02/01/22 Happy new year!

① $E_\theta(L(\theta, \delta(x))) = \int L(\theta, \delta(x)) f(x|\theta)\, dx = R(\theta, \delta(x))$

$E^\pi(L(\theta, d) | x) = \int L(\theta, d)\, \pi(\theta | x)\, d\theta = \rho(\pi, d | x)$

② Midterm 1

③ HW#2 solution : corrected

† Hierarchical Bayes

- A hierarchical model is simply a special case of Bayesian model.

$\theta_1 \sim \pi_2(\ \mid \theta_2)$, $\theta_2 \sim \pi_3(\cdot\cdot \mid \theta_3)$

$$\underbrace{x \sim f(x \mid \theta)}_{\text{sampling model}}, \quad \underbrace{\theta \sim \pi_1(\theta \mid \theta_1)}_{\text{stage 1 prior}}, \ldots, \quad \underbrace{\theta_n \sim \pi_{n+1}(\theta_n)}_{\text{stage } n+1 \text{ prior}}.$$

Then we recover the usual Bayes model

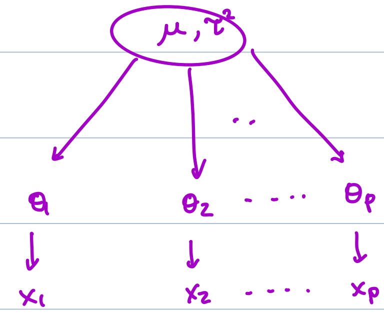
$f(x \mid \theta_1) = \int \underbrace{f(x \mid \theta)\pi_1(\theta \mid \theta_1)}_{f(x,\theta \mid \theta_1)} d\theta$

$$x \sim f(x \mid \theta), \theta \sim \pi(\theta),$$

for the prior

$$\pi(\theta) = \int_{\Theta_1 \times \ldots \times \Theta_n} \pi_1(\theta \mid \theta_1)\pi_2(\theta_1 \mid \theta_2) \ldots \pi_{n+1}(\theta_n) d\theta_1 \ldots d\theta_n.$$

⋆⋆ Most of time θ is of the primary interest, less interest for hyperparameters, $\theta_1, \ldots, \theta_n$.



← stage 2 : $\pi_2(\mu, \tau^2) = \pi_{21}(\mu \mid \tau^2)\, \pi_{22}(\tau^2)$

$= N(\mu_0, \kappa\tau^2)\, IG(a, b)$

with κ, a, b are fixed.

← stage 1 : $\theta_i \overset{iid}{\sim} N(\mu, \tau^2)$

← $x_i \mid \theta_i \overset{indep}{\sim} N(\theta_i, \sigma^2)$; σ^2 known

- $$f(x_1, \dots, x_p \mid \mu, \tau^2) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^{p} f(x_i \mid \theta_i, \sigma^2) \cdot \pi_1(\theta_i \mid \mu, \tau^2)\, d\theta_1 \cdots d\theta_p$$

- Unknown parameters :

① Random : $\theta_1, \dots, \theta_p$, μ, τ^2

② fixed : σ^2, κ, a, b : use prior information & specify their values.

① joint posterior distr.

$$\pi(\theta_1, \dots, \theta_p, \mu, \tau^2 \mid x) \propto \prod_{i=1}^{p} f(x_i \mid \theta_i, \sigma^2) \prod_{i=1}^{p} \pi_1(\theta_i \mid \mu, \tau^2) \cdot \pi_{21}(\mu \mid \mu_0, \kappa\tau^2)\, \pi_{22}(\tau^2 \mid a, b)$$

$$= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta_i)^2}{2\sigma^2}\right) \cdot \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\theta_i-\mu)^2}{2\tau^2}\right)$$

$$\times \frac{1}{\sqrt{2\pi\kappa\tau^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\kappa\tau^2}\right) \cdot (\tau^2)^{-a-1} \exp\left(-\frac{\tau^2}{b}\right)$$

ex1 $$\pi(\mu \mid \theta_1, \dots, \theta_p, \tau^2, x) \propto \exp\left(-\sum_{i=1}^{p} \frac{(\theta_i-\mu)^2}{2\tau^2} - \frac{(\mu-\mu_0)^2}{2\kappa\tau^2}\right)$$

$$\Rightarrow \mu \mid \theta_1, \dots, \theta_p, \tau^2, x \sim N\left(\left(\frac{p}{\tau^2} + \frac{1}{\kappa\tau^2}\right)^{-1}\left(\frac{\sum\theta_i}{\tau^2} + \frac{\mu_0}{\kappa\tau^2}\right), \left(\frac{p}{\tau^2} + \frac{1}{\kappa\tau^2}\right)^{-1}\right)$$

$\pi(\tau^2 \mid \theta_1, \dots, \theta_p, \mu, x)$ IG. $(\tau^2)^{-p/2 - 1/2 - a - 1}$

$\exp($

$\pi(\theta_1 \mid \theta_2, \dots, \theta_p, \mu, \tau^2, x)$ N

- **BJ Result 7, p180** Supposing all densities below exist and are nonzero, we have

$$\pi(\theta \mid \boldsymbol{x}) = \int_{\Theta_1 \times \ldots \times \Theta_n} \pi(\theta, \theta_1, \ldots, \theta_n \mid \boldsymbol{x}) \; d\theta_1 \ldots d\theta_n.$$

⋆⋆ *Implication?* Recall the posterior of θ is of main interest. Our strategy is

∗∗ Find the joint posterior of $\theta, \theta_1, \ldots, \theta_n$.

∗∗ Then integrate out $\theta_1, \ldots, \theta_n$ to obtain the marginal posterior of θ.

⋆⋆ Analytically impossible most of time, so numerically evaluate using posterior simulation.

⋆⋆ See CR Chapter 10 for more on Empirical Bayes and Hierarchical Bayes.

- A simple example of *Hierarchical Bayes* with two levels:

JB 4.5.2 (contd) Recall that we have $X_i \mid \theta_i \overset{indep}{\sim} \mathrm{N}(\theta_i, \sigma^2)$ with known σ^2, $i = 1, \ldots, p$ and $\theta_i \overset{iid}{\sim} \mathrm{N}(\mu, \tau^2)$, where hyperparameters$(\mu, \tau^2) \in \Theta_2 = \mathbb{R} \times \mathbb{R}^+$ are unknown.

⋆⋆ Sampling model: $X_i \mid \theta_i \overset{indep}{\sim} \mathrm{N}(\theta_i, \sigma^2)$.

⋆⋆ The first-level prior: $\theta_i \overset{iid}{\sim} \pi(\theta) = \mathrm{N}(\mu, \tau^2)$

⋆⋆ The second-level prior $\pi_2(\mu, \tau^2)$:

$$\pi_2(\mu, \tau^2) = \pi_{21}(\mu \mid \tau^2)\ \pi_{22}(\tau^2).$$

⋆⋆ π_2 is called a *hyperprior*.

⋆⋆ The parameters of π_2 are called *hyperparameters*.

JB 4.5.2 (contd)

⋆⋆ Let $\pi_2(\mu, \tau^2) = \mathrm{N}(\mu_0, \kappa\tau^2)\ \mathrm{IG}(a_\tau, b_\tau)$. Now we need to specify values of μ_0, κ, a_τ and b_τ.

⋆⋆ May use subjective beliefs to choose the values.

Say,

μ_0

∗∗ "mean true ability" is near 100 with a "standard error" of $\pm$ 20 $\sqrt{k}\,\tau = 20$

∗∗ "variance of true abilities", τ^2 is about 200 with "standard error" of $\pm$ 100.

$$E(\tau^2) = \frac{b}{a-1} = 200$$

$$Var(\tau^2) = \frac{b^2}{(a-1)^2(a-2)} = 100^2$$

† **Comments** *on Hierarchical Bayes*

- A full Bayesian approach using hierarchical priors
- A hierarchical Bayesian model compares very favorably with empirical Bayes analysis in practical and theoretical senses.
- A hierarchical modeling of the prior information decomposes the prior distribution into several conditional levels of distributions.
- According to the Bayesian paradigm, uncertainty at any of these levels is incorporated into additional prior distributions.
- The hierarchical model improves the robustness of the resulting Bayes estimator: while still incorporating prior information, the estimators are also well performing from a frequentist point of view.

† Conjugate Priors (Sec 3.3)

- **Example 3.2.6** Let $x \sim \mathrm{N}(\theta, 1)$. For Case 2, we considered the prior, $\theta \sim \mathrm{Cauchy}(0, 1)$. In the case, $\pi(\theta \mid x)$ and $m(x)$ are not easily calculable.

- **Def 3.3.1:** A family $\mathcal{F}$ of probability distributions on Θ is said to be *conjugate* (or closed under sampling) for a likelihood function $f(x \mid \theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta \mid x)$ also belong to $\mathcal{F}$.

- The main motivation for using conjugate priors is their tractability

- Also, when limited prior input is available, they are easy to specify since only the determination of a few parameters are needed.

† Examples: Conjugate Priors

e.g1 Assume $x \mid \theta \sim \mathrm{N}(\theta, \sigma^2)$ and $\theta \sim \mathrm{N}(\mu, \tau^2)$.

$$\Rightarrow \quad \theta \mid x \sim \mathrm{N}\left(\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2}\right), \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right).$$

⋆⋆ Normal priors are a conjugate family for normal sampling distributions.

e.g2 Assume $X \mid \theta \sim \mathrm{Bin}(n, \theta)$ and $\theta \sim \mathrm{Be}(\alpha, \beta)$.

$$\Rightarrow \theta \mid x \sim \mathrm{Be}(\alpha + x, \beta + n - x).$$

⋆⋆ Beta priors are a conjugate family for binomial sampling distributions.

† **Comments** on conjugate priors

- Sometimes called *objective* because the sampling model entirely determines the class of priors.

- Can be a reasonable approximation to the true prior

- Updating parameters provides an easy way of seeing the effect of prior and sample information

  $\Rightarrow$ easily calculate $\pi(\theta \mid x)$ (computationally convenient)

- *However,* possibly limited modeling capacity since it is not justified for its proper fitting of the available prior information (so, sometimes resulting in unappealing conclusions)

† Extension: The class of finite mixtures of natural conjugate priors (CR 3.4)

- Recall: One disadvantage of conjugate priors – limiting modeling capacity, but a big advantage – computational convenience.
- One possible extension to overcome the disadvantage while keeping the advantage is using a mixture model.
- Mixtures can be used as a basis to approximate any prior distribution.
- **Example 3.4.1** When a coin is spun on its edge, instead of being thrown in the air, the proportion of *heads* is rarely close to $1/2$, but is rather $1/3$ and $2/3$ because of irregularities in the edge that causes the game to favor one side or the other.

$$\int_0^1 \pi_2(\theta)\, d\theta = \int_0^1 \frac{1}{2} \underline{Be(10, 20)} + \frac{1}{2} Be(20, 10)\, d\theta$$

$$= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1$$

- **Example 3.4.1** (contd): When spinning, n times, a given coin on its edge, we observe the number of heads, $x \sim \text{Be}(n, p)$. The prior distribution on p is then likely to be bimodal. $Be(\alpha, \beta)$ $\alpha + \beta$

Let's consider three different priors.

★★ π_1: Be(1, 1)

mixture weights

$10/(10+20) = 1/3$

$2/3$

★★ π_2: a mixture prior distribution, $\underline{1/2}\underbrace{\text{Be}(10, 20)} + \underline{1/2}\underbrace{\text{Be}(20, 10)}$

mixture components.

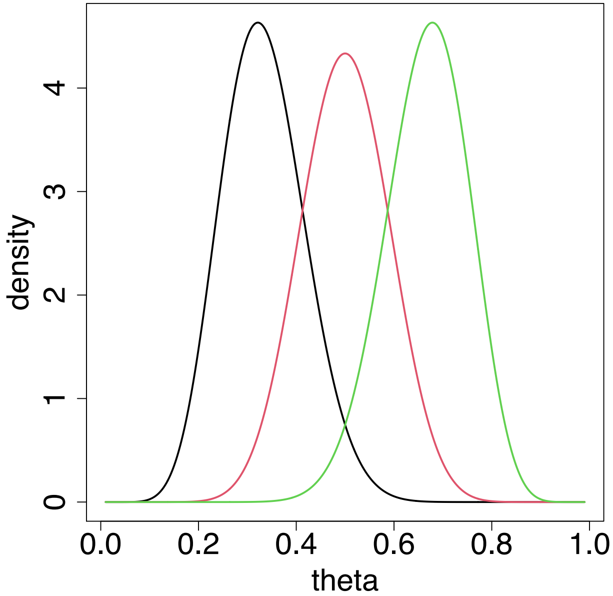
★★ π_3: previous experiments with the same coin have already hinted at a bias toward *head* and they lead to the following alternative, $0.5\text{Be}(10, 20) + 0.2\text{Be}(15, 15) + 0.3\text{Be}(20, 10)$.

$$\theta \sim 0.5\, \underline{Be(10, 20)} + 0.2\, \underline{Be(15, 15)} + 0.3\, \underline{Be(20, 10)}$$

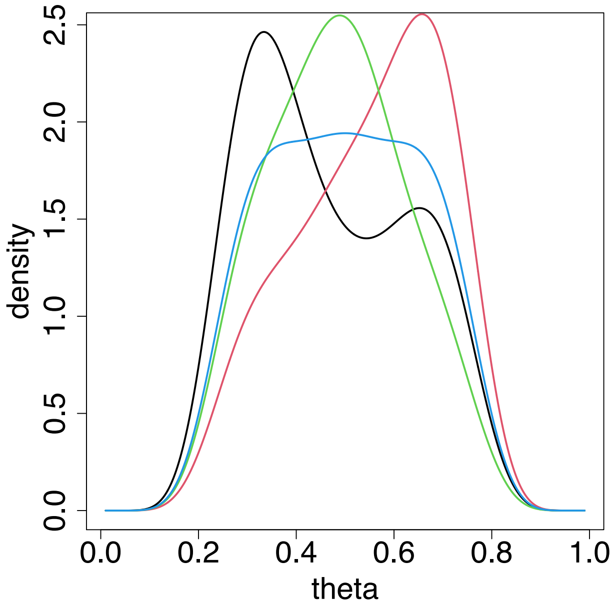
$$\delta \sim \text{Multi}(1, (0.5, 0.2, 0.3))$$

$$\begin{cases} \delta = 1 \Rightarrow \theta \sim Be(10, 20) \\ \delta = 2 \Rightarrow \theta \sim Be(15, 15) \\ \delta = 3 \Rightarrow \theta \sim Be(20, 10) \end{cases}$$

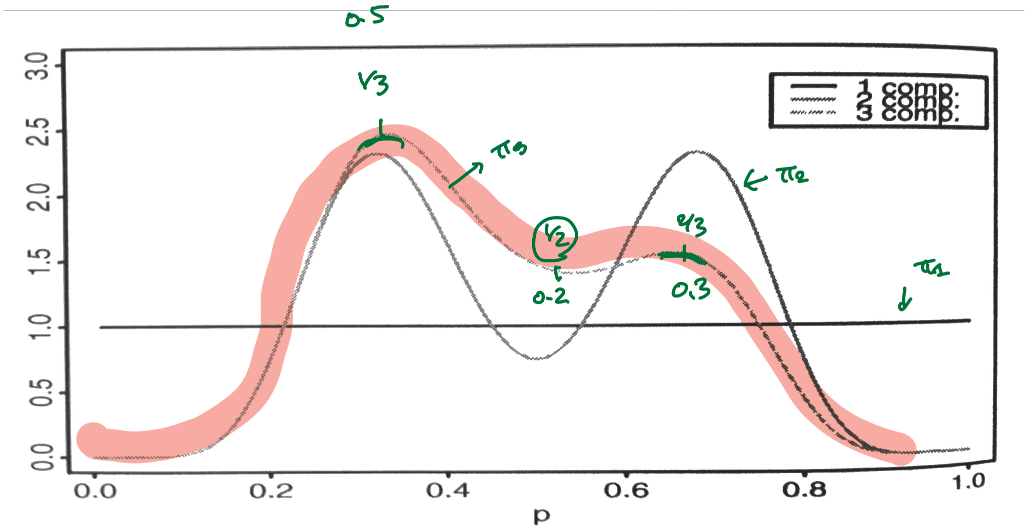
♣ Densities of $\mathrm{Be}(10, 20)$ (black), $\mathrm{Be}(15, 15)$ (red), and $\mathrm{Be}(20, 10)$ (green).



♣ The mixture $w_1\text{Be}(10, 20) + w_2\text{Be}(15, 15) + w_3\text{Be}(20, 10)$ with different weights.



- **Example 3.4.1** (contd): Three prior distributions



gure 3.4.1. *Three prior distributions for a spinning-coin experiment.*

- **Example 3.4.1** (contd): Suppose $x = 3$ for $n = 10$ is observed. The corresponding posterior distributions are

$Be(\alpha, \beta)$

$Be(\alpha + x, \beta + n - x)$

$x=3$, $n-x=7$

$3/10$

★★ π_1: Be(4, 8)

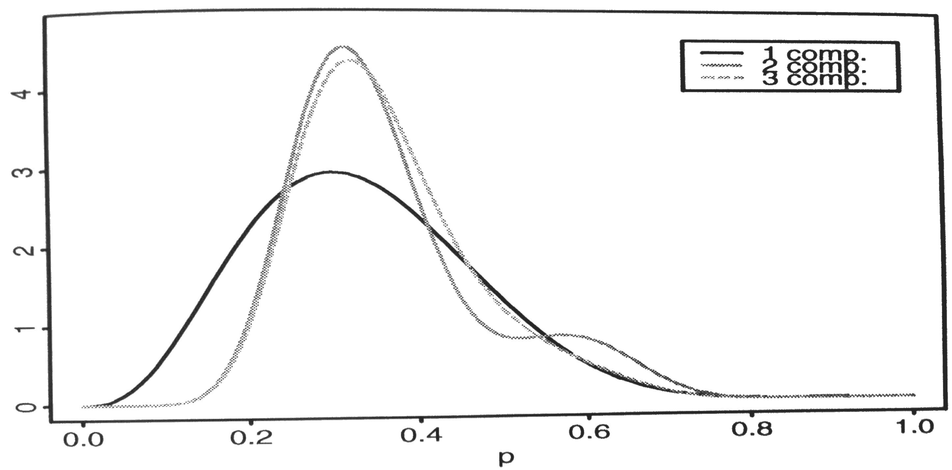
1 1 10, 20

★★ π_2: 0.84Be(13, 27) + 0.16Be(23, 17)

0.5 0.5

★★ π_3: 0.77Be(13, 27) + 0.16Be(18, 22) + 0.07Be(23, 17).

0.5 0.2 0.3



3.4.2. *Posterior distributions for the spinning model for 10 observations.*

- **Example 3.4.1** (contd): Suppose $x = 14$ for $n = 50$ is observed. The corresponding posterior distributions are

  ⋆⋆ π_1: Be$(15, 37)$

  ⋆⋆ π_2: $0.997\text{Be}(24, 56) + 0.003\text{Be}(34, 46)$

  ⋆⋆ π_3: $0.95\text{Be}(24, 56) + 0.047\text{Be}(29, 51) + 0.003\text{Be}(34, 46)$.

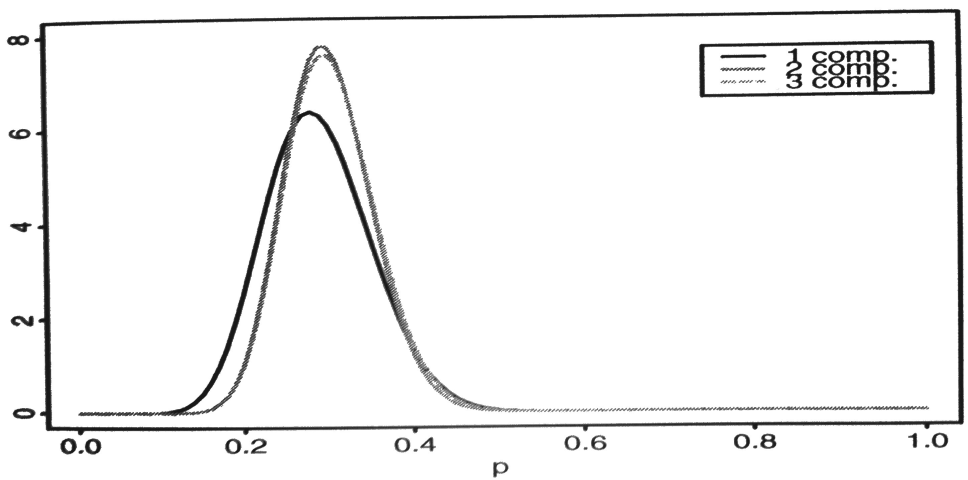


Figure 3.4.3. *Posterior distributions for 50 observations.*

sampling	prior	marginal	posterior
$Bin(n, \theta)$	$Be(\alpha, \beta)$	Beta-Binomial (n, α, β)	$Be(\alpha+x, \beta+n-x)$
Poi	Ga	NB	Ga

$$\pi(\theta) = \sum_{i=1}^{N} w_i \, \pi(\theta \mid \alpha_i, \beta_i), \quad 0 < w_i < 1, \quad \sum_{i=1}^{N} w_i = 1$$

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{m(x)}$$

$$= \frac{\sum_{i=1}^{N} w_i \left(\overbrace{\pi(\theta \mid \alpha_i, \beta_i) f(x \mid \theta)}^{m(x \mid \alpha_i, \beta_i)\, \pi(\theta \mid x, \alpha_i, \beta_i)} \right)}{\underbrace{\int \sum_{i=1}^{N} w_i \left(\pi(\theta \mid \alpha_i, \beta_i) f(x \mid \theta) \right) d\theta}_{= \underbrace{\sum_{i=1}^{N} w_i \, m(x \mid \alpha_i, \beta_i)}_{= m(x)}}}$$

$$= \sum_{i=1}^{N} \underbrace{\frac{w_i \, m(x \mid \alpha_i, \beta_i)}{\sum_{i=1}^{N} w_i \, m(x \mid \alpha_i, \beta_i)}}_{w'(x)} \cdot \underbrace{\pi(\theta \mid x, \alpha_i, \beta_i)}_{Be(\alpha_i + x, \beta_i + n - x)}$$

- Use a mixture of priors and find the posterior distribution
  - ⋆⋆ Consider the set of mixtures of N distributions,

$$\pi(\theta) = \sum_{i=1}^{N} w_i \pi(\theta \mid \mu_i),$$

where μ_i is hyperparameters.

  - ⋆⋆ Then the posterior distribution is a mixture

$$\pi(\theta \mid x) = \sum_{i=1}^{N} w_i'(x) \pi(\theta \mid \mu_i, x),$$

with

$$w_i'(x) = \frac{w_i m(x \mid \mu_i)}{m(x)} = \frac{w_i m(x \mid \mu_i)}{\sum_{j=1}^{N} w_j m(x \mid \mu_j)}.$$

- Finite mixtures of natural conjugate priors.
  - ⋆⋆ See **Lemma 3.4.2** for the case where the prior is the natural conjugate family of an exponential family.
  - ⋆⋆ Mixture models approximate bimodal or more complicated subjective prior distributions ($\Rightarrow$ flexibility); see Theorem 3.4.3.
  - ⋆⋆ Also, they preserve much of the calculational simplicity of natural conjugate priors.
  - ⋆⋆ In general, mixture models can be useful when the population of sampling units consists of a number of subpoplulations within each of which a relatively simple model applies.

- Finite mixtures of natural conjugate priors (contd)

  ⋆⋆ Possible extensions.

  $**$ unknown number of mixture components (random N)

  $**$ random mixture weights (random w_i).

  e.g. $(w_1, \ldots, w_N) \mid N \sim \text{Dir}(\alpha_1, \ldots, \alpha_N)$.

† Noninformative Prior Distributions (CR 3.5 & JB 3.3)

- When no (or minimal) prior information is available, we may use noninformative prior distributions:

  ** Priors which contain "no" information about θ (*roughly* favor no possible values of θ over others!)

  ** A mathematical expression of the state of ignorance about a parameter in a statistical model

- Noninformative priors cannot be expected to represent exactly total ignorance about the problem at hand. A choice of noninformative priors affects the posterior inference.

- Noninformative priors: Laplace priors, invariant priors, Jeffreys priors, reference priors...

† Laplace's Priors (uniform priors or flat priors)

- The principles of insufficient reason: Assign the equiprobability to elementary events

- When Θ is a finite set, consisting of n elements, the obvious noninformative prior is to give each element of Θ probability $1/n$.

  JB Sec 3.3.1 in testing between two simple hypotheses, the prior gives probability $\frac{1}{2}$ to each of the hypothesis.

- *Improper priors:* a prior probability distribution which has infinite mass (i.e., $\int_\Theta \pi(\theta)d\theta = \infty$)

  JB Ex4, p82 Suppose the parameter of interest is a normal mean θ, so $\Theta = (-\infty, \infty)$. It seems reasonable that a natural noninformative prior gives equal weight to all possible values of θ, uniform density on $\mathbb{R}$. Thus, $\pi(\theta) = c > 0$. Since a choice of the value of c is not important, typical $\pi(\theta) = 1$.

  $**$ Observe π has infinite mass!

  $**$ The posterior distribution $\pi(\theta \mid x)$ can be given by Bayes formula when the pseudo marginal distribution $\int_\Theta f(x \mid \theta)\pi(\theta)d\theta < \infty$ for every x in the support of $f(x \mid \theta)$.

  $**$ Since $\pi(\theta \mid x)$ is proper, $\cancel{\rho(\pi(\theta \mid x), a)}$ is finite and so we can find a Bayes action! $\rho(\pi, d \mid x)$

$\theta \in \mathbb{R}$ $\qquad \pi(\theta) = c$, $c > 0$

$\eta = e^{\theta} \in \mathbb{R}^+$ $\qquad \pi_\eta(\eta) = c \cdot \frac{1}{\eta}$

$\theta = \log \eta$

- Invariance under Reparameterization

  ⋆⋆ Consider a reparameterization $\eta = g(\theta)$, where $g(\cdot)$ is monotone over the domain of θ.

  ⋆⋆ Find the induced prior for η

$$\pi_\eta(\eta) = \pi_\theta(g^{-1}(\eta))|dg^{-1}(\eta)/d\theta|.$$

  ⋆⋆ A more intrinsic and more acceptable notion of noninformative priors should satisfy *invariance under reparameterization*.

  i.e., $\pi_\eta(\eta)$ is also a flat prior for η.

- JB Ex4, p82 (contd) Consider $\eta = \exp(\theta)$ by a one-to-one transformation.

  ⋆⋆ It is reasonable to assume that $\pi^\star(\eta)$ is also a noninformative prior for η.

  ⋆⋆ We can find

$$\pi(\theta) = 1 \quad \Rightarrow \quad \pi^\star(\eta) = |\frac{d}{d\eta} g^{-1}(\eta)| = \eta^{-1}.$$

  Observe $\pi^\star(\eta) = \eta^{-1}$ is not constant. $\Rightarrow$ Not invariant under reparameterization.

  ⋆⋆ Do Ex 3.5.1 for more example.

† Invariant Priors

- priors invariant under transformation of x. **Ex 3.5.2** (location parameter) and **Ex 3.5.3** (scale parameter)

  ⋆⋆ (intuition) Consider $x \sim \mathrm{N}(\theta, \sigma^2)$, σ^2 fixed. Assume instead of observing x, we observe $y = x + c$ with a constant $c \in \mathbb{R}$. Defining $\eta = \theta + c$, the problems of (x, θ) and (y, η) are identical so θ and η should have the same noninofrmative prior.

- For a location parameter θ, $\pi(\theta) = c$

- For a scale parameter σ, $\pi(\sigma) = c/\sigma$

† Fisher Information (CB p338 or 203 Textbook §8.8)

- (Def: Fisher Information in a Random Variable) Let X be a random variable whose distribution depends on a parameter θ that takes values in an open interval Θ of the real line. Let the pf or pdf of X be $f(x \mid \theta)$. Assume that the set of x such that $f(x \mid \theta) > 0$ is the same for all θ and that $\log(f(x \mid \theta))$ is twice differentiable as a function of θ. The Fisher information $I(\theta)$ in the random variable X is defined as

$$I(\theta) = \mathrm{E}_\theta \left[\left(\frac{\partial \log f(x \mid \theta)}{\partial \theta} \right)^2 \right].$$