STAT 206B
Chapter 3: From Prior Information to Prior Distributions

Winter 2022

- A fundamental basis of Bayesian decision theory is that statistical inference should start with the rigorous determination of three factors.

  $\star\star$ the distribution family for the observations (sampling distribution), $f(x \mid \theta)$ for $x \in \mathcal{X}$

  $\star\star$ the prior distribution for the parameter $\pi(\theta)$, $\theta \in \Theta$

  $\star\star$ the loss association with the decisions, $L(\theta, \delta) \in [0, +\infty)$.

- In this chapter, we will discuss prior distributions – CR Chapter 3 and JB Chapters 3 & 4.

† Priors!

- Priors are carriers of external knowledge (outside the data being modeled and analyzed) that is coherently incorporated via Bayes theorem to the inference.

- Parameters (θ) are unobservable.

  $\Rightarrow$ Prior specification is **subjective** in nature.

- There is no unique way of choosing a prior distribution.

  $\Rightarrow$ There is no such a thing as *the* prior distribution.

- The choice of the prior distribution has an influence on the resulting inference.

  $\Rightarrow$ Ungrounded prior distributions produce unjustified posterior inference.

† *Is using a prior a problem?*

- The elicitation of a model (likelihood) and loss function is highly subjective, and Bayesians merely divide the necessary subjectivity to two sources - that from the model and from the prior.

- Vast amount scientific information coming from theoretical and physical models is guiding specification of priors and merging such information with the data for better inference.

- Being subjective $\neq$ Being nonscientific

- If complete information is given, an exact prior can be elicited. **However**, it is very rare!

- How to specify priors?

  $\star\star$ Subjective determination and approximations (Sec 3.2)

  $\star\star$ Conjugate priors (Sec 3.3)

  $\star\star$ Noninformative prior distributions (Sec 3.5): *have* **little influence** *on the posterior distribution*

- *criticism* Bayesian inference is overly sensitive to the choice of a prior

  $\Rightarrow$ the development of non-informative and robust priors (so change in the prior distribution does not change the posterior inference much)

† Subjective Determination (Sec 3.2)

- Subjective prior distributions exist as a consequence of an ordering of relative likelihoods.

- Approximations to the prior distribution. e.g.

  ⋆⋆ When the parameter space Θ is finite, obtain a subjective evaluation of the probabilities of the different values of θ.

  ⋆⋆ When Θ is noncountable (e.g. an interval of the real line), use the histogram approach.

  - Divide Θ into intervals
  - Determine the subjective probability of each interval
  - Plot a probability histogram
  - If needed, a smooth density $\pi(\theta)$ can be sketched.

- Approximations to the prior distribution. (contd)

JB Example 1 Assume that $\Theta = [0, 1]$. Suppose that

- ⋆⋆ the parameter point $\theta = 3/4$ is felt to be the most likely, while $\theta = 0$ is the least likely.
- ⋆⋆ $3/4$ is estimated to be three times as likely to be the true value of θ as is 0.
- ⋆⋆ $\theta = 1/2$ and $\theta = 1$ are twice likely as $\theta = 0$ while $\theta = 1/4$ is 1.5 times as likely as $\theta = 0$.

- Approximations to the prior distribution. (contd)

  $\star\star$ So far we have seen "histogram approach" and "relative likelihood approach".

  $\star\star$ JB discusses using a subjective construction of CDF in Section 3.2.

- When Θ is not bounded, the subjective determination of π is complicated due to the difficulty of subjectively evaluating the probabilities of the extreme regions of Θ (will see this from Example 3.2.6).

- Using marginal distribution to determine the prior (JB 3.5)

- Parametric Approximations

  ⋆⋆ *How?* Assume that $\pi(\theta)$ is of a given functional form and then choose the density of this given form which most closely matches prior beliefs (through the *moments* or the *quantiles*).

  ⋆⋆ Most used (and misused)

  ⋆⋆ Very useful when a density of a standard functional form gives a good match to the prior information.

  ⋆⋆ Also useful when only vague prior information is available.

  ⋆⋆ Considerably different functional forms can often be chosen for the prior density (as seen in the example).

  ⋆⋆ *drawback:* The choice of the parameterized family is often based on ease in the mathematical treatment. The resulting posterior inference is affected by the choice.

- **Ex 3.2.5** Let $X_i \sim \text{Bin}(n_i, p_i)$ be the number of passing students in a freshman calculus course of n_i students. Over the previous years, the average of the p_i is 0.70, with variance 0.1. If we assume that the p_i's are all generated according to the same beta distribution, $\text{Be}(\alpha, \beta)$, then we choose the values of α and β which most closely matches the prior beliefs. That is, set

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \tau^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)},$$

and solve for α and β.

• **Example 3.2.6** Let $x \sim \mathrm{N}(\theta, 1)$. Assume that the prior median of θ is 0, the first quartile is -1, and the third quartile is $+1$. Use the quadratic loss function.

⋆⋆ Case 1: Assume $\theta \sim \mathrm{N}(\mu, \tau^2)$ and set $\mu = 0$ and $\tau^2 = 2.19$.

$\Rightarrow \delta_1^{\pi}(x) = x - \frac{x}{3.19}$

⋆⋆ Suppose $x = 4$ is observed, and have $\delta_1^{\pi}(x) = 2.75$.

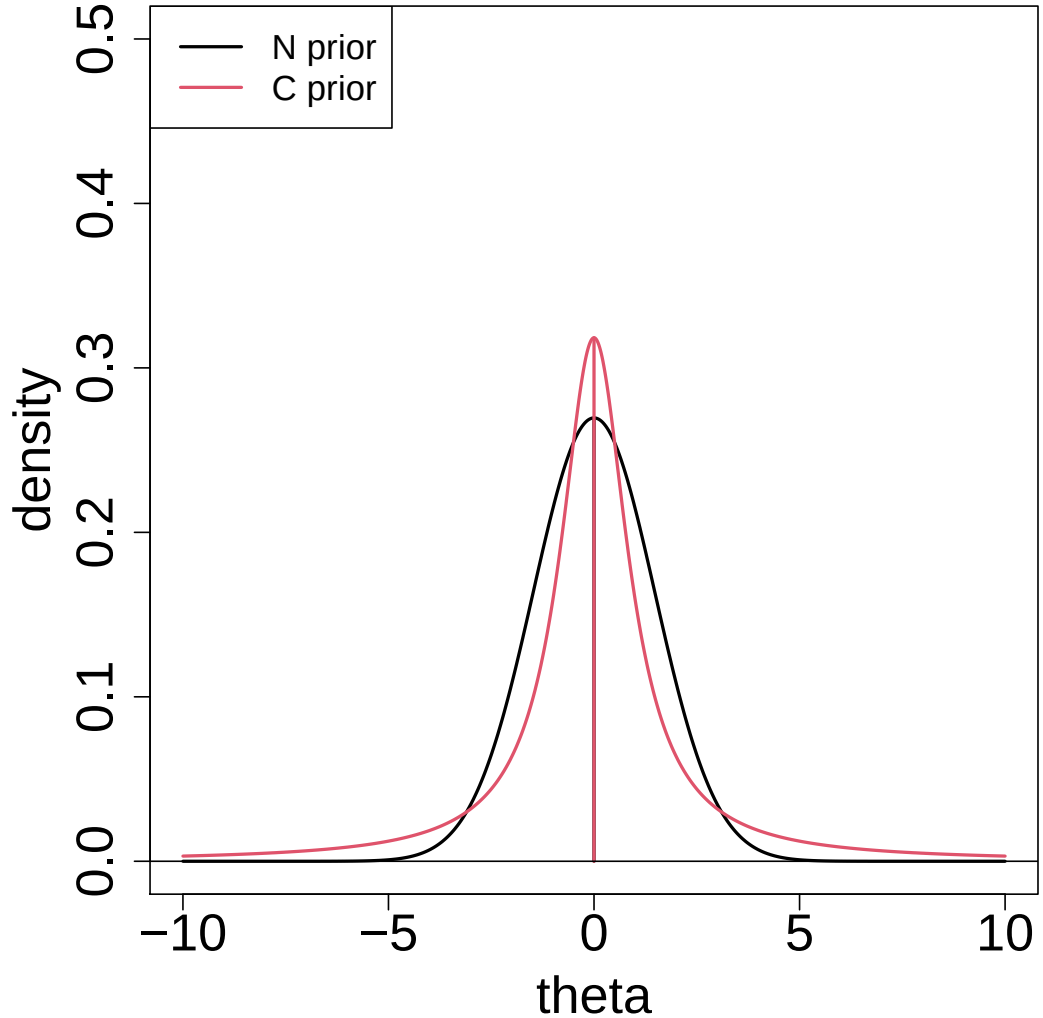
- **Example 3.2.6** (contd)

  ⋆⋆ Case 2: Assume θ has a Cauchy distribution and set $\theta \sim$ Cauchy$(0, 1)$.

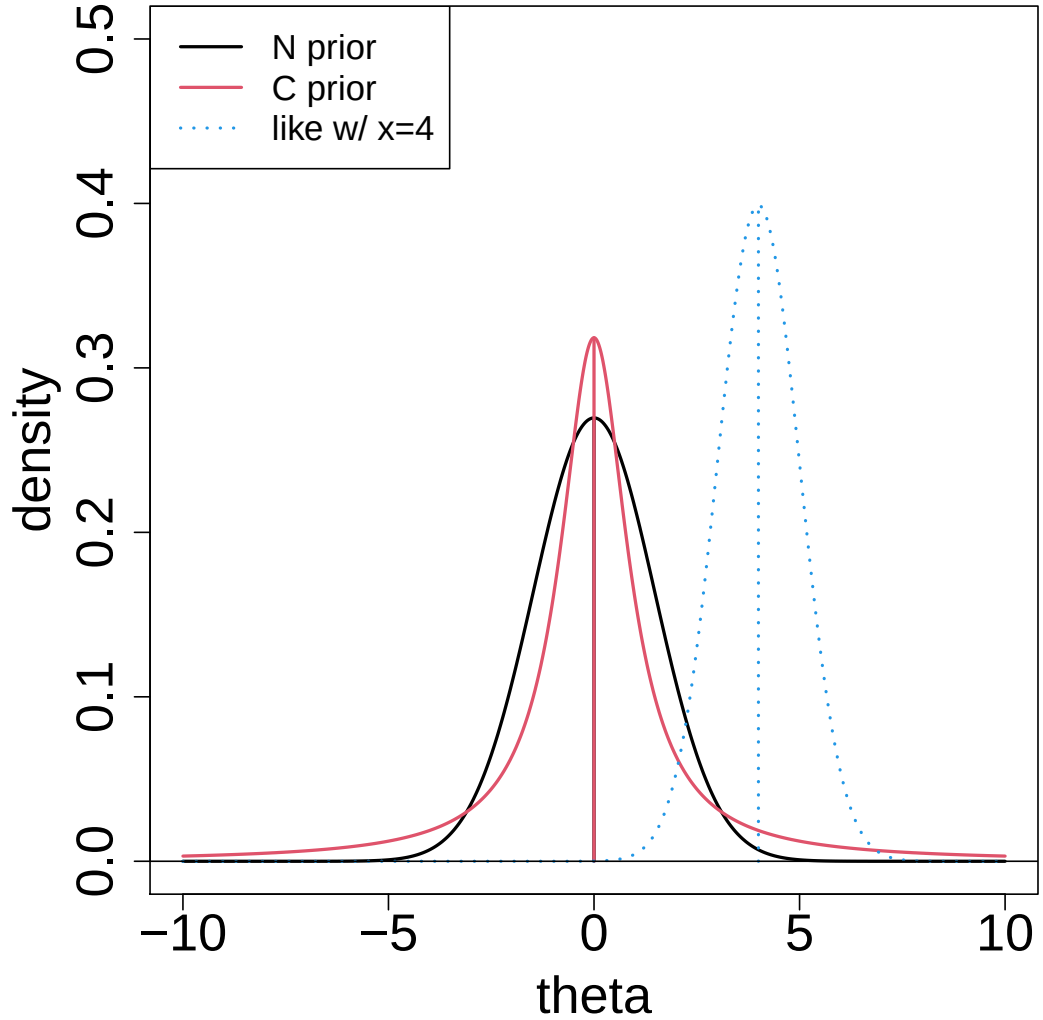
  $$\Rightarrow \delta_1^\pi(x) \approx x - \frac{x}{1+x^2} \text{ for } |x| \geq 4$$

  ⋆⋆ For $x = 4$, we have $\delta_2^\pi(x) = 3.76$.

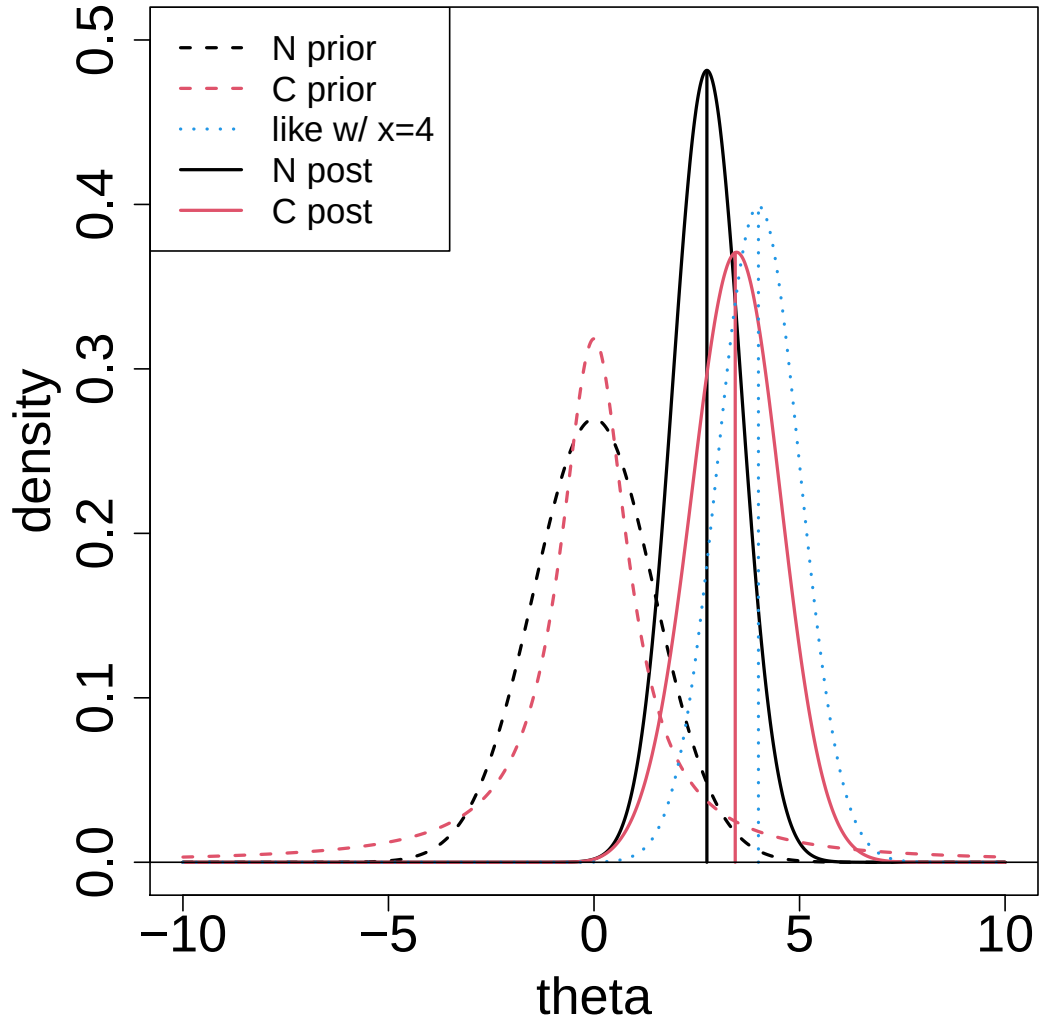
- **Example 3.2.6** (contd)



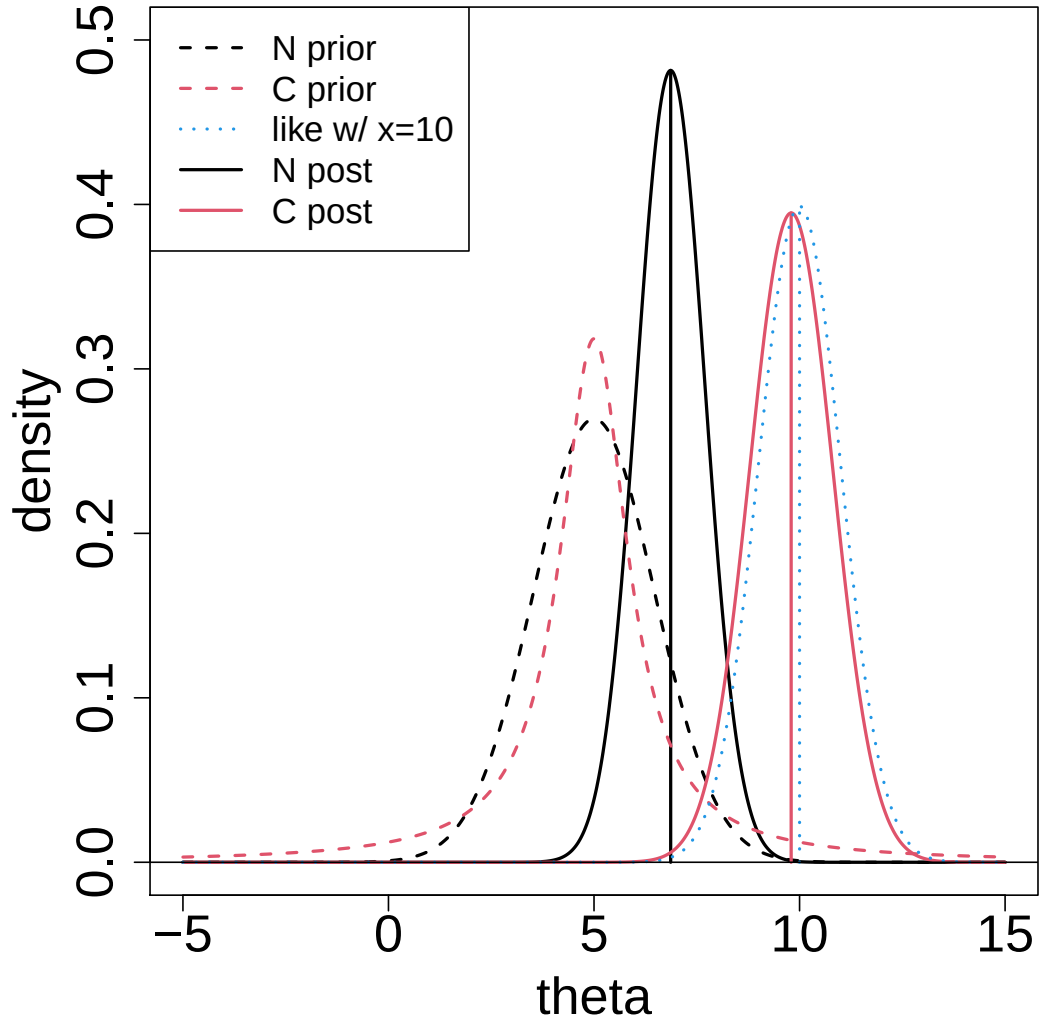
- **Example 3.2.6** (contd)



- **Example 3.2.6** (contd)



- **Example 3.2.6**(contd) If $x = 10$ is observed,



- **Example 3.2.6** (contd) Take-home message;

  ⋆⋆ The selection of the parameterized family greatly affects the inference about θ, especially due to the tail of the chosen prior where prior information is scarce.

  ⋆⋆ These posterior discrepancies call for some tests on the validity (or robustness) of the selected priors.

† *Empirical Bayes*

- **Use data** to estimate some features of the prior distribution
- Choose a *prior* distribution a posteriori! $\Rightarrow$ It does not belong to the Bayesian paradigm.
- Parametric empirical Bays:

  $\star\star$ Assume that the prior distribution of θ is in some parametric class with unknown parameters.

  $\star\star$ Use data to specify the unknown parameters.

JB in Section 4.5.2 Assume that $X_i \mid \theta_i \overset{indep}{\sim} \mathrm{N}(\theta_i, \sigma^2)$ with known σ^2, $i = 1, \ldots, p$ and θ_i are from a common prior distribution. Specify the prior distribution for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ using data.

Assume $\theta_i \overset{iid}{\sim} \mathrm{N}(\mu, \tau^2)$. The hyperparameters μ and τ^2 are unknown.

X_i is the test score of individual i, random about his/her true ability θ_i with known "reliability" σ^2. True abilities θ_i, $i = 1, \ldots, p$ are from an unknown normal population.

JB 4.5.2 (contd) How do we specify values for μ and τ^2?

$\star\star$ We use the data to estimate μ and τ^2.

$\star\star$ One way is to consider $m(\boldsymbol{x} \mid \pi)$ as a likelihood function for π as follows;

$\star\star$ Intuition $m(\boldsymbol{x} \mid \pi)$ is the density according to which X will actually occur.

If X_i is a test score of individual i which was normally distributed about "true ability" θ_i, and the true ability in the population varied according to a normal distribution with mean μ and τ^2, then $m(x_i)$ would be the actual distribution of observed test scores.

$**$ Recall we called $m(x \mid \pi)$ the predictive distribution for x.

JB 4.5.2 (contd)

⋆⋆ Seek to maximize $m(\boldsymbol{x} \mid \pi)$ over the hyperparameters μ and τ^2 by maximum likelihood.

Intuition If $m(x \mid \pi_1) > m(x \mid \pi_2)$, we can conclude that the data provides more support for π_1 than for π_2.

⋆⋆ Recall that

$$\begin{aligned} m(\boldsymbol{x} \mid \pi) &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left\{-\frac{(x_i-\mu)^2}{2(\sigma^2+\tau^2)}\right\} \\ &= \{2\pi(\sigma^2+\tau^2)\}^{-p/2} \exp\left\{-\frac{s^2}{2(\sigma^2+\tau^2)}\right\} \exp\left\{-\frac{p(\bar{x}-\mu)^2}{2(\sigma^2+\tau^2)}\right\}, \end{aligned}$$

where $\bar{x} = \sum_{i=1}^{p} x_i/p$ and $s^2 = \sum_{i=1}^{p}(x_i-\bar{x})^2$.

JB 4.5.2 (contd)

⋆⋆ We find the MLEs

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\tau}^2 = \max\left\{0, \frac{1}{p}s^2 - \sigma^2\right\}.$$

⋆⋆ We can pretend that the θ_i are iid from $\mathrm{N}(\hat{\mu}, \hat{\tau}^2)$ and proceed with a Bayesian analysis.

⋆⋆ **Or** we can use the moment method by matching the first two moments, $\hat{\mu} = \bar{x}$ and $\hat{\tau}^2 = \sum_{i=1}^{p}(x_i - \bar{x})^2/(p-1) - \sigma^2$.

- *Drawbacks of Empirical Bayes*

  $\star\star$ Seems paradoxical since it uses data twice.

  $\star\star$ It ignores the fact that μ and τ^2 were estimated; the errors undoubtedly introduced in the hyperparameter estimation will not be reflected in any of the conclusions.

  $\Rightarrow$ do not fully enjoy the optimality properties of the true Bayes estimators although asymptotically equivalent.

  $\star\star$ Too many choices are possible for the estimation techniques, e.g. MLE, MoM.

  $\Rightarrow$ arbitrariness in the selection of the prior.

† Hierarchical Bayes

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

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

Then we recover the usual Bayes model

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

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, $\theta_1, \ldots, \theta_n$.

- **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 $X_i \mid \theta_i \overset{indep}{\sim} \mathsf{N}(\theta_i, \sigma^2)$ with known σ^2, $i = 1, \ldots, p$ and $\theta_i \overset{iid}{\sim} \mathsf{N}(\mu, \tau^2)$, where hyperparameters$(\mu, \tau^2) \in \Theta_2 = \mathbb{R} \times \mathbb{R}^+$ are unknown.

- ⋆⋆ Sampling model: $X_i \mid \theta_i \overset{indep}{\sim} \mathsf{N}(\theta_i, \sigma^2)$.
- ⋆⋆ The first-level prior: $\theta_i \overset{iid}{\sim} \pi(\theta) = \mathsf{N}(\mu, \tau^2)$
- ⋆⋆ 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,

  ** "mean true ability" is near 100 with a "standard error" of $\pm$ 20

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

† **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 $x \mid \theta \sim \mathsf{N}(\theta, \sigma^2)$ and $\theta \sim \mathsf{N}(\mu, \tau^2)$.

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

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

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

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

⋆⋆ 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)

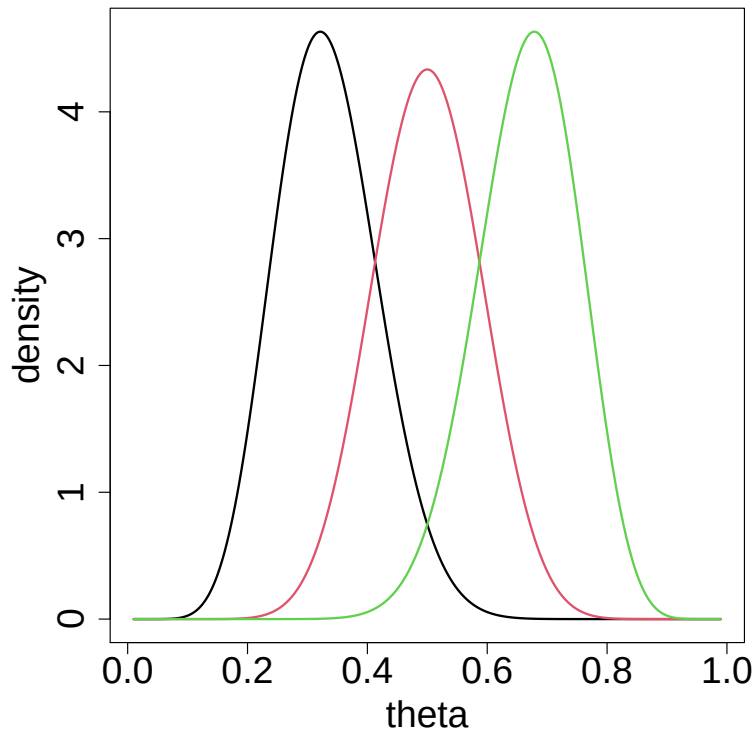
- Recall: One disadvantage of conjugate priors – limiting modeling capacity, but a big advantage – computational convenience.
- One possible extension to overcome the disadvantage while keeping the advantage is using a mixture model.
- Mixtures can be used as a basis to approximate any prior distribution.
- **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.

• **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.

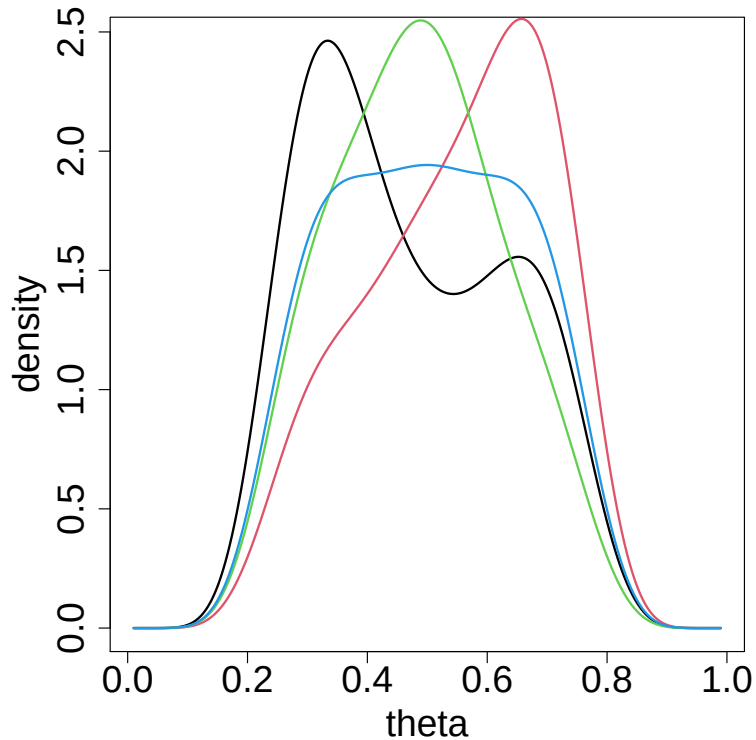
Let's consider three different priors.

- ⋆⋆ π_1: $\text{Be}(1, 1)$
- ⋆⋆ π_2: a mixture prior distribution, $1/2\text{Be}(10, 20) + 1/2\text{Be}(20, 10)$
- ⋆⋆ π_3: previous experiments with the same coin have already hinted at a bias toward *head* and they lead to the following alternative, $0.5\text{Be}(10, 20) + 0.2\text{Be}(15, 15) + 0.3\text{Be}(20, 10)$.

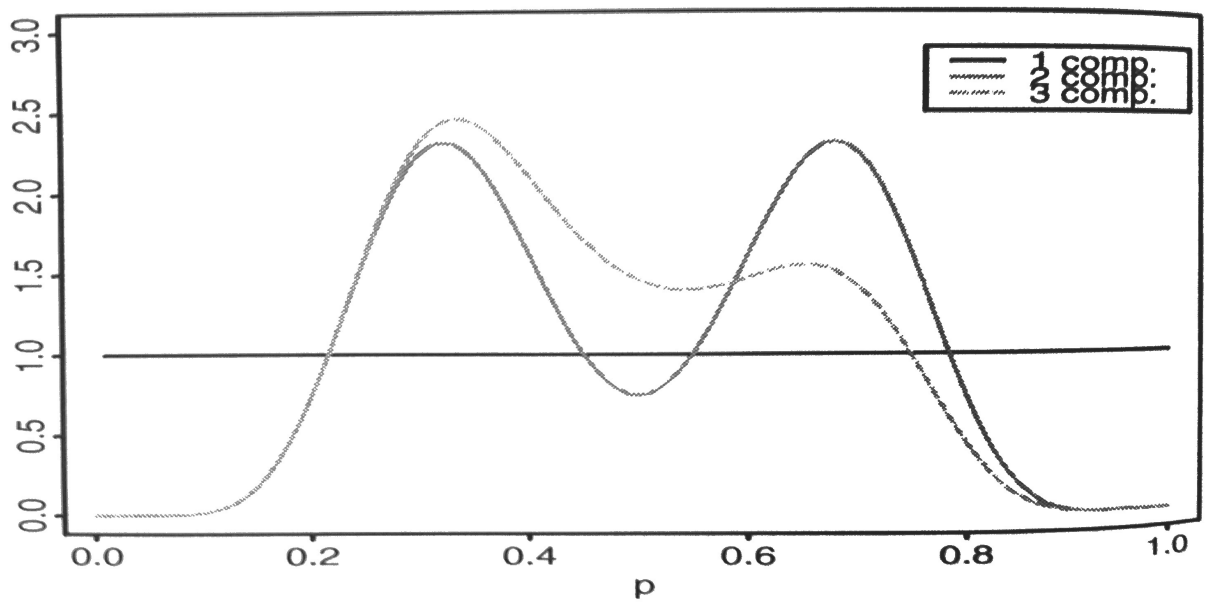
♣ Densities of Be(10, 20) (black), Be(15, 15) (red), and Be(20, 10) (green).



♣ 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

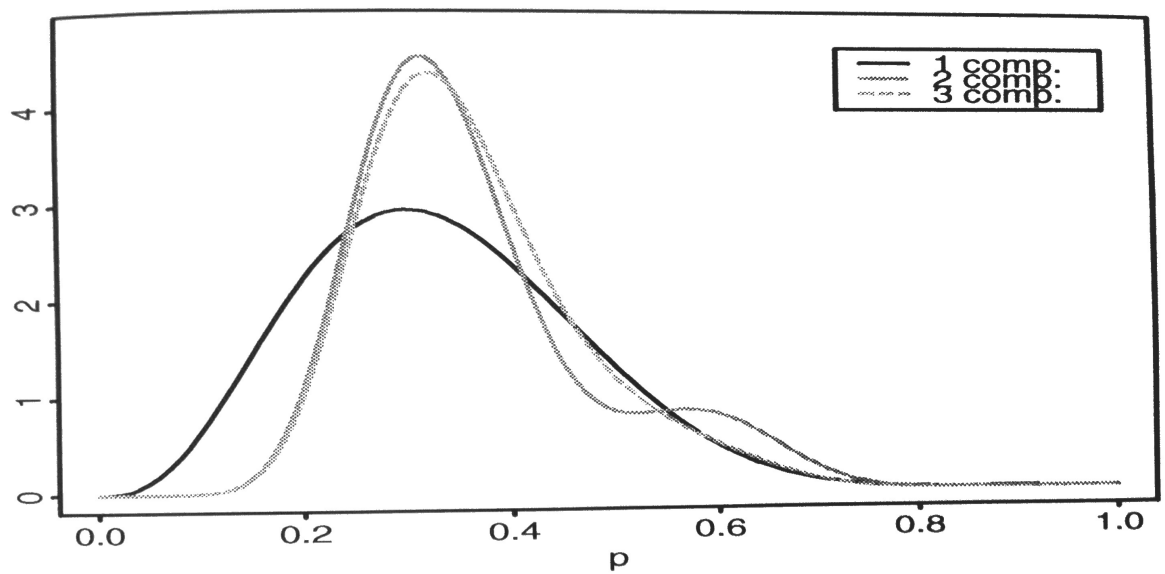


gure 3.4.1. *Three prior distributions for a spinning-coin experiment.*

- **Example 3.4.1** (contd): Suppose $x = 3$ for $n = 10$ is observed.

The corresponding posterior distributions are

- π_1: $\text{Be}(4, 8)$
- π_2: $0.84\text{Be}(13, 27) + 0.16\text{Be}(23, 17)$
- π_3: $0.77\text{Be}(13, 27) + 0.16\text{Be}(18, 22) + 0.07\text{Be}(23, 17)$.



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

- $\star\star$ π_1: $\text{Be}(15, 37)$
- $\star\star$ π_2: $0.997\text{Be}(24, 56) + 0.003\text{Be}(34, 46)$
- $\star\star$ π_3: $0.95\text{Be}(24, 56) + 0.047\text{Be}(29, 51) + 0.003\text{Be}(34, 46)$.

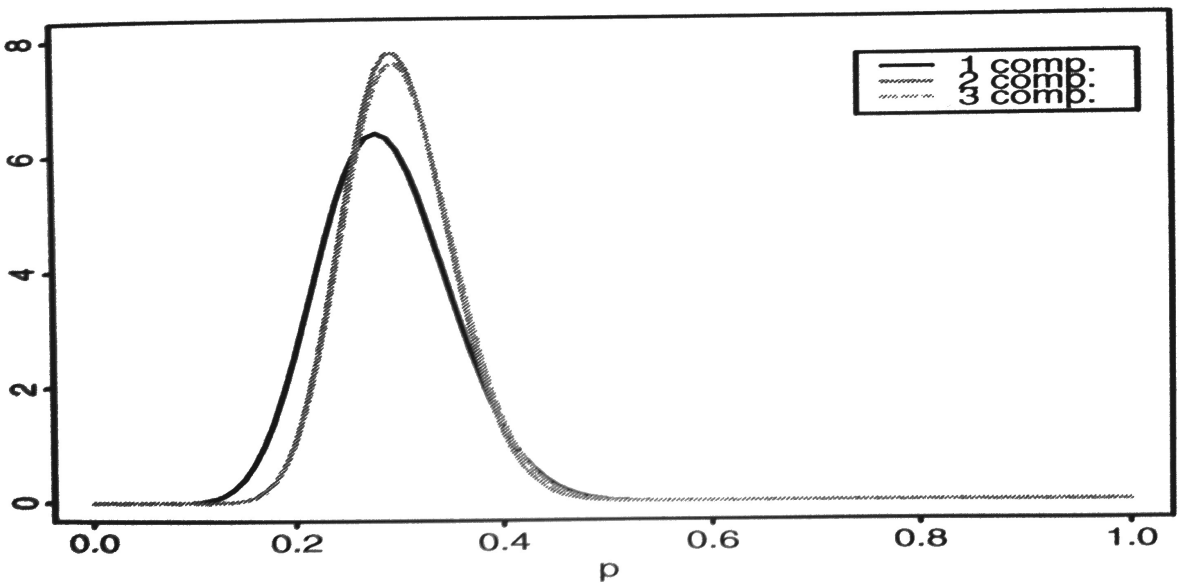


Figure 3.4.3. *Posterior distributions for 50 observations.*

- 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

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

- 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, $\rho(\pi(\theta \mid x), a)$ is finite and so we can find a Bayes action!

- Invariance under Reparameterization

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

  $\star\star$ Find the induced prior for η

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

  $\star\star$ 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.

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

  $\star\star$ 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.

  $\star\star$ 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

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

- The Fisher information $I(\theta)$ in the random variable X is defined as

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

  $\Rightarrow$ the expected slope of $\log f(x \mid \theta)$

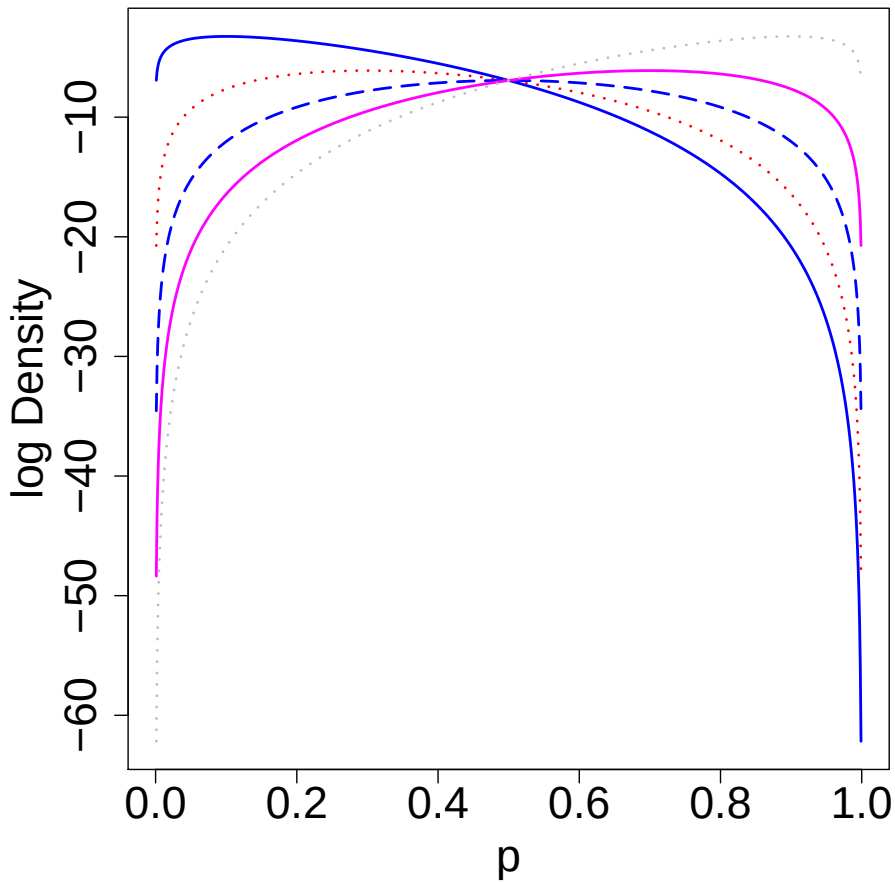
  $\Rightarrow$ measure the amount of information that a sample of data contains about unknown parameters.

- Under commonly satisfied conditions (true for exponential families),

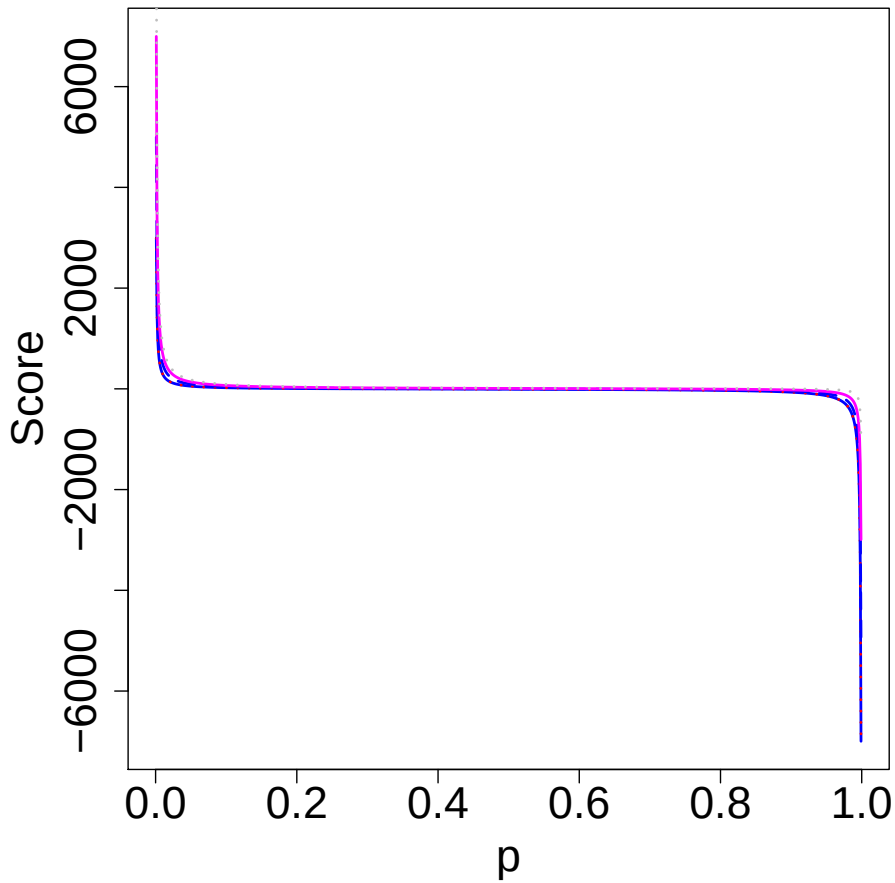
$$\mathrm{E}_\theta \left[\left(\frac{\partial \log f(x \mid \theta)}{\partial \theta} \right)^2 \right] = -\mathrm{E}_\theta \left[\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} \right].$$

- **Ex 3.5.4** Consider $x \sim \text{Bin}(n, p)$. Find the Fisher information, $I(p)$.

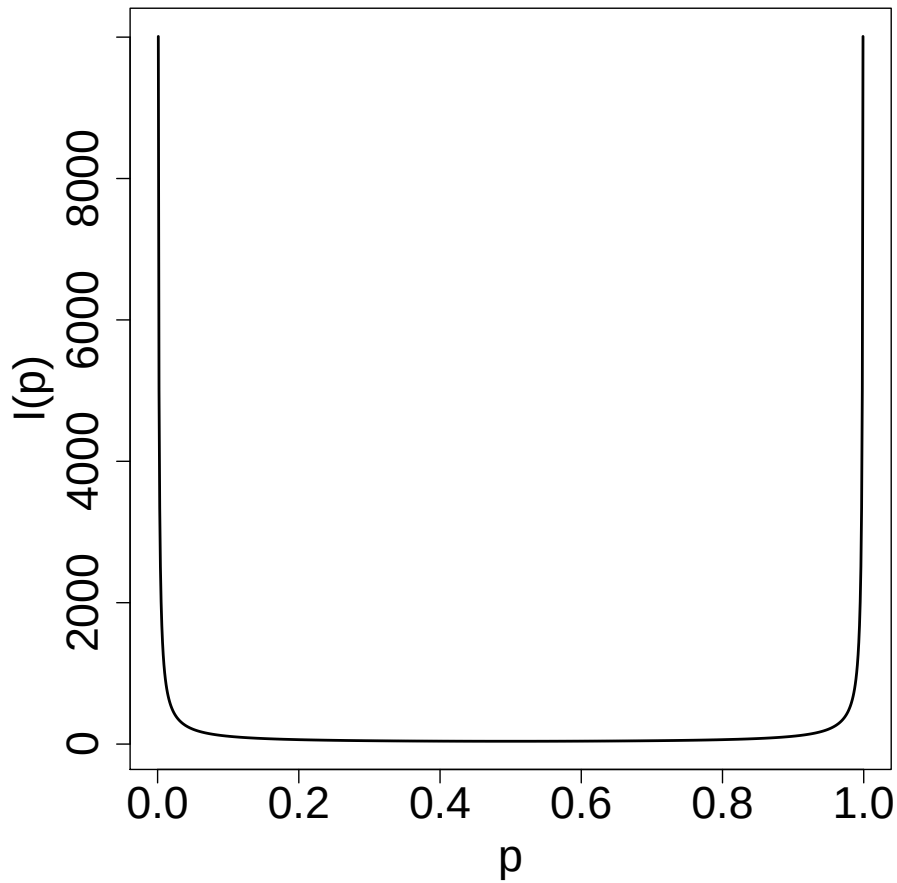
♣ **Ex 3.5.4**: $\log(f(x \mid p))$ with $n = 10$ and $x = 1, 3, \ldots, 9$.



♣ **Ex 3.5.4**: $d \log(f(x \mid p))/dp$ with $n = 10$ and $x = 1, 3, \ldots, 9$.



♣ Ex 3.5.4: $I(p)$



† The Jeffreys Prior

- Jeffreys Prior: noninformative priors in general settings based on Fisher information.

$$\pi^{\star}(\theta) \propto [I(\theta)]^{1/2}.$$

- Recall $I(\theta)$: an indicator of the amount of information brought by the sampling model (or the observation) about θ.

  i.e. large $I(\theta) \Rightarrow$ more sample info to discriminate between θ and $\theta + d\theta$.

  $\Rightarrow$ Assign more prior probability to the values that have large $I(\theta)$ so that the influence of the prior distribution is minimized.

- Jeffreys Prior:

$$\pi(\theta) \propto [I(\theta)]^{1/2}.$$

- defined up to a normalizing constant when $\pi^{\star}(\theta)$ is proper.
- Observe for any one-to-one transform $h(\theta)$

$$I(\theta) = I(h(\theta))(h'(\theta))^2$$

$\Rightarrow$ The invariant reparameterization requirement is satisfied.

- **Ex 3.5.4** Consider $x \sim \text{Bin}(n, p)$. Find the Jeffreys prior for this model.

† The Jeffreys Prior (contd)

- For multidimensional $\boldsymbol{\theta} \in \mathbb{R}^p$,

$$\pi(\boldsymbol{\theta}) = \{\det(\boldsymbol{I}(\boldsymbol{\theta}))\}^{1/2},$$

where $\boldsymbol{I}$: $p \times p$ Fisher information matrix.

- Under commonly satisfied conditions (true for exponential families),

$$I_{ij}(\boldsymbol{\theta}) = -\mathsf{E}_\theta \left[\frac{\partial^2}{\partial\, \theta_i \partial\, \theta_j} \log f(x \mid \boldsymbol{\theta}) \right].$$

- **Ex 3.5.6** Consider $x \sim \mathrm{N}(\mu, \sigma^2)$ with μ and σ unknown. Find the Jeffreys prior for this model.

$\star\star$ Jeffreys (1961) was mainly emphasizing the use of the Jeffreys priors in the one-dimensional case.

- **Any drawback?** The use of Jeffreys priors does not satisfy the Likelihood Principle!

- **Ex 3.5.7** Consider $n \mid p \sim \text{NB}(x, p)$ where n: total # of trials, x: # of successes (pre-determined), p: prob. of success. Find the Jeffreys prior for this model.

† Reference Priors: Read CR §3.5.4.

- For multidimensional $\theta = (\theta_1, \theta_2)$, distinguish between θ_1 (parameters of primary interest) and θ_2 (nuisance parameters)
  - ⋆⋆ Conditional on θ_1, define prior $\pi(\theta_2 \mid \theta_1)$ as the Jeffreys prior associated with $f(x \mid \theta)$ when θ_1 fixed.
  - ⋆⋆ Integrate out θ_2 according to $\pi(\theta_2 \mid \theta_1)$,

$$p(x \mid \theta_1) = \int_{\Theta_2} p(x \mid \theta_1, \theta_2)\pi(\theta_2 \mid \theta_1)d\theta_2.$$

  - ⋆⋆ Find the Jeffreys prior $\pi(\theta_1)$ based on $p(x \mid \theta_1)$.
  - ⋆⋆ Set $\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2 \mid \theta_2)$.

  Note: Choosing different nuisance parameters generates different priors.

- For one dimensional θ, the reference prior is identical to the Jeffreys prior

† Objectivity (JB §3.7)

- To most non-Bayesians, classical statistics is "objective", while Bayesian statistics is "subjective".
- No method of analysis is fundamentally "objective": e.g. the choice of a loss function is almost always subjective.

  ⋆⋆ The use of a prior distribution introduces another subjective feature into the analysis.
- When there is an overwhelming amount of data, virtually *any* analysis would yield the same conclusion.

  ⋆⋆ In such cases, choices of feature as $f(x \mid \theta)$ will have a serious bearing on the conclusion.

† Posterior Validation and Robustness (CR 3.6, JB 4.7.1 and 4.7.2)

- Do slight changes in the prior distribution cause significant changes in the decision and, if so, what should be done?

- JB Example 2, p111 We observe $X \sim \mathrm{N}(\theta, 1)$ and subjectively specify a prior median of 0 and prior quantiles of ± 1.

  $\star\star$ Either the $C(0,1)$ (π_C) or $\mathrm{N}(0, 2.19)(\pi_N)$ densities are thought to be reasonable matches to prior beliefs.

  $\star\star$ The difference between the two priors is mainly in the functional form (which is difficult to determine).

  $\star\star$ Suppose θ is estimated under the squared-error loss, so that the posterior means will be used. Does it matter whether we use π_C or π_N?

- JB Ex, p 111 & p 195(contd)

Table 4.7. Posterior Means.

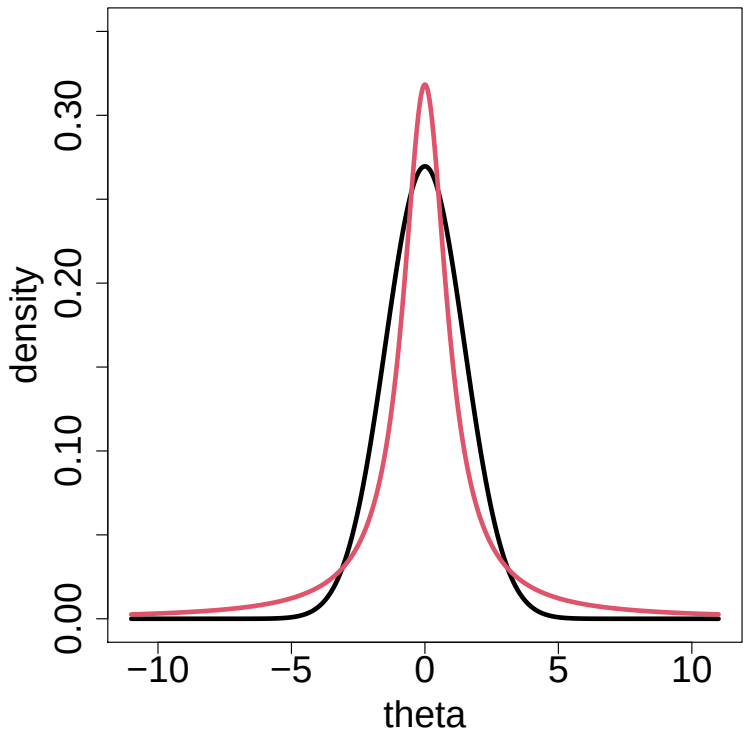
x	0	1	2	4.5	10
$\delta^C(x)$	0	0.52	1.27	4.09	9.80
$\delta^N(x)$	0	0.69	1.37	3.09	6.87

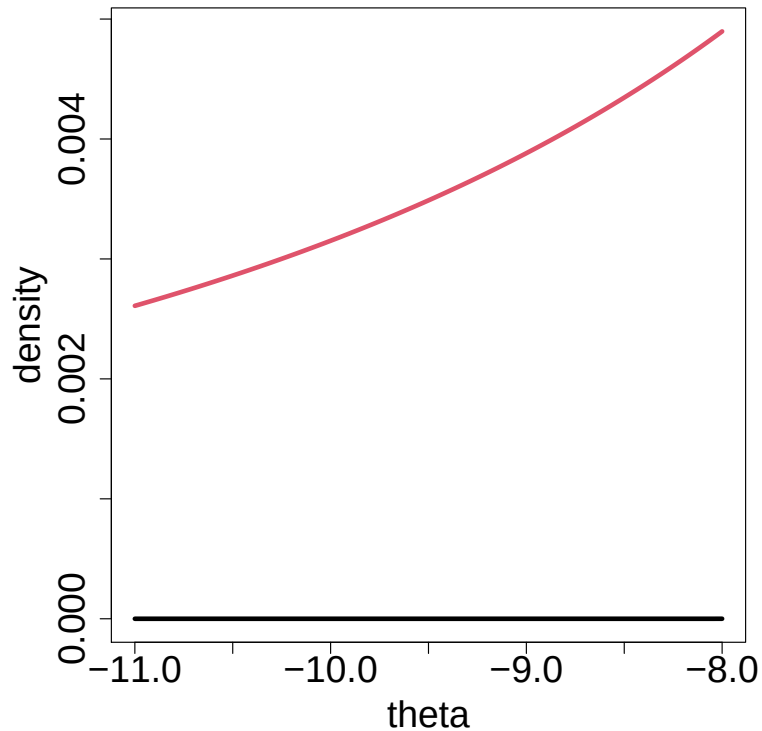
- ⋆⋆ For small x, $\delta^C(x)$ and $\delta^N(x)$ are quite close $\Rightarrow$ indicate some degree of robustness with respect to choice of the prior
- ⋆⋆ For moderate or large x, substantial difference between $\delta^C(x)$ and $\delta^N(x)$ $\Rightarrow$ not robust to reasonable variation in the prior

• JB Ex, p 111 & p 195(contd): Compare priors via $m(x)$ and eliminate from considerations priors which seem to be ruled out by data.

x	0	4.5	6.0	10
$m(x \mid \pi_N)$	0.22	0.0093	0.00079	3.5×10^{-8}
$m(x \mid \pi_C)$	0.21	0.018	0.0094	0.0032

• JB Ex, p 111 & p 195(contd): One of the main sources of nonrobustness in estimation (in specific for moderate or large x), will be seen to be the **degree of flatness of the prior tail**.





- Sensitivity of Bayesian analysis to possible misspecification of the prior distribution (assuming the likelihood is known).

  $\star\star$ Try different reasonable priors and evaluate how a change in the prior changes the inference about the parameter of interest.

- How to robustify our priors?

  $\star\star$ Robust prior distributions: Parameterized distributions as insensitive as possible to small variations in the prior information.

  e.g. $t-$distributions are preferable to normal priors in the normal case.

  - Robustify the conjugate priors by *hierarchical modeling*. e.g.;

$$\begin{aligned} \lambda &\sim \pi_2(\lambda), \\ \theta \mid \lambda &\sim \pi_1(\theta \mid \lambda), \\ x \mid \theta &\sim f(x \mid \theta), \\ \Rightarrow \quad \pi(\theta) &= \int \pi_1(\theta \mid \lambda)\pi_2(\lambda)d\lambda. \end{aligned}$$

    An additional level in the prior modeling increases the robustness of the prior distributions.