( \rho(\pi, d|x) \) is well-defined for every \( x \).

**Warning:** Generalized Bayes \( \neq \) Improper Bayes

---

Bayes estimator (2)

**Definition (Bayes optimal procedure)**

A Bayes estimator associated with a prior distribution \( \pi \) and a loss function \( L \) is

\[
\arg \min_\delta r(\pi, \delta)
\]

The value \( r(\pi) = r(\pi, \delta^\pi) \) is called the Bayes risk

---

Admissibility

Reduction of the set of acceptable estimators based on “local” properties

**Definition (Admissible estimator)**

An estimator \( \delta_0 \) is inadmissible if there exists an estimator \( \delta_1 \) such that, for every \( \theta \),

\[
R(\theta, \delta_0) \geq R(\theta, \delta_1)
\]

and, for at least one \( \theta_0 \)

\[
R(\theta_0, \delta_0) > R(\theta_0, \delta_1)
\]

Otherwise, \( \delta_0 \) is admissible

---

Theory of Probability revisited
Asymptotics & DT & ...
Loss functions

Theory of Probability revisited
Asymptotics & DT & ...
Loss functions

Theory of Probability revisited
Admissibility
The Bayesian perspective

Admissibility strongly related to the Bayes paradigm: Bayes estimators often constitute the class of admissible estimators

- If $\pi$ is strictly positive on $\Theta$, with

$$r(\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta) \, d\theta < \infty$$

and $R(\theta, \delta)$, is continuous, then the Bayes estimator $\delta^\pi$ is admissible.

- If the Bayes estimator associated with a prior $\pi$ is unique, it is admissible.

Regular (≠ generalized) Bayes estimators always admissible

Example (Normal mean)

Consider $x \sim \mathcal{N}(\theta, 1)$ and the test of $H_0: \theta \leq 0$, i.e. the estimation of $I_{H_0}(\theta)$

Under the loss

$$(I_{H_0}(\theta) - \delta(x))^2,$$

the estimator ($p$-value)

$$p(x) = P_0(X > x) \quad (X \sim \mathcal{N}(0, 1))$$

$$= 1 - \Phi(x),$$

is Bayes under Lebesgue measure.

Example (Normal mean (2))

Indeed

$$p(x) = \mathbb{E}_{\pi}[I_{H_0}(\theta)|x] = P^\pi(\theta < 0|x)$$

$$= P^\pi(\theta - x < -x|x) = 1 - \Phi(x).$$

The Bayes risk of $p$ is finite and $p(s)$ is admissible.

Example (Normal mean (3))

Consider $x \sim \mathcal{N}(\theta, 1)$. Then $\delta_0(x) = x$ is a generalised Bayes estimator, is admissible, but

$$r(\pi, \delta_0) = \int_{-\infty}^{+\infty} R(\theta, \delta_0) \, d\theta$$

$$= \int_{-\infty}^{+\infty} 1 \, d\theta = +\infty.$$
Example (Normal mean (4))

Consider $x \sim \mathcal{N}_p(\theta, I_p)$. If

$$L(\theta, d) = (d - ||\theta||^2)^2$$

the Bayes estimator for the Lebesgue measure is

$$\delta_{\pi}(x) = ||x||^2 + p.$$ 

This estimator is not admissible because it is dominated by

$$\delta_0(x) = ||x||^2 - p.$$ 

The reason for using the expectation of the square of the error as the criterion is that, given a large number of observations, the probability of a set of statistics given the parameters, and that of the parameters given the statistics, are usually distributed approximately on a normal correlation surface (IV, §4.3)

Explan.: Asymptotic normal distribution of MLE

Proper loss

The Bayes estimator $\delta_{\pi}$ associated with the prior $\pi$ and with the quadratic loss is the posterior expectation

$$\delta_{\pi}(x) = \mathbb{E}_{\pi}[\theta|x] = \frac{\int_{\Theta} \theta f(x|\theta)\pi(\theta) d\theta}{\int_{\Theta} f(x|\theta)\pi(\theta) d\theta}.$$ 

Orthogonal parameters

Interesting disgression: reparameterise the parameter set so that Fisher information is (nearly) diagonal.

...the quadratic term in $E(\log L)$ will reduce to a sum of squares (IV, §4.3)

But this is local orthogonality: the diagonal terms in $I(\theta)$ may still depend on all parameters and Jeffreys distinguishes global orthogonality where each diagonal term only depends on one $\beta_i$ and thus induces an independent product for the Jeffreys prior.

Generally impossible, even though interesting for dealing with nuisance parameters...
The absolute error loss

Alternatives to the quadratic loss:

\[ L(\theta, d) = |\theta - d|, \]

or

\[ L_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{if } \theta > d, \\ k_1(d - \theta) & \text{otherwise}. \end{cases} \]

\textit{L}_1 \text{ estimator}

The Bayes estimator associated with \( \pi \) and \( L_{k_1,k_2} \) is a \( \left( \frac{k_2}{k_1 + k_2} \right) \) fractile of \( \pi(\theta|x) \).

\text{Posterior median}

Relates to Jeffreys'

\[ \ldots \text{we can use the median observation as a statistic for the median of the law (IV, \S 4.4)} \]

even though it lacks DT justification

\text{Example (Median law)}

If

\[ P(dx|m, a, H) \propto \frac{1}{2} \exp \left( -\frac{|x - m|}{a} \right) \frac{dx}{a} \]

the likelihood is maximum if \( m \) is taken equal to the median observation and if \( a \) is the average residual without regard to sign.

\[ \text{[\'tis Laplace's law]} \]

It is only subject to that law that the average residual leads to the best estimate of uncertainty, and then the best estimate of location is provided by the median observation and not by the mean (IV, \S 4.4)

\text{ToP 4 bric-à-brac!}

Sequence of remarks and cases

- Difficulty when no sufficient statistics
- Model/law misspecification (departure from normality)
- Random effect
- Randomization (contradiction to later Bayesian persp's)
- Rank tests (Spearman: \textit{It is an estimate but what is it an estimate of?})

with very little relevance to either Bayesian methodology or DT...

Maybe a reflection on computational difficulties?

No trace whatsoever of Bayesian estimation???
Chapter 5: Significance tests: one new parameter

Fundamental setting

"Is the new parameter supported by the observations or is any variation expressible by it better interpreted as random? Thus we must set two hypotheses for comparison, the more complicated having the smaller initial probability (V, §5.0)"

[Occam’s rule again!]

...compare a specially suggested value of a new parameter, often 0 \([q]\), with the aggregate of other possible values \([q']\). We shall call \(q\) the null hypothesis and \(q'\) the alternative hypothesis [and] we must take

\[P(q|H) = P(q'|H) = 1/2.\]

Construction of Bayes tests

Definition (Test)

Given an hypothesis \(H_0 : \theta \in \Theta_0\) on the parameter \(\theta \in \Theta_0\) of a statistical model, a test is a statistical procedure that takes its values in \(\{0, 1\}\).

Example (Normal mean)

For \(x \sim \mathcal{N}(\theta, 1)\), decide whether or not \(\theta \leq 0\).
The 0 − 1 loss

Neyman–Pearson loss for testing hypotheses

Test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$.
Then
$$D = \{0, 1\}$$

The Bayes estimator associated with $\pi$ and with the 0 − 1 loss is
$$\delta^{\pi}(x) = \begin{cases} 1 & \text{if } P(\theta \in \Theta_0|x) > P(\theta \notin \Theta_0|x), \\ 0 & \text{otherwise,} \end{cases}$$

Jeffreys’ example (§5.0)

Testing whether the mean $\alpha$ of a normal observation is zero:

$$P(q|aH) \propto \exp \left( -\frac{a^2}{2s^2} \right)$$
$$P(q|d\alpha|aH) \propto \exp \left( -\frac{(a - \alpha)^2}{2s^2} \right) f(\alpha)d\alpha$$
$$P(q'|aH) \propto \int \exp \left( -\frac{(a - \alpha)^2}{2s^2} \right) f(\alpha)d\alpha$$

Type–one and type–two errors

Associated with the risk
$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(x))]$$
$$= \begin{cases} P_0(\delta(x) = 0) & \text{if } \theta \in \Theta_0, \\ P_0(\delta(x) = 1) & \text{otherwise}, \end{cases}$$

A point of contention

Jeffreys asserts

Suppose that there is one old parameter $\alpha$; the new parameter is $\beta$ and is 0 on $q$. In $q'$ we could replace $\alpha$ by $\alpha'$, any function of $\alpha$ and $\beta$: but to make it explicit that $q'$ reduces to $q$ when $\beta = 0$ we shall require that $\alpha' = \alpha$ when $\beta = 0$ (V, §5.0).

This amounts to assume identical parameters in both models, a controversial principle for model choice (see Chapter 6) or at the very best to make $\alpha$ and $\beta$ dependent a priori, a choice contradicted by the following paragraphs!
Orthogonal parameters

If

\[ I(\alpha, \beta) = \begin{bmatrix} g_{\alpha\alpha} & 0 \\ 0 & g_{\beta\beta} \end{bmatrix}, \]

\( \alpha \) and \( \beta \) orthogonal, but not [a posteriori] independent, contrary to ToP assertions

...the result will be nearly independent on previous information on old parameters (V, §5.01).

and

\[ K = \frac{1}{f(b, a)} \sqrt{\frac{ng_{\beta\beta}}{2\pi}} \exp \left( -\frac{1}{2} ng_{\beta\beta} b^2 \right) \]

[where] \( h(\alpha) \) is irrelevant (V, §5.01)

Acknowledgement in ToP

In practice it is rather unusual for a set of parameters to arise in such a way that each can be treated as irrelevant to the presence of any other. More usual cases are (...) where some parameters are so closely associated that one could hardly occur without the others (V, §5.04).

Bound comparison

Determination of \( a_0/a_1 \) depends on consequences of “wrong decision” under both circumstances

Often difficult to assess in practice and replacement with “golden” bounds like .05, biased towards \( H_0 \)

Example (Binomial probability)

Consider \( x \sim B(n, p) \) and \( \Theta_0 = [0, 1/2] \). Under the uniform prior \( \pi(p) = 1 \), the posterior probability of \( H_0 \) is

\[ P^\pi(p \leq 1/2|x) = \int_0^{1/2} \frac{(1/2)^{n+1}}{B(x+1, n-x+1)} \left\{ \frac{1}{x+1} + \ldots + \frac{(n-x)!x!}{(n+1)!} \right\} dp \]

\[ \frac{B(x+1, n-x+1)}{(n+1)!} \left\{ \frac{1}{x+1} + \ldots + \frac{(n-x)!x!}{(n+1)!} \right\} \]

Theory of Probability revisited
Significance tests: one new parameter
Bayesian tests
Loss/prior duality

Decomposition

\[ \Pr^R(\theta \in \Theta_0|x) = \int_{\Theta_0} \pi(\theta|x) \, d\theta \]

\[ = \int_{\Theta_0} f(x|\theta_0)\pi(\theta) \, d\theta \]

suggests representation

\[ \pi(\theta) = \pi(\Theta_0)\pi_0(\theta) + (1 - \pi(\Theta_0))\pi_1(\theta) \]

and decision

\[ \delta^R(x) = 1 \text{ iff } \frac{\pi(\Theta_0)}{(1 - \pi(\Theta_0))} \int_{\Theta_0} f(x|\theta_0)\pi_0(\theta) \, d\theta \geq a_0/\ a_1 \]

What matters is \((\pi(\Theta_0)/a_0, (1 - \pi(\Theta_0))/a_1)\)

Self-contained concept

Outside decision-theoretic environment:

- eliminates choice of \(\pi(\Theta_0)\)
- but depends on the choice of \((\pi_0, \pi_1)\)
- Bayesian/marginal equivalent to the likelihood ratio

Jeffreys’ scale of evidence (Appendix B):

- if \(\log_{10}(B_{10}) < 0\) null \(H_0\) supported
- if \(\log_{10}(B_{10})\) between 0 and 0.5, evidence against \(H_0\) weak,
- if \(\log_{10}(B_{10})\) 0.5 and 1, evidence substantial,
- if \(\log_{10}(B_{10})\) 1 and 1.5, evidence strong,
- if \(\log_{10}(B_{10})\) 1.5 and 2, evidence very strong and
- if \(\log_{10}(B_{10})\) above 2, evidence decisive

A function of posterior probabilities

Definition (Bayes factors)

For hypotheses \(H_0 : \theta \in \Theta_0\) vs. \(H_a : \theta \notin \Theta_0\)

\[ B_{01} = \frac{\pi(\Theta_0|x)}{\pi(\Theta_0)} \int_{\Theta_0} f(x|\theta_0)\pi(\theta) \, d\theta \]

\[ \left/ \int_{\Theta_0} f(x|\theta)\pi_1(\theta) \, d\theta \right. \]

[Good, 1958 & ToP, V, §5.01]

Equivalent to Bayes rule: acceptance if

\[ B_{01} > \{(1 - \pi(\Theta_0))/a_1\}/\{\pi(\Theta_0)/a_0\} \]

Example (Binomial homogeneity)

Consider \(H_0 : y_i \sim \text{Po}(n_i, \omega)\) \((i = 1, \ldots, G)\) vs. \(H_1 : y_i \sim \text{Be}(\alpha = \xi/\omega, \beta = (1 - \xi)/\omega)), \) with a uniform prior on \(E[p_i|\xi, \omega] = \xi\) and on \(p (\omega \text{ is fixed})\)

\[ B_{10} = \int_{0}^{1} \cdots \int_{0}^{1} \prod_{i=1}^{G} p_i^{y_i} (1 - p_i)^{n_i - y_i} p_i^{\alpha - 1}(1 - p_i)^{\beta - 1} \, dp_i \]

\[ \times \Gamma(1/\omega) / \Gamma(\xi/\omega)\Gamma((1 - \xi)/\omega)) \, d\xi \]

\[ \int_{0}^{1} p \sum_i y_i (1 - p) \prod_i (n_i - y_i) \, dp \]

For instance, \(\log_{10}(B_{10}) = -0.79\) for \(\omega = 0.005\) and \(G = 138\) slightly favours \(H_0\).

[Kass & Raftery, 1995]
Multiple alternatives (§5.03)

If \( q' = q_1 \cup \cdots \cup q_m \), and if \( P(q'|H) = 1/2 \), then, if \( P(q_i|H) = \kappa \),

\[
(1 - \kappa)^m = \frac{1}{2}
\]

and

\[
\frac{P(q|H)}{P(q_i|H)} \approx \frac{m}{2 \log 2} = 0.7m
\]

© If testing for a separate hypothesis \( q_i \), Bayes factor \( B_{0i} \) multiplied by 0.7m

A major modification

When the null hypothesis is supported by a set of measure 0,
\( \pi(\Theta_0) = 0 \)

[End of the story?!]

Suppose we are considering whether a location parameter \( \alpha \) is 0. The estimation prior probability for it is uniform

and we should have to take \( f(\alpha) = 0 \) and \( K = B_{10} \)

would always be infinite (V, §5.02)

Contingency table (§5.11)

Then the alternatives, the sampling numbers, and the chances may be shown as follows:

\[
\begin{pmatrix}
\phi \psi & \tilde{\phi} \tilde{\psi} \\
\tilde{\phi} \tilde{\psi} & \phi \psi
\end{pmatrix},
\begin{pmatrix}
x & y \\
\tilde{x} \tilde{y}
\end{pmatrix},
\begin{pmatrix}
p_{11} & p_{12} \\
p_{21} & p_{22}
\end{pmatrix}.
\]

If \( \phi \) and \( \psi \) are in proportion we have hypothesis \( q \), that

\[
p_{11}p_{22} = p_{12}p_{21}.
\]
Under $q$, 
\[
\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} \alpha \beta & \alpha(1 - \beta) \\ (1 - \alpha)\beta & (1 - \alpha)(1 - \beta) \end{pmatrix}
\]
and under $q'$, 
\[
\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} \alpha \beta + \gamma & \alpha(1 - \beta) - \gamma \\ (1 - \alpha)\beta - \gamma & (1 - \alpha)(1 - \beta) + \gamma \end{pmatrix}
\]
If $\alpha \leq \beta \leq \frac{1}{2}$, then $-\alpha \beta \leq \gamma \leq \alpha(1 - \beta)$.

In general, it should be 
\[-\{\alpha \beta \land (1 - \alpha)(1 - \beta)\} \leq \gamma \leq \{\alpha(1 - \beta) \land (1 - \alpha)\beta\}\]
Then 
\[
\pi_1(\alpha, \beta, \gamma) = \frac{1}{\min(\alpha, 1 - \alpha, \beta, 1 - \beta)} \times I_{\{1 - \alpha \beta \land (1 - \alpha)(1 - \beta)\}}(\gamma)
\]
and 
\[
P(q|\theta H) \propto \frac{(x + y)!(x' + y')!(x + x')!(y + y')!}{\{(x + y + x' + y')^2\}}
\]

...but ToP gets it wrong when integrating in $P(q'|\theta H)$ since it keeps dividing by $\alpha$ rather than by $\min(\alpha, 1 - \alpha, \beta, 1 - \beta)$...

Obvious ToP difficulty in computing 
\[
\int (\alpha \beta + \gamma)^x(\alpha(1 - \beta) - \gamma)^y((1 - \alpha)\beta - \gamma)^x'((1 - \alpha)(1 - \beta) + \gamma)^y' \pi_1(d\alpha, d\beta, d\gamma)
\]

MC resolution

1. Simulate $(\alpha, \beta, \gamma) \sim \pi_1(\alpha, \beta, \gamma)$
2. Average 
\[
(\alpha \beta + \gamma)^x(\alpha(1 - \beta) - \gamma)^y((1 - \alpha)\beta - \gamma)^x'((1 - \alpha)(1 - \beta) + \gamma)^y'
\]

...data on the conviction of twin brothers or sisters (of like sex) of convicted criminals, according as the twins were monozygotic (identical) or dizygotic (no more alike physically than ordinary brothers or sisters)

<table>
<thead>
<tr>
<th></th>
<th>Monogozytic</th>
<th>Dizygotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convicted</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Not convicted</td>
<td>3</td>
<td>15</td>
</tr>
</tbody>
</table>

Then 
\[
K = \frac{1}{171}
\]

(...) we can assert on the data that the odds on the existence of a difference are about 171 to 1 (V, §5.14)
Point null hypotheses

Particular case $H_0 : \theta = \theta_0$
Take $\rho_0 = \Pr(\theta = \theta_0)$ and $g_1$ prior density under $H_a$.
Posterior probability of $H_0$

$$\pi(\Theta_0 | x) = \frac{f(x|\theta_0)\rho_0}{\int f(x|\theta)\pi(\theta) \, d\theta} = \frac{f(x|\theta_0)\rho_0}{f(x|\theta_0)\rho_0 + (1 - \rho_0)m_1(x)}$$

and marginal under $H_a$

$$m_1(x) = \int_{\Theta_1} f(x|\theta)g_1(\theta) \, d\theta.$$ 

Dual representation

$$\pi(\Theta_0 | x) = \left[1 + \frac{1 - \rho_0}{\rho_0} \frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$ 

and

$$B_{01}^\pi(x) = \frac{f(x|\theta_0)\rho_0}{m_1(x)(1 - \rho_0)} \frac{\rho_0}{1 - \rho_0} = \frac{f(x|\theta_0)}{m_1(x)}.$$ 

Connection

$$\pi(\Theta_0 | x) = \left[1 + \frac{1 - \rho_0}{\rho_0} \frac{1}{B_{01}^\pi(x)}\right]^{-1}.$$ 

Example (Normal mean)

Test of $H_0 : \theta = 0$ when $x \sim \mathcal{N}(\theta, 1)$: we take $\pi_1$ as $\mathcal{N}(0, \tau^2)$

$$m_1(x) = \frac{\sigma}{\sqrt{\sigma^2 + \tau^2}} \frac{e^{-x^2/2(\sigma^2 + \tau^2)}}{e^{-x^2/2\sigma^2}}$$

$$= \sqrt{\frac{\sigma^2}{\sigma^2 + \tau^2}} \exp \left\{ -\frac{\tau^2 x^2}{2\sigma^2(\sigma^2 + \tau^2)} \right\}$$

and

$$\pi(\theta = 0 | x) = \left[1 + \frac{1 - \rho_0}{\rho_0} \sqrt{\frac{\sigma^2}{\sigma^2 + \tau^2}} \exp \left( \frac{\tau^2 x^2}{2\sigma^2(\sigma^2 + \tau^2)} \right) \right]^{-1}.$$ 

Influence of $\tau$:

<table>
<thead>
<tr>
<th>$\tau/x$</th>
<th>0</th>
<th>0.68</th>
<th>1.28</th>
<th>1.96</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.586</td>
<td>0.557</td>
<td>0.484</td>
<td>0.351</td>
</tr>
<tr>
<td>10</td>
<td>0.768</td>
<td>0.729</td>
<td>0.612</td>
<td>0.366</td>
</tr>
</tbody>
</table>
### A fundamental difficulty

**Improper priors are not allowed here**

If

\[ \int_{\Theta_1} \pi_1(d\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(d\theta_2) = \infty \]

then either \( \pi_1 \) or \( \pi_2 \) cannot be coherently normalised but the normalisation matters in the Bayes factor. Recall Bayes factor.

### Constants matter

**Example (Poisson versus Negative binomial)**

If \( \mathcal{M}_1 \) is a \( \mathcal{P}(\lambda) \) distribution and \( \mathcal{M}_2 \) is a \( \mathcal{NB}(m, p) \) distribution, we can take

\[
\begin{align*}
\pi_1(\lambda) & = \frac{1}{\lambda} \\
\pi_2(m, p) & = \frac{1}{M} \sum_{\{1, \ldots, M\}} (m - x)！p^x(1 - p)^{m - x} dp
\end{align*}
\]

### Constants matter (cont’d)

**Example (Poisson versus Negative binomial (2))**

then

\[
B_{12}^p = \frac{\int_0^\infty \frac{\lambda^{x-1}}{x!} e^{-\lambda} d\lambda}{\frac{1}{M} \sum_{m=1}^M \int_0^\infty \left( \frac{m}{x - 1} \right) p^x(1 - p)^{m - x} dp}
\]

\[
= \frac{1}{\frac{1}{M} \sum_{m=x}^{M} \left( \frac{m}{x - 1} \right)x!(m - x)！}{m!}
\]

\[
= \frac{1}{\frac{1}{M} \sum_{m=x}^{M} x/(m - x + 1)}
\]

### Constants matter (cont’d)

**Example (Poisson versus Negative binomial (3))**

- does not make sense because \( \pi_1(\lambda) = 10/\lambda \) leads to a different answer, ten times larger!
- same thing when both priors are improper

Improper priors on common (nuisance) parameters do not matter (so much)
Take $x \sim \mathcal{N}(\theta, 1)$ and $H_0 : \theta = 0$

<table>
<thead>
<tr>
<th>Influence of the constant $\pi(\theta)/x$</th>
<th>0.0</th>
<th>1.0</th>
<th>1.65</th>
<th>1.96</th>
<th>2.58</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi(\theta)/x$</td>
<td>0.285</td>
<td>0.195</td>
<td>0.089</td>
<td>0.055</td>
<td>0.014</td>
</tr>
<tr>
<td>10</td>
<td>0.0384</td>
<td>0.0236</td>
<td>0.0101</td>
<td>0.00581</td>
<td>0.00143</td>
</tr>
</tbody>
</table>

Example of testing for a zero normal mean:

If $\sigma$ is the standard error and $\lambda$ the true value, $\lambda$ is 0 on $q$. We want a suitable form for its prior on $q'$. (...) Then we should take

$$P(q\sigma|H) \propto d\sigma/\sigma$$
$$P(q'|\sigma d\lambda|H) \propto f\left(\frac{\lambda}{\sigma}\right) d\sigma/\sigma d\lambda/\lambda$$

where $f$ [is a true density] (V, §5.2).

**Fallacy of the “same” $\sigma$!**

If $s' = 0$ [!!!], then [for $\sigma = |\bar{x}|/\tau$, $\lambda = \sigma v$]

$$P(q|\theta H) \propto \int_0^\infty \left(\frac{\tau}{|\bar{x}|}\right)^n \exp\left(-\frac{1}{2}n\tau^2\right) d\tau,$$
$$P(q'|\theta H) \propto \int_0^\infty \frac{d\tau}{\tau} \int_{-\infty}^\infty \left(\frac{\tau}{|\bar{x}|}\right)^n f(v) \exp\left(-\frac{1}{2}n(v-\tau)^2\right).$$

If $n = 1$ and $f(v)$ is any even [density],

$$P(q'|\theta H) \propto \frac{1}{2} \sqrt{2\pi \over |\bar{x}|}$$
and $P(q|\theta H) \propto \frac{1}{2} \sqrt{2\pi \over |\bar{x}|}$

and therefore $K = 1$ (V, §5.2).

This is the prior recommended by Jeffreys hereafter.

**But**, first, many other families of densities satisfy this constraint and a scale of 1 cannot be universal!

Second, $s' = 0$ is a zero probability event...
Further puzzlements!

When taking two normal sample $x_{11}, \ldots, x_{1n}$ and $x_{21}, \ldots, x_{2n}$ with means $\lambda_1$ and $\lambda_2$ and same variance $\sigma$, testing for $H_0: \lambda_1 = \lambda_2$ suddenly gets outwordly:

...we are really considering four hypotheses, not two as in the test for agreement of a location parameter with zero; for neither may be disturbed, or either, or both may.

**ToP** then uses parameters $(\lambda, \sigma)$ in all versions of the alternative hypotheses, with

$$
\pi_0(\lambda, \sigma) \propto 1/\sigma
$$

$$
\pi_1(\lambda, \sigma, \lambda_1) \propto 1/\pi\{\sigma^2 + (\lambda_1 - \lambda)^2\}
$$

$$
\pi_2(\lambda, \sigma, \lambda_2) \propto 1/\pi\{\sigma^2 + (\lambda_2 - \lambda)^2\}
$$

$$
\pi_{12}(\lambda, \sigma, \lambda_1, \lambda_2) \propto \sigma/\pi^2\{\sigma^2 + (\lambda_1 - \lambda)^2\}\{\sigma^2 + (\lambda_2 - \lambda)^2\}
$$

**ToP misses the points that**

1. $\lambda$ not have the same meaning under $q$, under $q_1 (= \lambda_2)$ and under $q_2 (= \lambda_1)$
2. $\lambda$ has no precise meaning under $q_{12}$ [hyperparameter?]
3. On $q_{12}$, since $\lambda$ does not appear explicitly in the likelihood we can integrate it (V, §5.41).
4. even $\sigma$ has a varying meaning over hypotheses
5. integrating over measures is meaningless!

$$
P(q_{12|\theta} \sigma \lambda_1 \lambda_2|H) \propto \frac{2}{\pi} \frac{d\sigma d\lambda_1 d\lambda_2}{4\sigma^2 + (\lambda_1 - \lambda_2)^2}
$$

simply defines a new prior...

Further, erases the fake complexity in the end:

*But there is so little to choose between the alternatives that we may as well combine them (V, §5.41).*

**Vague proper priors are not the solution**

Similar confusion in following sections (§5.42 — §5.45): the use of improper priors in testing settings simply does not make sense because ... constants matter!

Note also the aggravating effect of the multiple alternatives (e.g., §5.46):

$$
P(q'|\theta H) = P(q_1|\theta H) + P(q_2|\theta H) + P(q_{12}|\theta H)
$$

which put more weight on $q'$
Lindley’s paradox

Example (Normal case)
If testing $H_0 : \theta = 0$
when observing $x \sim \mathcal{N}(\theta, 1)$,
under a normal $\mathcal{N}(0, \alpha)$ prior
$B_{01}(x) \xrightarrow{\alpha \to \infty} 0$

Often dubbed Jeffreys–Lindley paradox...

In terms of
$$t = \sqrt{n - 1} \bar{x} / s', \quad \nu = n - 1$$
$$K \sim \sqrt{\frac{\pi \nu}{2}} \left(1 + \frac{t^2}{\nu}\right)^{-1/2\nu+1/2}$$

(... The variation of $K$ with $t$ is much more important than the variation with $\nu$ (V, §5.2).

But ToP misses the point that under $H_0$ $t \sim T_\nu$ so does not vary much with $\nu$ while $\nu$ goes to $\infty$...

Vague proper priors are not the solution (cont’d)

Example (Poisson versus Negative binomial (4))

$B_{12} = \int_0^1 \frac{\lambda^{\alpha+x-1} e^{-\lambda \beta}}{x! \Gamma(\alpha)} d\lambda$
if $\lambda \sim \mathcal{G}(\alpha, \beta)$

$= \frac{\Gamma(\alpha + x)}{x! \Gamma(\alpha)} \beta^{-x} \left/ \frac{1}{M} \sum_m \frac{x}{m - x + 1} \right.$
$= \frac{(x + \alpha - 1) \cdots \alpha}{x(x-1) \cdots 1} \beta^{-x} \left/ \frac{1}{M} \sum_m \frac{x}{m - x + 1} \right.$
$
$\text{depends on choice of } \alpha(\beta) \text{ or } \beta(\alpha) \to 0$

Learning from the sample

Definition (Learning sample)
Given an improper prior $\pi$, $(x_1, \ldots, x_n)$ is a learning sample if
$\pi(\cdot | x_1, \ldots, x_n)$ is proper and a minimal learning sample if none of its subsamples is a learning sample

There is just enough information in a minimal learning sample to make inference about $\theta$ under the prior $\pi$
Pseudo-Bayes factors

**Idea**

Use one part $x_{[i]}$ of the data $x$ to make the prior proper:
- $\pi_i$ improper but $\pi_i(\cdot|x_{[i]})$ proper
- and
  $$\frac{\int f_i(x_{[n/i]}|\theta_i) \pi_i(\theta_i|x_{[i]})d\theta_i}{\int f_j(x_{[n/i]}|\theta_j) \pi_j(\theta_j|x_{[i]})d\theta_j}$$

independent of normalizing constant
- Use remaining $x_{[n/i]}$ to run test as if $\pi_j(\theta_j|x_{[i]})$ is the true prior

**Details**

Since $\pi_1(\theta_1|x_{[i]}) = \frac{\pi_1(\theta_1) f^1_1(x_{[i]}|\theta_1)}{\int \pi_1(\theta_1) f^1_1(x_{[i]}|\theta_1) d\theta_1}$

$$B_{12}(x_{[n/i]}) = \frac{\int f^1_{[n/i]}(x_{[n/i]}|\theta_1) \pi_1(\theta_1|x_{[i]})d\theta_1}{\int f^2_{[n/i]}(x_{[n/i]}|\theta_2) \pi_2(\theta_2|x_{[i]})d\theta_2}$$

$$= \int f_1(x|\theta_1) \pi_1(\theta_1)d\theta_1 \int \pi_2(\theta_2) f^2_1(x_{[i]}|\theta_2)d\theta_2$$

$$= \int f_2(x|\theta_2) \pi_2(\theta_2)d\theta_2 \int \pi_1(\theta_1) f^1_1(x_{[i]}|\theta_1)d\theta_1$$

$$= B^N_{12}(x) B_{21}(x_{[i]})$$

© Independent of scaling factor!

**Motivation**

- Provides a working principle for improper priors
- Gather enough information from data to achieve properness
- and use this properness to run the test on remaining data
- does not use $x$ twice as in Aitkin’s (1991)

**Unexpected problems!**

- depends on the choice of $x_{[i]}$
- many ways of combining pseudo-Bayes factors
  - $AIBF = B^N_{ji} \frac{1}{\sum_{\ell} B_{ij}(x_{[i]})}$
  - $MIBF = B^N_{ji} \text{med}[B_{ij}(x_{[i]})]$
  - $GIBF = B^N_{ji} \exp \frac{1}{\sum_{\ell} \log B_{ij}(x_{[i]})}$
- not often an exact Bayes factor
- and thus lacking inner coherence

$$B_{12} \neq B_{10} B_{02} \quad \text{and} \quad B_{01} \neq 1/B_{10}.$$  

[Berger & Pericchi, 1996]
Unexpec'd problems (cont'd)

Example (Mixtures)

There is no sample size that proper-ises improper priors, except if a training sample is allocated to each component.

**Reason**

If

\[ x_1, \ldots, x_n \sim \sum_{i=1}^{k} p_i f(x|\theta_i) \]

and

\[ \pi(\theta) = \prod_{i} \pi_i(\theta_i) \text{ with } \int \pi_i(\theta_i) d\theta_i = +\infty, \]

the posterior is never defined, because

\[ \Pr(\text{"no observation from } f(\cdot|\theta_i)\text{"}) = (1 - p_i)^n \]

Intrinsic priors

There may exist a true prior that provides the same Bayes factor.

Example (Normal mean)

Take \( x \sim \mathcal{N}(\theta, 1) \) with either \( \theta = 0 \) (\( \mathcal{M}_1 \)) or \( \theta \neq 0 \) (\( \mathcal{M}_2 \)) and \( \pi_2(\theta) = 1 \).

Then

\[ B_{21}^{\text{IBF}} = B_{21} \frac{1}{\sqrt{2\pi}} \frac{1}{n} \sum_{i=1}^{n} e^{-x_i^2/2} \approx B_{21} \text{ for } \mathcal{N}(0, 2) \]
\[ B_{21}^{\text{MIBF}} = B_{21} \frac{1}{\sqrt{2\pi}} e^{-\text{med}(x_i^2)/2} \approx 0.93B_{21} \text{ for } \mathcal{N}(0, 1.2) \]

When such a prior exists, it is called an intrinsic prior.

Intrinsic priors (cont'd)

Example (Exponential scale)

Take

\( x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \exp(\theta - x) \mathbb{I}_{x \geq \theta} \)

and \( H_0 : \theta = \theta_0, H_1 : \theta > \theta_0 \), with \( \pi_1(\theta) = 1 \).

Then

\[ B_{10}^{\text{IBF}} = B_{10}(x) \frac{1}{n} \sum_{i=1}^{n} \left[ e^{x_i - \theta_0} - 1 \right]^{-1} \]

is the Bayes factor for

\[ \pi_2(\theta) = e^{\theta_0 - \theta} \left\{ 1 - \log \left( 1 - e^{\theta_0 - \theta} \right) \right\} \]

Most often, however, the pseudo-Bayes factors do not correspond to any true Bayes factor.

Fractional Bayes factor

**Idea**

use directly the likelihood to separate training sample from testing sample

\[ B_{12}^{\text{F}} = B_{12}(x) \frac{\int L_2(\theta_2) \pi_2(\theta_2) d\theta_2}{\int L_1(\theta_1) \pi_1(\theta_1) d\theta_1} \]

[O'Hagan, 1995]

Proportion \( b \) of the sample used to gain proper-ness.

[Berger and Pericchi, 2001]
Fractional Bayes factor (cont’d)

Example (Normal mean)

\[ B_{12}^F = \frac{1}{\sqrt{b}} e^{n(b-1)\bar{x}_n^2/2} \]

This corresponds to the exact Bayes factor for the prior \( \mathcal{N}(0, \frac{1}{nb}) \):

- If \( b \) constant, prior variance goes to 0.
- If \( b = \frac{1}{n} \), prior variance stabilises around 1.
- If \( b = n^{-\alpha}, \alpha < 1 \), prior variance goes to 0 too.

Random effect models

In ToP, systematic errors (V, §5.6) correspond to random effect models:

\[ x_{ij} = y_{ij} + \epsilon_i \]

with \( y_{ij} \sim \mathcal{N}(\lambda, \sigma^2) \) and \( \epsilon_i \sim \mathcal{N}(0, \tau^2) \).

Test of a systematic error is then equivalent to testing \( \tau = 0 \).

But use of \( \chi^2 \) test and MLE’s....!

Comparison with classical tests

\( p \)-value

Definition (\( p \)-value)

The \( p \)-value \( p(x) \) associated with a test is the largest significance level for which \( H_0 \) is rejected.

Note

An alternative definition is that a \( p \)-value is distributed uniformly under the null hypothesis.

Example (Normal mean)

Since the UUMP test is \( \{|x| > k\} \), standard \( p \)-value:

\[ p(x) = \inf \{ \alpha; |x| > k_\alpha \} \]

\[ = P^X(|X| > |x|), \quad X \sim \mathcal{N}(0, 1) \]

\[ = 1 - \Phi(|x|) + \Phi(|x|) = 2[1 - \Phi(|x|)] \]

Thus, if \( x = 1.68 \), \( p(x) = 0.10 \) and, if \( x = 1.96 \), \( p(x) = 0.05 \).
Problems with $p$-values

- Evaluation of the **wrong** quantity, namely the probability to exceed the observed quantity. (Wrong condition in)
  
  *What the use of $P$ implies, therefore, is that a hypothesis that may be true may be rejected because it had not predicted observable results that have not occurred* (VII, §7.2)

- No transfer of the UMP optimality
- No decisional support (occurrences of inadmissibility)
- Evaluation only under the null hypothesis
- Huge numerical difference with the Bayesian range of answers

Bayesian lower bounds

For illustration purposes, consider a class $\mathcal{G}$ of prior distributions

\[
B(x, \mathcal{G}) = \inf_{g \in \mathcal{G}} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta_0)g(\theta)d\theta},
\]

\[
P(x, \mathcal{G}) = \inf_{g \in \mathcal{G}} \frac{f(x|\theta_0) + \int_\Theta f(x|\theta_0)g(\theta)d\theta}{\int_\Theta f(x|\theta_0)g(\theta)d\theta}
\]

when $\theta_0 = 1/2$ or

\[
B(x, \mathcal{G}) = \frac{f(x|\theta_0)}{\sup_{g \in \mathcal{G}} \int_\Theta f(x|\theta)g(\theta)d\theta}, \quad P(x, \mathcal{G}) = \left[1 + \frac{1}{B(x, \mathcal{G})}\right]^{-1}.
\]

Resolution

**Lemma**

*If there exists a mle for $\theta$, $\hat{\theta}(x)$, the solutions to the Bayesian lower bounds are*

\[
B(x, \mathcal{G}) = \frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))}, \quad PP(x, \mathcal{G}) = \left[1 + \frac{f(x|\hat{\theta}(x))}{f(x|\theta_0)}\right]^{-1}
\]

*respectively*

Normal case

When $x \sim \mathcal{N}(\theta, 1)$ and $H_0 : \theta = 0$, the lower bounds are

\[
B(x, G_A) = e^{-x^2/2} \quad \text{and} \quad P(x, G_A) = \left(1 + e^{x^2/2}\right)^{-1},
\]

i.e.

<table>
<thead>
<tr>
<th>$p$-value</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>0.205</td>
<td>0.128</td>
<td>0.035</td>
<td>0.004</td>
</tr>
<tr>
<td>$B$</td>
<td>0.256</td>
<td>0.146</td>
<td>0.036</td>
<td>0.004</td>
</tr>
</tbody>
</table>

[Quite different!]
Unilateral case

Different situation when $H_0 : \theta \leq 0$
- Single prior can be used both for $H_0$ and $H_a$
- Improper priors are therefore acceptable
- Similar numerical values compared with $p$-values

Unilateral agreement

**Theorem**

When $x \sim f(x - \theta)$, with $f$ symmetric around 0 and endowed with the monotone likelihood ratio property, if $H_0 : \theta \leq 0$, the $p$-value $p(x)$ is equal to the lower bound of the posterior probabilities, $P(x, \mathcal{G}_{SU})$, when $\mathcal{G}_{SU}$ is the set of symmetric unimodal priors and when $x > 0$.

Reason:

$$p(x) = P_{\theta=0}(X > x) = \int_x^{+\infty} f(t) \, dt = \inf_K \frac{1}{1 + \left[ \frac{\int_{-K}^{0} f(x-\theta) \, d\theta}{\int_{-K}^{K} f(x-\theta) \, d\theta} \right]^{-1}}$$

Cauchy example

When $x \sim \mathcal{C}(\theta, 1)$ and $H_0 : \theta \leq 0$, lower bound inferior to $p$-value:

<table>
<thead>
<tr>
<th>$p$-value</th>
<th>0.437</th>
<th>0.102</th>
<th>0.063</th>
<th>0.013</th>
<th>0.004</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>0.429</td>
<td>0.077</td>
<td>0.044</td>
<td>0.007</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Comments

- **ToP** very imprecise about choice of priors in the setting of tests
- **ToP** misses the difficulty of improper priors [coherent with earlier stance]
- but this problem still generates debates within the B community
- Some degree of goodness-of-fit testing but against fixed alternatives
- Persistence of the form

$$K \approx \sqrt{\frac{\pi n}{2}} \left( 1 + \frac{t^2}{\nu} \right)^{-1/2\nu+1/2}$$

but $\nu$ not so clearly defined...
Chapter 6: Significance tests: various complications

1. Fundamental notions
2. Direct Probabilities
3. Estimation problems
4. Asymptotics & DT& ...
5. Significance tests: one new parameter
6. Significance tests: various complications
7. Frequency definitions and direct methods

Contents of Chapter 6

Certainly not a foundational chapter!!!

Some elementary remarks like

Anything that alters the prior probability of [the alternative parameter] will alter the inferences about \( q' \) (VI, §6.0)

and

One further possibility is that \( q \) and \( q' \) may not be initially equally probable (VI, §6.0)

Repetition of the multiple alternatives

With several parameters under \( H_1 \), there are several embedded alternatives:

If the parameters are \( \alpha, \beta \), we can write \( q \) for the proposition \( \alpha = \beta = 0 \), \( q_\alpha \) for \( \alpha \neq 0, \beta = 0 \), \( q_\beta \) for \( \alpha =, \beta \neq 0 \), and \( q_{\alpha\beta} \) for \( \alpha \neq 0, \beta \neq 0 \) (VI, §6.1).

Difficulty to order \( q_\alpha \) and \( q_\beta \) reminiscent of Bernardo’s reference priors but things get worse...

There is a best order of procedures, which is to assert the [hypothesis] that is most strongly supported, reject those that are denied and proceed to consider further combinations (VI, §6.12)
Re-enter Ockham explicitly!

### Pluralitas non est ponenda sine neccesitate

*Variation is random until the contrary is shown; and new parameters in laws, when they are suggested, must be tested one at a time, unless there is specific reason to the contrary. (...) This principle is workable and is a complete reversal of the usual notion of a 'principle of causality' (VI, §6.12)*

---

---

### Case of two location parameters (VI, §6.2)

*ToP suggests to use a Cauchy prior $C(0, \sigma^2)$ on the radius $\rho$ and a uniform prior on the angle

Similar shape of the Bayes factor

$$K \approx n^{1/2} \pi^{-2} t \left(\nu + t^2\right)^{-\nu/2+1}$$

---

---

### Interesting extensions

Some hints at

- Hierarchical modelling (§6.3)

$$x_s \sim f(x - \alpha \ell), \quad \alpha \ell \sim \tau g(\{\alpha \ell - \alpha\}/\tau)$$

- Hidden Markov model (§6.4)

$$P = \begin{bmatrix}
\alpha + (1 - \alpha)p_1 & (1 - \alpha)p_2 & \cdots & (1 - \alpha)p_r \\
(1 - \alpha)p_1 & \alpha + (1 - \alpha)p_2 & \cdots & (1 - \alpha)p_r \\
\vdots & \vdots & \ddots & \vdots \\
(1 - \alpha)p_1 & (1 - \alpha)p_2 & \cdots & \alpha + (1 - \alpha)p_r
\end{bmatrix}$$

---

---

### Un-interesting disgressions

Section §6.5 very windy about the nature of deduction and the approximation of point null hypotheses by interval representations

...by extending the meaning of $q$ so as to say that the new parameter is not 0 but may be anywhere in some finite range. (...) I think, however, that it is both impossible and undesirable. (...) If there is anything to suggest a range of possible values it should go into the statement of $q'$, not of $q$ (VI, §6.5).
Model choice and model comparison

Choice of models
Several models available for the same observation

\[ M_i : x \sim f_i(x|\theta_i), \quad i \in I \]

where \( I \) can be finite or infinite

Example (Galaxy normal mixture)
Set of observations of radial speeds of 82 galaxies possibly modelled as a mixture of normal distributions

\[ M_i : x_j \sim \sum_{\ell=1}^{i} p_{\ell i} N(\mu_{\ell i}, \sigma_{\ell i}^2) \]

Bayesian resolution

B Framework
Probabilises the entire model/parameter space
This means:
- allocating probabilities \( p_i \) to all models \( M_i \)
- defining priors \( \pi_i(\theta_i) \) for each parameter space \( \Theta_i \)

Resolution

1. Compute

\[
p(M_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i) \pi_i(\theta_i) d\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j) \pi_j(\theta_j) d\theta_j}
\]

2. Take largest \( p(M_i|x) \) to determine “best” model, or use averaged predictive

\[
\sum_j p(M_j|x) \int_{\Theta_j} f_j(x'|\theta_j) \pi_j(\theta_j|x) d\theta_j
\]
Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences
  - representation of parsimony/sparcity (Ockham’s rule)
  - how to fight overfitting for nested models

  **Which loss?**

Several types of problems (2)

- **Choice of prior structures**
  - adequate weights $p_i$:
    - if $\mathcal{M}_1 = \mathcal{M}_2 \cup \mathcal{M}_3$, $p(\mathcal{M}_1) = p(\mathcal{M}_2) + p(\mathcal{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathcal{I}$
    - $\pi_i(\theta_i)$ proper (Jeffreys)
    - $\pi_i(\theta_i)$ coherent (?) for nested models

**Warning**
Parameters common to several models must be treated as separate entities!

Several types of problems (3)

- Computation of predictives and marginals
  - infinite dimensional spaces
  - integration over parameter spaces
  - integration over different spaces
  - summation over many models ($2^k$)

Compatibility principle

Difficulty of finding simultaneously priors on a collection of models $\mathcal{M}_i$ ($i \in \mathcal{I}$)
Easier to start from a single prior on a “big” model and to derive the others from a coherence principle

[Dawid & Lauritzen, 2000]
**Projection approach**

For $M_2$ submodel of $M_1$, $\pi_2$ can be derived as the distribution of $\theta_2^{\perp}(\theta_1)$ when $\theta_1 \sim \pi_1(\theta_1)$ and $\theta_2^{\perp}(\theta_1)$ is a projection of $\theta_1$ on $M_2$, e.g.

$$d(f(\cdot | \theta_1), f(\cdot | \theta_1^{\perp})) = \inf_{\theta_2 \in \Theta_2} d(f(\cdot | \theta_1), f(\cdot | \theta_2)),$$

where $d$ is a divergence measure

[McCulloch & Rossi, 1992]

Or we can look instead at the posterior distribution of

$$d(f(\cdot | \theta_1), f(\cdot | \theta_1^{\perp}))$$

[Goutis & Robert, 1998]

**Operational principle for variable selection**

**Selection rule**

Among all subsets $A$ of covariates such that

$$d(M_g, M_A) = \mathbb{E}_x[d(f_g(\cdot | x, \alpha), f_A(\cdot | x_A, \alpha^{\perp}))] < \epsilon$$

select the submodel with the smallest number of variables.

[Dupuis & Robert, 2001]

**Kullback proximity**

Alternative to above

**Definition (Compatible prior)**

Given a prior $\pi_1$ on a model $M_1$ and a submodel $M_2$, a prior $\pi_2$ on $M_2$ is *compatible* with $\pi_1$ when it achieves the minimum Kullback divergence between the corresponding marginals:

$m_1(x; \pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)d\theta$ and $m_2(x; \pi_2) = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)d\theta$,

$$\pi_2 = \arg \min_{\pi_2} \int \log \left( \frac{m_1(x; \pi_1)}{m_2(x; \pi_2)} \right) m_1(x; \pi_1) dx$$

**Difficulties**

- Does not give a working principle when $M_2$ is not a submodel of $M_1$
- Depends on the choice of $\pi_1$
- Prohibits the use of improper priors
- Worse: useless in unconstrained settings...
Case of exponential families

Models

\[ \mathcal{M}_1 : \{ f_1(x|\theta), \theta \in \Theta \} \]

and

\[ \mathcal{M}_2 : \{ f_2(x|\lambda), \lambda \in \Lambda \} \]

sub-model of \( \mathcal{M}_1 \),

\[ \forall \lambda \in \Lambda, \exists \theta(\lambda) \in \Theta, f_2(x|\lambda) = f_1(x|\theta(\lambda)) \]

Both \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \) are natural exponential families

\[
\begin{align*}
n_1(x|\theta) &= h_1(x) \exp(\theta^T t_1(x) - M_1(\theta)) \\
n_2(x|\lambda) &= h_2(x) \exp(\lambda^T t_2(x) - M_2(\lambda))
\end{align*}
\]

Conjugate compatible priors

\( (Q.) \) Existence and unicity of Kullback-Leibler projection

\[
(s_2^*, n_2^*) = \arg \min_{(s_2, n_2)} \mathcal{R}(m_1(\cdot; s_1, n_1), m_2(\cdot; s_2, n_2))
\]

\[
= \arg \min_{(s_2, n_2)} \int \log \left( \frac{m_1(x; s_1, n_1)}{m_2(x; s_2, n_2)} \right) m_1(x; s_1, n_1)dx
\]

A sufficient condition

Sufficient statistic \( \psi = (\lambda, -M_2(\lambda)) \)

Theorem (Existence)

If, for all \((s_2, n_2)\), the matrix

\[
\mathbb{V}^{s_2, n_2}_{s_2, n_2}[\psi] - \mathbb{E}^{m_1}_{s_1, n_1}[\mathbb{V}^{s_2, n_2}_{s_2, n_2}(\psi|x)]
\]

is semi-definite negative, the conjugate compatible prior exists, is unique and satisfies

\[
\begin{align*}
\mathbb{E}^{s_2, n_2}_{s_2, n_2}[\Lambda] - \mathbb{E}^{m_1}_{s_1, n_1}[\mathbb{E}^{s_2, n_2}_{s_2, n_2}(\Lambda|x)] &= 0 \\
\mathbb{E}^{s_2, n_2}_{s_2, n_2}(M_2(\lambda)) - \mathbb{E}^{m_1}_{s_1, n_1}[\mathbb{E}^{s_2, n_2}_{s_2, n_2}(M_2(\lambda)|x)] &= 0.
\end{align*}
\]
An application to linear regression

\( \mathcal{M}_1 \) and \( \mathcal{M}_2 \) are two nested Gaussian linear regression models with Zellner’s \( g \)-priors and the same variance \( \sigma^2 \sim \pi(\sigma^2) \):

- **\( \mathcal{M}_1 \)**:
  \[
  y | \beta_1, \sigma^2 \sim N(X_1 \beta_1, \sigma^2), \quad \beta_1 | \sigma^2 \sim N\left(s_1, \sigma^2 n_1 (X_1^T X_1)^{-1}\right)
  \]
  where \( X_1 \) is a \((n \times k_1)\) matrix of rank \( k_1 \leq n \)

- **\( \mathcal{M}_2 \)**:
  \[
  y | \beta_2, \sigma^2 \sim N(X_2 \beta_2, \sigma^2), \quad \beta_2 | \sigma^2 \sim N\left(s_2, \sigma^2 n_2 (X_2^T X_2)^{-1}\right)
  \]
  where \( X_2 \) is a \((n \times k_2)\) matrix with \( \text{span}(X_2) \subseteq \text{span}(X_1) \)

For a fixed \((s_1, n_1)\), we need the projection \((s_2, n_2) = (s_1, n_1)^\perp\)

### Compatible \( g \)-priors

Since \( \sigma^2 \) is a nuisance parameter, we can minimize the Kullback-Leibler divergence between the two marginal distributions conditional on \( \sigma^2 \): \( m_1(y | \sigma^2; s_1, n_1) \) and \( m_2(y | \sigma^2; s_2, n_2) \)

**Theorem**

**Conditional on \( \sigma^2 \), the conjugate compatible prior of \( \mathcal{M}_2 \) wrt \( \mathcal{M}_1 \) is**

\[
\beta_2 | X_2, \sigma^2 \sim N\left(s_2^*, \sigma^2 n_2^* (X_2^T X_2)^{-1}\right)
\]

with

\[
s_2^* = (X_2^T X_2)^{-1} X_2^T X_1 s_1 \\
n_2^* = n_1
\]

### Variable selection

Regression setup where \( y \) regressed on a set \( \{x_1, \ldots, x_p\} \) of \( p \) potential explanatory regressors (plus intercept)

Corresponding \( 2^p \) submodels \( \mathcal{M}_\gamma \), where \( \gamma \in \{0, 1\}^p \) indicates inclusion/exclusion of variables by a binary representation, e.g. \( \gamma = 101001011 \) means that \( x_1, x_3, x_5, x_7 \) and \( x_8 \) are included.

For model \( \mathcal{M}_\gamma \),

- \( q_\gamma \) variables included
- \( t_1(\gamma) = \{t_{1,1}(\gamma), \ldots, t_{1,q_\gamma}(\gamma)\} \) indices of those variables and \( t_0(\gamma) \) indices of the variables not included
- For \( \beta \in \mathbb{R}^{p+1} \),
  \[
  \beta_{t_1(\gamma)} = \left[\beta_0, \beta_{t_1,1}(\gamma), \ldots, \beta_{t_1,q_\gamma}(\gamma)\right]
  \]
  \[
  X_{t_1(\gamma)} = \left[1_n | x_{t_1,1}(\gamma) | \ldots | x_{t_1,q_\gamma}(\gamma)\right].
  \]

Submodel \( \mathcal{M}_\gamma \) is thus

\[
y | \beta, \gamma, \sigma^2 \sim N(X_{t_1(\gamma)} \beta_{t_1(\gamma)} , \sigma^2 I_n)
\]
Global and compatible priors

Use Zellner's $g$-prior, i.e. a normal prior for $\beta$ conditional on $\sigma^2$,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^TX)^{-1})$$

and a Jeffreys prior for $\sigma^2$,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^TX_{t_1(\gamma)}\right)^{-1}X_{t_1(\gamma)}^T\tilde{\beta}, c\sigma^2(X_{t_1(\gamma)}^TX_{t_1(\gamma)})^{-1}\right)$$

Posterior model probability

Can be obtained in closed form:

$$\pi(\gamma|y) \propto (c+1)^{-(q+1)/2} \left[ y^Ty - \frac{cy^TP_1y}{c+1} + \tilde{\beta}^TX^TP_1X\tilde{\beta} - \frac{2y^TP_1X\tilde{\beta}}{c+1} \right]^{-n/2}.$$ 

Conditionally on $\gamma$, posterior distributions of $\beta$ and $\sigma^2$:

$$\beta_{t_1(\gamma)}|\sigma^2, y, \gamma \sim \mathcal{N}\left[\frac{c}{c+1}(U_1y + U_1X\tilde{\beta}/c), \frac{\sigma^2c}{c+1}(X_{t_1(\gamma)}^TX_{t_1(\gamma)})^{-1}\right],$$

$$\sigma^2|y, \gamma \sim \mathcal{I}\mathcal{G}\left[\frac{n}{2}, \frac{y^Ty}{2(c+1)} - \frac{cy^TP_1y}{2(c+1)} + \frac{\tilde{\beta}^TX^TP_1X\tilde{\beta}}{2(c+1)} - \frac{y^TP_1X\tilde{\beta}}{c+1}\right].$$

Noninformative case

Use the same compatible informative $g$-prior distribution with $\tilde{\beta} = 0_{p+1}$ and a hierarchical diffuse prior distribution on $c$,

$$\pi(c) \propto c^{-1}\mathbb{I}_{(N^*)}(c)$$

The choice of this hierarchical diffuse prior distribution on $c$ is due to the model posterior sensitivity to large values of $c$:

Taking $\tilde{\beta} = 0_{p+1}$ and $c$ large does not work
Influence of $c$

Consider the 10-predictor full model

$$y|\beta, \sigma^2 \sim N\left(\beta_0 + \sum_{i=1}^{10} \beta_i x_i + \sum_{i=1}^{10} \beta_{i+1} x_i^2 + \beta_{i+2} x_i x_{i+1} + \beta_{i+3} x_{i+2} x_{i+3} + \beta_{i+4} x_{i+2} x_{i+3} + \beta_{i+5} x_{i+1} x_{i+2} x_{i+3}, \sigma^2 \mathbf{I}_{10}\right)$$

where the $x_i$s are iid $\mathcal{U}(0,10)$

True model: two predictors $x_1$ and $x_2$, i.e. $\gamma^* = 110 \ldots 0$, $(\beta_0, \beta_1, \beta_2) = (5, 1, 3)$, and $\sigma^2 = 4$.

Casella & Moreno's example

In the noninformative setting,

$$\pi(\gamma|y) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q+1)/2} \left[ y^T y - \frac{c}{c+1} y^T \mathbf{P}_1 y \right]^{-n/2}$$

converges for all $y$'s

<table>
<thead>
<tr>
<th>$t_1(\gamma)$</th>
<th>$c = 10$</th>
<th>$c = 100$</th>
<th>$c = 10^2$</th>
<th>$c = 10^4$</th>
<th>$c = 10^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,1</td>
<td>0.04062</td>
<td>0.35368</td>
<td>0.65858</td>
<td>0.85895</td>
<td>0.98222</td>
</tr>
<tr>
<td>0,1,2</td>
<td>0.01326</td>
<td>0.06142</td>
<td>0.08395</td>
<td>0.04434</td>
<td>0.00524</td>
</tr>
<tr>
<td>0,1,2,4</td>
<td>0.01299</td>
<td>0.05310</td>
<td>0.05805</td>
<td>0.02868</td>
<td>0.00336</td>
</tr>
<tr>
<td>0,2,4</td>
<td>0.02927</td>
<td>0.03962</td>
<td>0.00409</td>
<td>0.00246</td>
<td>0.00254</td>
</tr>
<tr>
<td>0,1,2,8</td>
<td>0.01240</td>
<td>0.03833</td>
<td>0.01100</td>
<td>0.00126</td>
<td>0.00126</td>
</tr>
</tbody>
</table>

Noninformative case (cont’d)

In the noninformative setting,

$$\pi(\gamma|y) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q+1)/2} \left[ y^T y - \frac{c}{c+1} y^T \mathbf{P}_1 y \right]^{-n/2}$$

converges for all $y$'s
When \( p \) large, impossible to compute the posterior probabilities of the \( 2^p \) models.

Use of a Monte Carlo approximation of \( \pi(\gamma|y) \)

**Gibbs sampling**

- At \( t = 0 \), draw \( \gamma^0 \) from the uniform distribution on \( \Gamma \)
- At \( t \), for \( i = 1, \ldots, p \), draw
  \[
  \gamma_t^i \sim \pi(\gamma_i|y, \gamma_t^{1}, \ldots, \gamma_t^{i-1}, \ldots, \gamma_t^{i+1}, \ldots, \gamma_t^{p-1})
  \]

**Example (Simulated data)**

Severe multicolinearities among predictors for a 20-predictor full model

\[
y|\beta, \sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{20} \beta_i x_i, \sigma^2 \mathbf{I}_n\right)
\]

where \( x_i = z_i + 3z \), the \( z_i \)‘s and \( z \) are iid \( \mathcal{N}(0, \mathbf{I}_n) \).

True model with \( n = 180 \), \( \sigma^2 = 4 \) and seven predictor variables

\[
x_1, x_3, x_5, x_6, x_{12}, x_{18}, x_{20},
\]

\[
(\beta_0, \beta_1, \beta_3, \beta_5, \beta_6, \beta_{12}, \beta_{18}, \beta_{20}) = (3, 4, 1, -3, 12, -1, 5, -6)
\]

**Example (Simulated data (2))**

| \( \gamma \) | \( \pi(\gamma|y) \) | \( \hat{\pi}(\gamma|y) \) |
|---------------|----------------|------------------|
| 0,1,3,5,6,12,18,20 | 0.1893 | 0.1822 |
| 0,1,3,5,6,18,20 | 0.0588 | 0.0598 |
| 0,1,3,5,6,9,12,18,20 | 0.0223 | 0.0236 |
| 0,1,3,5,6,12,14,18,20 | 0.0220 | 0.0193 |
| 0,1,2,3,5,6,12,18,20 | 0.0216 | 0.0222 |
| 0,1,3,5,6,7,12,18,20 | 0.0212 | 0.0233 |
| 0,1,3,5,6,10,12,18,20 | 0.0199 | 0.0222 |
| 0,1,3,4,5,6,12,18,20 | 0.0197 | 0.0182 |
| 0,1,3,5,6,12,15,18,20 | 0.0196 | 0.0196 |

Gibbs (\( T = 100,000 \)) results for \( \tilde{\beta} = 0_{21} \) and \( c = 100 \)

**Processionary caterpillar**

Influence of some forest settlement characteristics on the development of caterpillar colonies

Response \( y \) log-transform of the average number of nests of caterpillars per tree on an area of 500 square meters (\( n = 33 \) areas)
Processionary caterpillar (cont’d)

Potential explanatory variables

\begin{align*}
  &x_1 \text{ altitude (in meters)}, \ x_2 \text{ slope (in degrees)}, \\
  &x_3 \text{ number of pines in the square}, \\
  &x_4 \text{ height (in meters) of the tree at the center of the square}, \\
  &x_5 \text{ diameter of the tree at the center of the square}, \\
  &x_6 \text{ index of the settlement density}, \\
  &x_7 \text{ orientation of the square (from 1 if southb’d to 2 ow)}, \\
  &x_8 \text{ height (in meters) of the dominant tree}, \\
  &x_9 \text{ number of vegetation strata}, \\
  &x_{10} \text{ mix settlement index (from 1 if not mixed to 2 if mixed)}. \\
\end{align*}

Bayesian regression output

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>BF</th>
<th>log10(BF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>9.2714</td>
<td>26.334</td>
<td>1.4205</td>
</tr>
<tr>
<td>X1</td>
<td>-0.0037</td>
<td>7.0839</td>
<td>0.8502</td>
</tr>
<tr>
<td>X2</td>
<td>-0.0454</td>
<td>3.6850</td>
<td>0.5664</td>
</tr>
<tr>
<td>X3</td>
<td>0.0573</td>
<td>0.4356</td>
<td>-0.3609</td>
</tr>
<tr>
<td>X4</td>
<td>-1.0905</td>
<td>2.8314</td>
<td>0.4520</td>
</tr>
<tr>
<td>X5</td>
<td>0.1953</td>
<td>2.5157</td>
<td>0.4007</td>
</tr>
<tr>
<td>X6</td>
<td>-0.3008</td>
<td>0.3621</td>
<td>-0.4412</td>
</tr>
<tr>
<td>X7</td>
<td>-0.2002</td>
<td>0.3627</td>
<td>-0.4404</td>
</tr>
<tr>
<td>X8</td>
<td>0.1526</td>
<td>0.4589</td>
<td>-0.3383</td>
</tr>
<tr>
<td>X9</td>
<td>-1.0835</td>
<td>0.9069</td>
<td>-0.0424</td>
</tr>
<tr>
<td>X10</td>
<td>-0.3651</td>
<td>0.4132</td>
<td>-0.3838</td>
</tr>
</tbody>
</table>

Evidence against H0: (****) decisive, (*** ) strong, (**) substantial, (*) poor

Bayesian variable selection

|     | t₁(γ) | π(γ|y,X) | ̂π(γ|y,X) |
|-----|-------|----------|-----------|
| 0,1,2,4,5 | 0.0929 | 0.0929   |
| 0,1,2,4,5,9 | 0.0325 | 0.0326   |
| 0,1,2,4,5,10 | 0.0295 | 0.0272   |
| 0,1,2,4,5,7 | 0.0231 | 0.0231   |
| 0,1,2,4,5,8 | 0.0228 | 0.0229   |
| 0,1,2,4,5,6 | 0.0228 | 0.0226   |
| 0,1,2,3,4,5 | 0.0224 | 0.0220   |
| 0,1,2,3,4,5,9 | 0.0167 | 0.0182   |
| 0,1,2,4,5,6,9 | 0.0167 | 0.0171   |
| 0,1,2,4,5,8,9 | 0.0137 | 0.0130   |

Noninformative G-prior model choice and Gibbs estimations
Postulate

Previous principle requires embedded models (or an encompassing model) and proper priors, while being hard to implement outside exponential families. Now we determine prior measures on two models $M_1$ and $M_2$, $\pi_1$ and $\pi_2$, directly by a compatibility principle.

Generalised expected posterior priors

**EPP Principle**

Starting from reference priors $\pi_1^N$ and $\pi_2^N$, substitute by prior distributions $\pi_1$ and $\pi_2$ that solve the system of integral equations

$$\pi_1(\theta_1) = \int_{\mathcal{X}} \pi_1^N(\theta_1 | x)m_2(x)dx$$

and

$$\pi_2(\theta_2) = \int_{\mathcal{X}} \pi_2^N(\theta_2 | x)m_1(x)dx,$$

where $x$ is an imaginary minimal training sample and $m_1$, $m_2$ are the marginals associated with $\pi_1$ and $\pi_2$ respectively.

Motivations

- Eliminates the “imaginary observation” device and proper-isation through part of the data by integration under the “truth”
- Assumes that both models are equally valid and equipped with ideal unknown priors $\pi_i$, $i = 1, 2,$ that yield “true” marginals balancing each model wrt the other
- For a given $\pi_1$, $\pi_2$ is an expected posterior prior

Using both equations introduces symmetry into the game.

Dual properness

**Theorem (Proper distributions)**

If $\pi_1$ is a probability density then $\pi_2$ solution to

$$\pi_2(\theta_2) = \int_{\mathcal{X}} \pi_2^N(\theta_2 | x)m_1(x)dx$$

is a probability density

© Both EPPs are either proper or improper
Bayesian coherence

**Theorem (True Bayes factor)**

If $\pi_1$ and $\pi_2$ are the EPPs and if their marginals are finite, then the corresponding Bayes factor

$$B_{1,2}(x)$$

is either a (true) Bayes factor or a limit of (true) Bayes factors.

Obviously only interesting when both $\pi_1$ and $\pi_2$ are improper.

Existence/Unicity

**Theorem (Recurrence condition)**

When both the observations and the parameters in both models are continuous, if the Markov chain with transition

$$Q(\theta'_1 | \theta_1) = \int g(\theta_1, \theta'_1, \theta_2, x, x') \, dx \, dx' \, d\theta_2$$

where

$$g(\theta_1, \theta'_1, \theta_2, x, x') = \pi_N(\theta'_1 | x) f_2(x | \theta_2) \pi_N(\theta_2 | x') f_1(x' | \theta_1),$$

is recurrent, then there exists a solution to the integral equations, unique up to a multiplicative constant.

Consequences

- If the M chain is positive recurrent, there exists a unique pair of proper EPPs.
- The transition density $Q(\theta'_1 | \theta_1)$ has a dual transition density on $\Theta_2$.
- There exists a parallel M chain on $\Theta_2$ with identical properties; if one is (Harris) recurrent, so is the other.
- **Duality property** found both in the MCMC literature and in decision theory
  
  [Diebolt & Robert, 1992; Eaton, 1992]

- When Harris recurrence holds but the EPPs cannot be found, the Bayes factor can be approximated by MCMC simulation.

Point null hypothesis testing

Testing $H_0 : \theta = \theta^*$ versus $H_1 : \theta \neq \theta^*$, i.e.

- $\mathcal{M}_1 : f(x | \theta^*)$,
- $\mathcal{M}_2 : f(x | \theta), \theta \in \Theta$.

Default priors

$$\pi_N(\theta) = \delta_{\theta^*}(\theta) \text{ and } \pi_N(\theta) = \pi_N(\theta)$$

For $x$ minimal training sample, consider the proper priors

$$\pi_1(\theta) = \delta_{\theta^*}(\theta) \text{ and } \pi_2(\theta) = \int \pi_N(\theta | x) f(x | \theta^*) \, dx$$
Point null hypothesis testing (cont’d)

Then

$$
\int \pi_1^N (\theta | x) m_2 (x) \, dx = \delta_{\theta^*} (\theta) \int m_2 (x) \, dx = \delta_{\theta^*} (\theta) = \pi_1 (\theta)
$$

and

$$
\int \pi_2^N (\theta | x) m_1 (x) \, dx = \int \pi_N^N (\theta | x) f (x | \theta^*) \, dx = \pi_2 (\theta)
$$

© $\pi_1 (\theta)$ and $\pi_2 (\theta)$ are integral priors

Note

Uniqueness of the Bayes factor
Integral priors and intrinsic priors coincide

[Moreno, Bertolino and Racugno, 1998]

Location models

Two location models

$\mathcal{M}_1 : f_1 (x | \theta_1) = f_1 (x - \theta_1)$
$\mathcal{M}_2 : f_2 (x | \theta_2) = f_2 (x - \theta_2)$

Default priors

$$
\pi_i^N (\theta_i) = c_i, \ i = 1, 2
$$

with minimal training sample size one

Marginal densities

$$
m_i^N (x) = c_i, \ i = 1, 2
$$

Location models (cont’d)

In that case, $\pi_1^N (\theta_1)$ and $\pi_2^N (\theta_2)$ are integral priors when $c_1 = c_2$:

$$
\int \pi_1^N (\theta_1 | x) m_2^N (x) \, dx = \int c_2 f_1 (x - \theta_1) \, dx = c_2
$$
$$
\int \pi_2^N (\theta_2 | x) m_1^N (x) \, dx = \int c_1 f_2 (x - \theta_2) \, dx = c_1.
$$

© If the associated Markov chain is recurrent,

$$
\pi_1^N (\theta_1) = \pi_2^N (\theta_2) = c
$$

are the unique integral priors and they are intrinsic priors

[Cano, Kessler & Moreno, 2004]
Transition $\theta \to \theta'$ of the Markov chain made of steps:

1. $x' = \theta + \varepsilon_1, \varepsilon_1 \sim \mathcal{N}(0, 1)$
2. $\lambda = x' + \varepsilon_2, \varepsilon_2 \sim \mathcal{DE}(0, 1)$
3. $x = \lambda + \varepsilon_3, \varepsilon_3 \sim \mathcal{DE}(0, 1)$
4. $\theta' = x + \varepsilon_4, \varepsilon_4 \sim \mathcal{N}(0, 1)$

i.e. $\theta' = \theta + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$

random walk in $\theta$ with finite second moment, null recurrent

© Resulting Lebesgue measures $\pi_1(\theta) = 1 = \pi_2(\lambda)$ invariant and unique solutions to integral equations

In a dubious battle...

First, discussion against mathematical formalism that tries to build on intuition for finite state spaces

For continuous distributions there are an infinite number of possible cases, and the definition makes the probability, in the face of it, the ratio of two infinite numbers and therefore meaningless. (...) On the infinite population definition, any finite probability is the ratio of two infinite numbers and therefore is indeterminate (VII, §7.0)

Not worth much except as an historical perspective
Dual representation

Next, discussion of dual meaning of Student’s $T$ distribution:

$$P(dz|x, \sigma, H) \propto (1 + z^2)^{-1/2}dz \quad (1)$$

where $$z = \frac{x - \bar{x}}{s}.$$ 

My result is

$$P(dz|\bar{x}, s, H) \propto (1 + z^2)^{-1/2}dz \quad (4)$$

This is not the same thing as (1) since the data is different.

Explanation

While (1) is the (sampling) distribution of $z$ as a transform of the data $(\bar{x}, s)$, (4) is the (posterior) distribution of the mean parameter $x$ given the data.

Instance of a (Fisherian) pivotal quantity

Warnin! Dependence on the prior distribution

there is only one distribution of the prior probability that can lead to it, namely

$$P(dxds|H) \propto dxds/\sigma$$

Criticism of frequentist tests

Rejection of Student’s $t$ test:

...we should reject a suggested value of $x$ by such rule as this, but applying this in practice would imply that if $x$ was known to be always the same we must accept it in 95 per cent. and reject it in 5 per cent. of the cases which hardly seems a satisfactory state of affairs. There is no positive virtue in rejecting a hypothesis in 5 per cent. of the cases where it is true, though it may be inevitable if we are to have any rule at all for rejecting it when it is false, that we shall sometimes reject it when it is true. In practice nobody would use the rule in this way if $x$ was always the same; samples would always be combined (VII, §7.1).

Missing [degree of] freedom

Same criticism of Pearson’s $\chi^2$ test [if acceptance of Pearson’s $\chi^2$ estimation method...]

if there were $n$ groups of observations [and if] $m$ parameters had been found from the data, [Pearson] would form the integral (VII, §7.2)

$$P(\chi^2) = \int_{\chi}^{\infty} \chi^{n-m-2}e^{-1/2\chi} d\chi \int_{0}^{\infty} \chi^{n-m-1}e^{-1/2\chi} d\chi$$

Should be $n - m - 2$ to correspond to the standard $\chi^2_{n-m-1}$ approximation...
Criticism of $p$-values

One of ToP most quoted sentences:

*What the use of $P$ implies, therefore, is that a hypothesis that may be true may be rejected because it had not predicted observable results that have not occurred* (VII, §7.2)

Even more to the point:

*If $P$ is small that means that there have been unexpectedly large departures from prediction. But why should these be stated in terms of $P$?* (VII, §7.2)

Jeffreys defends the use of likelihood ratios [or inverse probability] versus $p$ values (VII, §7.2)

...if the actual value is unknown the value of the power function is also unknown (...) [and] if we must choose between two definitely stated alternatives we should naturally take the one that gives the larger likelihood (VII, §7.5)

Criticism of $p$-values (cont’d)

Acceptance of posterior probability statements related to confidence assessments:

...several of the $P$ integrals have a definitive place in the present theory, in problems of pure estimation. For the normal law with a known standard error, the total area of the tail represents the probability, given the data, that the estimated difference has the wrong sign (VII, §7.21)

As for instance in design:

...the $P$ integral found from the difference between the mean yields of two varieties gives correctly the probability on the data that the estimates are in the wrong order, which is what is required (VII, §7.21)

But does not make sense for testing point null hypotheses

*If some special value has to be excluded before we can assert any other value, what is the best rule on the data available for deciding to retain it or adopt a new one?* (VII, §7.21)

And ToP finds no justification in the .05 golden rule

*In itself it is fallacious [and] there is not the slightest reason to suppose that it gives the best standard* (VII, §7.21)
Another fundamental issue

Why are \( p \) values so bad?
Because they do not account for the alternative:

\[ \text{Is it of the slightest use to reject an hypothesis unless we have some idea of what to put in its place? (VII, §7.22)} \]

...and for the consequences of rejecting the null:

\[ \text{The test required, in fact, is not whether the null hypothesis is altogether satisfactory, but whether any suggested alternative is likely to give an improvement in representing future data (VII, §7.22).} \]

The disagreement with Fisher

Main points of contention (VII, §7.4)

...general agreement between Fisher and myself...
- ...hypothetical infinite population...
- lack of conditioning
- ...use of the \( P \) integrals...

Oooops!

...at that time, to my regret, I had not read ‘Student’s’ papers and it was not till considerably later that I saw the intimate relation between [Fisher’s] methods and mine.

Overall risk

Criticism of power as parameter dependent
Use of average risk

...the expectation of the total fraction of mistakes will be

\[
2 \int_{a_c}^{\infty} P(q|a|H) + 2 \int_{0}^{a_c} \int P(q'|a|a_{\alpha}|H).
\]

Hence the total number of mistakes will be made a minimum if the line is drawn at the critical value that makes \( K = 1 \) (VII, §7.4).

But bound becomes data-dependent!
Priors are not frequencies

First part (§8.0) focussing on the concept of prior distribution and the differences with a frequency based probability

The essence of the present theory is that no probability, direct, prior, or posterior is simply a frequency (VIII, §8.0).

Extends this perspective to sampling distributions too [with hairy arguments!].

Common criticism

Next, discussion of the subjective nature of priors

Critics (...) usually say that the prior probability is ‘subjective’ (...) or refer to the vagueness of previous knowledge as an indication that the prior probability cannot be assessed (VIII, §8.0).

Conditional features of probabilities

Long argument about the subjective nature of knowledge

What the present theory does is to resolve the problem by making a sharp distinction between general principles, which are deliberately designed to say nothing about what experience is possible, and, on the other hand, propositions that do concern experience and are in the first place merely considered among alternatives (VIII, §8.1).

and definition of probability

The probability of a proposition irrespective of the data has no meaning and is simply an unattainable ideal (VIII, §8.1).
Noninformative priors

ToP then advances the use of Jeffreys’ priors as the answer to missing prior information.

A prior probability used to express ignorance is merely the formal statement of ignorance (VIII, §8.1).

Overlooks the lack of uniqueness of such priors

Missing alternatives

Next section §8.2 fairly interesting in that ToP discusses the effect of a missing alternative.

We can never rule out the possibility that some new explanation may be suggested of any set of experimental facts (VIII, §8.2).

Seems partly wrong though...

Missing alternatives (cont’d)

Indeed, if $H_0$ tested against $H_1$, Bayes factor is

$$B_{01}^\pi = \frac{\int f_0(x|\theta_0)\pi_0(d\theta_0)}{\int f_1(x|\theta_1)\pi_1(d\theta_1)}$$

while if another (exclusive) alternative $H_2$ is introduced, it would be

$$B_{01}^\pi = \frac{\int f_0(x|\theta_0)\pi_0(d\theta_0)}{\omega_1 \int f_1(x|\theta_1)\pi_1(d\theta_1) + (1 - \omega_1) \int f_2(x|\theta_2)\pi_2(d\theta_2)}$$

where $\omega_1$ relative prior weight of $H_1$ vs $H_2$

Basically biased in favour of $H_0$

Marginaliae

The remaining sections are not very interesting from a Bayesian point of view [but may be so from an epistemological point of view (quantum theory, relativity, “rejection of unobservables”, realism vs. idealism)...]
Jeffreys’ priors at the center of this theory:

*The prior probabilities needed to express ignorance of the value of a quantity to be estimated, where there is nothing to call special attention to a particular value are given by an invariance theory (VIII, §8.6).*

...with adequate changes for testing hypotheses:

*Where a question of significance arises, that is, where previous considerations call attention to some particular value, half, or possibly some smaller fraction, of the prior probability is concentrated at that value (VIII, §8.6).*

---

Main results

1. A proof independent of limiting processes that the whole information is contained in the likelihood
2. A development of pure estimation processes without further hypothesis
3. A general theory of significance tests
4. An account of how in certain conditions a law can reach a high probability

Corresponding remaining problems in ToP

1. Information also contained in prior distribution
2. Choice of estimation procedure never explicited
3. Complete occultation of the infinite mass problem
4. No true theory of goodness of fit tests

**but...**

*...it is enough (VIII, §8.8).*