An Introduction to Objective Bayesian Statistics

José M. Bernardo
Universitat de València, Spain
jose.m.bernardo@uv.es

XXVI Winter School of Astrophysics

Tenerife (Islas Canarias, Spain), November, 2014

(i) Concept of Probability

- *Introduction.* Notation. Statistical models.
- *Intrinsic discrepancy.* Intrinsic convergence of distributions.
- *Foundations.* Probability as a rational degree of belief.

(ii) Basics of Bayesian Analysis

- *Parametric inference.* The learning process.
- *Reference analysis.* No relevant initial information.
- *Inference summaries.* Point and region estimation.
- *Prediction.* Regression.
- *Hierarchical models.* Exchangeability.

(iii) **Integrated Reference Analysis**

- *Structure of a decision problem.* Axiomatics.
- *Decision structure of inference summaries.* Estimation and Hypothesis testing.
- *Loss functions in inference problems.* The intrinsic discrepancy loss.
- *Objective Bayesian methods.* Reference priors.
- *An integrated approach to objective Bayesian inference.* Intrinsic posterior analysis.

(iv) **Basic References**

Concept of Probability

Introduction

- One tentatively accepts a *formal* statistical model, typically suggested by informal descriptive evaluation.

  Conclusions are obviously conditional on the assumption that the model is correct.

- The Bayesian approach is firmly based on *axiomatic foundations*. These imply that *all* uncertainties *must* be described by probabilities.

- In particular, parameters *must* have a (*prior*) distribution describing available information about their values.

  This is *not* a description of their variability (since they are *fixed unknown* quantities), but a description of the *uncertainty* about their true values.

- An important *particular* case arises when there is *no relevant objective initial information*, and the use of subjective information is not desired.

  This typically occurs in scientific and industrial reporting, public decision making, auditing, and many other situations

- Bayesian analysis then requires a prior distributions *exclusively* based on model assumptions and available, well-documented data.

  This is the subject of *Objective Bayesian Statistics*.

□ Notation

- Under conditions C, $p(\boldsymbol{x}\,|\,C)$, $\pi(\boldsymbol{\theta}\,|\,C)$ are, respectively, *probability* densities (or mass) functions of *observables* $\boldsymbol{x}$ and *parameters* $\boldsymbol{\theta}$.

$$p(\boldsymbol{x}\,|\,C) \geq 0,\ \int_{\mathcal{X}} p(\boldsymbol{x}\,|\,C)\,d\boldsymbol{x} = 1,\ \mathrm{E}[\boldsymbol{x}\,|\,C] = \int_{\mathcal{X}} \boldsymbol{x}\,p(\boldsymbol{x}\,|\,C)\,d\boldsymbol{x},$$

$$\pi(\boldsymbol{\theta}\,|\,C) \geq 0,\ \int_{\Theta} \pi(\boldsymbol{\theta}\,|\,C)\,d\boldsymbol{\theta} = 1,\ \mathrm{E}[\boldsymbol{\theta}\,|\,C] = \int_{\Theta} \boldsymbol{\theta}\,\pi(\boldsymbol{\theta}\,|\,C)\,d\boldsymbol{\theta}.$$

- Special densities (or mass) functions use specific notation, as $\mathrm{N}(x\,|\,\mu,\sigma)$, $\mathrm{Bi}(x\,|\,n,\theta)$, or $\mathrm{Po}(x\,|\,\lambda)$.

  In particular,

$$\mathrm{Be}(x\,|\,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}, \quad 0<x<1,\ \alpha>0,\ \beta>0.$$

$$\mathrm{Ga}(x\,|\,\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\,x^{\alpha-1}e^{-\beta x}, \quad x>0,\ \alpha>0,\ \beta>0.$$

$$\mathrm{St}(x\,|\,\mu,\sigma,\alpha) = \frac{\Gamma\{(\alpha+1)/2)\}}{\Gamma(\alpha/2)}\frac{1}{\sigma\sqrt{\alpha\pi}}\left[1+\frac{1}{\alpha}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\alpha+1)/2},$$

$$x\in\Re,\ \mu\in\Re, \sigma>0,\ \alpha>0.$$

□ Statistical Models

- *Statistical model* generating $\boldsymbol{x} \in \mathcal{X}$, $\{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$
  Parameter vector $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\} \in \boldsymbol{\Theta}$.
  Parameter space $\boldsymbol{\Theta} \subset \Re^k$.
  Data set $\boldsymbol{x} \in \mathcal{X}$. *Sampling space* $\mathcal{X}$, of arbitrary structure.
- *Likelihood function* of $\boldsymbol{x}$, $l(\boldsymbol{\theta} \,|\, \boldsymbol{x})$.
  $l(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, as a function of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
- *Maximum likelihood estimator (mle)* of $\boldsymbol{\theta}$
  $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \; l(\boldsymbol{\theta} \,|\, \boldsymbol{x})$
- Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ *random sample* (iid) from model if
  $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{n} p(x_j \,|\, \boldsymbol{\theta}),\; x_j \in \mathcal{X}, \quad \mathcal{X} = \mathcal{X}^{\mathbf{n}}$.

- Behaviour under repeated sampling (general, not necessarily iid data) is obtained by considering $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$, a (possibly infinite) sequence of possible replications of the *complete* data set $\boldsymbol{x}$.

Denote by $\boldsymbol{x}^{(m)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ a finite set of m such replications. Asymptotic results are obtained as $m \to \infty$.

- Data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are often assumed to be *exchangeable* in that their joint distribution $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ may be considered to be invariant under permutations (order of observation inmaterial)
- Under exchangeability, parameters may formally be *defined*as asymptotic limits of specific data functions.

Thus, If data $\{x_1, \ldots, x_n\}$ are exchangeable 0-1 random quantities, then $p(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$, with $\theta = \lim_{m\to\infty}(1/m)\sum_{i=1}^{m} x_i$. Or, if data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are exchangeable normal vectors $\mathrm{N}(\boldsymbol{x}_i \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{\mu} = \lim_{m\to\infty}(1/m)\sum_{i=1}^{m} \boldsymbol{x}_i$.

Intrinsic Divergence

□ Logarithmic divergences

- The logarithmic divergence (Kullback-Leibler) $k\{\hat{p}\,|\,p\}$ of a density $\hat{p}(\boldsymbol{x})$, $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ from its true density $p(\boldsymbol{x})$, is

$$\kappa\{\hat{p}\,|\,p\} = \int_{\mathcal{X}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{\hat{p}(\boldsymbol{x})}\, d\boldsymbol{x}, \quad \text{(provided this exists)}$$

- The functional $\kappa\{\hat{p}\,|\,p\}$ is non-negative; it is zero iff $\hat{p}(\boldsymbol{x}) = p(\boldsymbol{x})$ (a.e.), and it is *invariant* under one-to-one transformations of $\boldsymbol{x}$.
- But $\kappa\{p_1\,|\,p_2\}$ is *not symmetric* and diverges if, strictly, $\boldsymbol{\mathcal{X}_2} \subset \boldsymbol{\mathcal{X}_1}$.

□ Intrinsic discrepancy between distributions

- $\delta\{p_1, p_2\} = \min \left\{ \int_{\mathcal{X}_1} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\, d\boldsymbol{x}, \int_{\mathcal{X}_2} p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})}\, d\boldsymbol{x} \right\}$

The *intrinsic discrepancy* $\delta\{p_1, p_2\}$ is non-negative; it is zero iff, $p_1 = p_2$ a.e., and it is *invariant* under one-to-one transformations of $\boldsymbol{x}$,

- The intrinsic discrepancy is defined even if either $\boldsymbol{\mathcal{X}_2} \subset \boldsymbol{\mathcal{X}_1}$ or $\boldsymbol{\mathcal{X}_1} \subset \boldsymbol{\mathcal{X}_2}$. It has an operative interpretation as the minimum amount of information (in *nits*) required to discriminate between the two distributions.

□ Interpretation of the intrinsic discrepancy

- Let $\{p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1), \boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1\}$ or $\{p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2), \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2\}$ be two alternative statistical models for $\boldsymbol{x} \in X$, one of which is assumed to be true. The intrinsic divergence $\delta\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\} = \delta\{p_1, p_2\}$ is then *minimum expected log-likelihood ratio in favour of the true model*.

Indeed, if $p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)$ true model, the expected log-likelihood ratio in its favour is $\mathrm{E}_1[\log\{p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)/p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)\}] = \kappa\{p_2 \,|\, p_1\}$. If the true model is $p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2)$, the expected log-likelihood ratio in favour of the true model is $\kappa\{p_2 \,|\, p_1\}$. But $\delta\{p_2 \,|\, p_1\} = \min[\kappa\{p_2 \,|\, p_1\}, \kappa\{p_1 \,|\, p_2\}]$.

□ Calibration of the intrinsic discrepancy

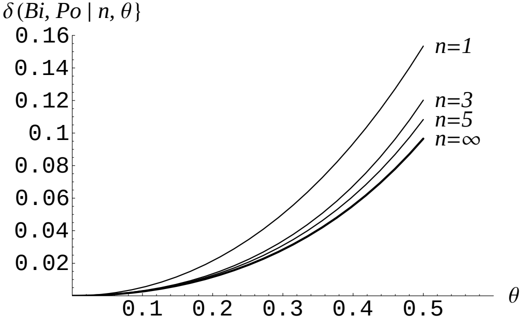
- If, say, $\delta = \log[100] \approx 4.6$ *nits*, likelihood ratios for the true model are expected to be larger than 100 making *discrimination relatively easy*.
- For small $\varepsilon > 0$ values if $\delta = \log(1+\varepsilon) \approx \varepsilon$ *nits*, likelihood ratios for the true model are expected to be about $1 + \epsilon$ making *discrimination very hard*.
- Some conventional values:

Intrinsic Discrepancy δ	0.01	0.69	2.3	4.6	6.9
Min Expected Likelihood Ratio for **true** model	1.01	2	10	100	1000

Thus, discrepancy values of, say, $\log[100] \approx 4.6$, corresponding to minimum expected likelihood ratios of about 100 for the true model, may be conventionally regarded as evidence that the two models are too wide apart to be used indistinctly.

Example. Conventional Poisson approximation $\mathrm{Po}(r \,|\, n\theta)$ of Binomial probabilities $\mathrm{Bi}(r \,|\, n, \theta)$.

$$\begin{aligned}\delta\{\mathrm{Bi}, \mathrm{Po} \,|\, n, \theta\} &= \min[k\{\mathrm{Bi} \,|\, \mathrm{Po}\}, k\{\mathrm{Po} \,|\, \mathrm{Bi}\}] = k\{\mathrm{Bi} \,|\, \mathrm{Po}\} \\ &= \sum_{r=0}^{n} \mathrm{Bi}(r \,|\, n,\ \theta) \log\left[\frac{\mathrm{Bi}(r \,|\, n,\ \theta)}{Po(r \,|\, n\theta)}\right] = \delta\{n, \theta\}\end{aligned}$$



$$\delta\{3, 0.05\} = 0.00074$$

$$\delta\{5000, 0.05\} = 0.00065$$

$$\delta\{\infty, \theta\} = \tfrac{1}{2}[-\theta - \log(1-\theta)]$$

- Good Poisson approximations to Binomial probabilities are *not possible* if θ is not small, however large n might be.

□ Intrinsic Convergence of Distributions

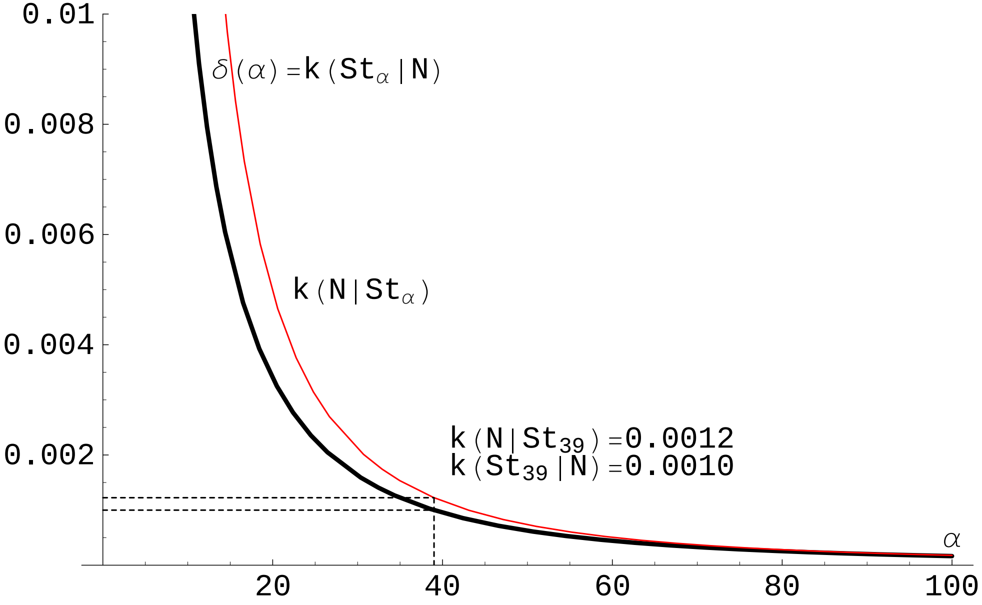
- *Intrinsic convergence.* A sequence of probability densities (or mass) functions $\{p_i(\boldsymbol{x})\}_{i=1}^{\infty}$ converges *intrinsically* to $p(\boldsymbol{x})$ if (and only if) the intrinsic divergence between $p_i(x)$ and $p(x)$ converges to zero. *i.e.*, iff $\lim_{i\to\infty} \delta(p_i, p) = 0$.
- *Example*. Normal approximation to a Student distribution.

$$\begin{aligned} \delta(\alpha) &= \delta\{\mathrm{St}(x\,|\,\mu,\sigma,\alpha), \mathrm{N}(x\,|\,\mu,\sigma)\} = \min[k\{\mathrm{St}_\alpha\,|\,\mathrm{N}\}, k\{\mathrm{N}\,|\,\mathrm{St}_\alpha\}] \\ &= k\{\mathrm{St}_\alpha\,|\,\mathrm{N}\} = \int_{\Re} \mathrm{N}(x\,|\,0,1) \log \frac{\mathrm{N}(x\,|\,0,1)}{\mathrm{St}(x\,|\,0,1,\alpha)}\, dx \approx \frac{7}{\alpha(22+4\alpha)} \end{aligned}$$

$k\{\mathrm{N}\,|\,\mathrm{St}_\alpha\}$ diverges for $\alpha \le 2$

$k\{\mathrm{St}_\alpha\,|\,\mathrm{N}\}$ is finite for all $\alpha > 0$.

$\delta(18) \approx 0.04 \quad \delta(25) \approx 0.02$



Expected log-density ratios when approximating Student by Normal smaller than 0.001 when $\alpha \geq 40$. This suggests that the approximation may be reasonable if the degrees of freedom are larger than 40.

Foundations

□ Foundations of Statistics

- Axiomatic foundations on rational description of uncertainty imply that the uncertainty about all unknown quantities should be measured with *probability* distributions $\{\pi(\boldsymbol{\theta} \mid C), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ describing their plausibility given available conditions C.
- Axioms have a strong intuitive appeal; examples include:
- *Transitivity of plausibility*
  If $E_1 > E_2 \mid C$, and $E_2 > E_3 \mid C$, then $E_1 > E_3 \mid C$
- *The sure-thing principle*
  If $E_1 > E_2 \mid A, C$ and $E_1 > E_2 \mid \overline{A}, C$, then $E_1 > E_2 \mid C)$.
- Axioms are not a *description* of actual human activity, but a *normative* set of principles for those aspiring to rational behaviour.

- "Absolute" probabilities do not exist. Typical applications produce $\Pr(E \mid \boldsymbol{x}, A, K)$, a measure of rational belief in the occurrence of the *event* E, given data $\boldsymbol{x}$, assumptions A and available knowledge K.

□ Probability as a Measure of Conditional Uncertainty

- Axiomatic foundations imply that $\Pr(E \mid C)$, the *probability* of an event E given C is *always* a conditional measure of the (presumably rational) uncertainty, on a $[0, 1]$ scale, about the occurrence of E in conditions C.

- *Probabilistic diagnosis*. V is the event that a person carries a virus and $+$ a positive test result. *All* related probabilities, *e.g.*,
  $\Pr(+ \mid V) = 0.98$, $\Pr(+ \mid \overline{V}) = 0.01$, $\Pr(V \mid K) = 0.002$,
  $\Pr(+ \mid K) = \Pr(+ \mid V)\Pr(V \mid K) + \Pr(+ \mid \overline{V})\Pr(\overline{V} \mid K) = 0.012$
  $\Pr(V \mid +, A, K) = \frac{\Pr(+ \mid V)\Pr(V \mid K)}{\Pr(+ \mid K)} = 0.164$ (Bayes' Theorem)
  are conditional uncertainty measures (and proportion estimates).

- *Estimation of a proportion*. Survey conducted to estimate the proportion θ of positive individuals in a population. Random sample of size n with r positive. $\Pr(a < \theta < b \,|\, r, n, A, K)$, a conditional measure of the uncertainty about the event that θ belongs to $[a, b]$ *given* assumptions A, initial knowledge K and data $\{r, n\}$.
- *Measurement of a physical constant*. Measuring the unknown value of a physical constant μ, with data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, considered to be measurements of μ subject to error.

Desired to find $\Pr(a < \mu < b \,|\, \boldsymbol{x} = \{x_1, \ldots, x_n\}, A, K)$, the *probability* that the unknown value of μ (fixed in nature, but unknown to the scientists) belongs to $[a, b]$, given the information provided by the data $\boldsymbol{x}$, any assumptions A made, and available knowledge K.

□ Nuisance parameters

• The statistical model usually include *nuisance* parameters, unknown quantities , which have to be eliminated in the statement of the final results.

For instance, the precision of the measurements described by unknown standard deviation σ in a $\mathrm{N}(x \mid \mu, \sigma)$ normal model.

□ Restrictions

• Relevant scientific information may impose *restrictions* on the admissible values of the quantities of interest. These must be taken into account.

For instance, in measuring the value of the gravitational field g in a laboratory, it is known that it must lie between 9.7803 m/sec^2 (average value at the Equator) and 9.8322 m/sec^2 (average value at the poles).

- *Future discrete observations*. Counting the number r of times that an event E takes place in each of n replications. Desired to forecast the number of times r that E will take place in a future, similar situation, $\Pr(r \,|\, r_1, \ldots, r_n, A, K)$. For instance, no accidents in each of $n = 10$ consecutive months may yield $\Pr(r = 0 \,|\, \boldsymbol{x}, A, K) = 0.953$.
- *Future continuous observations*. Data $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$. Desired $p(\boldsymbol{y} \,|\, \boldsymbol{x}, A, K)$, to forecast a future observation $\boldsymbol{y}$. For instance, from breaking strengths $\boldsymbol{x} = \{y_1, \ldots, y_n\}$ of n randomly chosen safety belt webbings, the engineer may find $\Pr(y > y^* \,|\, \boldsymbol{x}, A, K) = 0.9987$.
- *Regression*. Data set consists of pairs $\boldsymbol{x} = \{(\boldsymbol{y}_1, \boldsymbol{v}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{v}_n)\}$ of quantity $\boldsymbol{y}_j$ observed in conditions $\boldsymbol{v}_j$. Desired to forecast the value of $\boldsymbol{y}$ in conditions $\boldsymbol{v}$, $p(\boldsymbol{y} \,|\, \boldsymbol{v}, \boldsymbol{x}, A, K)$. For instance, with y contamination levels and v wind speed from source, environment authorities may be interested in $\Pr(y > y^* \,|\, v, \boldsymbol{x}, A, K)$.

Basics of Bayesian Analysis

Parametric Inference

□ Bayes Theorem

- Let $M = \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be an statistical model, let $\pi(\boldsymbol{\theta} \,|\, K)$ be a probability density for $\boldsymbol{\theta}$ given prior knowledge K and let $\boldsymbol{x}$ be some available data.

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, M, K) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, K)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, K)\, d\boldsymbol{\theta}},$$

encapsulates all information about $\boldsymbol{\theta}$ given data and prior knowledge.

- Simplifying notation, Bayes' theorem may be expressed as

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$$

The posterior is proportional to the likelihood times the prior.

- The missing proportionality constant in $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \propto p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$, $c(\boldsymbol{x}) = [\int_{\Theta} p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}]^{-1}$ may be deduced from the fact that $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ must integrate to one. To identify a posterior distribution it suffices to identify a *kernel* $k(\boldsymbol{\theta}, \boldsymbol{x})$ such that $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = c(\boldsymbol{x})\,k(\boldsymbol{\theta}, \boldsymbol{x})$. This is a very common technique.

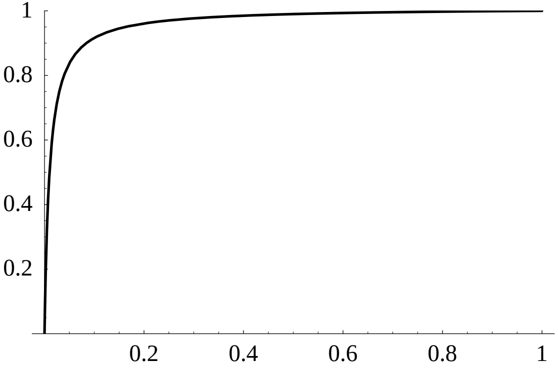
□ Bayesian Inference with a Finite Parameter Space

- Model $\{p(\boldsymbol{x}\,|\,\theta_i), \boldsymbol{x} \in \mathcal{X}, \theta_{\mathbf{i}} \in \boldsymbol{\Theta}\}$, with $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_m\}$, so that θ may only take a *finite* number m of different values. Using the finite form of Bayes' theorem,

$$\Pr(\theta_i\,|\,\boldsymbol{x}) = \frac{p(\boldsymbol{x}\,|\,\theta_i)\ \Pr(\theta_i)}{\sum_{j=1}^{m} p(\boldsymbol{x}\,|\,\theta_j)\ \Pr(\theta_j)}\,, \quad i = 1, \ldots, m.$$

- *Example: Probabilistic diagnosis*. A test to detect a virus, is known from laboratory research to give a positive result in 98% of the infected people and in 1% of the non-infected. The posterior probability that a person who tested positive is infected as a function of $p = \Pr(V)$ is

$$\Pr(V \mid +) = \frac{0.98\, p}{0.98\, p + 0.01\, (1 - p)}\,.$$

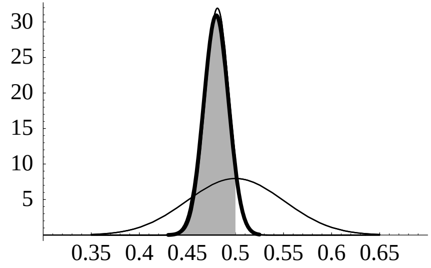


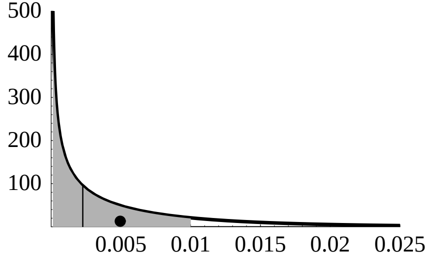
Notice the sensitivity of posterior probability $\Pr(V \mid +)$ to changes in the prior $p = \Pr(V)$.

□ Example: Inference about a binomial parameter.

- Let data $\boldsymbol{x}$ consist of n Bernoulli observations with parameter θ which contain r positives, so that $p(\boldsymbol{x} \mid \theta, n) = \theta^r (1-\theta)^{n-r}$.

- If $\pi(\theta) = \mathrm{Be}(\theta \mid \alpha, \beta)$, then
  $\pi(\theta \mid \boldsymbol{x}) \propto \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}$
  kernel of $\mathrm{Be}(\theta \mid r+\alpha, n-r+\beta)$.
- Prior information (K)
  $P(0.4 < \theta < 0.6) = 0.95$,
  and symmetric, yields $\alpha = \beta = 47$;
- No prior information $\alpha = \beta = 1/2$
- $n = 100, r = 0$
  $P(\theta < 0.01 \mid \boldsymbol{x}) = 0.844$
  Notice: $\hat{\theta} = 0$, but $\mathrm{Me}[\theta \mid \boldsymbol{x}] = 0.0023$





□ Sufficiency

- Given a model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$, is a *sufficient* statistic if it encapsulates all information about $\boldsymbol{\theta}$ available in $\boldsymbol{x}$. Formally, $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is *sufficient* if (and only if), for any prior $\pi(\boldsymbol{\theta})$ $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \pi(\boldsymbol{\theta} \,|\, \boldsymbol{t})$. Hence, $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \pi(\boldsymbol{\theta} \,|\, \boldsymbol{t}) \propto p(\boldsymbol{t} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$. This is equivalent to the frequentist definition; thus $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is sufficient iff $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = f(\boldsymbol{\theta}, \boldsymbol{t}) g(\boldsymbol{x})$.
- A sufficient statistic always exists, for $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{x}$ is obviously sufficient, but a much simpler sufficient statistic, with fixed dimensionality independent of the sample size, exists whenever the statistical model belongs to the *generalized exponential family*.
- Bayesian methods are independent of the possible existence of a sufficient statistic of fixed dimensionality. For instance, if data come from an Student distribution, there is *no sufficient statistic* of fixed dimensionality: *all data are needed.*

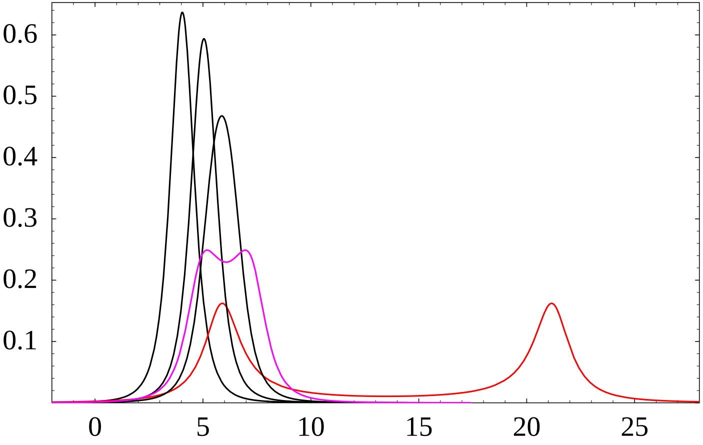
□ Example: Inference from Cauchy observations

- Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ random from $\text{Ca}(x \,|\, \mu, 1) = \text{St}(x \,|\, \mu, 1, 1)$.
- Objective prior for the location parameter μ is $\pi(\mu) = 1$, and

$$\pi(\mu \,|\, \boldsymbol{x}) \propto \prod_{j=1}^{n} \text{Ca}(x_j \,|\, \mu, 1)\pi(\mu) \propto \prod_{j=1}^{n} \frac{1}{1 + (x_j - \mu)^2}\,.$$

- Five samples of size $n = 2$, simulated from $\text{Ca}(x \,|\, 5, 1)$

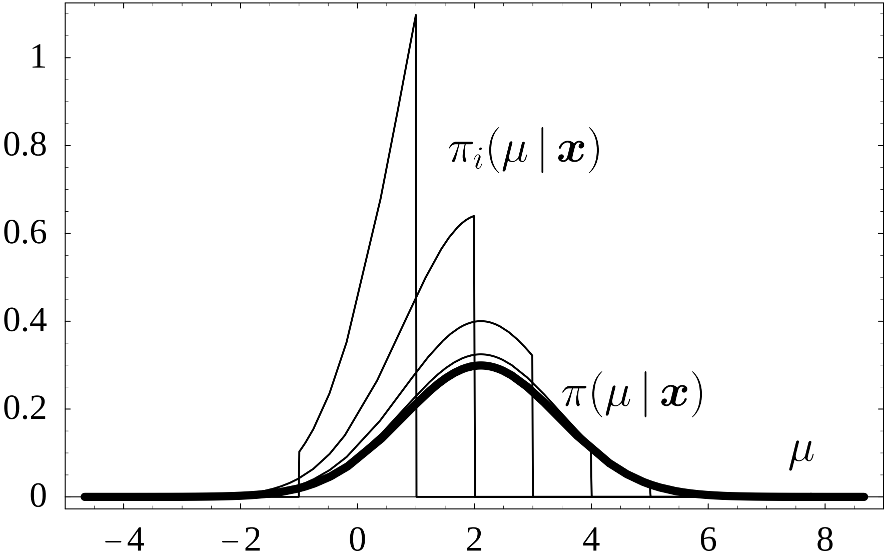
x_1	x_2
4.034	4.054
21.220	5.831
5.272	6.475
4.776	5.317
7.409	4.743



□ Improper prior functions

- Objective Bayesian methods often use functions which play the role of prior distributions but are *not* probability distributions. An *improper prior function* is an non-negative function $\pi(\boldsymbol{\theta})$ such that $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ is not finite. The Cauchy example uses the improper prior function $\pi(\mu) = 1$, $\mu \in \Re$.
- Let $\pi(\boldsymbol{\theta})$ be an improper prior function, $\{\boldsymbol{\Theta}_i\}_{i=1}^{\infty}$ an increasing sequence approximating $\boldsymbol{\Theta}$, such that $\int_{\boldsymbol{\Theta}_i} \pi(\boldsymbol{\theta}) < \infty$, and let $\{\pi_i(\boldsymbol{\theta})\}_{i=1}^{\infty}$ be the proper priors obtained by *renormalizing* $\pi(\boldsymbol{\theta})$ within the $\boldsymbol{\Theta}_i$'s. Then, For any data $\boldsymbol{x}$ with likelihood $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, the sequence of posteriors $\pi_i(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ *converges intrinsically* to $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$.
- This procedure allows a systematic use of improper prior functions, whenever this is required, as it is often the case in objective Bayesian statistics.

- *Example*. Normal $\mathrm{N}(x \,|\, \mu, \sigma)$ data, σ known, $\pi(\mu) = 1$.
  $\pi(\mu \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \mu, \sigma)\pi(\mu) \propto \exp[-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2]$.
  Hence, $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{N}(\mu \,|\, \bar{x}, \sigma/\sqrt{n})$.
  For instance, with $n = 9$, $\bar{x} = 2.11$, $\sigma = 4$.



□ Sequential updating

- Prior and posterior are only terms *relative* to a particular set of data.
- If data $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are sequentially presented, the final result will be the same whether data are globally or sequentially processed.

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1}) \propto p(\boldsymbol{x}_{i+1} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_i).$$

The "posterior" at a given stage becomes the "prior" at the next.

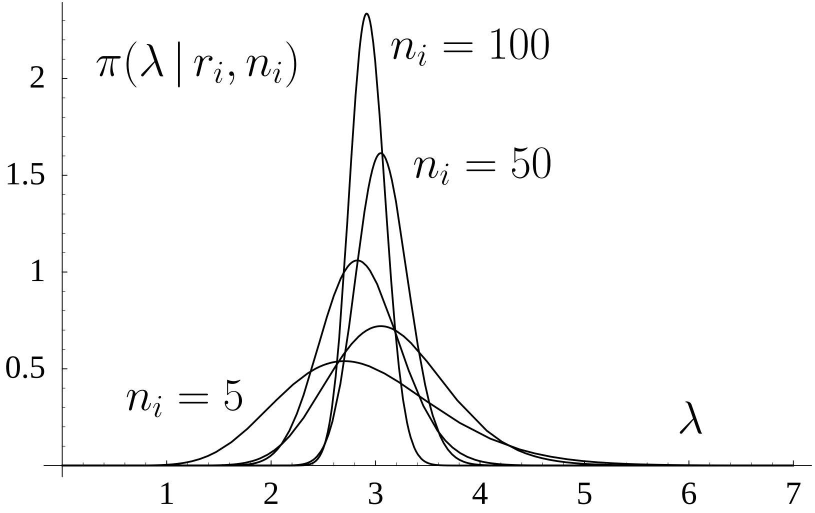
- As one should certainly expect, Bayesian procedures with exchangeable data are *always* independent of the particular order or grouping in which the data are processed. This is often *not* the case with conventional statistical procedures.
- Typically (but not always), the new posterior, $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1})$, is more concentrated around the true value than $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$.

- *Example*. Poisson $\text{Po}(x \mid \lambda)$ data, with $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Objective prior $\pi(\lambda) =\propto \lambda^{-1/2}$. Sufficient statistic $r = \sum_{j=1}^{n} x_j$.

$$\pi(\lambda \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \lambda)\pi(\lambda) \propto \exp(-n\lambda)\lambda^{r-1/2} \propto \text{Ga}(\lambda \mid r + 1/2, n).$$

- Posteriors $\text{Ga}(\lambda \mid r_i + 1/2, n_i)$ from increasingly large simulated data from $\text{Po}(x \mid \lambda = 3)$.

  The posteriors concentrate around the true value $\lambda = 3$.



□ Nuisance parameters

- In general the *vector of interest* is not the whole parameter vector $\boldsymbol{\theta}$, but some function $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ of possibly lower dimension.
- By Bayes' theorem $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$. Let $\boldsymbol{\omega} = \boldsymbol{\omega}(\boldsymbol{\theta}) \in \boldsymbol{\Omega}$ be another function of $\boldsymbol{\theta}$ such that $\boldsymbol{\psi} = \{\boldsymbol{\phi}, \boldsymbol{\omega}\}$ is a bijection of $\boldsymbol{\theta}$, and let $J(\boldsymbol{\psi}) = (\partial\boldsymbol{\theta}/\partial\boldsymbol{\psi})$ be the Jacobian of the inverse function $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$. From probability theory, $\pi(\boldsymbol{\psi} \,|\, \boldsymbol{x}) = |J(\boldsymbol{\psi})| \big[\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\big]_{\boldsymbol{\theta}=\boldsymbol{\theta}(\boldsymbol{\psi})}$ and $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x}) = \int_{\boldsymbol{\Omega}} \pi(\boldsymbol{\phi}, \boldsymbol{\omega} \,|\, \boldsymbol{x})\, d\boldsymbol{\omega}$. Any valid conclusion on $\boldsymbol{\phi}$ will be contained in $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x})$.
- Particular case: *marginal posteriors*. If model is expressed in terms of vector of interest $\boldsymbol{\phi}$, and vector of nuisance parameters $\boldsymbol{\omega}$, so that $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})$, specify the prior $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\phi})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\phi})$; get the joint posterior $\pi(\boldsymbol{\phi}, \boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\phi})\, \pi(\boldsymbol{\phi})$, and integrate out $\boldsymbol{\omega}$, so that $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x}) \propto \pi(\boldsymbol{\phi}) \int_{\Omega} p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\phi})\, d\boldsymbol{\omega}$.

□ Inferences about a Normal mean

- Data $\boldsymbol{x} = \{x_1, \dots\, x_n\}$ random from $\mathrm{N}(x \mid \mu, \sigma)$ with likelihood

$$p(\boldsymbol{x} \mid \mu, \sigma) \propto \sigma^{-n} \exp[-n\{s^2 + (\bar{x} - \mu)^2\}/(2\sigma^2)],$$

where $n\bar{x} = \sum_i x_i$, and $ns^2 = \sum_i (x_i - \bar{x})^2$.

- The objective prior is $\pi(\mu, \sigma) = \sigma^{-1}$, uniform in both μ and $\log(\sigma)$; the corresponding joint posterior is

$$\pi(\mu, \sigma \mid \boldsymbol{x}) \propto \sigma^{-(n+1)} \exp[-n\{s^2 + (\bar{x} - \mu)^2\}/(2\sigma^2)].$$

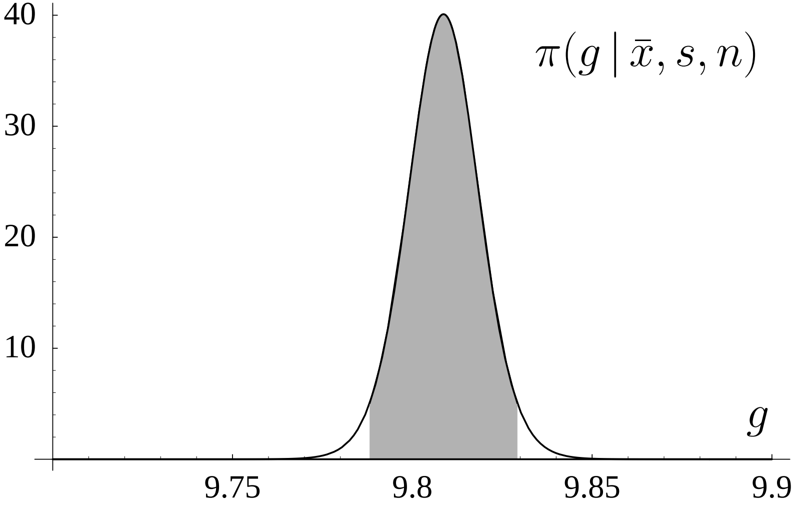
- Hence, the marginal posterior for μ is

$$\pi(\mu \mid \boldsymbol{x}) \propto \int_0^\infty \pi(\mu, \sigma \mid \boldsymbol{x})\, d\sigma \propto [s^2 + (\bar{x} - \mu)^2]^{-n/2}$$

which is the kernel of the Student density

$$pi(\mu \mid \boldsymbol{x}) = \mathrm{St}(\mu \mid \bar{x}, s/\sqrt{n-1}, n-1).$$

- *Example*. Classroom experiment to measure gravity g, with $n = 20$ observations from a pendulum yielded $\bar{x} = 9.8087$ and $s = 0.0428$.



$$\pi(g \,|\, \bar{x}, s, n) = \mathrm{St}(g \,|\, 9.9087, 0.0001, 19)$$

$$\Pr(9.788 < g < 9.829 \,|\, \boldsymbol{x}) = 0.95 \text{ (shaded area)}$$

□ Restricted parameter space

- The range of $\boldsymbol{\theta}$ values is often restricted by contextual considerations. If $\boldsymbol{\theta}$ is known to belong to $\Theta_c \subset \Theta$, so that $\pi(\boldsymbol{\theta}) > 0$ iff $\boldsymbol{\theta} \in \boldsymbol{\Theta}_c$, the use of Bayes' theorem yields

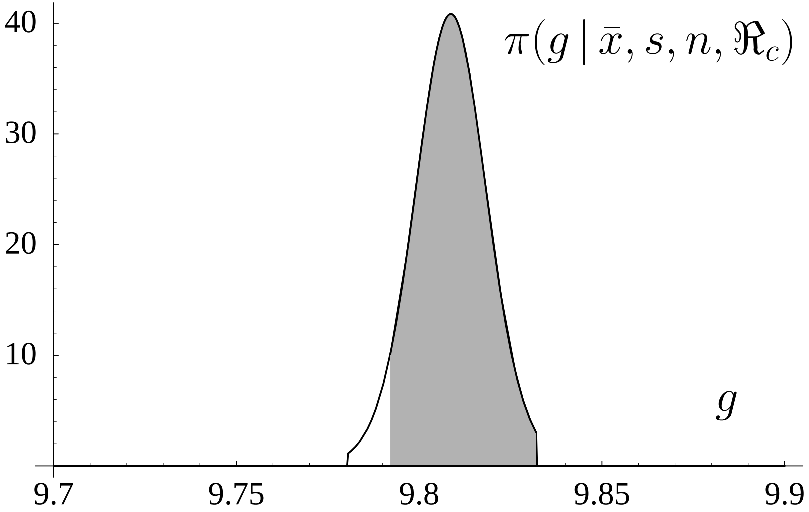
$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \boldsymbol{\theta} \in \boldsymbol{\Theta}_c) = \frac{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{\int_{\Omega_c} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}}\,, \quad \text{if} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_c,$$

and 0 otherwise.

- Thus, to incorporate a restriction, it suffices to *renormalize* the unrestricted posterior distribution to the set $\Theta_c \subset \Theta$ of admissible parameter values.

- This is often very important in applications. Yet, incorporation of parameter restrictions is often not possible in conventional, frequentist statistics.

- *Example (cont)*. Measuring gravity g with restrictions to lie between $g_0 = 9.7803$ (equator) and $g_1 = 9.8322$ (poles).



$$\Pr(9.7803 < g < 9.8322 \mid \boldsymbol{x}) = 0.95 \quad \text{(shaded area)}.$$

□ Asymptotic behaviour, discrete case

- If the parameter space $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots\}$ is *countable* and the true parameter value θ_t is *distinguishable* from all the others, so that

$$\delta\{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_t), p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_i)) > 0, \quad \text{for all} \quad i \neq t,$$

then the posterior converges to a degenerate distribution with probability one on the true value:

$$\begin{aligned} \lim_{n\to\infty} \pi(\theta_t \,|\, \boldsymbol{x}_1 \ldots, \boldsymbol{x}_n) &= 1 \\ \lim_{n\to\infty} \pi(\theta_i \,|\, \boldsymbol{x}_1 \ldots, \boldsymbol{x}_n) &= 0, \quad i \neq t. \end{aligned}$$

- To prove this, take logarithms is Bayes' theorem, define

$$z_i = \log[p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_i)/p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_t)],$$

and use the strong law of large numbers on the n i.i.d. random variables $z_1, \ldots, z_n$.

- For instance, in *probabilistic diagnosis* the posterior probability of the true disease converges to one as new relevant information accumulates, *provided* the model distinguishes the probabilistic behaviour of data under the true disease from its behaviour under the other alternatives.
- If the true value of the parameter id not in $\boldsymbol{\Theta}$ the posterior concentrates on the *closest* model in the intrinsic discrepancy sense, *i.e.*, in that value θ^* such that

$$\theta^* = \arg\min_{\theta_i \in \boldsymbol{\Theta}} \delta\{p(\boldsymbol{x} \mid \boldsymbol{\theta}_t), p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)\}.$$

□ Asymptotic behaviour, continuous case

- If the parameter θ is *one-dimensional and continuous*, so that $\boldsymbol{\Theta} \subset \Re$, and the model $\{p(\boldsymbol{x}\,|\,\theta),\ \boldsymbol{x} \in \mathcal{X}\}$ is *regular* (basically, $\mathcal{X}$ does not depend on θ, and $p(\boldsymbol{x}\,|\,\theta)$ is twice θ differentiable)
- Then, as $n \to \infty$, $\pi(\theta\,|\,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ converges intrinsically to a *normal* distribution with mean at the mle estimator $\hat{\theta}$, and with variance $v(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\hat{\theta})$, where $v^{-1}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\hat{\theta}) = -\sum_{j=1}^{n}\frac{\partial^2}{\partial\theta^2}\log[p(\boldsymbol{x}_j\,|\,\theta]$
- To prove this, write Bayes' theorem as $\pi(\theta\,|\,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \propto \exp[\log\pi(\theta) + \sum_{j=1}^{n}\log p(\boldsymbol{x}_j\,|\,\theta)]$, and expand $\sum_{j=1}^{n}\log p(\boldsymbol{x}_j\,|\,\theta)]$ about its maximum, the mle $\hat{\theta}$.
- The result is easily extended to the multivariate case $\boldsymbol{\theta} = \{\theta_1,\ldots,\theta_k\}$, to obtain a limiting k-variate normal centered at $\hat{\boldsymbol{\theta}}$, and with a dispersion matrix $\boldsymbol{V}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\hat{\boldsymbol{\theta}})$ which generalizes $v(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\hat{\theta})$.

□ Asymptotic behaviour, continuous case. Simpler form

- Using the strong law of large numbers on the sums above a simpler, less precise approximation is obtained:
- If the parameter $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$ is continuous, so that $\boldsymbol{\Theta} \subset \Re^k$ and the model $\{p(\boldsymbol{x} \,|\, \theta),\ \boldsymbol{x} \in \mathcal{X}\}$ is *regular*, so that $\mathcal{X}$ does not depend on $\boldsymbol{\theta}$ and $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ is twice differentiable with respect to each of the θ_i's, then, as $n \to \infty$, $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ converges intrinsically to a *multivariate normal* distribution with mean the mle $\hat{\boldsymbol{\theta}}$ and precision matrix (inverse of the dispersion or variance-covariance matrix) $n\,\boldsymbol{F}(\hat{\boldsymbol{\theta}})$, where $\boldsymbol{F}(\boldsymbol{\theta})$ is Fisher's matrix, of general element

$$\boldsymbol{F}_{ij}(\boldsymbol{\theta}) = -\mathrm{E}_{\boldsymbol{x} \,|\, \boldsymbol{\theta}}[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\boldsymbol{x} \,|\, \boldsymbol{\theta})].$$

- The properties of the multivariate normal easily yield from the last result the asymptotic forms for the *marginal* and the *conditional* posterior distributions of any subgroup of the θ_j's.
- In one dimension,

$$\pi(\theta \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \approx \mathrm{N}(\theta \,|\, \hat{\theta},\; [nF(\hat{\theta})]^{-1/2})$$

a normal density centered at the mle $\hat{\theta}$ with precision $nF(\hat{\theta})$, where $F(\theta)$ is Fisher's function,

$$F(\theta) = -\mathrm{E}_{\boldsymbol{x} \,|\, \theta}[\partial^2 \log p(\boldsymbol{x} \,|\, \theta)/\partial\theta^2].$$

□ Example: Asymptotic approximation with Poisson data

- Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ a random sample from

$$\mathrm{Po}(x \,|\, \lambda) \propto e^{-\lambda} \lambda^x / x!$$

Hence, $p(\boldsymbol{x} \,|\, \lambda) \propto e^{-n\lambda} \lambda^r$, $r = \Sigma_j\, x_j$, and $\hat{\lambda} = r/n$.

- Fisher's function is $F(\lambda) = -\mathrm{E}_{x \,|\, \lambda} \left[\frac{\partial^2}{\partial \lambda^2} \log \mathrm{Po}(x \,|\, \lambda) \right] = \frac{1}{\lambda}$
- The objective prior function is $\pi(\lambda) = F(\lambda)^{1/2} = \lambda^{-1/2}$.

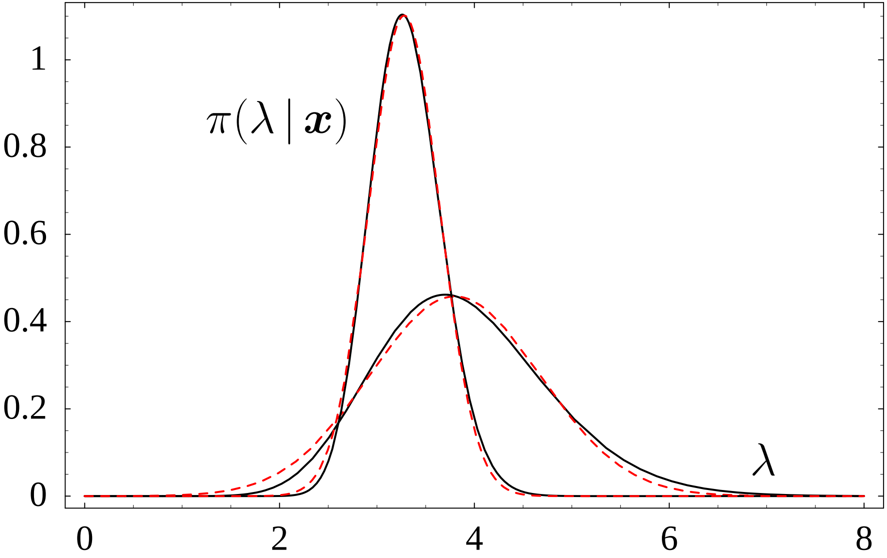
Hence $\pi(\lambda \,|\, \boldsymbol{x}) \propto e^{-n\lambda} \lambda^{r-1/2}$ which is the kernel of the Gamma distribution

$$\pi(\lambda \,|\, \boldsymbol{x}) = \mathrm{Ga}(\lambda \,|\, r + 1/2, n).$$

- The Normal approximation is

$$\pi(\lambda \,|\, \boldsymbol{x}) \approx \mathrm{N}\{\lambda \,|\, \hat{\lambda}, (n\, F(\hat{\lambda}))^{-1/2}\} = \mathrm{N}\{\lambda \,|\, r/n,\, \sqrt{r}/n\}.$$

- Numerical illustration. Samples with $n = 5$ and $n = 25$, simulated from a Poisson distribution with $\lambda = 3$, yielded $r = 19$ and $r = 82$.



The exact posteriors (solid black) are well approximated (dashed red) even for relatively small samples.

Reference Analysis

□ No Relevant Initial Information

- Identify the mathematical form of a “noninformative” prior: One with *minimal effect, relative to the data, on the posterior distribution of the quantity of interest*. Intuitive basis:

(i) Use *information theory* to measure the amount on information about the quantity of interest to be expected from the data, which depends on prior knowledge.

(ii) Define the *missing information* about the quantity of interest as that which infinite independent replications of the experiment could possible provide.

(iii) Define the *reference prior* as that which *maximizes the missing information about the quantity if interest*.

□ Expected information from the data

- Given model $\{p(\boldsymbol{x}\,|\,\theta), \boldsymbol{x} \in \mathcal{X}, \theta \in \boldsymbol{\Theta}\}$, the *amount of information* $I^\theta\{\mathcal{X}, p(\theta)\}$ which may be expected to be provided by $\boldsymbol{x}$, about the value of θ is defined by

$$I^\theta\{\mathcal{X}, p(\theta)\} = \delta\{p(\boldsymbol{x}, \theta), p(\boldsymbol{x})p(\theta)\},$$

the *intrinsic discrepancy* between the joint distribution $p(\boldsymbol{x}, \theta)$ and the product of their marginals $p(\boldsymbol{x})p(\theta)$, which is the *instrinsic association* between the random quantities $\boldsymbol{x}$ and θ.

- This is related to Shannon mutual information between $\boldsymbol{x}$ and θ, $S\{p(\boldsymbol{x}, \theta)\} = \int_{\mathcal{X}} \int_\Theta p(\boldsymbol{x}, \theta) \log \frac{p(\boldsymbol{x},\theta)}{p(\boldsymbol{x})p(\theta)}\, d\boldsymbol{x}\, d\theta$, while $\delta\{p(\boldsymbol{x}, \theta), p(\boldsymbol{x})p(\theta)\}$ is $\min[\int \int p(\boldsymbol{x}, \theta) \log \frac{p(\boldsymbol{x},\theta)}{p(\boldsymbol{x})p(\theta)}\, d\boldsymbol{x}\, d\theta, \int \int p(\boldsymbol{x})p(\theta) \log \frac{p(\boldsymbol{x})p(\theta)}{p(\boldsymbol{x},\theta)}\, d\boldsymbol{x}\, d\theta]$.

  Often, but not always, the minimum is attained by integrating with the joint density, in which case, $I^\theta\{\mathcal{X}, \pi(\theta)\} = S\{p(\boldsymbol{x}, \theta)\}$.

- Consider $I^{\theta}\{\mathcal{X}^k, p(\theta)\}$ the information about θ which may be expected from k conditionally independent replications of the original setup.

As $k \to \infty$, this would provide any *missing information* about θ. Hence, as $k \to \infty$, the functional $I^{\theta}\{\mathcal{X}^k, \pi(\theta)\}$ will approach the missing information about θ associated with the prior $p(\theta)$.

- Let $\pi_k(\theta)$ be the prior which maximizes $I^{\theta}\{\mathcal{X}^k, p(\theta)\}$ in the class $\mathcal{P}$ of strictly positive prior distributions compatible with accepted assumptions on the value of θ (which could be the class of *all* strictly positive priors).

The *reference prior* $\pi^*(\theta)$ is the limit as $k \to \infty$ (in a sense to be made precise) of the sequence of priors $\{\pi_k(\theta), k = 1, 2, \ldots\}$.

□ Reference priors in the finite case

- If θ may only take a *finite* number m of different values $\{\theta_1, \ldots, \theta_m\}$ and $p(\theta) = \{p_1, \ldots, p_m\}$, with $p_i = \Pr(\theta = \theta_i)$, then $\lim_{k\to\infty} I^\theta\{\mathcal{X}^k, p(\theta)\} = H(p_1, \ldots, p_m) = -\sum_{i=1}^m p_i \log(p_i)$, that is, the *entropy* of the prior distribution $\{p_1, \ldots, p_m\}$.
- In the finite case, with no additional structure,the reference prior is that with *maximum entropy* within the class $\mathcal{P}$ of priors compatible with accepted assumptions (cf. Statistical Physics).
- If, in particular, $\mathcal{P}$ contains *all* priors over $\{\theta_1, \ldots, \theta_m\}$, the reference prior is the *uniform* prior, $\pi(\theta) = \{1/m, \ldots, 1/m\}$ (cf. Bayes-Laplace postulate of insufficient reason) .
- Example. Prior $\{p_1, p_2, p_3, p_4\}$ in genetics problem with $p_1 = 2p_2$. Reference prior (maximum entropy given restriction) is $\pi(\boldsymbol{p}) = \{0.324, 0.162, 0.257, 0.257\}$.

□ Reference priors in one-dimensional continuous case

- Let $\pi_k(\theta)$ be the prior which maximizes $I^\theta\{\mathcal{X}^k, \pi(\theta)\}$ in the class $\mathcal{P}$ of acceptable priors. For any data $\boldsymbol{x} \in \mathcal{X}$, let $\pi_k(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, \pi_k(\theta)$ be the corresponding posterior.

- The *reference posterior density* $\pi^*(\theta \,|\, \boldsymbol{x})$ is defined to be the limit of the sequence $\{\pi_k(\theta \,|\, \boldsymbol{x}), k = 1, 2, \ldots\}$. A *reference prior function* $\pi^*(\theta)$ is any positive function such that, for all $\boldsymbol{x} \in \mathcal{X}$, $\pi^*(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, \pi^*(\theta)$. This is defined up to an (irrelevant) arbitrary constant.

- Let $\boldsymbol{x}^{(k)} \in \mathcal{X}^{\mathbf{k}}$ be the result of k independent replications of $\boldsymbol{x} \in \mathcal{X}$. The exact expression for $\pi_k(\theta)$ (cf. calculus of variations) is
$\pi_k(\theta) = \exp\left[\,\mathrm{E}_{\boldsymbol{x}^{(k)} \,|\, \theta}\{\log \pi_k(\theta \,|\, \boldsymbol{x}^{(k)})\}\right]$ (a geometric average).

- This formula may be used, by repeated simulation from $p(\boldsymbol{x} \,|\, \theta)$ for different θ values, to obtain a *numerical approximation* to the reference prior.

□ Reference priors under regularity conditions

- Let $\tilde\theta_k = \tilde\theta(x^{(k)})$ be a consistent, asymptotically sufficient estimator of θ. In regular problems this is often the case with the mle $\hat\theta$.
- The exact expression for $\pi_k(\theta)$ then becomes, for large k,

  $\pi_k(\theta) \approx \exp[\mathrm{E}_{\tilde\theta_k \,|\, \theta}\{\log \pi_k(\theta \,|\, \tilde\theta_k)\}]$.

  As $k \rightarrow \infty$ this converges to $\pi_k(\theta \,|\, \tilde\theta_k)|_{\tilde\theta_k=\theta}$.
- Let $\tilde\theta_k = \tilde\theta(x^{(k)})$ be a consistent, asymptotically sufficient estimator of θ. Let $\pi(\theta \,|\, \tilde\theta_k)$ be any asymptotic approximation to $\pi(\theta \,|\, x^{(k)})$, the posterior distribution of θ. The reference prior may then be analytically computed as $\pi^*(\theta) = \pi(\theta \,|\, \tilde\theta_k)|_{\tilde\theta_k=\theta}$.
- Under regularity conditions, the posterior distribution of θ is asymptotically Normal, with mean $\hat\theta$ and precision $n\,F(\hat\theta)$, where $F(\theta) = -\mathrm{E}_{\boldsymbol{x} \,|\, \theta}[\partial^2 \log p(\boldsymbol{x} \,|\, \theta)/\partial\theta^2]$ is Fisher's information function. Hence, using the expression above, $\pi^*(\theta) = F(\theta)^{1/2}$ (Jeffreys' rule).

□ One nuisance parameter

- *Two parameters*: reduce the problem to a *sequential* application of the one parameter case. Model is $\{p(\boldsymbol{x}\,|\,\theta, \lambda, \theta \in \Theta, \lambda \in \Lambda\}$ and a θ-reference prior $\pi^*_\theta(\theta, \lambda)$ is required. Two steps:

  (i) Conditional on θ, $p(\boldsymbol{x}\,|\,\theta, \lambda)$ only depends on λ, and it is possible to obtain the *conditional* reference prior $\pi^*(\lambda\,|\,\theta)$.

  (ii) If $\pi^*(\lambda\,|\,\theta)$ is a proper distribution, integrate out λ to get the one-parameter model $p(\boldsymbol{x}\,|\,\theta) = \int_\Lambda p(\boldsymbol{x}\,|\,\theta, \lambda)\,\pi^*(\lambda\,|\,\theta)\,d\lambda$, and use the one-parameter solution to obtain $\pi^*(\theta)$.
- The θ-reference prior is then $\pi^*_\theta(\theta, \lambda) = \pi^*(\lambda\,|\,\theta)\,\pi^*(\theta)$.
- The required reference posterior is $\pi^*(\theta\,|\,\boldsymbol{x}) \propto p(\boldsymbol{x}\,|\,\theta)\pi^*(\theta)$.

- If $\pi^*(\lambda \,|\, \theta)$ is an *improper* prior function, proceed within an increasing sequence $\{\Lambda_i\}$ over which $\pi^*(\lambda \,|\, \theta)$ is integrable and, for given data $\boldsymbol{x}$, obtain the corresponding sequence of reference posteriors $\{\pi_i^*(\theta \,|\, \boldsymbol{x}\}$, defined over the $\{\Lambda_i\}$'s.
- The required reference posterior $\pi^*(\theta \,|\, \boldsymbol{x})$ is their intrinsic limit.

$$\pi^*(\theta \,|\, \boldsymbol{x}) = \lim_{i \to \infty} \pi_i^*(\theta \,|\, \boldsymbol{x}).$$

- A θ-reference prior is any positive function such that, for any data $\boldsymbol{x}$,

$$\pi^*(\theta \,|\, \boldsymbol{x}) \propto \int_\Lambda p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi_\theta^*(\theta, \lambda)\, d\lambda.$$

□ The regular two-parameter continuous case

- Model $p(\boldsymbol{x}\,|\,\theta,\lambda)$. If the joint posterior of (θ,λ) is asymptotically normal, the θ-reference prior may be derived in terms of the corresponding Fisher's information matrix,

$$\boldsymbol{F}(\theta,\lambda)=\begin{pmatrix} F_{\theta\theta}(\theta,\lambda) & F_{\theta\lambda}(\theta,\lambda) \\ F_{\theta\lambda}(\theta,\lambda) & F_{\lambda\lambda}(\theta,\lambda)\end{pmatrix},\quad \boldsymbol{S}(\theta,\lambda)=\boldsymbol{F}^{-1}(\theta,\lambda)$$

- The θ-reference prior is

$$\pi^*_\theta(\theta,\lambda)=\pi^*(\lambda\,|\,\theta)\,\pi^*(\theta),$$

$$\pi^*(\lambda\,|\,\theta)\propto F^{1/2}_{\lambda\lambda}(\theta,\lambda),\ \ \lambda\in\Lambda,$$

and, if $\pi^*(\lambda\,|\,\theta)$ is proper,

$$\pi^*(\theta)\propto\exp\Big\{\int_\Lambda\pi^*(\lambda\,|\,\theta)\,\log[S^{-1/2}_{\theta\theta}(\theta,\lambda)]\,d\lambda\Big\},\ \ \theta\in\Theta.$$

- If $\pi^*(\lambda \mid \theta)$ is not proper, integrations are to be performed within an approximating sequence $\{\Lambda_i\}$ to obtain a sequence $\{\pi_i^*(\lambda \mid \theta)\, \pi_i^*(\theta)\}$, and the θ-reference prior $\pi_\theta^*(\theta, \lambda)$ is defined as its intrinsic limit,

$$\pi_\theta^*(\theta, \lambda) = \lim_{i \to \infty} \pi_i^*(\lambda \mid \theta)\, \pi_i^*(\theta).$$

- Even if $\pi^*(\lambda \mid \theta)$ is improper, if

(i) θ and λ are variation independent, and

(ii) the functions $S_{\theta\theta}$ and $F_{\lambda\lambda}$ factorize such that

$$S_{\theta\theta}^{-1/2}(\theta, \lambda) \propto f_\theta(\theta)\, g_\theta(\lambda),$$
$$F_{\lambda\lambda}^{1/2}(\theta, \lambda) \propto f_\lambda(\theta)\, g_\lambda(\lambda),$$

Then the joint reference prior when θ is the quantity of interest is

$$\pi_\theta^*(\theta, \lambda) = f_\theta(\theta)\, g_\lambda(\lambda).$$

□ Example: Inference on normal parameters

- The information matrix for the normal model $\mathrm{N}(x \,|\, \mu, \sigma)$ is

$$\boldsymbol{F}(\mu,\sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}, \quad \boldsymbol{S}(\mu,\sigma) = \begin{pmatrix} \sigma^{2} & 0 \\ 0 & \sigma^{2}/2 \end{pmatrix}$$

Since μ and σ are variation independent, and both $F_{\sigma\sigma}$ and $S_{\mu\mu}$ factorize, $\pi^*(\sigma \,|\, \mu) \propto F_{\sigma\sigma}^{1/2} \propto \sigma^{-1}$, $\pi^*(\mu) \propto S_{\mu\mu}^{-1/2} \propto 1$.

- The μ-reference prior is $\pi^*_\mu(\mu,\sigma) = \pi^*(\sigma \,|\, \mu)\, \pi^*(\mu) = \sigma^{-1}$, *i.e.*, uniform on both μ and $\log \sigma$.
- Since $\boldsymbol{F}(\mu,\sigma)$ is diagonal the σ-reference prior is $\pi^*_\sigma(\mu,\sigma) = \pi^*(\mu \,|\, \sigma)\pi^*(\sigma) = \sigma^{-1}$, the same as $\pi^*_\mu(\mu,\sigma)$.
- In fact, it may be shown that for all *location-scale* models, so that $p(x \,|\, \mu,\sigma) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$, the reference priors for the location and scale parameters are $\pi^*_\mu(\mu,\sigma) = \pi^*_\sigma(\mu,\sigma) = \sigma^{-1}$.

- Within any given model $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ the ϕ-reference prior $\pi^*_\phi(\boldsymbol{\theta})$ maximizes the missing information about $\phi = \phi(\boldsymbol{\theta})$ and, in multiparameter problems, that prior *may change with the quantity of interest* ϕ.
- For instance, consider that the *standardized mean* $\phi = \mu/\sigma$ within a normal $\mathrm{N}(x \mid \mu, \sigma)$ model, is the quantity of interest. Fisher's information matrix in terms of ϕ and σ is $\boldsymbol{F}(\phi, \sigma) = J^t\, \boldsymbol{F}(\mu, \sigma)\, J$, where $J = (\partial(\mu, \sigma)/\partial(\phi, \sigma))$, and this yields

$$\boldsymbol{F}(\phi, \sigma) = \begin{pmatrix} 1 & \phi/\sigma \\ \phi/\sigma & (2+\phi^2)/\sigma^2 \end{pmatrix}, \quad \boldsymbol{S}(\phi, \sigma) = \begin{pmatrix} 1+\phi^2/2 & -\phi\,\sigma/2 \\ -\phi\,\sigma/2 & \sigma^2/2 \end{pmatrix}.$$

- Hence, the ϕ-reference prior is, $\pi^*_\phi(\phi, \sigma) = (1 + \phi^2/2)^{-1/2}\sigma^{-1}$. In the original parametrization, $\pi^*_\phi(\mu, \sigma) = (1 + (\mu/\sigma)^2/2)^{-1/2}\sigma^{-2}$, which is different from $\pi^*_\mu(\mu, \sigma) = \pi^*_\sigma(\mu, \sigma)$. The prior $\pi^*_\phi(\mu, \sigma)$ leads to a posterior for ϕ with *consistent marginalization properties*.

□ Many parameters

- The reference algorithm generalizes to any number of parameters. If the model is $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = p(\boldsymbol{x} \,|\, \theta_1, \ldots, \theta_m)$, a joint reference prior $\pi^*(\phi_m \,|\, \phi_{m-1}, \ldots, \phi_1) \times \ldots \times \pi^*(\phi_2 \,|\, \phi_1) \times \pi^*(\phi_1)$ may sequentially be obtained for each *ordered parametrization*, $\{\phi_1(\boldsymbol{\theta}), \ldots, \phi_m(\boldsymbol{\theta})\}$.

  Reference priors are *invariant* under reparametrization of the $\phi_i(\boldsymbol{\theta})$'s.

- The choice of the ordered parametrization $\{\phi_1, \ldots, \phi_m\}$ describes the particular prior required, namely that which *sequentially* maximizes the missing information about each of the ϕ_i's, conditional on $\{\phi_1, \ldots, \phi_{i-1}\}$, for $i = m, m-1, \ldots, 1$.

- In many problems, the results are equal or similar for many parameterizations, but this is not necessarily the case. As a consequence, an approximate *overall* reference prior is often pragmatically required.

- An *overall* reference prior, one leading to marginal posteriors which are not far from the appropriate reference posteriors for all quantities of interest, may be obtained by minimizing the average intrinsic discrepancy between the marginal and the reference posteriors. For details sees Berger, Bernardo and Sun (2014).

- *Example: Stein's paradox*. Data random from a m-variate normal $\mathrm{N}_m(\boldsymbol{x}\,|\,\boldsymbol{\mu}, \boldsymbol{I})$. The reference prior for any permutation of the μ_i's is uniform, and leads to appropriate posterior distributions for any of the μ_i's, but cannot be used if the quantity of interest is $\theta = \sum_i \mu_i^2$, the distance of $\boldsymbol{\mu}$ to the origin, for this leads to an inconsistent estimation of θ.

However, the reference prior for θ, $\pi_\theta(\boldsymbol{\mu})\}$ produces, an appropriate, consistent marginal reference posterior for θ.

Inference Summaries

□ Summarizing the posterior distribution

- *The* Bayesian final *outcome* of a problem of inference about any unknown quantity $\boldsymbol{\theta}$ *is* precisely the *posterior density* $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, C)$.
- Bayesian inference may be described as the problem of stating a probability distribution for the quantity of interest encapsulating all available information about its value.
- In one or two dimensions, a *graph of the posterior probability density* of the quantity of interest conveys an intuitive summary of the main conclusions. This is greatly appreciated by users, and is an important asset of Bayesian methods.
- But graphical methods are not easily extended to more than two dimensions and elementary *quantitative* conclusions are often required.

The simplest forms to *summarize* the information contained in the posterior distribution are closely related to the conventional concepts of point estimation and interval estimation.

□ Point Estimation: Posterior mean and posterior mode

- It is often required to provide *point estimates* of relevant quantities. Bayesian point estimation is best described as a *decision problem* where one has to *choose* a particular value $\tilde{\boldsymbol{\theta}}$ as an approximate proxy for the actual, unknown value of $\boldsymbol{\theta}$.
- Intuitively, any location measure of the posterior density $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ may be used as a point estimator. When they exist, either

  $\mathrm{E}[\boldsymbol{\theta} \mid \boldsymbol{x}] = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta}\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta}$ *(posterior mean)*, or

  $\mathrm{Mo}[\boldsymbol{\theta} \mid \boldsymbol{x}] = \arg\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ *(posterior mode)*

  are often regarded as natural choices.

- *Lack of invariance*. Neither the posterior mean not the posterior mode are invariant under reparametrization. The point estimator $\tilde{\boldsymbol{\psi}}$ of a bijection $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ will generally not be equal to $\boldsymbol{\psi}(\tilde{\boldsymbol{\theta}})$.
- In pure "inferential" applications, where one is requested to provide a point estimate of the vector of interest without an specific application in mind, it is difficult to justify a non-invariant solution:

  The best estimate of, say, $\phi = \log(\theta)$ should be $\phi^* = \log(\theta^*)$, but this is not the case if the posterior mean or the posterior mode are used as point estimators.
- Notice that most estimation procures in conventional statistics also suffer from this lack of (intuitively necessary) invariance under reparameterization. However, general invariant multivariate definitions of point estimators is possible using Bayesian *decision theory*.

□ Point Estimation: Posterior median

- In *one-dimensional continuous* problems the *posterior median*, is easily defined and computed as

$\mathrm{Me}[\theta \,|\, \boldsymbol{x}] = q\,; \quad \int_{\{\theta \le q\}} \pi(\theta \,|\, \boldsymbol{x})\, d\theta = 1/2.$

The one-dimensional posterior median has attractive properties:

(i) it is *invariant* under bijections, $\mathrm{Me}[\phi(\theta) \,|\, \boldsymbol{x}] = \phi(\mathrm{Me}[\theta \,|\, \boldsymbol{x}])$.

(ii) it *exists* and it is *unique* under very wide conditions.

(iii) it is rather *robust* under moderate perturbations of the data.

- The posterior median is often considered to be the best 'automatic' Bayesian point estimator in *one-parameter* continuous problems, but its definition is *not* easily extended to a multiparameter setting.

□ General Credible Regions

- To describe $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ it is often convenient to quote regions $\boldsymbol{\Theta}_p \subset \boldsymbol{\Theta}$ of given probability content p under $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x})$. This is the intuitive basis of graphical representations like boxplots.
- A subset $\boldsymbol{\Theta}_p$ of the parameter space $\boldsymbol{\Theta}$ such that
  $\int_{\boldsymbol{\Theta}_p} \pi(\boldsymbol{\theta}\,|\,\boldsymbol{x})\, d\boldsymbol{\theta} = p$ and, hence, $\Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta}_p\,|\,\boldsymbol{x}) = p$,
  is a *posterior p-credible region* for $\boldsymbol{\theta}$.
- A credible region is invariant under reparametrization:
  If $\boldsymbol{\Theta}_p$ is p-credible for $\boldsymbol{\theta}$, $\boldsymbol{\phi}(\boldsymbol{\Theta}_p)$ is a p-credible for $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$.
- For any given p there are generally infinitely many credible regions
- Credible regions may be selected to have minimum size (length, area, volume), resulting in *highest probability density* (HPD) regions, where all points in the region have larger probability density than all points outside.

- Like their related modal estimates, HPD regions are *not invariant*. Thus, the image $\boldsymbol{\phi}(\boldsymbol{\Theta}_p)$ of an HPD region $\boldsymbol{\Theta}_p$ will be a credible region for $\boldsymbol{\phi}$, but will not generally be HPD.

  There is no reason to restrict attention to HPD credible regions.

□ Credible Intervals

- In *one-dimensional continuous* problems, posterior quantiles are often used to derive credible intervals.
- If $\theta_q = Q_q[\theta \,|\, \boldsymbol{x}]$ is the q-quantile of the posterior distribution of θ,
  the interval $\Theta_p = \{\theta;\ \theta \leq \theta_p\}$ is a p-credible region,
  and it is *invariant under reparameterization.*
- *Equal-tailed* p-credible intervals of the form
  $\Theta_p = \{\theta;\ \ \theta_{(1-p)/2} \leq \theta \leq \theta_{(1+p)/2}\}$
  are typically unique, and they invariant under reparametrization.

- Example: Credible intervals for the normal mean.
- With model $\mathrm{N}(x \mid \mu, \sigma)$, the reference posterior for μ given $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is $\pi(\mu \mid \boldsymbol{x}) = \mathrm{St}(\mu \mid \bar{x}, s/\sqrt{n-1}, n-1)$. Hence the reference *posterior* distribution of

$$\tau = \sqrt{n-1}(\mu - \bar{x})/s,$$

as a function of μ, is $\pi(\tau \mid \bar{x}, s, n) = \mathrm{St}(\tau \mid 0, 1, n-1)$.

- It then follows that the equal-tailed p-credible intervals for μ are

$$\{\mu;\ \mu \in \bar{x} \pm q_{n-1}^{(1-p)/2}\, s/\sqrt{n-1}\},$$

where $q_{n-1}^{(1-p)/2}$ is the $(1-p)/2$ quantile of a standard Student density with $n-1$ degrees of freedom.

- To study the long term behavior of these credible intervals, the expression $\sqrt{n-1}(\mu - \bar{x})/s$ may *also* be analyzed, for fixed μ, as a *function of the data*.

□ Calibration

The fact that the *sampling* distribution of the statistic

$$t = t(\bar{x}, s \,|\, \mu, n) = \sqrt{n-1}(\mu - \bar{x})/s$$

is *also* an standard Student $p(t \,|\, \mu, n) = \mathrm{St}(t \,|\, 0, 1, n-1)$ with the same degrees of freedom implies that, in this example, objective Bayesian credible intervals are *also* be *exact* frequentist confidence intervals.

- *Exact numerical agreement* (exact matching) between Bayesian credible intervals and frequentist confidence intervals is however the *exception, not the norm*.
- For *large samples*, convergence to normality implies *approximate numerical agreement*. This provides a frequentist *calibration* to objective Bayesian methods.

- Exact numerical *agreement* is obviously *impossible when the data are discrete*: Precise (non randomized) frequentist confidence intervals do not exist in that case for most confidence levels.

  The computation of Bayesian credible regions for continuous parameters is however *precisely the same* whether the data are *discrete or continuous*.

- The coverage properties of credible regions obtained form different objective priors is often computed to compare the behavior of the priors used to derive them. Although exact numerical agreement of posterior probabilities and average coverage is *not* to be expected, they should be close to order $O(n^{-1/2})$.

Prediction

□ Posterior predictive distributions

- Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, $x_i \in \mathcal{X}$, set of "homogeneous" observations. Desired to predict the value of a future observation $x \in \mathcal{X}$ generated by the same mechanism.
- From the foundations arguments the solution *must* be a probability distribution $p(x \,|\, \boldsymbol{x}, K)$ describing the uncertainty on the value that x will take, given data $\boldsymbol{x}$ and any other available knowledge K. This is called the (posterior) *predictive density* of x.
- To derive $p(x \,|\, \boldsymbol{x}, K)$ it is necessary to specify the *precise sense* in which the x_i's are judged to be *homogeneous*.
- It is often directly assumed that the data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of a *random sample* from some statistical model.

□ Posterior predictive distributions from random samples

- Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, $x_i \in \mathcal{X}$, a random sample of size n from the statistical model $\{p(x \,|\, \boldsymbol{\theta}), x \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and let $\pi(\boldsymbol{\theta})$ a prior distribution describing available knowledge (in any) about the value of the parameter vector $\boldsymbol{\theta}$.
- The *posterior predictive distribution* is

$$p(x \,|\, \boldsymbol{x}) = p(x \,|\, x_1, \ldots, x_n) = \int_{\boldsymbol{\Theta}} p(x \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}.$$

This encapsulates all available information about the outcome of any future observation $x \in \mathcal{X}$ from the same model.

- To prove this, make use the total probability theorem, to have $p(x \,|\, \boldsymbol{x}) = \int_{\boldsymbol{\Theta}} p(x \,|\, \boldsymbol{\theta}, \boldsymbol{x})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}$, and notice the new observation x has been assumed to be conditionally independent of the observed data $\boldsymbol{x}$, so that $p(x \,|\, \boldsymbol{\theta}, \boldsymbol{x}) = p(x \,|\, \boldsymbol{\theta})$.

- The observable values $x \in \mathcal{X}$ may be either *discrete* or *continuous* random quantities. In the discrete case, the predictive distribution will be described by its probability *mass* function; in the continuous case, by its probability *density* function. Both are denoted $p(x \mid \boldsymbol{x})$.

□ Prediction in a Poisson process

- Data $\boldsymbol{x} = \{r_1, \ldots, r_n\}$ random from $\mathrm{Po}(r \mid \lambda)$. The reference posterior density of λ is $\pi^*(\lambda \mid \boldsymbol{x}) = \mathrm{Ga}(\lambda \mid t + 1/2, n)$, where $t = \Sigma_j \, r_j$.

The (reference) posterior predictive distribution is

$$\begin{aligned} p(r \mid \boldsymbol{x}) &= \Pr[r \mid t, n] = \int_0^\infty \mathrm{Po}(r \mid \lambda) \, \mathrm{Ga}(\lambda \mid t + 1/2, n) \, d\lambda \\ &= \frac{n^{t+1/2}}{\Gamma(t+1/2)} \frac{1}{r!} \frac{\Gamma(r+t+1/2)}{(1+n)^{r+t+1/2}} \,, \end{aligned}$$

an example of a Poisson-Gamma probability mass function.

- *Example. Flash floods.* No flash floods have been recorded on a particular location in 10 consecutive years. Local authorities are interested in forecasting possible future flash floods. Using a Poisson model, and assuming that meteorological conditions remain similar, the probabilities that r flash floods will occur next year in that location are given by the Poisson-Gamma mass function above, with $t = 0$ and $n = 10$. This yields,

$$\{\Pr[0 \mid t, n] = 0.953,\ \Pr[1 \mid t, n] = 0.043,\ \Pr[2 \mid t, n] = 0.003, \ldots\}.$$

□ Prediction of Normal measurements

- Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ random from $N(x \mid \mu, \sigma)$. Reference prior $\pi^*(\mu, \sigma) = \sigma^{-1}$ or, in terms of the precision $\lambda = \sigma^{-2}$, $\pi^*(\mu, \lambda) = \lambda^{-1}$.
- The *joint* reference posterior, $\pi^*(\mu, \lambda \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \mu, \lambda)\, \pi^*(\mu, \lambda)$, is $\pi^*(\mu, \lambda \mid \boldsymbol{x}) = N(\mu \mid \bar{x}, (n\lambda)^{-1/2})\, \mathrm{Ga}(\lambda \mid (n-1)/2, ns^2/2)$.

- The corresponding predictive distribution is

$$\begin{aligned}\pi^*(x \,|\, \boldsymbol{x}) &= \int_0^\infty \int_{-\infty}^\infty \mathrm{N}(x \,|\, \mu, \lambda^{-1/2})\, \pi^*(\mu, \lambda \,|\, \boldsymbol{x})\, d\mu\, d\lambda \\ &\propto \{(1+n)s^2 + (\mu - \bar{x})^2\}^{-n/2},\end{aligned}$$

  a kernel of the *Student* density

$$\pi^*(x \,|\, \boldsymbol{x}) = \mathrm{St}(x \,|\, \bar{x}, s\sqrt{\frac{n+1}{n-1}}, n-1).$$

- *Example. Production of safety belts*. Observed breaking strengths of $n = 10$ randomly chosen webbings have mean $\bar{x} = 28.011$ kN and standard deviation $s = 0.443$ kN. Specification requires $x > 26$ kN.

  Reference posterior predictive $p(x \,|\, \boldsymbol{x}) = \mathrm{St}(x \,|\, 28.011, 0.490, 9)$.

  $\Pr(x > 26 \,|\, \boldsymbol{x}) = \int_{26}^\infty \mathrm{St}(x \,|\, 28.011, 0.490, 9)\, dx = 0.9987.$

□ Regression

● In prediction problems there is often *additional information* from relevant covariates. The data structure consists of set of pairs $\{(\boldsymbol{y}_i, \boldsymbol{v}_i)\}$, where both the observables $\boldsymbol{y}_i$ and the covariates $\boldsymbol{v}_i$ may be vectors of any dimension. Given a new observation, with $\boldsymbol{v}$ known, one wants to predict the corresponding value of $\boldsymbol{y}$. Formally, one has to compute the predictive distribution $p\{\boldsymbol{y} \,|\, \boldsymbol{v}, (\boldsymbol{y}_1, \boldsymbol{v}_1), \ldots (\boldsymbol{y}_n, \boldsymbol{v}_n)\}$.

● A model $\{p(\boldsymbol{y} \,|\, \boldsymbol{v}, \boldsymbol{\theta}), \boldsymbol{y} \in \boldsymbol{Y}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is needed which makes precise the probabilistic relationship between $\boldsymbol{y}$ and $\boldsymbol{v}$. The simplest option assumes a *linear dependency* of the form

$$p(\boldsymbol{y} \,|\, \boldsymbol{v}, \boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{y} \,|\, \boldsymbol{V}\boldsymbol{\beta}, \Sigma),$$

but far more complex structures are common in applications.

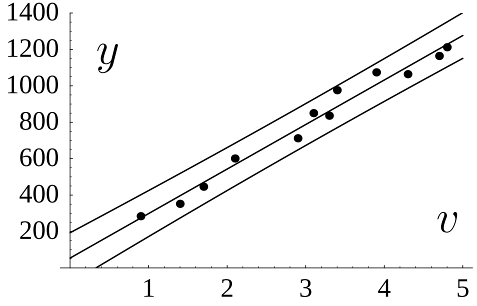
- *Univariate linear regression on k covariates.*
  $Y \subset \Re$, $\boldsymbol{v} = \{v_1, \ldots, v_k\}$.
  $p(y \,|\, \boldsymbol{v}, \boldsymbol{\beta}, \sigma) = \mathrm{N}(y \,|\, \boldsymbol{v}\boldsymbol{\beta}, \sigma^2)$, $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_k\}^t$.
  Data is $\boldsymbol{x} = \{\boldsymbol{y}, \boldsymbol{V}\}$, where $\boldsymbol{y} = \{y_1, \ldots, y_n\}^t$,
  and $\boldsymbol{V}$ the $n \times k$ matrix with the $\boldsymbol{v}_i$'s as rows.
  The likelihood is $p(\boldsymbol{y} \,|\, \boldsymbol{V}, \boldsymbol{\beta}, \sigma) = \mathrm{N}_n(\boldsymbol{y} \,|\, \boldsymbol{V}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$,
  and the reference prior is $\pi^*(\boldsymbol{\beta}, \sigma) = \sigma^{-1}$.
- The predictive posterior is the Student density

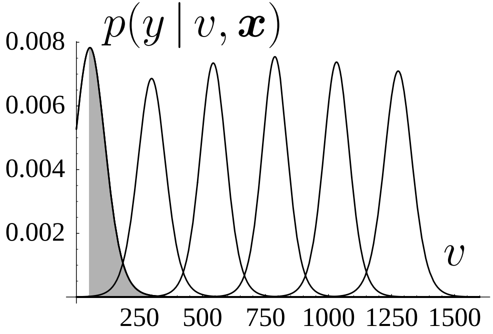
$$p(y \,|\, \boldsymbol{v}, \boldsymbol{y}, \boldsymbol{V}) = \mathrm{St}(y \,|\, \boldsymbol{v}\hat{\boldsymbol{\beta}}, f(\boldsymbol{v}, \boldsymbol{V})\,\frac{ns^2}{n-k}, n-k)$$

$\hat{\boldsymbol{\beta}} = (\boldsymbol{V}^t\boldsymbol{V})^{-1}\boldsymbol{V}^t\boldsymbol{y}$, $\quad ns^2 = (\boldsymbol{y} - \boldsymbol{v}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{v}\hat{\boldsymbol{\beta}})$ and
$f(\boldsymbol{v}, \boldsymbol{V}) = 1 + \boldsymbol{v}(\boldsymbol{V}^t\boldsymbol{V})^{-1}\boldsymbol{v}^t$.

- Example. Pollution control. Data: pairs $\{y_j, v_j\}$of pollution density (in $\mu gr/m^3$) and wind speed from source (in m/s).

y_j	1212	836	850	446	1164	601
v_j	4.8	3.3	3.1	1.7	4.7	2.1
y_j	1074	284	352	1064	712	976
v_j	3.9	0.9	1.4	4.3	2.9	3.4





The predictive distributions $p(y \,|\, v, \boldsymbol{x})$ allow forecasting pollution as a function of wind.

High levels ($y > 50$) are likely even with no wind (v=0) $\Pr[y > 50 \,|\, v = 0, \boldsymbol{x}] = 0.66$ (shaded area).

Hierarchical Models

□ Exchangeability

- Random quantities are often "homogeneous" in the precise sense that only their *values* matter, not the *order* in which they appear. The set of random vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is *exchangeable* iff their joint distribution is invariant under permutations. An infinite sequence $\{\boldsymbol{x}_j\}$ of random vectors is exchangeable if all its finite subsequences are exchangeable.
- *Any random sample is exchangeable.* The *representation theorem* establishes that if observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are exchangeable, they are a *a random sample* from some model $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\theta}\}$, labeled by a *parameter vector* $\boldsymbol{\theta}$, *defined* as the limit (as $n \to \infty$) of some function of the $\boldsymbol{x}_i$'s. Information about $\boldsymbol{\theta}$ in prevailing conditions C is *necessarily* described by *some* probability distribution $\pi(\boldsymbol{\theta} \mid C)$.

- Formally, the joint density of any finite set of exchangeable observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ has an *integral representation* of the form
  $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \,|\, C) = \int_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(\boldsymbol{x}_i \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, C)\, d\boldsymbol{\theta}.$
- Complex data structures may often be usefully described by partial exchangeability assumptions.

- *Example: Public opinion.* Sample k different regions in the country. Sample n_i citizens in region i, and record whether or not ($y_{ij} = 1$ or $y_{ij} = 0$) citizen j would vote A. Assuming *exchangeable* citizens within each region implies

$$p(y_{i1}, \ldots, y_{in_i} \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{n_i} p(y_{ij} \,|\, \theta_i) = \theta_i^{r_i}(1 - \theta_i)^{n_i - r_i},$$

where θ_i is the (unknown) proportion of citizens in region i voting A and $r_i = \Sigma_j y_{ij}$ is the number of citizens voting A in region i.

Assuming regions *exchangeable* within the country similarly yields

$$p(\theta_1, \ldots, \theta_k g \boldsymbol{\phi}) = \prod_{i=1}^{k} \pi(\theta_i \,|\, \boldsymbol{\phi})$$

for some probability distribution $\pi(\theta \,|\, \boldsymbol{\phi})$ describing the political variation within the regions.

Often a general Beta density $\pi(\theta \,|\, \boldsymbol{\phi}) = \mathrm{Be}(\theta \,|\, \alpha, \beta)$ is chosen to describe this variation.

- The resulting *two-stages hierarchical Binomial-Beta model*
  $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$, $\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{in_i}\}$, random from $\mathrm{Bi}(y \,|\, \theta_i)$,
  $\{\theta_1, \ldots, \theta_k\}$, random from $\mathrm{Be}(\theta \,|\, \alpha, \beta)$
  provides a far richer model than (the rather unrealistic, but too frequently used) simple binomial model.

- *Example: Biological response.* Sample k different animals of the same species in a specific environment. Control n_i times animal i and record his responses $\{\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{in_i}\}$ to prevailing conditions. Assuming exchangeable observations within each animal implies

$$p(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{in_i} \,|\, \boldsymbol{\theta}) = \prod\nolimits_{j=1}^{n_i} p(\boldsymbol{y}_{ij} \,|\, \boldsymbol{\theta}_i).$$

Often a normal model $p(\boldsymbol{y}_{ij} \,|\, \boldsymbol{\theta}_i) = \mathrm{N}_r(\boldsymbol{y} \,|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_1)$ is chosen, where the dimension r is the number of biological responses measured.

Assuming exchangeable animals within the environment leads to

$$p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k g \boldsymbol{\phi}) = \prod\nolimits_{i=1}^{k} \pi(\boldsymbol{\mu}_i \,|\, \boldsymbol{\phi})$$

for some probability distribution $\pi(\boldsymbol{\mu} \,|\, \boldsymbol{\phi})$ describing the biological variation within the species.

Often a normal model $\pi(\boldsymbol{\mu} \,|\, \boldsymbol{\phi}) = \mathrm{N}_r(\boldsymbol{\mu} \,|\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_2)$ is chosen.

- The *two-stages hierarchical multivariate Normal-Normal model* $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$, $\boldsymbol{y}_i = \{\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{in_i}\}$, random from $\mathrm{N}_r(\boldsymbol{y} \,|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_1)$, $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$, random from $\mathrm{N}_r(\boldsymbol{\mu} \,|\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_2)$ provides a far richer model than (again unrealistic) simple multivariate normal sampling.
- Finer subdivisions, such as subspecies within each species, similarly lead to hierarchical models with more stages.

□ Bayesian analysis of hierarchical models

- A *two-stages hierarchical model* has the general form $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$, $\boldsymbol{y}_i = \{\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i}\}$

  $\boldsymbol{y}_i$ random sample of size n_i from $p(\boldsymbol{z} \,|\, \boldsymbol{\theta}_i)$, $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$,

  $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$, random of size k from $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{\phi})$, $\boldsymbol{\phi} \in \boldsymbol{\Phi}$.

- To analyze a hierarchical model

(i) Specify a *prior distribution* (or a reference prior function) $\pi(\boldsymbol{\phi})$ for the *hyperparameter vector* $\boldsymbol{\phi}$.

(ii) Use *standard probability theory* to compute all desired *posterior distributions*:

$\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x})$ for inferences about the hyperparameters,

$\pi(\boldsymbol{\theta}_i \,|\, \boldsymbol{x})$ for inferences about the parameters,

$\pi(\psi \,|\, \boldsymbol{x})$ for inferences about the any function $\psi = \psi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$,

$\pi(\boldsymbol{y} \,|\, \boldsymbol{x})$ for predictions on future observations,

$\pi(t \,|\, \boldsymbol{x})$ for predictions on any function $t = t(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$

- *Markov Chain Monte Carlo* (MCMC) based *software* available for the necessary computations.

Integrated Reference Analysis

- The basic machinery of Baysian inference has been reviewed above. However, there is no obvious agreement on the appropriate Bayesian solution to even simple (textbook) stylized problems:

  Best point estimate for the normal variance?

  Inferences on the correlation coefficient of a bivariate normal?

  Comparing two normal means or two binomial proportions?

  Testing compatibility of the normal mean with a precise value?

- Let alone in problems within complex models with many parameters!

  Proposal: Return to basics and use decision-theoretic machinery.

Structure of a Decision Problem

□ Alternatives, consequences, relevant events

- A decision problem if two or more possible courses of action; $\mathcal{A}$ is the class of possible *actions*.

- For each $a \in \mathcal{A}$, $\boldsymbol{\Theta}_a$ is the set of *relevant events*, those may affect the result of choosing a.

- Each pair $\{a, \boldsymbol{\theta}\}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_a$, produces a consequence $c(a, \boldsymbol{\theta}) \in \mathcal{C}_a$. In this context, $\boldsymbol{\theta}$ if often referred to as the *parameter of interest*.

- The class of pairs $\{(\boldsymbol{\Theta}_a, \mathcal{C}_a), a \in \mathcal{A}\}$ describes the *structure* of the decision problem. Without loss of generality, it may be assumed that the possible actions are mutually exclusive, for otherwise the appropriate Cartesian product may be used.

- In many problems the class of relevant events $\boldsymbol{\Theta}_a$ is the same for all $a \in \mathcal{A}$. Even if this is not the case, a comprehensive *parameter space* Θ may be defined as the union of all the Θ_a.

□ Foundations of decision theory

- Different sets of principles capture a minimum collection of logical rules required for "rational" decision-making. These are axioms with strong intuitive appeal.
- Their basic qualitative structure consists of:

(i) The *transitivity* of preferences:

If $a_1 > a_2$ given C, and $a_2 > a_3$ given C, then $a_1 > a_3$ given C.

(ii) The *sure-thing principle*:

If $a_1 > a_2$ given C and E, and $a_1 > a_2$ given C and not E then $a_1 > a_2$ given C.

- To make possible quantitative statements, this is supplemented by

  (iii) The existence of *standard events*

  These are events of known plausibility, which may be used as a unit of measurement, and have the properties of a probability measure.

- These axioms are *not* a description of human decision-making, but a *normative* set of principles defining *coherent* decision-making.

- There are many different axiom sets published with different levels of generality, but they all basically lead to the same set of conclusions, namely:

- The consequences of making mistakes should be evaluated in terms of a real-valued *loss* function $\ell(a, \boldsymbol{\theta})$ which specifies, on a numerical scale, their degree of undesirability.

- The uncertainty about the parameter of interest $\boldsymbol{\theta}$ should be measured with a *probability distribution* $p(\boldsymbol{\theta} \,|\, C)$

$$p(\boldsymbol{\theta} \,|\, C) \geq 0, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}, \qquad \int_{\Theta} p(\boldsymbol{\theta} \,|\, C)\, d\boldsymbol{\theta} = 1,$$

describing all available knowledge about its value, given the conditions C under which the decision must be taken.

- The relative undesirability of available actions $a \in \mathcal{A}$ is measured by their *expected loss*

$$\ell[a \,|\, C] = \int_{\boldsymbol{\Theta}} \ell(a, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, C)\, d\boldsymbol{\theta}, \quad a \in \mathcal{A}.$$

- The best action is that with the smallest expected loss.

Decision structure of Inference Summaries

- Assume data $\boldsymbol{z}$ have been generated as one random observation form $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, where $\boldsymbol{\theta}$ is the vector of interest and $\boldsymbol{\lambda}$ a nuisance parameter vector, and let $p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = p(\boldsymbol{\lambda} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$ be the assumed joint prior.
- Given data $\boldsymbol{z}$ and assuming model $\mathcal{M}_{\boldsymbol{z}}$, the complete solution to all inference questions about $\boldsymbol{\theta}$ is contained in the marginal posterior $p(\boldsymbol{\theta} \mid \boldsymbol{z})$, derived by standard use of probability theory.
- As mentioned before, appreciation of $p(\boldsymbol{\theta} \mid \boldsymbol{z})$ may be enhanced by providing both point and region estimates of the vector of interest $\boldsymbol{\theta}$, and by declaring whether or not some context-suggested specific value $\boldsymbol{\theta}_0$ (or maybe a set of values $\boldsymbol{\Theta}_0$), is (are) compatible with the observed data $\boldsymbol{z}$. This provides useful (and often required) summaries of $p(\boldsymbol{\theta} \mid \boldsymbol{z})$.

- *All* these summaries may be framed as different decision problems which use precisely the same loss function $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ describing, as a function of the (unknown) $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ values which have generated the data, the loss to be suffered if, working with model $\mathcal{M}_{\boldsymbol{z}}$, the value $\boldsymbol{\theta}_0$ were used as a proxy for the unknown value of $\boldsymbol{\theta}$.

- The results dramatically depend on the choices made for both the prior and the loss functions but, given $\boldsymbol{z}$, only depend on those through the expected loss, $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}$.

- As a function of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ is a measure of the unacceptability of all possible values of the vector of interest. This provides a dual, complementary information on all $\boldsymbol{\theta}$ values (on a loss scale) to that provided by the posterior $p(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ (on a probability scale).

□ Point estimation

To choose a point estimate for $\boldsymbol{\theta}$ is a decision problem where the action space is the class $\boldsymbol{\Theta}$ of all possible $\boldsymbol{\theta}$ values.

Definition 1 *The* Bayes estimator $\boldsymbol{\theta}^*(\boldsymbol{z}) = \arg\inf_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}} \overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ *is that which minimizes the posterior expected loss.*

- Conventional examples include the ubiquitous quadratic loss $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = (\boldsymbol{\theta}_0 - \boldsymbol{\theta})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta})$, which yields the posterior mean as the Bayes estimator, and the zero-one loss on a neighborhood of the true value, which yields the posterior mode a a limiting result.
- Bayes estimators with conventional loss functions are typically not invariant under one to one transformations. Thus, the Bayes estimator under quadratic loss of a variance s not the square of the Bayes estimator of the standard deviation. This is rather difficult to explain when one merely wishes to report an estimate of some quantity of interest.

□ Region estimation

Bayesian region estimation is achieved by quoting posterior credible regions. To choose a q-credible region is a decision problem where the action space is the class of subsets of $\boldsymbol{\Theta}$ with posterior probability q.

Definition 2 *(Bernardo, 2005). A* Bayes q-credible region $\boldsymbol{\Theta}_q^*(\boldsymbol{z})$ *is a q-credible region where any value within the region has a smaller posterior expected loss than any value outside the region, so that* $\forall \boldsymbol{\theta}_i \in \boldsymbol{\Theta}_q^*(\boldsymbol{z}),\ \forall \boldsymbol{\theta}_j \notin \boldsymbol{\Theta}_q^*(\boldsymbol{z}), \quad \bar{\ell}(\boldsymbol{\theta}_i \,|\, \boldsymbol{z}) \leq \bar{\ell}(\boldsymbol{\theta}_j \,|\, \boldsymbol{z})$.

- The quadratic loss yields credible regions with those $\boldsymbol{\theta}$ values closest, in the Euclidean sense, to the posterior mean. A zero-one loss function leads to highest posterior density (HPD) credible regions.
- Conventional Bayes regions are often not invariant: HPD regions in one parameterization will not transform to HPD regions in another.

□ Precise hypothesis testing

- Consider a value $\boldsymbol{\theta}_0$ which deserves special consideration. Testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is as a decision problem where the action space $\mathcal{A} = \{a_0, a_1\}$ contains only two elements: to accept (a_0) or to reject (a_1) the hypothesis H_0.

- Foundations require to specify the loss functions $\ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ and $\ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ measuring the consequences of accepting or rejecting H_0 as a function of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. The optimal action is to reject H_0 iif

$$\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} [\ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}]\, p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta} d\boldsymbol{\lambda} > 0.$$

- Hence, only $\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which measures the conditional advantage of rejecting, must be specified.

- Without loss of generality, the function $\Delta\ell_h$ may be written as

$$\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_0$$

where (precisely as in estimation), $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ describes, as a function of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, the non-negative loss to be suffered if $\boldsymbol{\theta}_0$ were used as a proxy for $\boldsymbol{\theta}$, and the constant $\ell_0 > 0$ describes (in the same loss units) the context-dependent non-negative advantage of accepting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ when it is true.

Definition 3 *(Bernardo and Rueda, 2002). The* Bayes test criterion *to decide on the compatibility of $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with available data $\boldsymbol{z}$ is to reject $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ if (and only if), $\bar{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) > \ell_0$, where ℓ_0 is a context dependent positive constant.*

- The compound case may be analyzed by separately considering each of the values which make part of the compound hypothesis to test.

- Using a zero-one loss function, so that the loss advantage of rejecting $\boldsymbol{\theta}_0$ is equal to one whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and zero otherwise, leads to rejecting H_0 if (and only if) $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) < p_0$ for some context-dependent p_0. Use of this loss requires the prior probability $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0)$ to be *strictly positive*. If $\boldsymbol{\theta}$ is a continuous parameter this forces the use of a non-regular "sharp" prior, concentrating a positive probability mass at $\boldsymbol{\theta}_0$, the solution early advocated by Jeffreys.

This formulation (i) implies the use of radically different priors for hypothesis testing than those used for estimation, (ii) precludes the use of conventional, often improper, 'noninformative" priors, and (iii) may lead to the difficulties associated to Jeffreys-Lindley paradox.

- The quadratic loss function leads to rejecting a $\boldsymbol{\theta}_0$ value whenever its Euclidean distance to $\mathrm{E}[\boldsymbol{\theta} \,|\, \boldsymbol{z}]$, the posterior expectation of $\boldsymbol{\theta}$, is sufficiently large.

- The use of continuous loss functions (such as the quadratic loss) permits the use in hypothesis testing of precisely the same priors that are used in estimation.
- With conventional losses the Bayes test criterion is not invariant under one-to-one transformations. Thus, if $\boldsymbol{\phi}(\boldsymbol{\theta})$ is a one-to-one transformation of $\boldsymbol{\theta}$, rejecting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ does not generally imply rejecting $\boldsymbol{\phi}(\boldsymbol{\theta}) = \boldsymbol{\phi}(\boldsymbol{\theta}_0)$.
- The threshold constant ℓ_0, which controls whether or not an expected loss is too large, is part of the specification of the decision problem, and should be context-dependent. However a judicious choice of the loss function leads to calibrated expected losses, where the relevant threshold constant has an immediate, operational interpretation.

Loss Functions in Inference Problems

- A dissimilarity measure $\delta\{p_{\boldsymbol{z}}, q_{\boldsymbol{z}}\}$ between two probability densities $p_{\boldsymbol{z}}$ and $q_{\boldsymbol{z}}$ for a random vector $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ should be

  (i) non-negative, and zero if (and only if) $p_{\boldsymbol{z}} = q_{\boldsymbol{z}}$ a.e.,

  (ii) invariant under one-to-one transformations of $\boldsymbol{z}$,

  (iii) symmetric, so that $\delta\{p_{\boldsymbol{z}}, q_{\boldsymbol{z}}\} = \delta\{q_{\boldsymbol{z}}, p_{\boldsymbol{z}}\}$,

  (iv) defined for densities with strictly nested supports.

Definition 4 *The intrinsic discrepancy* $\delta\{p_1, p_2\}$ *is*

$$\delta\{p_1, p_2\} = \min\left[\,\kappa\{p_1 \,|\, p_2\},\ \kappa\{p_2 \,|\, p_1\}\,\right]$$

where $\kappa\{p_j \,|\, p_i\} = \int_{\boldsymbol{\mathcal{Z}}_i} p_i(\boldsymbol{z}) \log[p_i(\boldsymbol{z})/p_j(\boldsymbol{z})]\, d\boldsymbol{z}$ *is the* (KL) *divergence of* p_j *from* p_i*. The intrinsic discrepancy between* p *and a family* $\mathcal{F} = \{q_i, i \in I\}$ *is the intrinsic discrepancy between* p *and the closest of them,* $\delta\{p, \mathcal{F}\} = \inf_{q, \in \mathcal{F}} \delta\{p, q\}$.

The intrinsic discrepancy loss function

Definition 5 *Consider* $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$. *The intrinsic discrepancy loss from using* $\boldsymbol{\theta}_0$ *as a proxy for* $\boldsymbol{\theta}$ *is the intrinsic discrepancy between the true model and the class of models with* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\mathcal{M}_0 = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}\}$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \delta\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0)\}.$$

□ Invariance

- For any one-to-one reparameterization $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta}, \boldsymbol{\lambda})$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \ell_\delta\{\boldsymbol{\phi}_0, (\boldsymbol{\phi}, \boldsymbol{\psi}) \,|\, \mathcal{M}_{\boldsymbol{z}}\}.$$

This yields invariant Bayes point and region estimators, and invariant Bayes hypothesis testing procedures.

□ Reduction to sufficient statistics

- If $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z})$ is a sufficient statistic for model $\mathcal{M}_{\boldsymbol{z}}$, one may also work with marginal model $\mathcal{M}_{\boldsymbol{t}} = \{p(\boldsymbol{t} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{t} \in \boldsymbol{\mathcal{T}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ since

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{t}}\}.$$

□ Additivity

- If data consist of a random sample $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from some model $\mathcal{M}_{\boldsymbol{x}}$, so that $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}}^n$, and $p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \prod_{i=1}^n p(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = n\, \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{x}}\}.$$

This considerably simplifies frequent computations.

Objective Bayesian Methods

- The methods described may be used with any prior. However, an "objective" procedure, where the prior function is intended to describe a situation where there is no relevant information about the quantity of interest, is often required.
- Objectivity is an emotionally charged word, and it should be explicitly qualified. No statistical analysis is really objective (both the experimental design and the model have strong subjective inputs). However, frequentist procedures are branded as "objective" just because their conclusions are only conditional on the model assumed and the data obtained. Bayesian methods where the prior function is derived from the assumed model are objective is this limited, but precise sense.

□ Development of objective priors

- Vast literature devoted to the formulation of objective priors.
- Reference analysis, (Bernardo, 1979; Berger and Bernardo, 1992; Berger, Bernardo and Sun, 2009), has been a popular approach.

Very general, easily computable one-parameter result:

Theorem 1 *Let $\boldsymbol{z}^{(k)} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$ denote k conditionally independent observations from $\mathcal{M}_{\boldsymbol{z}}$. For sufficiently large k*

$$\pi_k(\theta) \propto \exp\left\{\mathrm{E}_{\boldsymbol{z}^{(k)}\,|\,\theta}[\log p_h(\theta \,|\, \boldsymbol{z}^{(k)})]\right\}$$

where $p_h(\theta \,|\, \boldsymbol{z}^{(k)}) \propto \prod_{i=1}^{k} p(\boldsymbol{z}_i \,|\, \theta)\, h(\theta)$ is the posterior which corresponds to some arbitrarily chosen prior function $h(\theta)$ which makes the posterior proper for any $\boldsymbol{z}^{(k)}$.

□ Approximate reference priors

- Reference priors are derived for an ordered parameterization. Given $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z}\,|\,\boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ with m parameters, the reference prior with respect to $\boldsymbol{\phi}(\boldsymbol{\omega}) = \{\phi_1, \ldots, \phi_m\}$ is sequentially obtained as $\pi(\boldsymbol{\phi}) = \pi(\phi_m\,|\,\phi_{m-1}, \ldots, \phi_1) \times \cdots \times \pi(\phi_2\,|\,\phi_1)\,\pi(\phi_1)$.
- One is often simultaneously interested in several functions of the parameters. Given $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z}\,|\,\boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega} \subset \Re^m\}$ with m parameters, consider a set $\boldsymbol{\theta}(\boldsymbol{\omega}) = \{\theta_1(\boldsymbol{\omega}), \ldots, \theta_r(\boldsymbol{\omega})\}$ of $r > 1$ functions of interest; Berger, Bernardo and Sun (2014) suggest a procedure to select a joint prior $\pi_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ whose corresponding marginal posteriors $\{\pi_{\boldsymbol{\theta}}(\theta_i\,|\,\boldsymbol{z})\}_{i=1}^r$ will be close, for all possible data sets $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, to the set of reference posteriors $\{\pi(\theta_i\,|\,\boldsymbol{z})\}_{i=1}^r$ yielded by the set of reference priors $\{\pi_{\theta_i}(\boldsymbol{\omega})\}_{i=1}^r$ derived under the assumption that each of the θ_i's is of interest.

Definition 6 *Consider model* $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\omega}), \boldsymbol{z} \in \mathcal{Z}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ *and* $r > 1$ *functions of interest,* $\{\theta_1(\boldsymbol{\omega}), \ldots, \theta_r(\boldsymbol{\omega})\}$*. Let* $\{\pi_{\theta_i}(\boldsymbol{\omega})\}_{i=1}^{r}$ *be the relevant reference priors, and* $\{\pi_{\theta_i}(\boldsymbol{z})\}_{i=1}^{r}$ *and* $\{\pi(\theta_i \,|\, \boldsymbol{z})\}_{i=1}^{r}$ *the corresponding prior predictives and marginal posteriors. Let* $\mathcal{F} = \{\pi(\boldsymbol{\omega} \,|\, \boldsymbol{a}), \boldsymbol{a} \in \mathcal{A}\}$ *be a family of prior functions. For each* $\boldsymbol{\omega} \in \boldsymbol{\Omega}$*, the best approximate joint reference prior within* $\mathcal{F}$ *is that which minimizes the average expected intrinsic loss*

$$d(\boldsymbol{a}) = \frac{1}{r} \sum_{i=1}^{r} \int_{\mathcal{Z}} \delta\{\pi_{\theta_i}(\cdot \,|\, \boldsymbol{z}),\, p_{\theta_i}(\cdot \,|\, \boldsymbol{z}, \boldsymbol{a})\}\, \pi_{\theta_i}(\boldsymbol{z})\, d\boldsymbol{z}, \qquad \boldsymbol{a} \in \mathcal{A}.$$

- Example. Use of the Dirichlet family in the m-multinomial model (with $r = m + 1$ cells) yields $\mathrm{Di}(\boldsymbol{\theta} \,|\, 1/r, \ldots, 1/r)$, with important applications to sparse multinomial data and contingency tables.

An Integrated Approach

- We suggest a systematic use of the intrinsic loss function, and an appropriate joint reference prior, for an integrated objective Bayesian solution to both estimation and hypothesis testing in pure inference problems.
- We have stressed foundations-like decision theoretic arguments, but a large collection of detailed, non-trivial examples prove that the procedures advocated lead to attractive, often novel solutions. Details in Bernardo (2011) and references therein.

Intrinsic point estimation

- Given the statistical model $\{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ the intrinsic discrepancy $\delta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ between two parameter values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is the intrinsic discrepancy $\delta\{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1), p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2)\}$ between the corresponding probability models. This is symmetric, non-negative (and zero iff $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$), invariant under both reparameterization and bijections of $\boldsymbol{x}$.
- The intrinsic estimator is the *reference* Bayes estimator which corresponds to the loss defined by the *intrinsic discrepancy*:
- The expected loss with respect to the reference posterior distribution

$$d(\tilde{\boldsymbol{\theta}}\,|\,\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}\, \pi(\boldsymbol{\theta}\,|\,\boldsymbol{x})\, d\boldsymbol{\theta}$$

is an objective measure, in information units, of the *expected* discrepancy between the model $p(\boldsymbol{x}\,|\,\tilde{\boldsymbol{\theta}})$ and the true (unknown) model $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$.

• The *intrinsic estimator* $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\boldsymbol{x})$ is the value which minimizes such expected discrepancy,

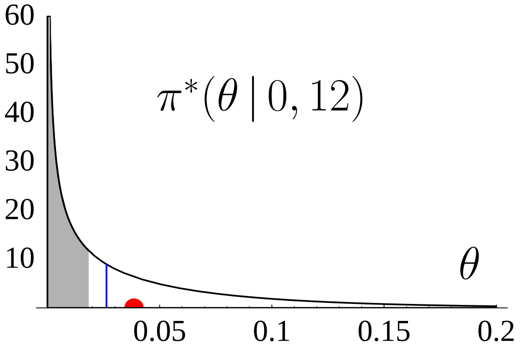
$$\boldsymbol{\theta}^* = \arg \inf_{\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}).$$

□ Example: Intrinsic estimation of the Binomial parameter

- Data are $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, a random sample of size n from $p(x \,|\, \theta) = \theta^x (1-\theta)^{1-x}$, $r = \Sigma x_j$.
- The intrinsic discrepancy is $\delta(\tilde{\theta}, \theta) = n \, \min\{k(\tilde{\theta} \,|\, \theta), k(\theta \,|\, \tilde{\theta})\}$, $k(\theta_1 \,|\, \theta_2) = \theta_2 \log \frac{\theta_2}{\theta_1} + (1-\theta_2) \log \frac{1-\theta_2}{1-\theta_1}$,
- Reference prior and posterior are $\pi^*(\theta) = \mathrm{Be}(\theta \,|\, \frac{1}{2}, \frac{1}{2})$ and $\pi^*(\theta \,|\, r, n) = \mathrm{Be}(\theta \,|\, r + \frac{1}{2}, n - r + \frac{1}{2})$.
- The expected reference discrepancy is

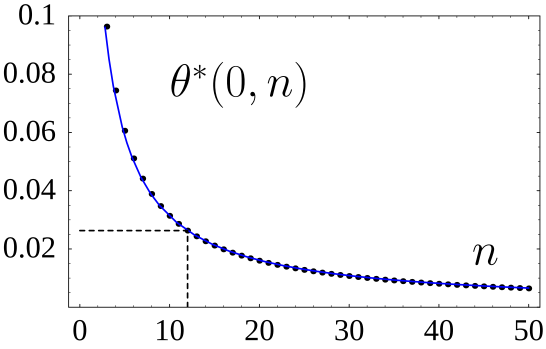
  $d(\tilde{\theta}, r, n) = \int_0^1 \delta(\tilde{\theta}, \theta) \, \pi^*(\theta \,|\, r, n) \, d\theta$

- The intrinsic point estimator is
  $\theta^*(r,n) = \arg\min_{0<\tilde\theta<1} d(\tilde\theta, r, n)$
  If $\phi = \phi(\theta)$, $\phi^* = \phi(\theta^*)$.



- Analytic approximation
  $\theta^*(r,n) \approx \frac{r+1/3}{n+2/3}, \quad n > 2$

- Numerical example
  $n = 12$, $r = 0$,
  $\theta^*(0, 12) = 0.026$
  $\mathrm{Me}[\theta \mid \boldsymbol{x}] = 0.018$
  $\mathrm{E}[\theta \mid \boldsymbol{x}] = 0.038$



□ Intrinsic estimation of the normal variance

- Given a random sample $\{x_1, \ldots, x_n\}$ from a normal distribution $\mathrm{N}(x \mid \mu, \sigma)$, the intrinsic (invariant) point estimator of the normal standard deviation σ is

$$\sigma^* \approx \frac{n}{n-1}\, s, \quad ns^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

- Hence, the intrinsic point estimator of the normal variance σ^2 is

$$\sigma^{2*} \approx \frac{n}{n-1} \frac{ns^2}{n-1}\,,$$

larger than both the mle s^2 and the unbiased estimator $ns^2/(n-1)$.

Intrinsic region (interval) estimation

- The *intrinsic q-credible region* $R^*(q) \subset \Theta$ is that q-credible reference region which corresponds to minimum expected intrinsic loss:

$$\text{(i)} \qquad \int_{R^*(q)} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = q$$

$$\text{(ii)} \qquad \forall \boldsymbol{\theta}_i \in R^*(q),\ \forall \boldsymbol{\theta}_j \notin R^*(q), \qquad d(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}) < d(\boldsymbol{\theta}_j \,|\, \boldsymbol{x})$$

- In one parameter problems, plotting together, with the same scale for θ, both the posterior distribution $\pi(\theta \,|\, \boldsymbol{x})$ and the posterior expected intrinsic loss $d(\theta_i \,|\, \boldsymbol{x})$ provides an illuminating comprehensive view, on a probability density and a loss scale respectively, of the inferential implications of the data $\boldsymbol{x}$ on the value of the parameter θ.

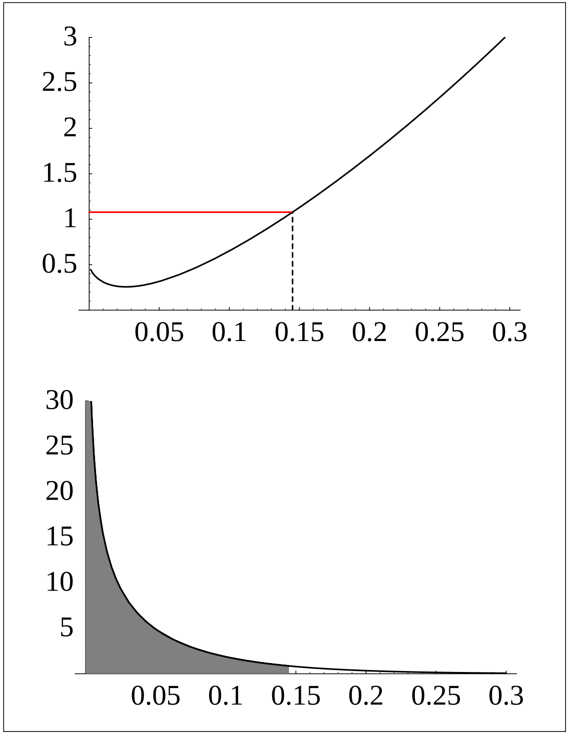
This is illustrated below with two examples from binomial data.

- Binomial example 1

  $r = 0,\ n = 12,$

  $\theta^* = 0.0263;$

  $R^*_{0.95} = [0, 0.145];$

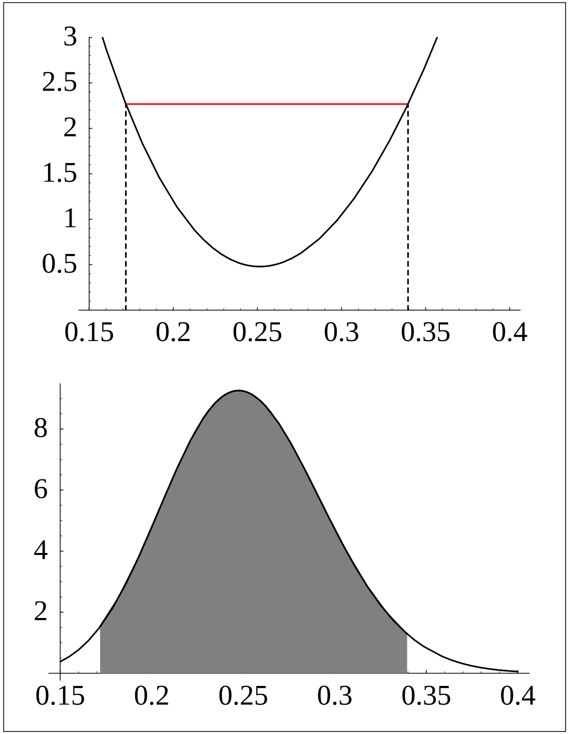


- Binomial example 2

  $r = 25,\ n = 100,$

  $\theta^* = 0.2514;$

  $R^*_{0.95} = [0.172, 0.340];$



Precise Hypothesis Testing

- The *intrinsic reference test criterion* is the Bayes test criterion which corresponds to the use of the intrinsic loss and the reference prior:

Reject $\boldsymbol{\theta}_0$ (and only if) the expected reference posterior intrinsic discrepancy $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ is too large,

$$d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x}) = \int_\Theta \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} > d^*, \text{ for some } d^* > 0.$$

□ Calibration of the test

- The intrinsic reference test statistic $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ is the reference posterior expected value of the intrinsic discrepancy between $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0)$ and $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, which is the minimum expected log-likelihood ratio against the hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Thus, A reference test statistic value $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ of, say, $\log(10) = 2.303$ implies that, given data $\boldsymbol{x}$, the *average* value of the likelihood ratio *against* the hypothesis, $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})/p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0)$, is expected to be about 10, suggesting some *mild evidence* against $\boldsymbol{\theta}_0$.

Similarly, a value $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ of $\log(100) = 4.605$ indicates an average value of the likelihood ratio against $\boldsymbol{\theta}_0$ of about 100, indicating rather *strong evidence* against the hypothesis, and $\log(1000) = 6.908$, a rather conclusive likelihood ratio against the hypothesis the of about 1000.

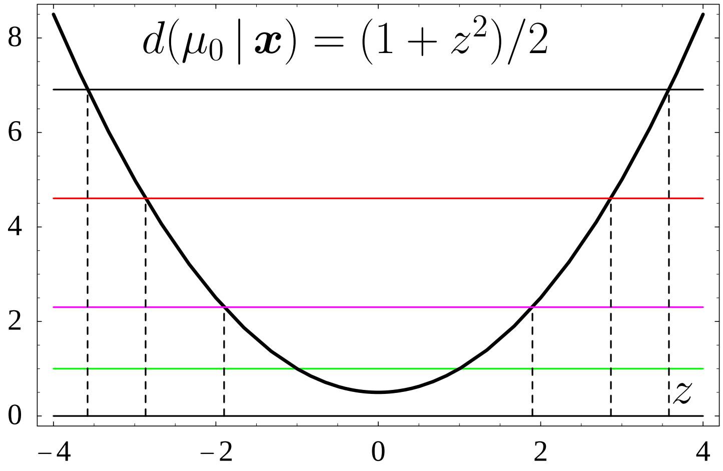
- Notice that the test statistic $d(\theta_0 \,|\, \boldsymbol{x})$ is deemed to be large because it directly indicates that given the data, the true model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ is probably rather different from $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0)$. This has nothing to do with the sampling properties of $d(\theta_0 \,|\, \boldsymbol{x})$ under $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0)$.

□ A canonical example: Testing a value for the Normal mean

- Consider $\mathrm{N}(x \mid \mu, \sigma)$ data. In the simplest case where σ is known, $\delta(\mu_0, \mu) = n(\mu - \mu_0)^2/(2\sigma^2), \qquad \pi(\mu \mid \boldsymbol{x}) = \mathrm{N}(\mu \mid \bar{x}, \sigma/\sqrt{n}),$ $d(\mu_0 \mid \boldsymbol{x}) = \frac{1}{2}(1 + z^2)$, with $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

  Thus rejecting $\mu = \mu_0$ if $d(\mu_0 \mid \boldsymbol{x}) > d^*$ is equivalent to rejecting if $|z| > \sqrt{2d^* - 1}$ and, hence, to a conventional two-sided frequentist test with significance level $\alpha = 2(1 - \Phi(|z|))$.
- The expected value of $d(\mu_0 \mid \boldsymbol{x})$ if the hypothesis is true is

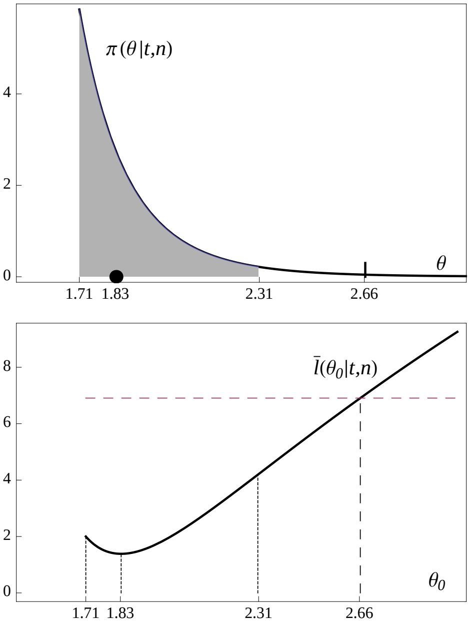
$$\int_{-\infty}^{\infty} \frac{1}{2}(1 + z^2)\mathrm{N}(z \mid 0, 1)\, dz = 1.$$



d^*	$\lvert z\rvert$	α
log(10)	1.8987	0.0576
log(100)	2.8654	0.0042
log(1000)	3.5799	0.0003

- Notice that the conventional $\alpha = 0.05$ significance level corresponds to an expected discrepancy of about log(10), indicating that the likelihood under the null is expected to be only about 10 times less likely that under the true model: a very mild evidence for rejection!

□ Uniform model $\mathrm{Un}(x\,|\,0,\theta)$



$$\ell_\delta\{\theta_0,\theta\,|\,\mathcal{M}_{\boldsymbol{z}}) = n\begin{cases}\log(\theta_0/\theta), & \text{if } \theta_0 \geq \theta,\\ \log(\theta/\theta_0, & \text{if } \theta_0 \leq \theta.\end{cases}$$

$\pi(\theta) = \theta^{-1}, \quad \boldsymbol{z} = \{x_1,\ldots,x_n\},$

$t = \max\{x_1,\ldots,x_n\}, \quad \pi(\theta\,|\,\boldsymbol{z}) = n\,t^n\theta^{-(n+1)}$

The q-quantile is $\theta_q = t\,(1-q)^{-1/n}$;

Exact probability matching.

$\theta^* = t\,2^{1/n}$ (posterior median)

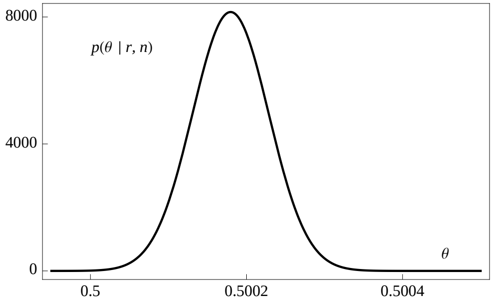
$\mathrm{E}[\overline{\ell}_\delta(\theta_0\,|\,t,n)\,|\,\theta] = (\theta/\theta_0)^n - n\log(\theta/\theta_0)$;

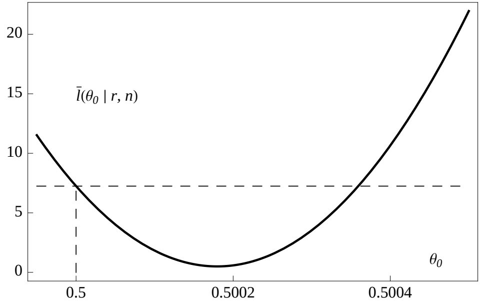
this is equal to 1 if $\theta = \theta_0$,

and increases with n otherwise.

- Simulation: $n = 10$ with $\theta = 2$ which yielded $t = 1.71$;
  $\theta^* = 1.83$, $\Pr[t < \theta < 2.31\,|\,\boldsymbol{z}] = 0.95$, $\overline{\ell}_\delta(2.66\,|\,\boldsymbol{z}) = \log 1000$.

□ Extra Sensory Power (ESP) testing





Jahn, Dunne and Nelson (1987)

Binomial data. Test $H_0 \equiv \{\theta = 1/2\}$ with $n = 104,490,000$ and $r = 52,263,471$.

For any sensible continuous prior $p(\theta)$, $p(\theta \mid \boldsymbol{z}) \approx \mathrm{N}(\theta \mid m_{\boldsymbol{z}}, s_{\boldsymbol{z}})$, with $m_{\boldsymbol{z}} = (r + 1/2)/(n + 1) = 0.50018$, $s_{\boldsymbol{z}} = [m_{\boldsymbol{z}}(1 - m_{\boldsymbol{z}})/(n+2)]^{1/2} = 0.000049$.

$\bar{\ell}(\theta_0 \mid \boldsymbol{z}) \approx \frac{n}{2} \log[1 + \frac{1}{n}(1 + t_{\boldsymbol{z}}(\theta_0)^2)]$,

$t_{\boldsymbol{z}}(\theta_0) = (\theta_0 - m_{\boldsymbol{z}})/s_{\boldsymbol{z}}$, $t_{\boldsymbol{z}}(1/2) = 3.672$.

$\bar{\ell}(\theta_0 \mid \boldsymbol{z}) = 7.24 = \log 1400$: Reject H_0

- Jeffreys-Lindley paradox: With any "sharp" prior, $\Pr[\theta = 1/2] = p_0$, $\Pr[\theta = 1/2 \mid \boldsymbol{z}] > p_0$ (Jefferys, 1990) suggesting data support H_0 !!!

□ More sophisticated examples

- **Two sample problems**: **Equality of two normal means**.
  $\bar{\ell}(H_0 \,|\, \boldsymbol{z}) \approx n \log[1 + \frac{1}{2n}(1 + t^2)], \quad t = \sqrt{n}(\bar{x} - \bar{y})/(s/\sqrt{2})$.
- **Trinomial data**: Testing for **Hardy-Weinberg equilibrium**.
  $\bar{\ell}(H_0 \,|\, \boldsymbol{z}) \approx \int_{\mathcal{A}} \ell_\delta\{H_0, (\alpha_1, \alpha_2)\}\, \pi(\alpha_1, \alpha_2 \,|\, \boldsymbol{z}) d\alpha_1 d\alpha_2$,
  where $\ell_\delta\{H_0, (\alpha_1, \alpha_2)\} = n\, \theta(\alpha_1, \alpha_2)$,
  $\theta(\alpha_1, \alpha_2)$ is the KL distance of H_0 from $\mathrm{Tri}(r_1, r_2, r_3 \,|\, \alpha_1, \alpha_2$ and
  $\pi(\alpha_1, \alpha_2 \,|\, \boldsymbol{z}) = \mathrm{Di}[\alpha_1, \alpha_2 \,|\, r_1 + 1/3, r_2 + 1/3, r_3 + 1/3]$.
- **Contingency tables**: Testing for **independence**.
  Data $\boldsymbol{z} = \{\{n_{11}, \ldots, n_{1b}\}, \ldots, \{n_{a1}, \ldots, n_{ab}\}\}$, $k = a \times b$,
  $\bar{\ell}(H_0 \,|\, \boldsymbol{z}) \approx \int_{\boldsymbol{\Theta}} n\, \phi(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta}, \quad \phi(\boldsymbol{\theta}) = \sum_{i=1}^{a} \sum_{j=1}^{b} \theta_{ij} \log[\frac{\theta_{ij}}{\alpha_i\, \beta_j}]$,
  where $\alpha_i = \sum_{j=1}^{b} \theta_{ij}$ and $\beta_j = \sum_{i=1}^{a} \theta_{ij}$ are the marginals, and
  $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{z}) = \mathrm{Di}_{k-1}(\boldsymbol{\theta} \,|\, n_{11} + 1/k, \ldots, n_{ab} + 1/k)$.

Basic References

(In chronological order)

Available on line from www.uv.es/bernardo

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion).

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35–60 (with discussion).

Bernardo, J. M. (1997). Noninformative priors do not exist *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.* **70**, 351–372.

Bernardo, J. M. (2005a). Reference analysis. *Bayesian Thinking: Modeling and Computation, Handbook of Statistics* **25** (Dey, D. K. and Rao, C. R., eds). Amsterdam: Elsevier, 17–90.

Bernardo, J. M. (2005b). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test* **14**, 317–384 (with discussion).

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 385–402 and 457–464, (with discussion).

Bernardo, J. M. (2007). Objective Bayesian point and region estimation in location-scale models. *Sort* **31**, 3–44, (with discussion).

Bernardo, J. M. (2009). Statistics: Bayesian methodology in statistics. *Comprehensive Chemometrics* (S. Brown, R. Tauler and R. Walczak eds.) Oxford: Elsevier, 213–245.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009a). The formal definition of reference priors. *Ann. Statist.* **37**, 905–938.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009b). Natural induction: An objective Bayesian approach. *Rev. Acad. Sci. Madrid, A* **103**, 125–159 (invited paper with discussion).

Bernardo, J. M. (2011). Integrated objective Bayesian estimation and hypothesis testing. *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds). Oxford: Oxford University Press, 1–68 (with discussion).

Bernardo, J. M. (2011). Bayes and discovery: Objective Bayesian hypothesis testing. *PHY- STAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding* (H. B. Prosper and L. Lyons, eds.). Geneva: CERN, 27-49 (with discussion).

Berger, J. O., Bernardo, J. M. and Sun, D. (2012). Objective priors for discrete parameter spaces. *J. Amer. Statist. Assoc.* **107**, 636–648

Berger, J. O., Bernardo, J. M. and Sun, D. (2014). Overall objective priors. *Bayesian Analysis* (forthcoming, with discussion).

jose.m.bernardo@uv.es

www.uv.es/bernardo