An Introduction to
Objective Bayesian Statistics

José M. Bernardo

Universitat de València, Spain

<jose.m.bernardo@uv.es>

http://www.uv.es/bernardo

Université de Neuchâtel, Switzerland

March 15th–March 17th, 2006

Summary

1. *Concept of Probability*

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

2. *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.

3. *Decision Making*

   Structure of a decision problem. Intrinsic loss functions.
   Point and region estimation. Intrinsic estimators and credible regions.
   Hypothesis testing. Bayesian reference criterion (BRC).

1. Concept of Probability

1.1. Introduction

- Tentatively accept a *formal* statistical model
  - Typically suggested by informal descriptive evaluation
  - Conclusions conditional on the assumption that model is correct
- Bayesian approach firmly based on *axiomatic foundations*
  - Mathematical need to describe by probabilities all uncertainties
  - Parameters *must* have a (*prior*) distribution describing available information about their values
  - *Not* a description of their variability (*fixed unknown* quantities), but a description of the *uncertainty* about their true values.
- Important particular case: no relevant (or subjective) initial information:
  - scientific and industrial reporting, public decision making, ...
  - Prior *exclusively* based on model assumptions and available, well-documented data: *Objective Bayesian Statistics*

- *Notation*

  - Under conditions C, $p(\boldsymbol{x} \mid C)$, $\pi(\boldsymbol{\theta} \mid C)$ are, respectively, *probability* densities (or mass) functions of *observables* $\boldsymbol{x}$ and *parameters* $\boldsymbol{\theta}$
    $p(\boldsymbol{x} \mid C) \geq 0$, $\int_{\mathcal{X}} p(\boldsymbol{x} \mid C)\, \mathrm{d}\boldsymbol{x} = 1$, $\mathrm{E}[\boldsymbol{x} \mid C] = \int_{\mathcal{X}} \boldsymbol{x}\, p(\boldsymbol{x} \mid C)\, \mathrm{d}\boldsymbol{x}$,
    $\pi(\boldsymbol{\theta} \mid C) \geq 0$, $\int_{\Theta} \pi(\boldsymbol{\theta} \mid C)\, \mathrm{d}\boldsymbol{\theta} = 1$, $\mathrm{E}[\boldsymbol{\theta} \mid C] = \int_{\Theta} \boldsymbol{\theta}\, \pi(\boldsymbol{\theta} \mid C)\, \mathrm{d}\boldsymbol{\theta}$.
  - Special densities (or mass) functions use specific notation, as $\mathrm{N}(x \mid \mu, \sigma)$, $\mathrm{Bi}(x \mid n, \theta)$, or $\mathrm{Pn}(x \mid \lambda)$. Other examples:

Beta $\quad \{\mathrm{Be}(x \mid \alpha, \beta), \quad 0 < x < 1, \quad \alpha > 0, \beta > 0\}$

$$\mathrm{Be}(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

Gamma $\quad \{\mathrm{Ga}(x \mid \alpha, \beta), \quad x > 0, \quad \alpha > 0, \beta > 0\}$

$$\mathrm{Ga}(x \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\beta x}$$

Student $\quad \{\mathrm{St}(x \mid \mu, \sigma, \alpha), \quad x \in \Re, \quad \mu \in \Re, \sigma > 0, \alpha > 0\}$

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

- *Statistical Models*

  - *Statistical model* generating $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \Theta\}$

    Parameter vector $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\} \in \Theta$. *Parameter space* $\Theta \subset \Re^k$.
    Data set $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. *Sampling (Outcome) space* $\boldsymbol{\mathcal{X}}$, of arbitrary structure.

  - *Likelihood function* of $\boldsymbol{x}$, $l(\boldsymbol{\theta} \mid \boldsymbol{x})$.

    $l(\boldsymbol{\theta} \mid \boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\theta})$, as a function of $\boldsymbol{\theta} \in \Theta$.

  - *Maximum likelihood estimator (mle)* of $\boldsymbol{\theta}$

    $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\sup_{\boldsymbol{\theta} \in \Theta} \; l(\boldsymbol{\theta} \mid \boldsymbol{x})$

  - Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ *random sample* (iid) from model if

    $p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} p(x_j \mid \boldsymbol{\theta}), \; x_j \in \mathcal{X}, \quad \boldsymbol{\mathcal{X}} = \mathcal{X}^n$

  - Behaviour under repeated sampling (general, not iid data)

    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 obtained as $m \to \infty$

1.2. 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

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

    The functional $\kappa\{\hat{p}\,|\,p\}$ is non-negative, (zero iff, $\hat{p}(\boldsymbol{x}) = p(\boldsymbol{x})$ a.e.) and *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*

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

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

  - Defined if $\boldsymbol{\mathcal{X}}_2 \subset \boldsymbol{\mathcal{X}}_1$ or $\boldsymbol{\mathcal{X}}_1 \subset \boldsymbol{\mathcal{X}}_2$, operative interpretation as the minimum amount of information (in *nits*) required to discriminate.

- *Interpretation and calibration of the intrinsic discrepancy*

  - Let $\{p_1(\boldsymbol{x}\,|\,\boldsymbol{\theta}_1), \boldsymbol{\theta}_1 \in \Theta_1\}$ or $\{p_2(\boldsymbol{x}\,|\,\boldsymbol{\theta}_2), \boldsymbol{\theta}_2 \in \Theta_2\}$ 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*. $\delta = \log[100] \approx 4.6$ *nits*, likelihood ratios for the true model larger than 100 making *discrimination very easy*.

    $\delta = \log(1+\varepsilon) \approx \varepsilon$ *nits*, likelihood ratios for the true model may about $1+\epsilon$ making *discrimination very hard*.

Intrinsic Discrepancy δ	0.01	0.69	2.3	4.6	6.9
Average Likelihood Ratio for **true** model $\exp[\delta]$	1.01	2	10	100	1000

- *Example.* Conventional Poisson approximation $\text{Pn}(r \mid n\theta)$ of Binomial probabilities $\text{Bi}(r \mid n, \theta)$

  Intrinsic discrepancy between Binomial and Poisson distributions

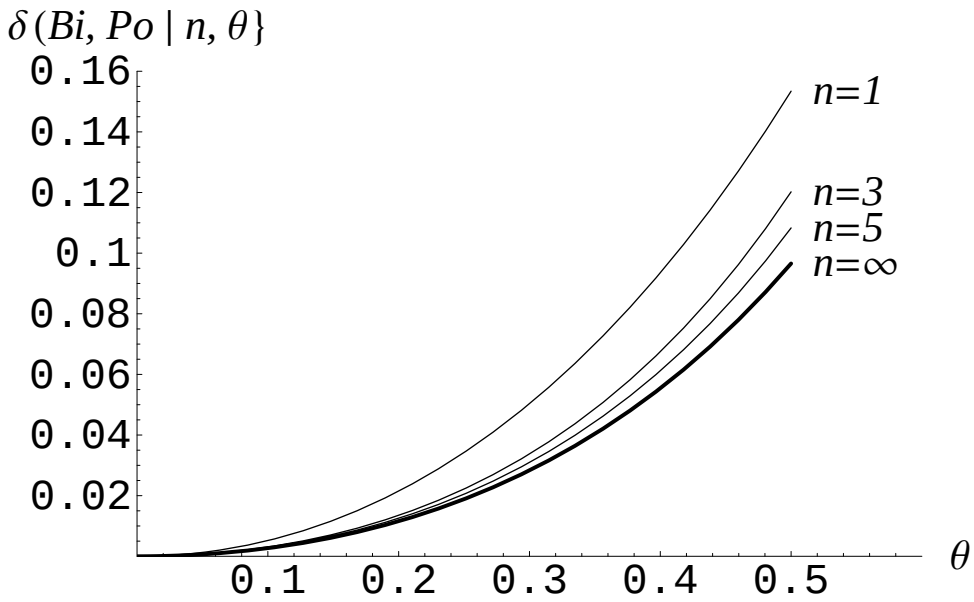
  $$\delta\{\text{Bi}(r \mid n, \theta), \text{Po}(r \mid n\theta\} = \min[k\{\text{Bi} \mid \text{Po}\}, k\{\text{Po} \mid \text{Bi}\}] = k\{\text{Bi} \mid \text{Po}\}$$
  $$= \sum_{r=0}^{n} \text{Bi}(r \mid n,\ \theta) \log[\text{Bi}(r \mid n,\ \theta)/\text{Po}(r \mid n\theta)] = \delta\{n, \theta\}$$

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

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

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

Good Poisson approximations are *impossible* if θ is not small, however large n might be.



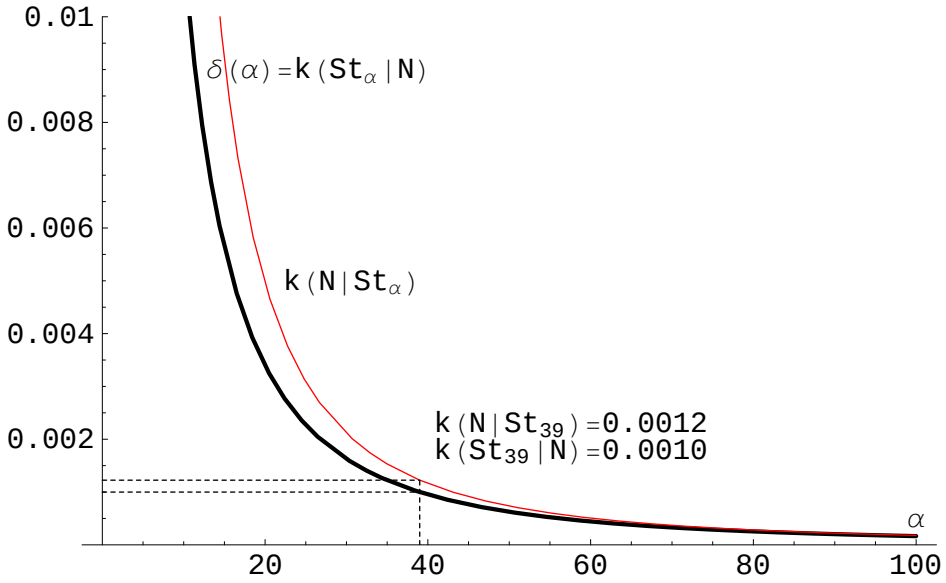
- *Intrinsic Convergence of Distributions*

  - *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.

$$\delta(\alpha) = \delta\{\mathrm{St}(x \,|\, \mu, \sigma, \alpha), \mathrm{N}(x \,|\, \mu, \sigma)\} = \min[k\{\mathrm{St}_\alpha \,|\, \mathrm{N}\}, k\{\mathrm{N} \,|\, \mathrm{St}_\alpha\}]$$

$$= k\{\mathrm{St}_\alpha \,|\, \mathrm{N}\} = \int_{\Re} \mathrm{N}(x \,|\, 0, 1) \log \frac{\mathrm{N}(x \,|\, 0, 1)}{\mathrm{St}(x \,|\, 0, 1, \alpha)} \,\mathrm{d}x \approx \frac{7}{\alpha(22 + 4\alpha)}$$



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

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

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

Expected log-density ratios at least 0.001 when $\alpha < 40$.

1.3. 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 $\{\pi(\boldsymbol{\theta} \,|\, C), \boldsymbol{\theta} \in \Theta\}$ describing the plausibility of their given available conditions C.
  - Axioms have a strong intuitive appeal; examples include
    - *Transitivity of plausibility*.
      If $E_1 \succ E_2 \,|\, C$, and $E_2 \succ E_3 \,|\, C$, then $E_1 \succ E_3 \,|\, C$
    - *The sure-thing principle*.
      If $E_1 \succ E_2 \,|\, A, C$ and $E_1 \succ E_2 \,|\, \overline{A}, C$, then $E_1 \succ E_2 \,|\, C$).
  - 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 $\Pr(E \,|\, \boldsymbol{x}, A, K)$, 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.*,
      $\Pr(+ \mid V) = 0.98$, $\Pr(+ \mid \overline{V}) = 0.01$, $\Pr(V \mid K) = 0.002$,
      $\Pr(+ \mid K) = \Pr(+ \mid V)\Pr(V \mid K) + \Pr(+ \mid \overline{V})\Pr(\overline{V} \mid K) = 0.012$
      $\Pr(V \mid +, A, K) = \dfrac{\Pr(+ \mid V)\Pr(V \mid K)}{\Pr(+ \mid K)} = 0.164$ (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.
      $\Pr(a < \theta < b \mid r, n, A, K)$, 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 physical constant μ, with data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, considered to be measurements of μ subject to error. Desired to find $\Pr(a < \mu < b \,|\, x_1, \ldots, x_n, A, K)$, 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 $\boldsymbol{x}$, assumptions A made, and available knowledge K.

- The statistical model may 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 $\mathrm{N}(x \,|\, \mu, \sigma)$ normal model

- 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*.Experiment 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, $\Pr(r \mid r_1, \ldots, r_n, A, K)$. For instance, no accidents in each of $n = 10$ consecutive months may yield $\Pr(r = 0 \mid \boldsymbol{x}, A, K) = 0.953$.
- *Future continuous observations*.Data $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$. Desired to forecast the value of a future observation $\boldsymbol{y}$, $p(\boldsymbol{y} \mid \boldsymbol{x}, A, K)$. For instance, from breaking strengths $\boldsymbol{x} = \{y_1, \ldots, y_n\}$ of n randomly chosen safety belt webbings, the engineer may find $\Pr(y > y^* \mid \boldsymbol{x}, A, K) = 0.9987$.
- *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 $\boldsymbol{v}$, $p(\boldsymbol{y} \mid \boldsymbol{v}, \boldsymbol{x}, A, K)$. For instance, y contamination levels, v wind speed from source; environment authorities interested in $\Pr(y > y^* \mid v, \boldsymbol{x}, A, K)$

2. Basics of Bayesian Analysis

2.1. Parametric Inference

- *Bayes Theorem*

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

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

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

  - Simplifying notation, Bayes' theorem may be expressed as

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

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

- *Bayesian Inference with a Finite Parameter Space*

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

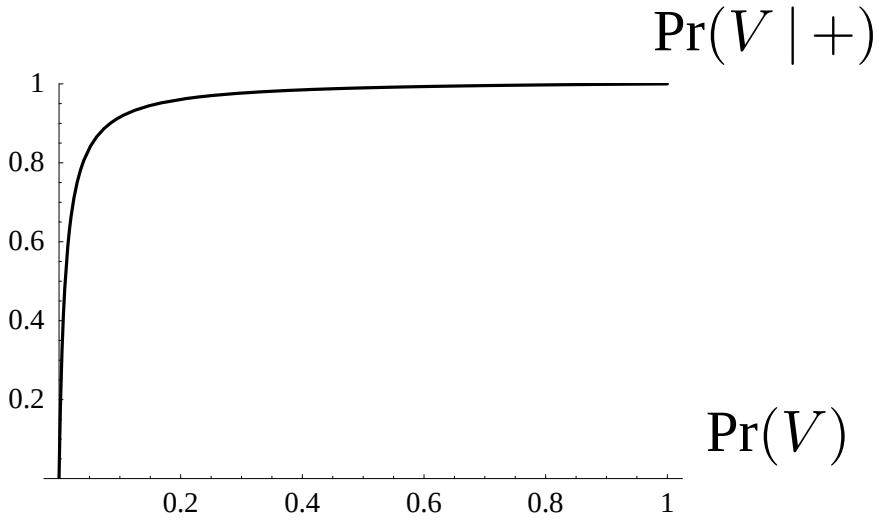
$$\Pr(\theta_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \theta_i)\Pr(\theta_i)}{\sum_{j=1}^{m} p(\boldsymbol{x} \mid \theta_j)\Pr(\theta_j)}, \quad i = 1, \ldots, m.$$

  - *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 is

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

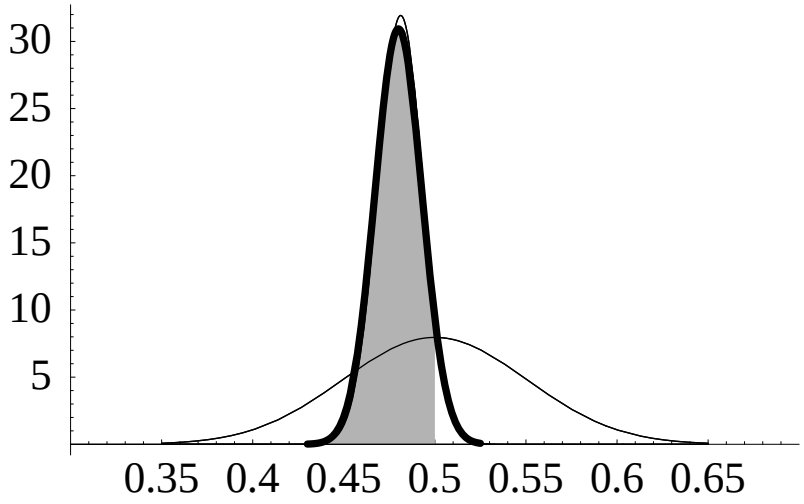
    as a function of $p = \Pr(V)$.

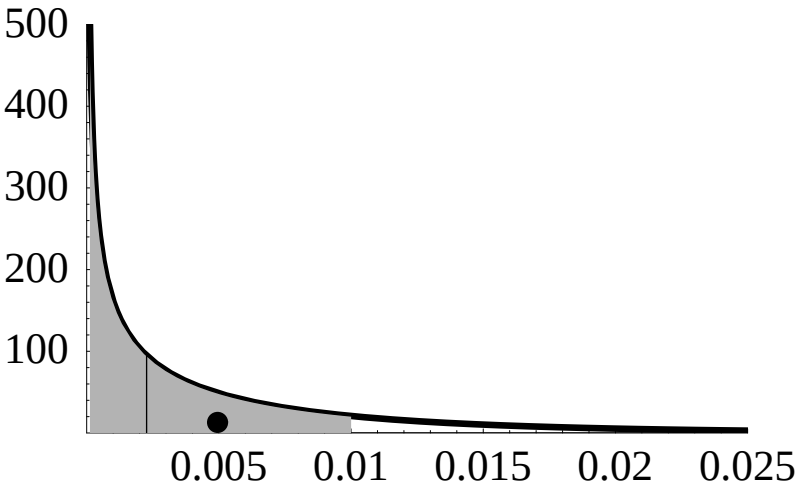
  - Notice sensitivity of posterior $\Pr(V \mid +)$ to changes in the prior $p = \Pr(V)$.



- *Example: Inference about a binomial parameter*

  - Let data $\boldsymbol{x}$ be 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 $\pi(\theta) = \text{Be}(\theta \mid \alpha, \beta)$, then $\pi(\theta \mid \boldsymbol{x}) \propto \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}$ kernel of $\text{Be}(\theta \mid r+\alpha, n-r+\beta)$.
  - Prior information (K) $P(0.4 < \theta < 0.6) = 0.95$, and symmetric, yields $\alpha = \beta = 47$;
  - No prior information $\alpha = \beta = 1/2$
  - $n = 1500, r = 720$ $P(\theta < 0.5 \mid \boldsymbol{x}, K) = 0.933$ $P(\theta < 0.5 \mid \boldsymbol{x}) = 0.934$
  - $n = 100, r = 0$ $P(\theta < 0.01 \mid \boldsymbol{x}) = 0.844$ Notice: $\hat{\theta} = 0$, but $\text{Me}[\theta \mid \boldsymbol{x}] = 0.0023$





- *Sufficiency*

  - Given a model $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, a function of the data $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$, 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 $\pi(\boldsymbol{\theta})$ $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{t})$. Hence, $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{t}) \propto p(\boldsymbol{t} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$.
  - This is equivalent to the frequentist definition; thus $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is sufficient iff $p(\boldsymbol{x} \mid \boldsymbol{\theta}) = f(\boldsymbol{\theta}, \boldsymbol{t}) g(\boldsymbol{x})$.
  - A sufficient statistic always exists, for $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{x}$ is obviously sufficient

    A much simpler sufficient statistic, with fixed dimensionality independent of the sample size, often exists.

    This is case whenever the statistical model belongs to the *generalized exponential family*, which includes many of the more frequently used statistical models.
  - In contrast to frequentist statistics, Bayesian methods are independent on 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*

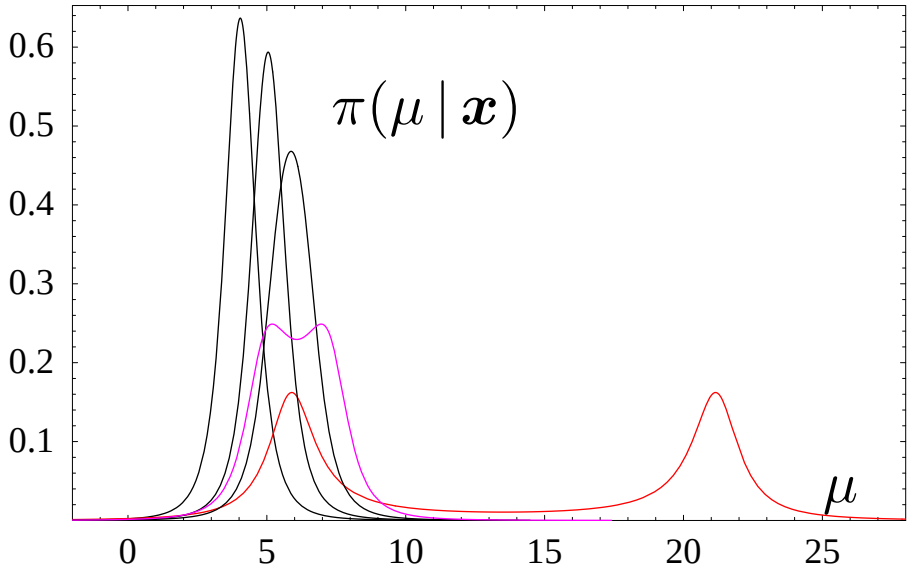
  - Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ random from $\text{Ca}(x \mid \mu, 1) = \text{St}(x \mid \mu, 1, 1)$.
  - Objective reference prior for the location parameter μ is $\pi(\mu) = 1$.
  - By Bayes' theorem,

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

    Proportionality constant easily obtained by numerical integration.

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

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 $\int_\Theta \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ is not finite.

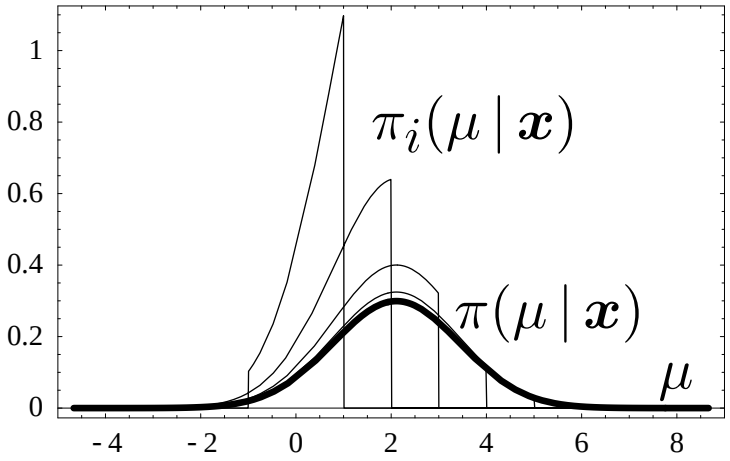
    The Cauchy example uses the improper prior function $\pi(\mu) = 1$, $\mu \in \Re$.
  - $\pi(\boldsymbol{\theta})$ is an improper prior function, $\{\Theta_i\}_{i=1}^{\infty}$ an increasing sequence approximating Θ, such that $\int_{\Theta_i} \pi(\boldsymbol{\theta}) < \infty$, and $\{\pi_i(\boldsymbol{\theta})\}_{i=1}^{\infty}$ the proper priors obtained by *renormalizing* $\pi(\boldsymbol{\theta})$ within the Θ_i's.
  - 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 $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$.
  - Normal data, σ known, $\pi(\mu) = 1$.

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

    $\propto \exp[-\frac{n}{2\sigma^2}(\overline{x} - \mu)^2]$

    $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{N}(\mu \,|\, \overline{x}, \sigma/\sqrt{n})$

    Example: $n = 9,\ \overline{x} = 2.11,\ \sigma = 4$



- *Sequential updating*

  - Prior and posterior are terms *relative* to a 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.

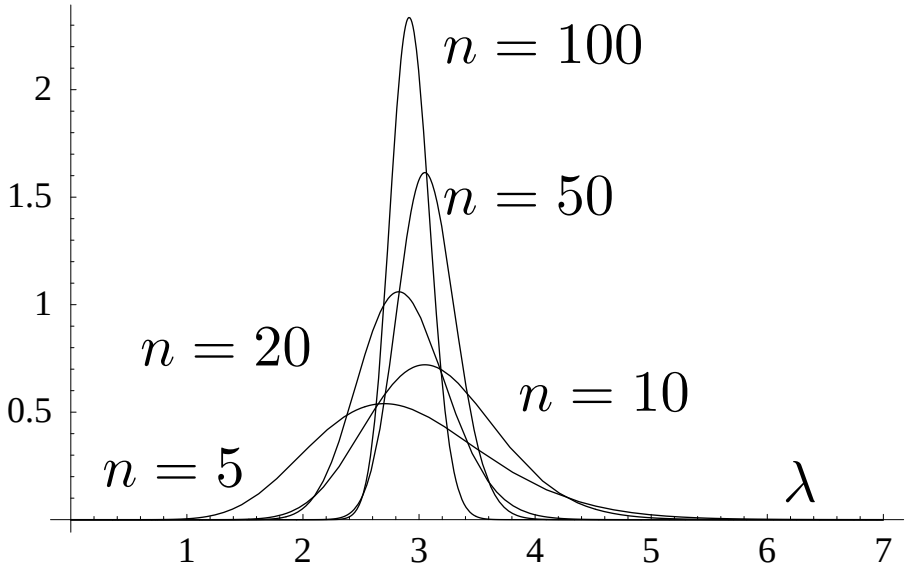
$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1}) \propto p(\boldsymbol{x}_{i+1} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_i).$$

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

  - 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 $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$.
  - Posteriors $\pi(\lambda \,|\, x_1, \ldots, x_i)$ from increasingly large simulated data from Poisson $\mathrm{Pn}(x \,|\, \lambda)$, with $\lambda = 3$

$$\pi(\lambda \,|\, x_1, \ldots, x_i) = \mathrm{Ga}(\lambda \,|\, r_i + 1/2, i)$$

$$r_i = \sum_{j=1}^{i} x_j$$



- *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 $\boldsymbol{\omega} = \boldsymbol{\omega}(\boldsymbol{\theta}) \in \Omega$ 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, $\pi(\boldsymbol{\psi} \,|\, \boldsymbol{x}) = |J(\boldsymbol{\psi})| [\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})]_{\boldsymbol{\theta}=\boldsymbol{\theta}(\boldsymbol{\psi})}$

    and $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x}) = \int_\Omega \pi(\boldsymbol{\phi}, \boldsymbol{\omega} \,|\, \boldsymbol{x})\, d\boldsymbol{\omega}$.
  - Any valid conclusion on $\boldsymbol{\phi}$ will be contained in $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x})$.
  - Particular case: *marginal posteriors*

    Often model directly expressed in terms of vector of interest $\boldsymbol{\phi}$, and vector of nuisance parameters $\boldsymbol{\omega}$, $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})$.

    Specify the prior $\quad \pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\phi})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\phi})$

    Get the joint posterior $\quad \pi(\boldsymbol{\phi}, \boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\phi})\, \pi(\boldsymbol{\phi})$

    Integrate out ω, $\quad \pi(\boldsymbol{\phi} \,|\, \boldsymbol{x}) \propto \pi(\boldsymbol{\phi}) \int_\Omega p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\omega}$

- *Example: Inferences about a Normal mean*

  - Data $\boldsymbol{x} = \{x_1, \ldots x_n\}$ random from $\mathrm{N}(x \,|\, \mu, \sigma)$. Likelihood function $p(\boldsymbol{x} \,|\, \mu, \sigma) \propto \sigma^{-n} \exp[-n\{s^2 + (\overline{x} - \mu)^2\}/(2\sigma^2)]$, with $n\overline{x} = \sum_i x_i$, and $ns^2 = \sum_i (x_i - \overline{x})^2$.

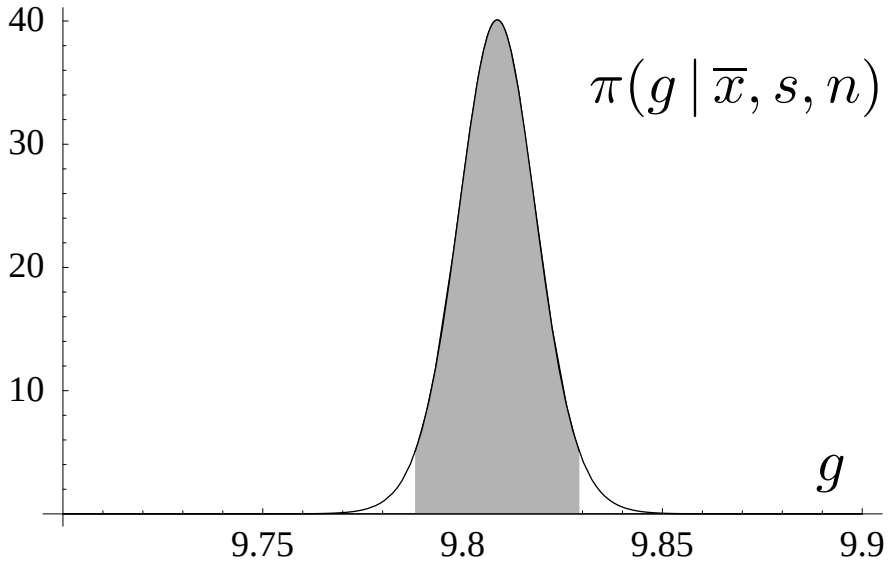
  - Objective prior is uniform in both μ and $\log(\sigma)$, *i.e.*, $\pi(\mu, \sigma) = \sigma^{-1}$. Joint posterior $\pi(\mu, \sigma \,|\, \boldsymbol{x}) \propto \sigma^{-(n+1)} \exp[-n\{s^2 + (\overline{x} - \mu)^2\}/(2\sigma^2)]$.

  - Marginal posterior $\pi(\mu \,|\, \boldsymbol{x}) \propto \int_0^\infty \pi(\mu, \sigma \,|\, \boldsymbol{x})\, \mathrm{d}\sigma \propto [s^2 + (\overline{x} - \mu)^2]^{-n/2}$, kernel of the Student density $\mathrm{St}(\mu \,|\, \overline{x}, s/\sqrt{n-1}, n-1)$

  - Classroom experiment to measure gravity g yields $\overline{x} = 9.8087$, $s = 0.0428$ with $n = 20$ measures.

    $\pi(g \,|\, \overline{x}, s, n)$
    $= \mathrm{St}(g \,|\, 9.8087, 0.0098, 19)$

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



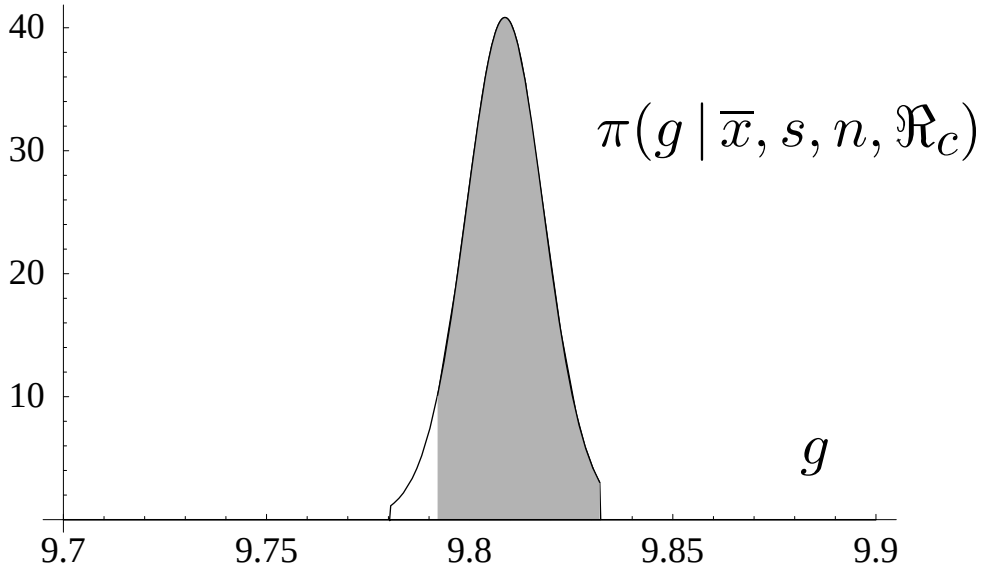
- *Restricted parameter space*

  - Range of values of $\boldsymbol{\theta}$ restricted by contextual considerations.
    If $\boldsymbol{\theta}$ known to belong to $\Theta_c \subset \Theta$, $\pi(\boldsymbol{\theta}) > 0$ iff $\boldsymbol{\theta} \in \Theta_c$
    By Bayes' theorem,

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \boldsymbol{\theta} \in \Theta_c) = \begin{cases} \dfrac{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{\int_{\Theta_c} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}}, & \text{if} \quad \boldsymbol{\theta} \in \Theta_c \\ 0 & \text{otherwise} \end{cases}$$

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

  - Classroom experiment to measure gravity g with restriction to lie between
    $g_0 = 9.7803$ (equator)
    $g_1 = 9.8322$ (poles).
    $\Pr(9.7921 < g < 9.8322 \,|\, \boldsymbol{x}) = 0.95$ (shaded area)



- *Asymptotic behaviour, discrete case*

  - If the parameter space $\Theta = \{\theta_1, \theta_2, \ldots\}$ is *countable* and
    The true parameter value θ_t is *distinguishable* from the others,*i.e.*,
    $\delta\{p(\boldsymbol{x} \mid \boldsymbol{\theta}_t), p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)) > 0$, $i \neq t$,

$$\lim_{n\to\infty} \pi(\theta_t \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 1$$
$$\lim_{n\to\infty} \pi(\theta_i \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = 0, \quad i \neq t$$

  - To prove this, take logarithms is Bayes' theorem,
    define $z_i = \log[p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)/p(\boldsymbol{x} \mid \boldsymbol{\theta}_t)]$,
    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.

- *Asymptotic behaviour, continuous case*

  - If the parameter θ is *one-dimensional and continuous*, so that $\Theta \subset \Re$, and the model $\{p(\boldsymbol{x} \,|\, \theta),\ \boldsymbol{x} \in \mathcal{X}\}$ is *regular*: basically,
    $\mathcal{X}$ does not depend on θ,
    $p(\boldsymbol{x} \,|\, \theta)$ is twice differentiable with respect to θ
  - 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, express is 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 $\Theta \subset \Re^k$ and the model $\{p(\boldsymbol{x} \mid \theta),\ \boldsymbol{x} \in \mathcal{X}\}$ is *regular*, so that $\boldsymbol{\mathcal{X}}$ does not depend on $\boldsymbol{\theta}$ and $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ is twice differentiable with respect to each of the θ_i's, then, as $n \to \infty$, $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ 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
    $$\boldsymbol{F}_{ij}(\boldsymbol{\theta}) = -\mathrm{E}_{\boldsymbol{x} \mid \boldsymbol{\theta}}[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\boldsymbol{x} \mid \boldsymbol{\theta})]$$
  - The properties of the multivariate normal yield from this result the asymptotic forms for the *marginal* and the *conditional* posterior distributions of any subgroup of the θ_j's.
  - In one dimension, $\pi(\theta \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \approx \mathrm{N}(\theta \mid \hat{\theta},\ (nF(\theta)^{-1/2})$, where $F(\theta) = -\mathrm{E}_{\boldsymbol{x} \mid \theta}[\partial^2 \log p(\boldsymbol{x} \mid \theta)/\partial\theta^2]$

- *Example: Asymptotic approximation with Poisson data*

  - Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ random from $\text{Pn}(x \,|\, \lambda) \propto e^{-\lambda}\lambda^x/x!$
    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 $F(\lambda) = -\text{E}_{x\,|\,\lambda}\left[\dfrac{\partial^2}{\partial\lambda^2}\log \text{Pn}(x \,|\, \lambda)\right] = \dfrac{1}{\lambda}$

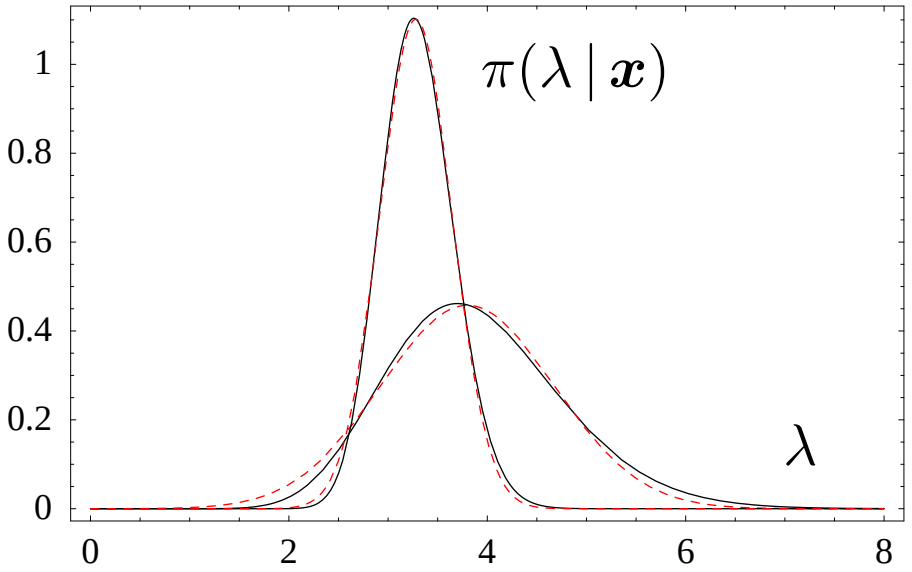
  - The objective prior function is $\pi(\lambda) = F(\lambda)^{1/2} = \lambda^{-1/2}$

    Hence $\pi(\lambda \,|\, \boldsymbol{x}) \propto e^{-n\lambda}\lambda^{r-1/2}$

    the kernel of $\text{Ga}(\lambda \,|\, r + \frac{1}{2}, n)$

  - The Normal approximation is
    $\pi(\lambda \,|\, \boldsymbol{x}) \approx \text{N}\{\lambda \,|\, \hat{\lambda}, (n\, F(\hat{\lambda}))^{-1/2}\}$
    $= \text{N}\{\lambda \,|\, r/n,\, \sqrt{r}/n\}$

  - Samples $n = 5$ and $n = 25$
    simulated from Poisson $\lambda = 3$
    yielded $r = 19$ and $r = 82$



2.2. 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:

    Use *information theory* to measure the amount on information about the quantity of interest to be expected from data. This depends on prior knowledge: the more it is known, the less the amount of information the data may be expected to provide.

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

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

- *Expected information from the data*

  - Given model $\{p(\boldsymbol{x}\,|\,\theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$, the *amount of information* $I^{\theta}\{\boldsymbol{\mathcal{X}}, \pi(\theta)\}$ which may be expected to be provided by $\boldsymbol{x}$, about the value of θ is defined by
    $I^{\theta}\{\boldsymbol{\mathcal{X}}, \pi(\theta)\} = \delta\{p(\boldsymbol{x}, \theta), p(\boldsymbol{x})\pi(\theta)\}$,
    the intrinsic discrepancy between the joint distribution $p(\boldsymbol{x}, \theta)$ and the product of their marginals $p(\boldsymbol{x})\pi(\theta)$, which is the *instrinsic association* between the random quantities $\boldsymbol{x}$ and θ.

  - Consider $I^{\theta}\{\boldsymbol{\mathcal{X}}^k, \pi(\theta)\}$ 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 $I^{\theta}\{\boldsymbol{\mathcal{X}}^k, \pi(\theta)\}$ will approach the missing information about θ associated with the prior $\pi(\theta)$.

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

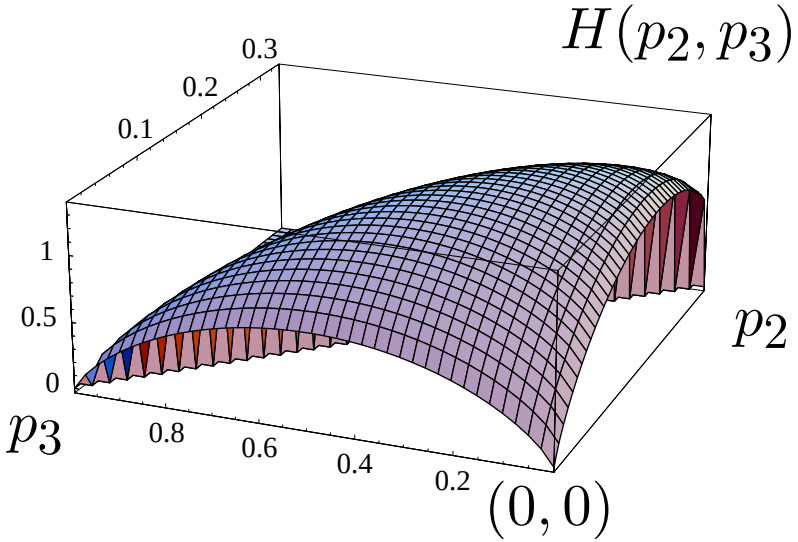
  - If θ may only take a *finite* number m of different values $\{\theta_1, \ldots, \theta_m\}$ and $\pi(\theta) = \{p_1, \ldots, p_m\}$, with $p_i = \Pr(\theta = \theta_i)$, then $\lim_{k\to\infty} I^{\theta}\{\boldsymbol{\mathcal{X}}^k, \pi(\theta)\} = H(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log(p_i)$, that is, the *entropy* of the prior distribution $\{p_1, \ldots, p_m\}$.

  - In the finite case, 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)

  - Prior $\{p_1, p_2, p_3, p_4\}$ in genetics problem where $p_1 = 2p_2$.

    Reference prior is $\{0.324, 0.162, 0.257, 0.257\}$



- *Reference priors in one-dimensional continuous case*

  - Let $\pi_k(\theta)$ be the prior which maximizes $I^{\theta}\{\boldsymbol{\mathcal{X}}^k, \pi(\theta)\}$ in the class $\mathcal{P}$ of acceptable priors.

    For any data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, let $\pi_k(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta)\, \pi_k(\theta)$ be the corresponding posterior.

  - The *reference posterior density* $\pi^*(\theta \mid \boldsymbol{x})$ is defined to be the intrinsic limit of the sequence $\{\pi_k(\theta \mid \boldsymbol{x}), k = 1, 2, \ldots\}$

    A *reference prior function* $\pi^*(\theta)$ is any positive function such that, for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $\pi^*(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta)\, \pi^*(\theta)$.
    This is defined up to an (irrelevant) arbitrary constant.

  - Let $\boldsymbol{x}^{(k)} \in \boldsymbol{\mathcal{X}}^k$ be the result of k independent replications of $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$.
    The exact expression for $\pi_k(\theta)$ (which may be obtained with calculus of variations) is

    $$\pi_k(\theta) = \exp\left[\mathrm{E}_{\boldsymbol{x}^{(k)} \mid \theta}\{\log \pi_k(\theta \mid \boldsymbol{x}^{(k)})\}\right]$$

  - This formula may be used, by repeated simulation from $p(\boldsymbol{x} \mid \theta)$ 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 estimator $\hat{\theta}$.

    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 $k \to \infty$ 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.

    Hence, $\pi^*(\theta) = \pi(\theta \,|\, \tilde{\theta}_k)|_{\tilde{\theta}_k = \theta}$

  - 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, $\pi^*(\theta) = F(\theta)^{1/2}$ (Jeffreys' rule).

- *One nuisance parameter*

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

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

    (ii) If $\pi^*(\lambda \mid \theta)$ is proper, integrate out λ to get the one-parameter model $p(\boldsymbol{x} \mid \theta) = \int_\Lambda p(\boldsymbol{x} \mid \theta, \lambda)\, \pi^*(\lambda \mid \theta)\, d\lambda$, and use the one-parameter solution to obtain $\pi^*(\theta)$.

    The θ-reference prior is then $\pi^*_\theta(\theta, \lambda) = \pi^*(\lambda \mid \theta)\, \pi^*(\theta)$.
    The required reference posterior is $\pi^*(\theta \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \theta) \pi^*(\theta)$.

  - If $\pi^*(\lambda \mid \theta)$ is an *improper* prior function, proceed within an increasing sequence $\{\Lambda_i\}$ over which $\pi^*(\lambda \mid \theta)$ is integrable and, for given data $\boldsymbol{x}$, obtain the corresponding sequence of reference posteriors $\{\pi^*_i(\theta \mid \boldsymbol{x}\}$.

    The required reference posterior $\pi^*(\theta \mid \boldsymbol{x})$ is their intrinsic limit.

    A θ-reference prior is any positive function such that, for any data $\boldsymbol{x}$, $\pi^*(\theta \mid \boldsymbol{x}) \propto \int_\Lambda p(\boldsymbol{x} \mid \theta, \lambda)\, \pi^*_\theta(\theta, \lambda)\, d\lambda$.

- *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, $\boldsymbol{F}(\theta, \lambda)$.

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

    The θ-reference prior is $\pi^*_\theta(\theta, \lambda) = \pi^*(\lambda \,|\, \theta)\, \pi^*(\theta)$, where

    $\pi^*(\lambda \,|\, \theta) \propto F_{\lambda\lambda}^{1/2}(\theta, \lambda),\ \lambda \in \Lambda$, and, if $\pi^*(\lambda \,|\, \theta)$ is proper,

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

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

  - Even if $\pi^*(\lambda \,|\, \theta)$ is improper, if θ and λ are variation independent, $S_{\theta\theta}^{-1/2}(\theta, \lambda) \propto f_\theta(\theta)\, g_\theta(\lambda)$, and $F_{\lambda\lambda}^{1/2}(\theta, \lambda) \propto f_\lambda(\theta)\, g_\lambda(\lambda)$,

    Then $\pi^*_\theta(\theta, \lambda) = f_\theta(\theta)\, g_\lambda(\lambda)$.

- *Examples: Inference on normal parameters*

  - The information matrix for the normal model $\mathrm{N}(x \mid \mu, \sigma)$ is
    $\boldsymbol{F}(\mu,\sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}$, $\quad \boldsymbol{S}(\mu,\sigma) = \boldsymbol{F}^{-1}(\mu,\sigma) = \begin{pmatrix} \sigma^{2} & 0 \\ 0 & \sigma^{2}/2 \end{pmatrix}$;
    Since μ and σ are variation independent, and both $F_{\sigma\sigma}$ and $S_{\mu\mu}$ factorize,
    $\pi^*(\sigma \mid \mu) \propto F_{\sigma\sigma}^{1/2} \propto \sigma^{-1}$, $\pi^*(\mu) \propto S_{\mu\mu}^{-1/2} \propto 1$.
    The μ-reference prior, as anticipated, is
    $\pi^*_\mu(\mu,\sigma) = \pi^*(\sigma \mid \mu)\, \pi^*(\mu) = \sigma^{-1}$,
    i.e., uniform on both μ and $\log\sigma$

  - Since $\boldsymbol{F}(\mu,\sigma)$ is diagonal the σ-reference prior is
    $\pi^*_\sigma(\mu,\sigma) = \pi^*(\mu \mid \sigma)\pi^*(\sigma) = \sigma^{-1}$, the same as $\pi^*_\mu(\mu,\sigma) = \pi^*_\sigma(\mu,\sigma)$.

  - In fact, it may be shown that, for location-scale models,
    $p(x \mid \mu, \sigma) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$,
    the reference prior for the location and scale parameters are always
    $\pi^*_\mu(\mu,\sigma) = \pi^*_\sigma(\mu,\sigma) = \sigma^{-1}$.

- Within any given model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ 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, within a normal $\mathrm{N}(x \,|\, \mu, \sigma)$ model, let the *standardized mean* $\phi = \mu/\sigma$. be the quantity of interest.

  Fisher's information matrix in terms of the parameters ϕ and σ is $\boldsymbol{F}(\phi, \sigma) = J^t\, \boldsymbol{F}(\mu, \sigma)\, J$, where $J = (\partial(\mu, \sigma)/\partial(\phi, \sigma))$ is the Jacobian of the inverse transformation; this yields

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

  with $F^{1/2}_{\sigma\sigma} \propto \sigma^{-1}$, and $S^{-1/2}_{\phi\phi} \propto (1+\phi^2/2)^{-1/2}$.

- 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 $\pi^*_\mu(\mu, \sigma) = \pi^*_\sigma(\mu, \sigma)$.

  This prior is shown to lead to a reference posterior for ϕ with *consistent marginalization properties*.

- *Many parameters*

  - The reference algorithm generalizes to any number of parameters. If the model is $p(\boldsymbol{x} \mid \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \theta_1, \ldots, \theta_m)$, a joint reference prior $\pi^*(\phi_m \mid \phi_{m-1}, \ldots, \phi_1) \times \ldots \times \pi^*(\phi_2 \mid \phi_1) \times \pi^*(\phi_1)$ 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 $i = m, m-1, \ldots, 1$.

  - Example: *Stein's paradox*. Data random from a m-variate normal $\mathrm{N}_m(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{I})$. The reference prior function 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.

    The reference prior for $\{\theta, \lambda_1, \ldots, \lambda_{m-1}\}$ produces, for any choice of the λ_i's, an appropriate the reference posterior for θ.

2.3. 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* $\pi(\boldsymbol{\theta} \mid \boldsymbol{x}, C)$.
  - 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.
  - However, graphical methods not easily extend 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
    $\mathrm{E}[\boldsymbol{\theta} \mid \boldsymbol{x}] = \int_{\Theta} \boldsymbol{\theta}\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}$ *(posterior mean)*, or
    $\mathrm{Mo}[\boldsymbol{\theta} \mid \boldsymbol{x}] = \arg\sup_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ *(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 $\phi^* = \log(\theta^*)$.

- *Point Estimation: Posterior median*

  - A summary of a multivariate density $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$, should contain summaries of:
    (i) each of the marginal densities $\pi(\theta_i \,|\, \boldsymbol{x})$,
    (ii) the densities $\pi(\phi \,|\, \boldsymbol{x})$ of other functions of interest $\phi = \phi(\boldsymbol{\theta})$.
  - In *one-dimensional continuous* problems the *posterior median*, is easily defined and computed as
    $\text{Me}[\theta \,|\, \boldsymbol{x}] = q\,; \quad \Pr[\theta \leq q \,|\, \boldsymbol{x}] = \int_{\{\theta \leq q\}} \pi(\theta \,|\, \boldsymbol{x})\, \text{d}\theta = 1/2$

    The one-dimensional posterior median has many attractive properties:
    (i) it is *invariant* under bijections, $\text{Me}[\phi(\theta) \,|\, \boldsymbol{x}] = \phi(\text{Me}[\theta \,|\, \boldsymbol{x}])$.
    (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-dimensional continuous problems.
  - The posterior median is not easily used to a multivariate setting. The natural extension of its definition produces *surfaces* (not points).

    General invariant multivariate definitions of point estimators is possible using Bayesian *decision theory*

- *General Credible Regions*

  - To describe $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ it is often convenient to quote regions $\Theta_p \subset \Theta$ of given probability content p under $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$. This is the intuitive basis of graphical representations like boxplots.
  - A subset Θ_p of the parameter space Θ such that
    $\int_{\Theta_p} \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta} = p, \quad$ so that $\Pr(\boldsymbol{\theta} \in \Theta_p \mid \boldsymbol{x}) = p,$
    is a *posterior p-credible region* for $\boldsymbol{\theta}$.
  - A credible region is invariant under reparametrization:
    If Θ_p is p-credible for $\boldsymbol{\theta}$, $\boldsymbol{\phi}(\Theta_p)$ 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.
  - HPD regions are *not invariant* : the image $\boldsymbol{\phi}(\Theta_p)$ of an HPD region Θ_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 reparametrization.
  - *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: Model $\mathrm{N}(x \,|\, \mu, \sigma)$. *Credible intervals for the normal mean*. The reference posterior for μ is $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{St}(\mu \,|\, \overline{x}, s/\sqrt{n-1}, n-1)$. Hence the reference *posterior* distribution of $\tau = \sqrt{n-1}(\mu - \overline{x})/s$, *a function of* μ, is $\pi(\tau \,|\, \overline{x}, s, n) = \mathrm{St}(\tau \,|\, 0, 1, n-1)$.

    Thus, the equal-tailed p-credible intervals for μ are $\{\mu;\ \mu \in \overline{x} \pm q_{n-1}^{(1-p)/2}\, s/\sqrt{n-1}\}$, where $q_{n-1}^{(1-p)/2}$ is the $(1-p)/2$ quantile of a standard Student density with $n-1$ degrees of freedom.

- *Calibration*

  - In the normal example above , the expression $t = \sqrt{n-1}(\mu - \overline{x})/s$ may *also* be analyzed, for fixed μ, as a *function of the data*.

    The fact that the *sampling* distribution of the statistic $t = t(\overline{x}, s \,|\, \mu, n)$ 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* between Bayesian credible intervals and frequentist confidence intervals is 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*.

2.4. 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 $p(x \mid \boldsymbol{x}, K)$ 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 $p(x \mid \boldsymbol{x}, K)$ 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 specified model, $\{p(x \mid \boldsymbol{\theta}), x \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$, so that $p(\boldsymbol{x} \mid \boldsymbol{\theta}) = p(x_1, \ldots, x_n \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} p(x_j \mid \boldsymbol{\theta})$.

    If this is the case, the solution to the prediction problem is immediate once a prior distribution $\pi(\boldsymbol{\theta})$ has been specified.

- *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 $\{p(x \mid \boldsymbol{\theta}), x \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$
    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

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

    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
    $p(x \mid \boldsymbol{x}) = \int_\Theta p(x \mid \boldsymbol{\theta}, \boldsymbol{x})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}$
    and notice the new observation x has been assumed to be conditionally independent of the observed data $\boldsymbol{x}$, so that $p(x \mid \boldsymbol{\theta}, \boldsymbol{x}) = p(x \mid \boldsymbol{\theta})$.

  - 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 $\text{Pn}(r \,|\, \lambda)$. The reference posterior density of λ is $\pi^*(\lambda \,|\, \boldsymbol{x}) = \text{Ga}(\lambda \,|\, t + 1/2, n)$, where $t = \Sigma_j \, r_j$.

    The (reference) posterior predictive distribution is

$$p(r \,|\, \boldsymbol{x}) = \Pr[r \,|\, t, n] = \int_0^\infty \text{Pn}(r \,|\, \lambda) \, \text{Ga}(\lambda \,|\, t + \tfrac{1}{2}, n) \, d\lambda$$

$$= \frac{n^{t+1/2}}{\Gamma(t + 1/2)} \, \frac{1}{r!} \, \frac{\Gamma(r + t + 1/2)}{(1 + n)^{r+t+1/2}} \,,$$

    an example of a Poisson-Gamma probability mass function.

  - For example, 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, $\Pr[0 \,|\, t, n] = 0.953$, $\Pr[1 \,|\, t, n] = 0.043$, and $\Pr[2 \,|\, t, n] = 0.003$.

    Many other situations may be described with the same model.

- *Prediction of Normal measurements*

  - Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ random from $N(x \,|\, \mu, \sigma)$. 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, $\pi^*(\mu, \lambda \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \mu, \lambda)\, \pi^*(\mu, \lambda)$, is $\pi^*(\mu, \lambda \,|\, \boldsymbol{x}) = N(\mu \,|\, \overline{x}, (n\lambda)^{-1/2})\, \mathrm{Ga}(\lambda \,|\, (n-1)/2, ns^2/2)$.

  - The predictive distribution is

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

    a kernel of the *Student* density $\pi^*(x \,|\, \boldsymbol{x}) = \mathrm{St}(x \,|\, \overline{x}, s\sqrt{\frac{n+1}{n-1}}, n-1)$.

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

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

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

- *Regression*

  - Often *additional information* from relevant covariates. Data structure, set of pairs $\boldsymbol{x} = \{(\boldsymbol{y}_1, \boldsymbol{v}_1), \ldots (\boldsymbol{y}_n, \boldsymbol{v}_n)\}$; $\boldsymbol{y}_i$, $\boldsymbol{v}_i$, both vectors. Given a new observation, with $\boldsymbol{v}$ known, predict the corresponding value of $\boldsymbol{y}$. Formally, compute $p\{\boldsymbol{y} \,|\, \boldsymbol{v}, (\boldsymbol{y}_1, \boldsymbol{v}_1), \ldots (\boldsymbol{y}_n, \boldsymbol{v}_n)\}$.
  - Need a model $\{p(\boldsymbol{y} \,|\, \boldsymbol{v}, \boldsymbol{\theta}), \boldsymbol{y} \in \boldsymbol{Y}, \boldsymbol{\theta} \in \Theta\}$ 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 $\boldsymbol{x} = \{\boldsymbol{y}, \boldsymbol{V}\}$, $\boldsymbol{y} = \{y_1, \ldots, y_n\}^t$, and $\boldsymbol{V}$ is the $n \times k$ matrix with the $\boldsymbol{v}_i$'s as rows. $p(\boldsymbol{y} \,|\, \boldsymbol{V}, \boldsymbol{\beta}, \sigma) = \mathrm{N}_n(\boldsymbol{y} \,|\, \boldsymbol{V}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$; reference prior $\pi^*(\boldsymbol{\beta}, \sigma) = \sigma^{-1}$.

    Predictive posterior is the Student density

    $$p(y \,|\, \boldsymbol{v}, \boldsymbol{y}, \boldsymbol{V}) = \mathrm{St}(y \,|\, \boldsymbol{v}\hat{\boldsymbol{\beta}},\ s\sqrt{f(\boldsymbol{v}, \boldsymbol{V})\,\tfrac{n}{n-k}},\ n-k)$$

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

    $$f(\boldsymbol{v}, \boldsymbol{V}) = 1 + \boldsymbol{v}(\boldsymbol{V}^t\boldsymbol{V})^{-1}\boldsymbol{v}^t$$

- *Example: Simple linear regression*

  - One covariate and a constant term; $p(y \mid v, \boldsymbol{\beta}, \sigma) = \mathrm{N}(y \mid \beta_1 + \beta_2 v, \sigma)$
    Sufficient statistic is $\boldsymbol{t} = \{\overline{v}, \overline{y}, s_{vy}, s_{vv}\}$, with $n\overline{v} = \Sigma v_j$, $n\overline{y} = \Sigma y_j$, $s_{yv} = \Sigma v_j y_j / n - \overline{v}\,\overline{y}$, $s_{vv} = \Sigma v_j^2 / n - \overline{v}^2$.

    $$p(y \mid v, \boldsymbol{t}) = \mathrm{St}(y \mid \hat{\beta}_1 + \hat{\beta}_2\, v,\ s\sqrt{f(v, \boldsymbol{t})\,\tfrac{n}{n-2}},\ n-2)$$

    $$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{v}, \quad \hat{\beta}_2 = \frac{s_{vy}}{s_{vv}},$$

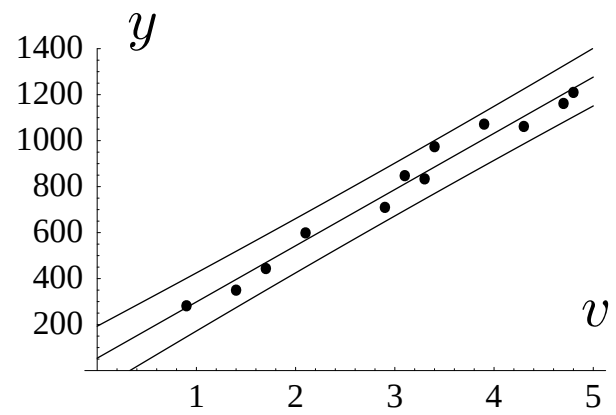
    $$n s^2 = \sum_{j=1}^{n} (y_j - \hat{\beta}_1 - \hat{\beta}_2 x_j)^2$$

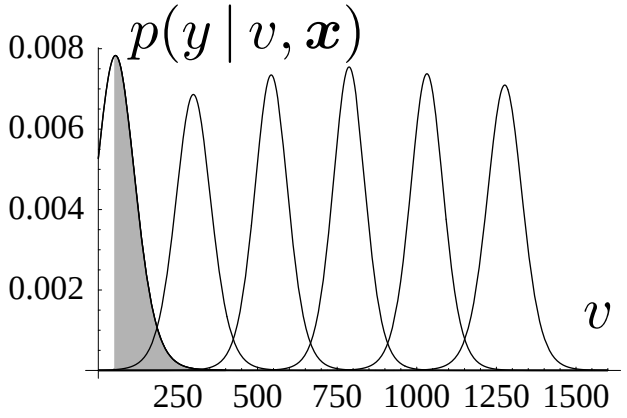
    $$f(v, \boldsymbol{t}) = 1 + \frac{1}{n}\frac{(v-\overline{v})^2 + s_{vv}}{s_{vv}}$$

  - Pollution density ($\mu gr/m^3$), and wind speed from source (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

$$\Pr[y > 50 \mid v = 0, \boldsymbol{x}] = 0.66$$





2.4. Hierarchical Models

- *Exchangeability*

  - Random quantities are often "homogeneous" in the precise sense that only their *values* matter, not the *order* in which they appear. Formally, this is captured by the notion of *exchangeability*. The set of random vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is exchangeable if 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 from any model 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 $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, 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 $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \mid C) = \int_\Theta \prod_{i=1}^n p(\boldsymbol{x}_i \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid C)\, d\boldsymbol{\theta}$.

- *Structured Models*

  - 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
    $p(y_{i1}, \ldots, y_{in_i}) = \prod_{j=1}^{n_i} p(y_{ij} \mid \theta_i) = \theta_i^{r_i}(1-\theta_i)^{n_i - r_i}$,
    where θ_i is the (unknown) proportion of citizens in region i voting A and $r_i = \Sigma_j y_{ij}$ the number of citizens voting A in region i.

    Assuming regions exchangeable within the country similarly leads to
    $p(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{k} \pi(\theta_i \mid \boldsymbol{\phi})$
    for some probability distribution $\pi(\theta \mid \boldsymbol{\phi})$ describing the political variation within the regions. Often choose $\pi(\theta \mid \boldsymbol{\phi}) = \text{Be}(\theta \mid \alpha, \beta)$.
  - 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 $\text{Bi}(y \mid \theta_i)$,
    $\{\theta_1, \ldots, \theta_k\}$, random from $\text{Be}(\theta \mid \alpha, \beta)$
    provides a far richer model than (unrealistic) simple binomial sampling.

- *Example: Biological response.* Sample k different animals of the same species in 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
  $p(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{in_i}) = \prod_{j=1}^{n_i} p(\boldsymbol{y}_{ij} \,|\, \boldsymbol{\theta}_i)$.
  Often choose $p(\boldsymbol{y}_{ij} \,|\, \boldsymbol{\theta}_i) = \mathrm{N}_r(\boldsymbol{y} \,|\, \boldsymbol{\mu}_i, \Sigma_1)$, where r is the number of biological responses measured.

  Assuming exchangeable animals within the environment leads to
  $p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \prod_{i=1}^{k} \pi(\boldsymbol{\mu}_i \,|\, \boldsymbol{\phi})$
  for some probability distribution $\pi(\boldsymbol{\mu} \,|\, \boldsymbol{\phi})$ describing the biological variation within the species. Often choose $\pi(\boldsymbol{\mu} \,|\, \boldsymbol{\phi}) = \mathrm{N}_r(\boldsymbol{\mu} \,|\, \boldsymbol{\mu}_0, \Sigma_2)$.
- 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 $\mathrm{N}_r(\boldsymbol{y} \,|\, \boldsymbol{\mu}_i, \Sigma_1)$,
  $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$, random from $\mathrm{N}_r(\boldsymbol{\mu} \,|\, \boldsymbol{\mu}_0, \Sigma_2)$
  provides a far richer model than (unrealistic) simple multivariate normal sampling.
- Finer subdivisions, *e.g.*, 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 $p(\boldsymbol{z} \mid \boldsymbol{\theta}_i)$, $\boldsymbol{\theta}_i \in \Theta$,
    $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$, random of size k from $\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi})$, $\boldsymbol{\phi} \in \Phi$.
  - Specify a *prior distribution* (or a reference prior function)
    $\pi(\boldsymbol{\phi})$ for the *hyperparameter vector* $\boldsymbol{\phi}$.
  - Use *standard probability theory* to compute all desired
    posterior distributions:
    $\pi(\boldsymbol{\phi} \mid \boldsymbol{x})$ for inferences about the hyperparameters,
    $\pi(\boldsymbol{\theta}_i \mid \boldsymbol{x})$ for inferences about the parameters,
    $\pi(\psi \mid \boldsymbol{x})$ for inferences about the any function $\psi = \psi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$
    of the parameters,
    $\pi(\boldsymbol{y} \mid \boldsymbol{x})$ for predictions on future observations,
    $\pi(t \mid \boldsymbol{x})$ for predictions on any function $t = t(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$
    of m future observations
  - *Markov Chain Monte Carlo* based *software* available for the necessary
    computations.

3. Decision Making

3.1 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 $a \in \mathcal{A}$, Θ_a is the set of *relevant events*, those may affect the result of choosing a.
  - Each pair $\{a, \boldsymbol{\theta}\}$, $\boldsymbol{\theta} \in \Theta_a$, 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 $\{(\Theta_a, \mathcal{C}_a), a \in \mathcal{A}\}$ 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 Θ_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 structure consists of:

    - The *Transitivity* of preferences:
      If $a_1 \succ a_2$ given C, and $a_2 \succ a_3$ given C,
      then $a_1 \succ a_3$ given C.
    - The *Sure-thing principle*:
      If $a_1 \succ a_2$ given C and E, and $a_1 \succ a_2$ given C and not E
      then $a_1 \succ a_2$ given C.
    - The existence of *Standard events*:
      There are events of known plausibility.
      These 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.

- *Decision making*

  - Many different axiom sets.
    All lead basically to the same set of conclusions, namely:

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

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

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

    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: *the optimal action minimizes the expected loss*.

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

    (alternatively, one may *maximize expected utility*)

- *Intrinsic loss functions: Intrinsic discrepancy*

  - The loss function is typically *context dependent*.
  - In mathematical statistics, *intrinsic* loss functions are used to measure the distance between between statistical models.

    They measure the *divergence between the models* $\{p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}\}$ and $\{p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}\}$ as some *non-negative* function of the form $\ell\{p_1, p_2\}$ which is zero if (and only if) the two distributions are equal almost everywhere.
  - The *intrinsic discrepancy* between two statistical models is simply the intrinsic discrepancy between their sampling distributions, *i.e.*,

$$\begin{aligned}\delta\{p_1, p_2\} &= \delta\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\} \\ &= \min\left\{\int_{\mathcal{X}_1} p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1) \log \frac{p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)}{p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2)} \, \mathrm{d}\boldsymbol{x}, \int_{\mathcal{X}_2} p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2) \log \frac{p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2)}{p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)} \, \mathrm{d}\boldsymbol{x}\right\}\end{aligned}$$

  - The intrinsic discrepancy is an *information-based, symmetric, invariant intrinsic loss*.

3.2 Point and Region Estimation

- *Point estimation as a decision problem*

  - Given statistical model $\{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \Omega\}$, quantity of interest $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \Theta$. A *point estimator* $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\boldsymbol{x})$ of $\boldsymbol{\theta}$ is some function of the data to be regarded as a proxy for the unknown value of $\boldsymbol{\theta}$.
  - To choose a point estimate for $\boldsymbol{\theta}$ is a *decision problem*, where the action space is $\mathcal{A} = \Theta$.
  - Given a *loss function* $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$, the posterior expected loss is

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

    The corresponding *Bayes estimator* is the function of the data,

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\boldsymbol{x}) = \arg \inf_{\tilde{\boldsymbol{\theta}} \in \Theta} \ell[\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}],$$

    which minimizes that expectation.

- *Conventional estimators*

  - The *posterior mean* and the *posterior mode* are the Bayes estimators which respectively correspond to a *quadratic* an a *zero-one* loss functions.

    • If $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$, then, assuming that the mean exists, the Bayes estimator is the *posterior mean* $E[\boldsymbol{\theta} \mid \boldsymbol{x}]$.

  - • If the loss function is a zero-one function, so that $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$ if $\tilde{\boldsymbol{\theta}}$ belongs to a ball of radius ε centered in $\boldsymbol{\theta}$ and $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 1$ otherwise then, assuming that a unique mode exists, the Bayes estimator converges to the *posterior mode* $\text{Mo}[\boldsymbol{\theta} \mid \boldsymbol{x}]$ as the ball radius ε tends to zero.

  - If θ is *univariate and continuous*, and the loss function is *linear*,

$$\ell(\tilde{\theta}, \theta) = \begin{cases} c_1(\tilde{\theta} - \theta) & \text{if} \quad \tilde{\theta} \geq \theta \\ c_2(\theta - \tilde{\theta}) & \text{if} \quad \tilde{\theta} < \theta \end{cases}$$

    then the Bayes estimator is the *posterior quantile* of order $c_2/(c_1 + c_2)$, so that $\Pr[\theta < \theta^*] = c_2/(c_1 + c_2)$.

    In particular, if $c_1 = c_2$, the Bayes estimator is the *posterior median*.

  - Any θ value may be optimal: *it all depends on the loss function*.

- *Intrinsic point estimation*

  - Given the statistical model $\{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \Theta\}$ 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 reparametrization and invariant under 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

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

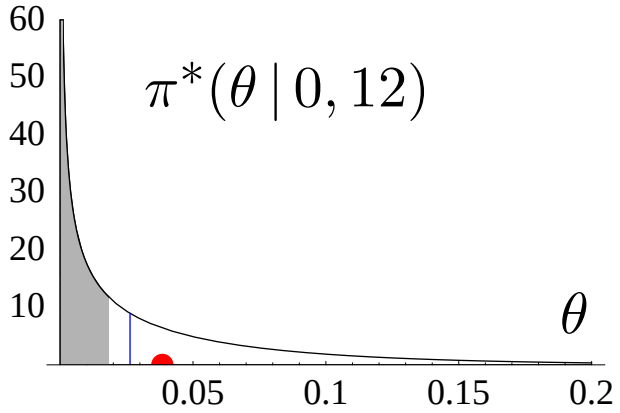
    is an objective measure, in information units, of the *expected* discrepancy between the model $p(\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}})$ and the true (unknown) model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$.

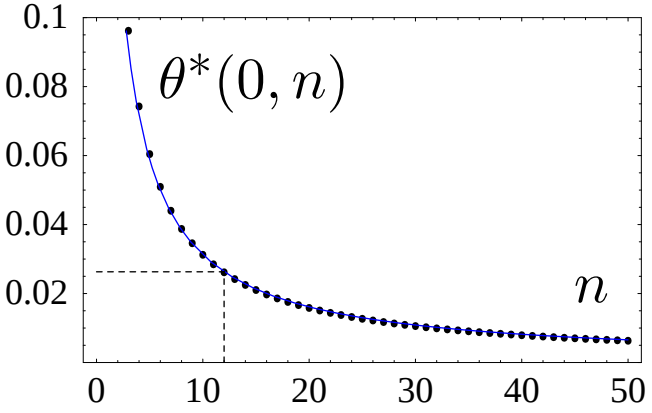
    - The *intrinsic estimator* $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\boldsymbol{x})$ is the value which minimizes such expected discrepancy,

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

- *Example: Intrinsic estimation of the Binomial parameter*

  - Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, random from $p(x \,|\, \theta) = \theta^x (1-\theta)^{1-x}$, $r = \Sigma x_j$. Intrinsic discrepancy $\delta(\tilde{\theta}, \theta) = n \, \min\{k(\tilde{\theta} \,|\, \theta), k(\theta \,|\, \tilde{\theta})\}$,
    $k(\theta_1 \,|\, \theta_2) = \theta_2 \log \frac{\theta_2}{\theta_1} + (1-\theta_2) \log \frac{1-\theta_2}{1-\theta_1}$, $\quad \pi^*(\theta) = \text{Be}(\theta \,|\, \frac{1}{2}, \frac{1}{2})$
    $\pi^*(\theta \,|\, r, n) = \text{Be}(\theta \,|\, r + \frac{1}{2}, n - r + \frac{1}{2})$.
  - Expected reference discrepancy
    $d(\tilde{\theta}, r, n) = \int_0^1 \delta(\tilde{\theta}, \theta) \, \pi^*(\theta \,|\, r, n) \, \text{d}\theta$
  - Intrinsic estimator
    $\theta^*(r, n) = \arg\min_{0<\tilde{\theta}<1} d(\tilde{\theta}, r, n)$
    From invariance, for any bijection $\phi = \phi(\theta)$, $\phi^* = \phi(\theta^*)$.
  - Analytic approximation
    $\theta^*(r, n) \approx \frac{r+1/3}{n+2/3}, \quad n > 2$
  - $n = 12, r = 0, \quad \theta^*(0, 12) = 0.026$
    $\text{Me}[\theta \,|\, \boldsymbol{x}] = 0.018$, $\text{E}[\theta \,|\, \boldsymbol{x}] = 0.038$





- *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:

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

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

  - Binomial examples: $d(\boldsymbol{\theta}_i \,|\, \boldsymbol{x}) = d(\theta_i \,|\, r, n)$

$r = 0, n = 12,$

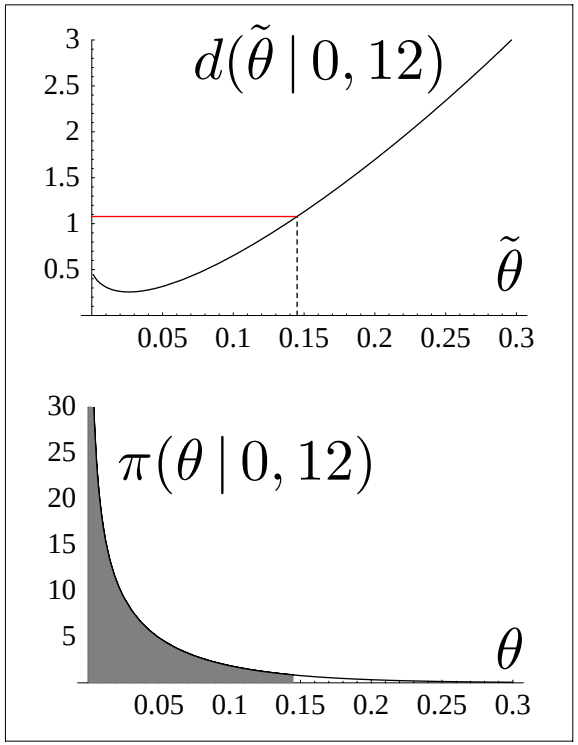
$\theta^* = 0.0263;$

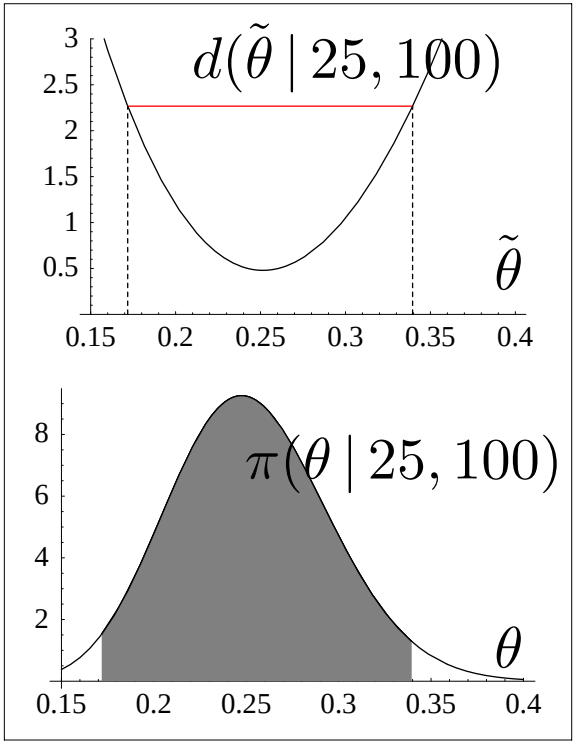
$R^*_{0.95} = [0, 0.145];$

$r = 25, n = 100,$

$\theta^* = 0.2514;$

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





3.3 Hypothesis Testing

- *Precise hypothesis testing as a decision problem*

  - The posterior $\pi(\boldsymbol{\theta} \mid D)$ conveys intuitive information on the values of $\boldsymbol{\theta}$ which are *compatible* with the observed data $\boldsymbol{x}$: those with a *relatively high probability density*.
  - Often a particular value $\boldsymbol{\theta}_0$ is suggested for special consideration:
    - Because $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ would greatly simplify the model
    - Because there are context specific arguments suggesting that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

    More generally, one may analyze the *restriction* of parameter space Θ to a subset Θ_0 which may contain more than one value.
  - Formally, testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is a *decision problem* with just two possible actions:
    - a_0: to *accept* H_0 and work with $p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)$.
    - a_1: to *reject* H_0 and keep the general model $p(\boldsymbol{x} \mid \boldsymbol{\theta})$.
  - To proceed, a *loss* function $\ell(a_i, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, describing the possible consequences of both actions, must be specified.

- *Structure of the loss function*

  - Given data $\boldsymbol{x}$, optimal action is to reject H_0 (action a_1) *iff* the expected posterior loss of accepting, $\int_\Theta \ell(a_0, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta}$, is *larger* than the expected posterior loss of rejecting, $\int_\Theta \ell(a_1, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta}$, *i.e.*, iff

    $$\int_\Theta [\ell(a_0, \boldsymbol{\theta}) - \ell(a_1, \boldsymbol{\theta})]\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta} = \int_\Theta \Delta\ell(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta} > 0.$$

    Therefore, only the loss difference $\Delta\ell(\boldsymbol{\theta}) = \ell(a_0, \boldsymbol{\theta}) - \ell(a_1, \boldsymbol{\theta})$, which measures the *advantage* of rejecting H_0 as a function of $\boldsymbol{\theta}$, has to be specified: The hypothesis should be rejected whenever the *expected* advantage of rejecting is positive.

  - The advantage $\Delta\ell(\boldsymbol{\theta})$ of rejecting H_0 as a function of $\boldsymbol{\theta}$ should be of the form $\Delta\ell(\boldsymbol{\theta}) = l(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - l^*$, for some $l^* > 0$, where

    - $l(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ measures the *discrepancy* between $p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)$ and $p(\boldsymbol{x} \mid \boldsymbol{\theta})$,
    - l^* is a positive *utility constant* which measures the advantage working with the simpler model when it is true.

  - The Bayes criterion will then be: *Reject* H_0 if (and only if)

    $\int_\Theta l(\boldsymbol{\theta}_0, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta} > l^*$, that is if (and only if)
    the *expected discrepancy* between $p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)$ and $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ is *too large*.

- *Bayesian Reference Criterion*

  - An good choice for the function $l(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is the *intrinsic discrepancy*,
    $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \min\{k(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}),\, k(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0)\}$,
    where $k(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}) = \int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \log\{p(\boldsymbol{x} \,|\, \boldsymbol{\theta})/p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0)\} d\boldsymbol{x}$.
    If $\boldsymbol{x} = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ is a random sample from $p(x \,|\, \boldsymbol{\theta})$, then
    $k(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}) = n \int_{\mathcal{X}} p(x \,|\, \boldsymbol{\theta}) \log \frac{p(x \,|\, \boldsymbol{\theta})}{p(x \,|\, \boldsymbol{\theta}_0)}\, dx$.
  - For objective results, exclusively based on model assumptions and data, the *reference* posterior distribution $\pi^*(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ should be used.
  - Hence, *reject if (and only if) the expected reference posterior intrinsic discrepancy $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ is too large*,
    $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x}) = \int_{\Theta} \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta})\, \pi^*(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta} > d^*$, for some $d^* > 0$.
    This is the *Bayesian reference criterion (BRC)*.
  - The *reference test statistic* $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ is nonnegative, it is invariant both under reparametrization and under sufficient transformation of the data, and it is a measure, in natural information units (nits) of the expected discrepancy between $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0)$ and the true model.

- *Calibration of the BRC*

  - The reference test statistic $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})$ is the posterior expected value of the intrinsic discrepancy between $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0)$ and $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$.

    • A reference test statistic value $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x}) \approx 1$ suggests that data are clearly compatible with the Hypotheis that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

    • A test statistic value $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x})\ \log(10) = 2.303$ nits 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: *mild evidence* against $\boldsymbol{\theta}_0$.

    • Similarly, $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{x}) \approx \log(100) = 4.605$ (expected likelihood ratio against $\boldsymbol{\theta}_0$ about 100), indicates *strong evidence* against $\boldsymbol{\theta}_0$, and $\log(1000) = 6.908$, *conclusive evidence* against $\boldsymbol{\theta}_0$.

  - Strong connections between BRC and intrinsic estimation:

    • The *intrinsic estimator* is the value of $\boldsymbol{\theta}$ with minimizes the reference test statistic: $\boldsymbol{\theta}^* = \arg\inf_{\boldsymbol{\theta} \in \Theta} d(\boldsymbol{\theta} \,|\, \boldsymbol{x})$.

    • The regions defined by $\{\boldsymbol{\theta};\ d(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \leq d^*\}$ are invariant *reference posterior $q(d^*)$-credible regions* for $\boldsymbol{\theta}$. For regular problems and large samples, $q(\log(10)) \approx 0.95$ and $q(\log(100)) \approx 0.995$.

- *A canonical example: Testing a value for the Normal mean*

  - In the simplest case where the variance σ^2 is known,

    $$\delta(\mu_0, \mu) = n(\mu - \mu_0)^2/(2\sigma^2), \qquad \pi^*(\mu \,|\, \boldsymbol{x}) = \mathrm{N}(\mu \,|\, \overline{x}, \sigma/\sqrt{n}),$$

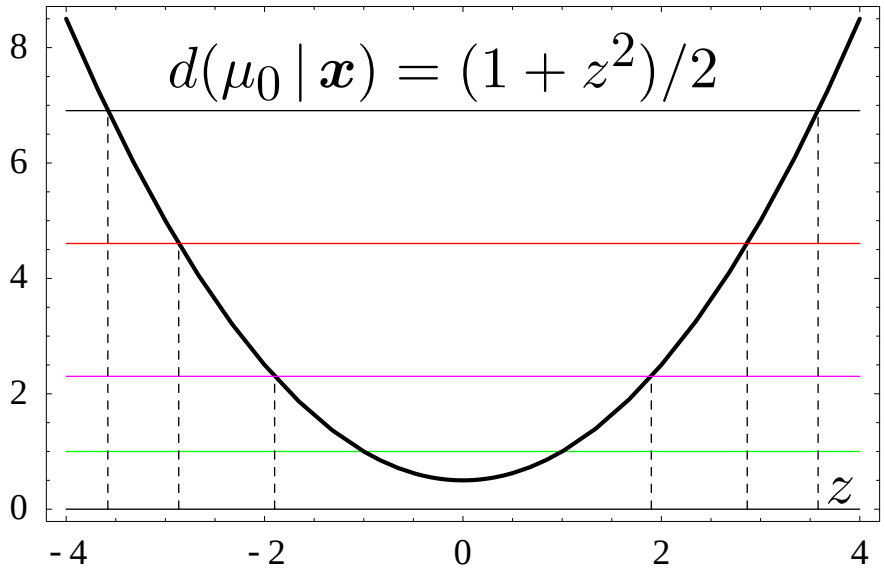
    $$d(\mu_0 \,|\, \boldsymbol{x}) = \tfrac{1}{2}(1 + z^2), \qquad z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

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

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

  - The expected value of $d(\mu_0 \,|\, \boldsymbol{x})$ if the hypothesis is true is

    $$\int_{-\infty}^{\infty} \tfrac{1}{2}(1 + z^2)\mathrm{N}(z \,|\, 0, 1)\, \mathrm{d}z = 1$$



- *Fisher's tasting tea lady*

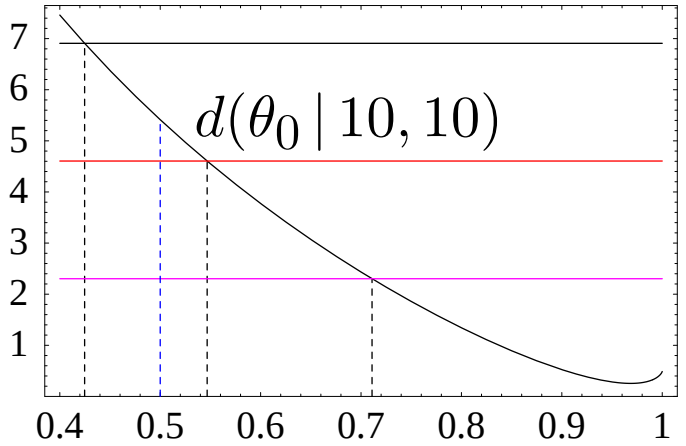
  - Data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, random from $p(x \,|\, \theta) = \theta^x (1-\theta)^{1-x}$, $r = \Sigma x_j$. Intrinsic discrepancy $\delta(\theta_0, \theta) = n \, \min\{k(\theta_0 \,|\, \theta), k(\theta \,|\, \theta_0)\}$,
    $k(\theta_1 \,|\, \theta_2) = \theta_2 \log \frac{\theta_2}{\theta_1} + (1-\theta_2) \log \frac{1-\theta_2}{1-\theta_1}$, $\quad \pi^*(\theta \,|\, r, n) = \text{Be}(\theta \,|\, r + \frac{1}{2}, n - r + \frac{1}{2})$

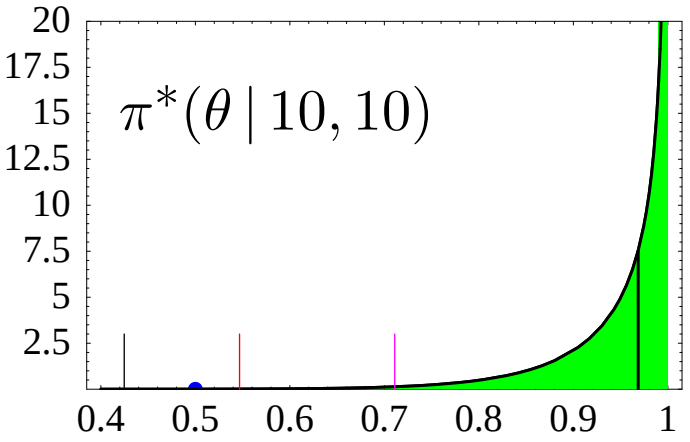
    Intrinsic test statistic
    $d(\theta_0 \,|\, r, n) = \int_0^1 \delta(\tilde{\theta}, \theta) \, \pi^*(\theta \,|\, r, n) \, \mathrm{d}\theta$

  - Fisher's example: $\boldsymbol{x} = \{10, 10\}$,
    Test $\theta_0 = 1/2$, $\theta^*(\boldsymbol{x}) = 0.9686$
    $d(\theta_0 \,|\, 10, 10) = 5.414 = \log[224]$

    Using $d^* = \log[100] = 4.61$,
    the value $\theta_0 = 1/2$ is rejected.
    $\Pr[\theta < 0.5 \,|\, \boldsymbol{x}] = 0.00016$

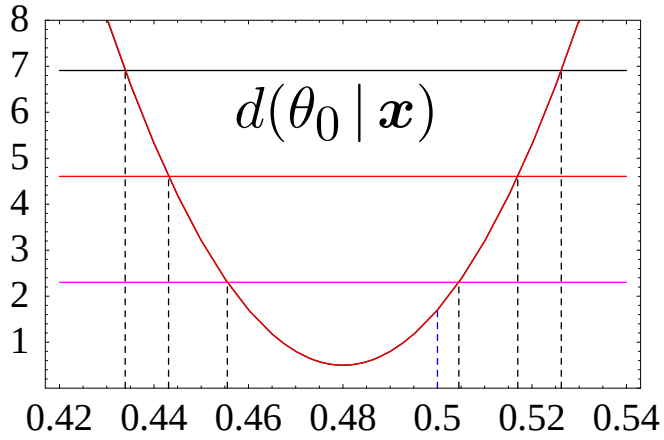
$d(\theta^* \,\|\, \boldsymbol{x})$	θ^*	$\Pr[\theta < \theta^* \,\|\, \boldsymbol{x}]$
$\log[10]$	0.711	0.00815
$\log[100]$	0.547	0.00043
$\log[1000]$	0.425	0.00003





- *Asymptotic approximation*

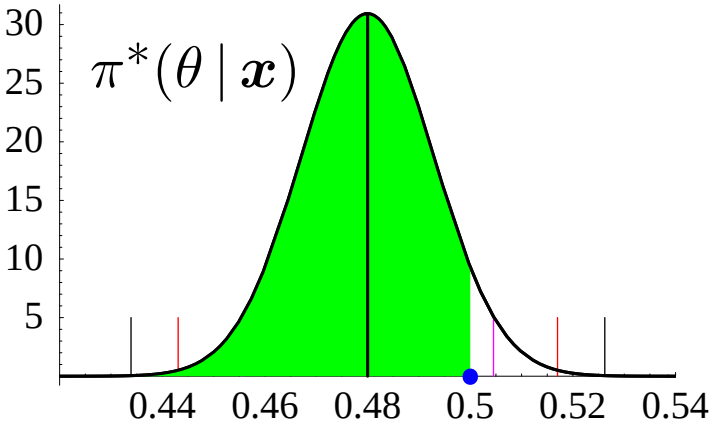
  - For large samples, the posterior approaches $\mathrm{N}(\theta \,|\, \hat{\theta}, (nF(\hat{\theta}))^{-1/2})$, where $F(\theta)$ is Fisher's function. Changing variables, the posterior distribution of $\phi = \phi(\theta) = \int F^{1/2}(\theta)\,\mathrm{d}\theta = 2 \arcsin \sqrt{\theta})$ is approximately normal $\mathrm{N}(\phi \,|\, \hat{\phi}, n^{-1/2})$. Since $d(\theta, \boldsymbol{x})$ is invariant, $d(\theta_0, \boldsymbol{x}) \approx \frac{1}{2}[1 + n\{\phi(\theta_0) - \phi(\hat{\theta})\}^2]$.



- *Testing for a majority ($\theta_0 = 1/2$)*

$$\boldsymbol{x} = \{720, 1500\}, \quad \theta^*(\boldsymbol{x}) = 0.4800$$

$d(\theta^* \,\|\, \boldsymbol{x})$	$R = (\theta_0^*, \theta_1^*)$	$\Pr[\theta \in R \,\|\, \boldsymbol{x}]$
log[10]	(0.456, 0.505)	0.9427
log[100]	(0.443, 0.517)	0.9959
log[1000]	(0.434, 0.526)	0.9997



Very mild evidence against $\theta = 0.5$:
$d(0.5 \,|\, 720, 1500) = 1.67$
$\Pr(\theta < 0.5 \,|\, 720, 1500) = 0.9393$

Basic References

Many available on line at **www.uv.es/bernardo**

- *Introductions*

Bernardo, J. M. and Ramón, J. M. (1998).
An introduction to Bayesian reference analysis. *The Statistician* **47**, 1–35.

Bernardo, J. M. (2003). Bayesian Statistics.
Encyclopedia of Life Support Systems (EOLSS): Probability and Statistics, (R. Viertl, ed). Oxford, UK: UNESCO.

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

- *Textbooks*

Gelman, A., Carlin, J. B., Stern, H. and Rubin, D. B. (2003).
Bayesian Data Analysis (2nd ed.) New York: CRC Press.

Bernardo, J. M. and Smith, A. F. M. (1994).
Bayesian Theory. Chichester: Wiley.
2nd ed. to appear in June 2006

- *Research papers on reference analysis (cronological order)*

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113—147, (with discussion).
Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.), Brookfield, VT: Edward Elgar, (1995), 229—263.

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. and Juárez, M. (2003). Intrinsic estimation. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, 465–476.

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

- Valencia International Meetings on Bayesian Statistics

  Sponsored by the University of Valencia. Held every four years in Spain. World forums on research and applications of Bayesian analysis.

8th Valencia International Meeting
on Bayesian Statistics

Benidorm (Alicante), June 1st – 6th 2006

www.uv.es/valenciameeting

- Valencia Mailing List

  The Valencia Mailing List contains about 2,000 entries of people interested in Bayesian Statistics. It sends information about the Valencia Meetings and other material of interest to the Bayesian community.

  If you do not belong to the list and want to be included,
  please send your e-mail to **<valenciameeting@uv.es>**

- José-Miguel Bernardo contact data

  <jose.m.bernardo@uv.es>

  www.uv.es/bernardo