Lecture 20: Bayesian Inference

MATH 697

Omidali Jazi

November 8, 2018

McGill University

Goals for this Chapter

- The Prior and Posterior Distributions
- Inferences Based on the Posterior
- Bayesian Computations
- Choosing Priors

Introduction

- The previous chapters dealt with the analysis of uncertainty based on the model and the data alone.
- At the heart of the theory of inference is the concept of the statistical model $\{f(s|\theta) : \theta \in \Omega\}$ for the data $s \in S$ that describes the statistician's uncertainty about how the observed data were produced.
- Many statisticians prefer to develop statistical theory without the additional ingredients necessary for a full probability description of the unknowns.

Introduction

- The Bayesian approach to inference is when the parameter is playing the role of the unobserved response.
- Bayesian model arises naturally from the statistician assuming more ingredients for the model. It is up to the statistician to decide what ingredients can be justified and then use appropriate methods.
- A natural question to ask is: Where do these ingredients come from in an application? An easy answer is to say that they come from previous experience with the random system under investigation or perhaps with related systems.

The Prior Distributions

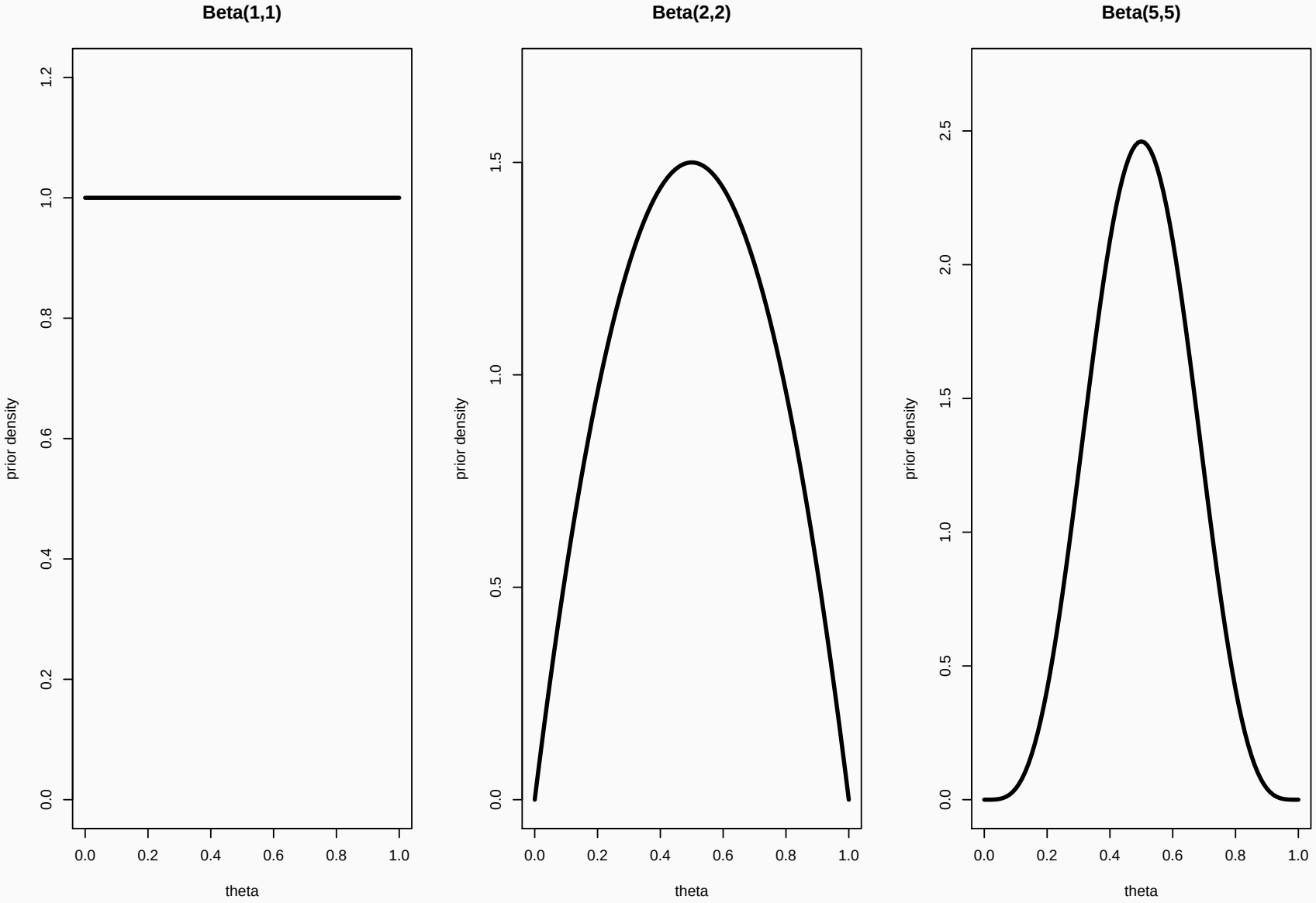
- The Bayesian model for inference contains the statistical model $\{f(s|\theta) : \theta \in \Omega\}$ for the data $s \in S$ and adds to this the prior probability measure Π for θ. The prior describes the statistician's beliefs about the true value of the parameter θ a priori, i.e., before observing the data.

- For example, if $\Omega = [0, 1]$ and θ equals the probability of getting a head on the toss of a coin, then the prior density π in the Figure (Left) indicates that the statistician knows nothing about the true value of θ, then using the uniform distribution on $[0, 1]$ might then be appropriate.

- Figure in the middle indicates that the statistician has some belief that the true value of θ is around 0.5. But this information is not very precise.

- Figure on the right indicates that the statistician has more precise information about the true value of θ.

Example: Prior Functions



The Prior Predictive Distribution

- We note that the ingredients of the Bayesian formulation for inference prescribe a marginal distribution for θ, namely, the prior Π, and a set of conditional distributions for the data s given θ, namely, $\{f(s|\theta) : \theta \in \Omega\}$.
- When the prior distribution is absolutely continuous, the marginal distribution for s is given by

$$m(s) = \int_{\Omega} \pi(\theta) f(s|\theta) d\theta$$

and is referred to as the prior predictive distribution of the data.

The Posterior Distribution

Definition 1

The posterior distribution of θ is the conditional distribution of θ given s. The posterior density, or posterior probability function (whichever is relevant), is given by

$$\pi(\theta|s) = \frac{\pi(\theta)f(s|\theta)}{m(s)}$$

Remark

- The prior predictive of the data s plays the role of the inverse normalizing constant for the posterior density. By this we mean that the posterior density of θ is proportional to $\pi(\theta)f(s|\theta)$, as a function of θ; to convert this into a proper density function, we need only divide by $m(s)$. In many examples, we do not need to compute the inverse normalizing constant.
- There are Monte Carlo methods that allow us to sample from $\pi(\theta|s)$ without knowing $m(s)$.

Example: Bernoulli Model

Suppose we observe a sample $y_1, ..., y_n$ from the $Bernoulli(p)$ distribution with $p \in [0, 1]$ is unknown. For the prior, we take π to be equal to a $Beta(\alpha, \beta)$ density.

$$\pi(p) = B(\alpha, \beta)^{-1} p^{\alpha-1}(1-p)^{\beta-1}$$

The likelihood function of the sample is

$$L(p|y_1, ..., y_n) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{n\bar{y}}(1-p)^{n(1-\bar{y})}$$

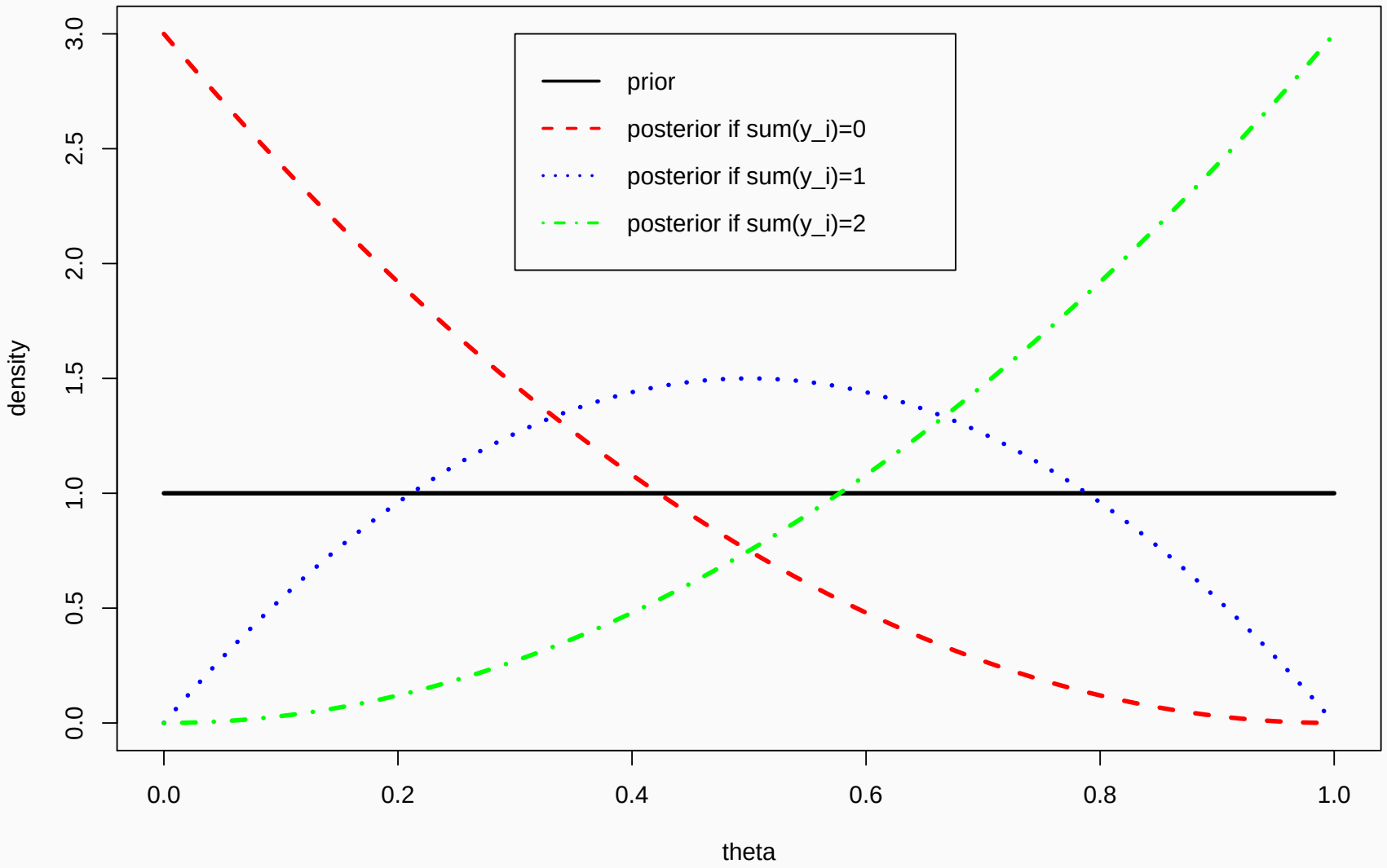
The posterior density function of p is then proportional to

$$\pi(p|y_1, ..., y_n) \propto p^{n\bar{y}+\alpha-1}(1-p)^{n(1-\bar{y})+\beta-1}$$

That is, the posterior distribution of p is

$$Beta(n\bar{y} + \alpha, n(1-\bar{y}) + \beta)$$

Prior and Posterior Curves (n=2, alpha=1,beta=1)



Example: Normal Model (σ known)

Suppose $y_1, ..., y_n$ is a random sample from $N(\mu, \sigma_0^2)$ distribution where $\mu \in \mathbb{R}$ is unknown and σ_0^2 is known. Suppose we take the prior distribution of μ to be $N(\mu_0, \tau_0^2)$ for some specified choice of μ_0 and τ_0^2. The likelihood function is then given by

$$L(\mu|y_1, ..., y_n) \propto e^{-\frac{n}{2\sigma_0^2}(\bar{y}-\mu)^2}$$

The posterior density of μ is then proportional to

$$\pi(\mu|y_1, ..., y_n) \propto e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2} e^{-\frac{n}{2\sigma_0^2}(\bar{y}-\mu)^2}$$

which can be shown that, the posterior distribution is

$$N\left(\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\tau_0^2}+\frac{n}{\sigma_0^2}\bar{y}\right), \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\right)$$

Remark

- Notice that the posterior mean is a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$ with weights

$$\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\frac{1}{\tau_0^2}=\frac{\sigma_0^2}{\sigma_0^2+n\tau_0^2} \qquad \text{and} \qquad \left(\frac{1}{\tau_0^2}+\frac{n}{\sigma_0^2}\right)^{-1}\frac{n}{\sigma_0^2}=\frac{n\tau_0^2}{n\tau_0^2+\sigma_0^2}$$

- In fact, if we define $k=\frac{n\tau_0^2}{n\tau_0^2+\sigma_0^2}$ then

$$\mu_*=(1-k)\mu_0+k\bar{y} \qquad \text{and} \qquad \sigma_*^2=k\frac{\sigma_0^2}{n}$$

- k is called the Credibility Factor. As $n\to\infty$, the credibility factor $k\to 1$ which means a closer posterior mean and variance to those of $\bar{y}$ (the MLE of μ).

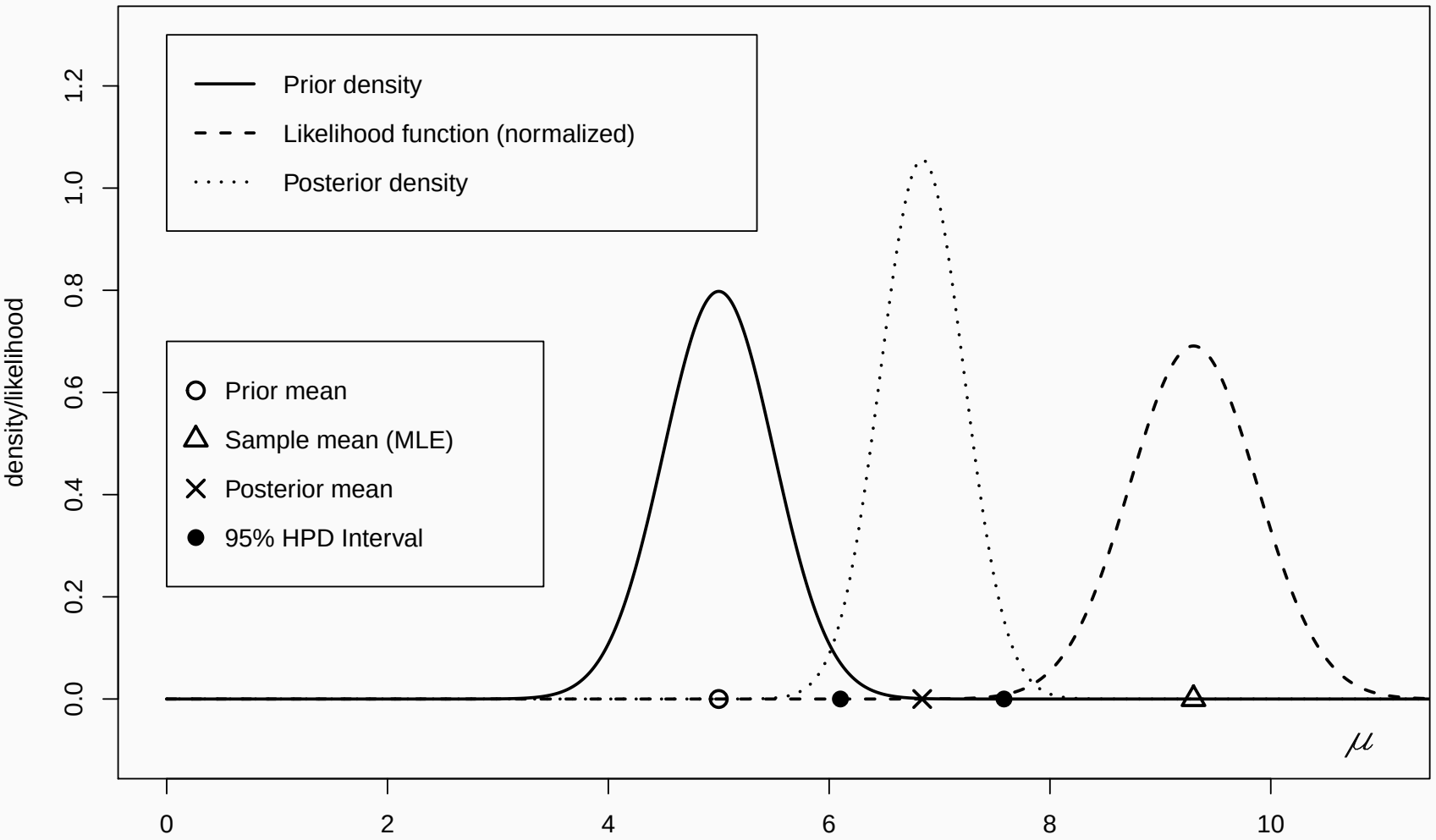
Remark

- Furthermore, the posterior variance is smaller than the variance of the sample mean. So if the information expressed by the prior is accurate, inferences about μ based on the posterior will be more accurate than those based on the sample mean alone. Note that the more diffuse the prior is (namely, the larger τ_0^2 is), the less influence the prior has.
- For example, when $n = 20$ and $\sigma_0^2 = 1$,$\tau_0^2 = 1$, then the ratio of the posterior variance to the sample mean variance (credibility factor) is $k = 20/21 \approx 0.95$. Thus, there has been a 5% improvement due to the use of prior information.
- The posterior is quite concentrated compared to the prior, so we have learned a lot from the data.

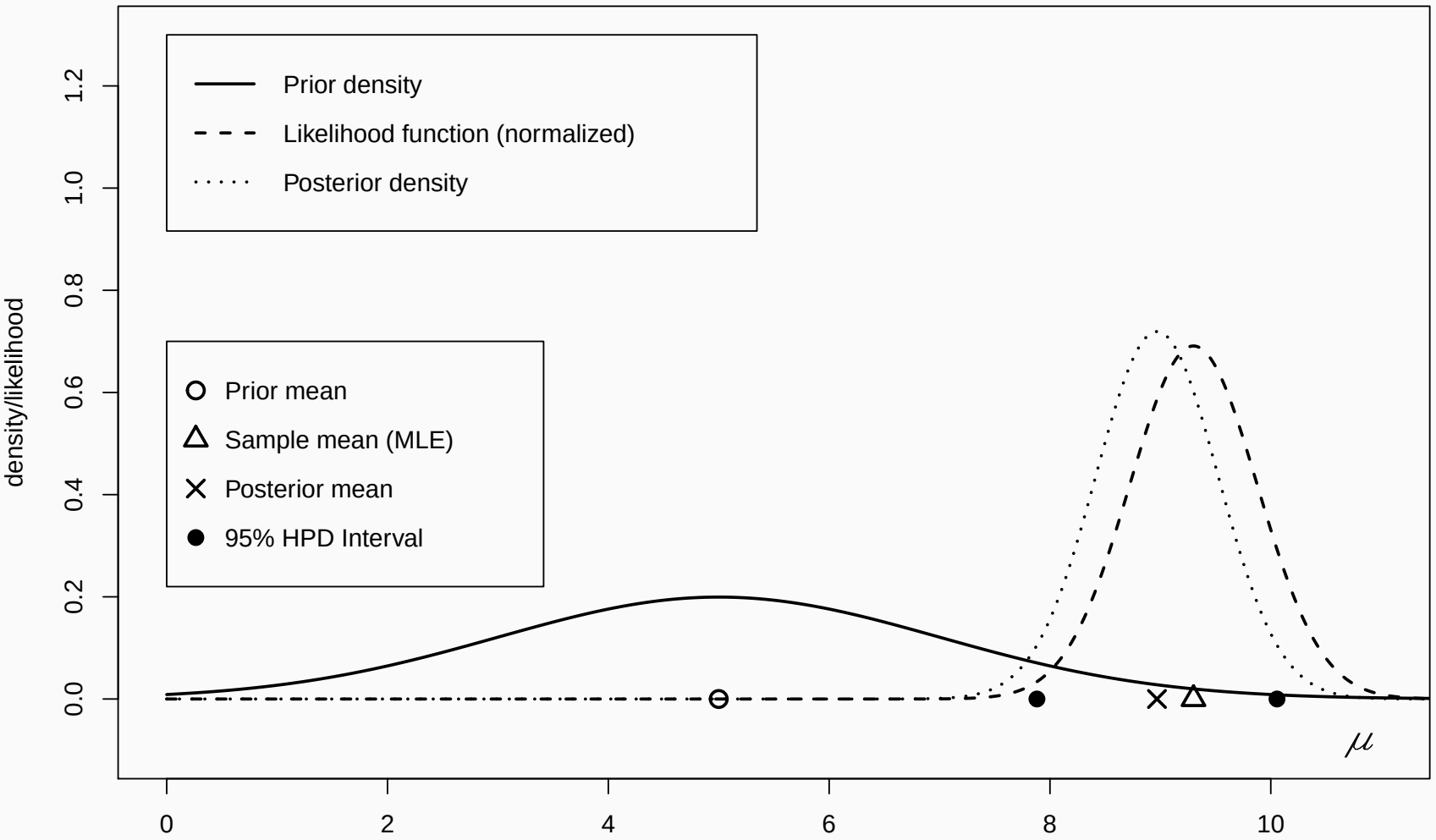
R code

```
X11(w=8,h=5); par(mfrow=c(1,1)); mu0=5; tau0=0.5; sig0=1
y = c(8.4, 10.1, 9.4); n = length(y);
k=n*tau0^2/(n*tau0^2+sig0^2); k # 0.4285714
ybar=mean(y); ybar # 9.3
mus = (1-k)*mu0 + k*ybar;  sigs2=k*sig0^2/n
c(mus,sigs2) # 6.8428571 0.1428571
muv=seq(0,15,0.01);prior = dnorm(muv,mu0,tau0);
post=dnorm(muv,mus,sqrt(sigs2))
like = dnorm(muv,ybar,sig0/sqrt(n))
HPD=mus+c(-1,1)*qnorm(0.975)*sqrt(sigs2)
HPD # 6.102060 7.583654
plot(c(0,11),c(-0.1,1.3),type="n",xlab="",ylab="density/likelihood")
lines(muv,prior,lty=1,lwd=2); lines(muv,like,lty=2,lwd=2)
lines(muv,post,lty=3,lwd=2)
points(c(mu0,ybar,mus),c(0,0,0),pch=c(1,2,4),cex=rep(1.5,3),lwd=2)
points(HPD,c(0,0),pch=rep(16,2),cex=rep(1.5,2))
legend(0,1.3,c("Prior density","Likelihood function (normalized)",
               "Posterior density"),lty=c(1,2,3),lwd=c(2,2,2))
legend(0,0.7,c("Prior mean","Sample mean (MLE)","Posterior mean",
"95% HPD Interval"), pch=c(1,2,4,16),pt.cex=rep(1.5,4),pt.lwd=rep(2,4))
text(10.8,-0.075,"m", vfont=c("serif symbol","italic"), cex=1.5)
```

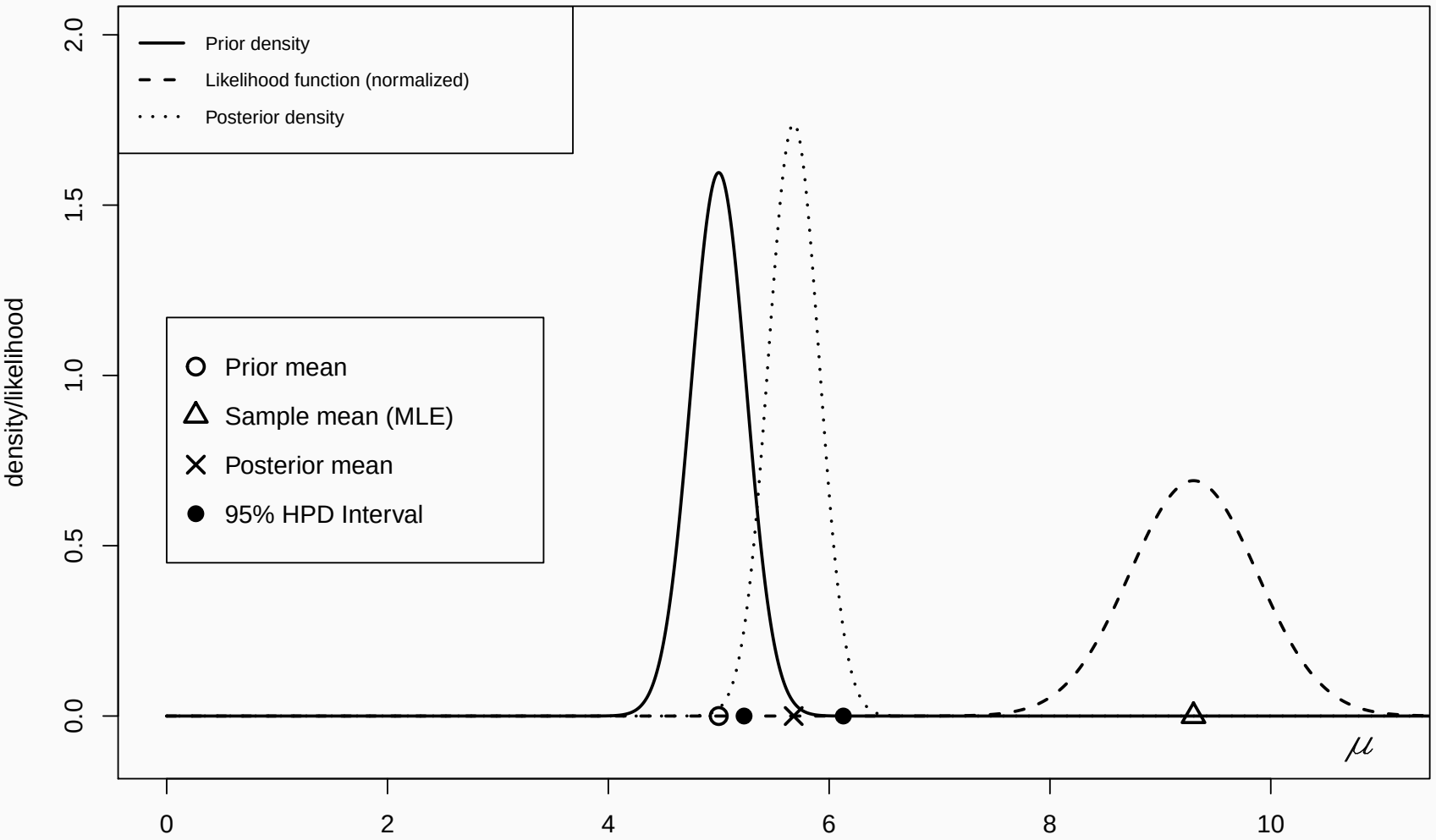
Plot : tau0=0.5; sig0=1; k=0.429



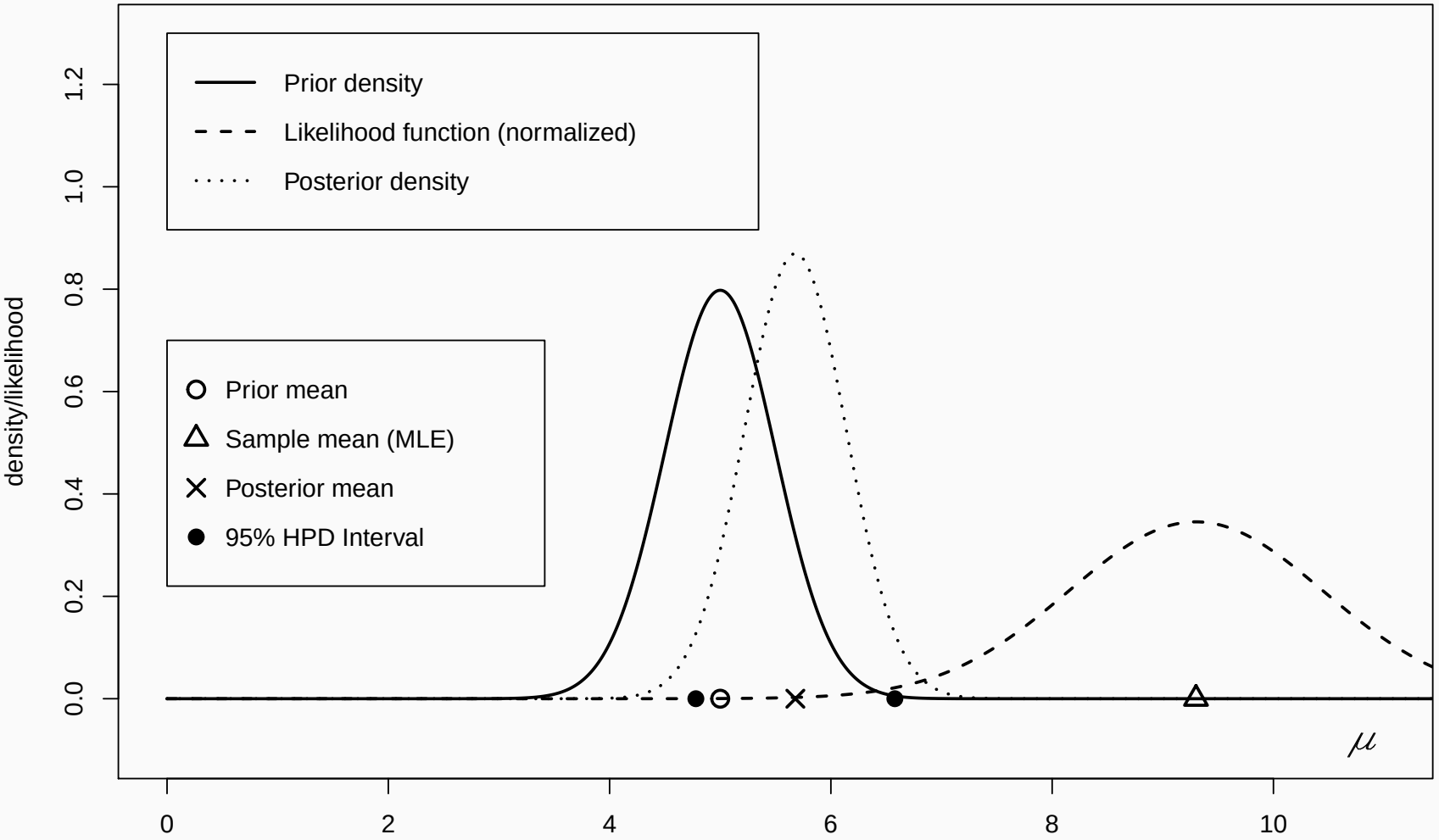
Plot : tau0=2; sig0=1; k=0.9223



Plot : tau0=0.25; sig0=1; k=0.158



Plot : tau0=0.5; sig0=2; k=0.158



Inferences Based on the Posterior

- We determined the posterior distribution of θ as a fundamental object of Bayesian inference. In essence, the principle of conditional probability asserts that the posterior distribution $\pi(\theta|s)$ contains all the relevant information in the sampling model $\{f(s|\theta) : \theta \in \Omega\}$, the prior π and the data s, about the unknown true value of θ.

- In particular, we must specify how to compute estimates, credible regions, and carry out hypothesis assessment.

- It turns out that there are often several plausible ways of proceeding, but they all have the common characteristic that they are based on the posterior.

Inferences Based on the Posterior

- In general, we are interested in specifying inferences about a real-valued characteristic of interest $\psi(\theta)$.
- One of the great advantages of the Bayesian approach is that inferences about ψ are determined in the same way as inferences about the full parameter θ, but with the marginal posterior distribution for ψ replacing the full posterior.
- When the posterior distribution of θ is discrete, the posterior probability function of ψ is given by

$$\omega(\psi_0|s) = \sum_{\{\theta:\psi(\theta)=\psi_0\}} \pi(\theta|s)$$

Bayesian Inference: Estimation

- Suppose we want to calculate an estimate of a characteristic of interest $\psi(\theta)$. We base this on the posterior distribution of this quantity.
- The most natural estimate is to obtain the posterior density (or probability function when relevant) of ψ and use the posterior mode $\hat{\psi}$, i.e., the point where the posterior probability or density function of ψ takes its maximum.
- To calculate the posterior mode, we need to maximize $\omega(\psi|s)$ as a function of ψ.

Bayesian Inference: Estimation

- An alternative estimate is commonly used and has a natural interpretation. This is given by the posterior mean, $E(\psi(\theta)|s)$, whenever this exists.
- When the posterior distribution of ψ is symmetrical about its mode, and the expectation exists, then the posterior expectation is the same as the posterior mode; otherwise, these estimates will be different.
- If we want the estimate to reflect where the central mass of probability lies, then in cases where $\omega(\cdot|s)$ is highly skewed, perhaps the mode is a better choice than the mean.

Example 1: Bernoulli Model

- Suppose we observe a sample $y_1, ..., y_n$ from the $Bernoulli(p)$ distribution with $p \in [0, 1]$ unknown and we place a $Beta(\alpha, \beta)$ prior on p. In the example, we determined the posterior distribution of p to be $Beta(n\bar{y} + \alpha, n(1 - \bar{y}) + \beta)$. Let us suppose that the characteristic of interest is $\psi(p) = p$.

- To determine the posterior mode, we need to maximize

$$\log p^{n\bar{y}+\alpha-1} p^{n(1-\bar{y})+\beta-1} = (n\bar{y} + \alpha - 1) \log p + (n(1 - \bar{y}) + \beta - 1) \log(1 - p)$$

  Setting the first derivative equal to 0 and solving gives the solution

$$\hat{p} = \frac{n\bar{y} + \alpha - 1}{n + \alpha + \beta - 2}$$

- Now, if $\alpha \geq 1$, $\beta \geq 1$, we see that the second derivative is always negative, and so $\hat{p}$ is the unique posterior mode.

Example 2: Normal Model (σ known)

- Suppose $y_1, ..., y_n$ is a random sample from $N(\mu, \sigma_0^2)$ distribution, where $\mu \in \mathbb{R}$ is unknown and σ_0^2 is known, and we take the prior distribution on μ to be $N(\mu, \tau_0^2)$.
- Let us suppose, that the characteristic of interest is $\psi(\mu) = \mu$.
- Because this distribution is symmetric about its mode, and the mean exists, the posterior mode and mean agree and equal

$$\mu_* = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1} \left(\frac{\mu_0}{\tau_0^2} + \frac{n}{\sigma_0^2}\bar{y}\right)$$

which is a weighted average of the prior mean and the sample mean and lies between these two values.

$$\mu_* = (1 - k)\mu_0 + k\bar{y}$$

Bayesian Inference: Credible Intervals

- A credible interval, for a real-valued parameter $\psi(\theta)$, is an interval $C(s) = [l(s), u(s)]$ that we believe will contain the true value of ψ. As with the sampling theory approach, we specify a probability γ and then find an interval $C(s)$ satisfying

$$\Pi(\psi(\theta) \in C(s)|s) = \Pi(\{\theta : l(s) \leq \psi(\theta) \leq u(s)\}|s) \geq \gamma \tag{1}$$

  We then refer to $C(s)$ as a γ-credible interval for ψ.

Bayesian Inference: Credible Intervals

- Naturally, we try to find a γ-credible interval $C(s)$ so that $\Pi(\psi(\theta) \in C(s)|s)$ is as close to γ as possible, and such that $C(s)$ is as short as possible. This leads to the consideration of highest posterior density (HPD) intervals, which are of the form

$$C(s) = \{\psi : \omega(\psi|s) \geq c\}$$

where $\omega(\cdot|s)$ is the marginal posterior density of ψ and where c is chosen as large as possible so that (1) is satisfied.

- Clearly, $C(s)$ contains the mode whenever $c \leq \max_\psi \omega(\psi|s)$. We can take the length of an HPD interval as a measure of the accuracy of the mode of $\omega(\cdot|s)$ as an estimator of $\psi(\theta)$. The length of a 0.95-credible interval for ψ will serve the same purpose as the margin of error does with confidence intervals.

Highest Posterior Density

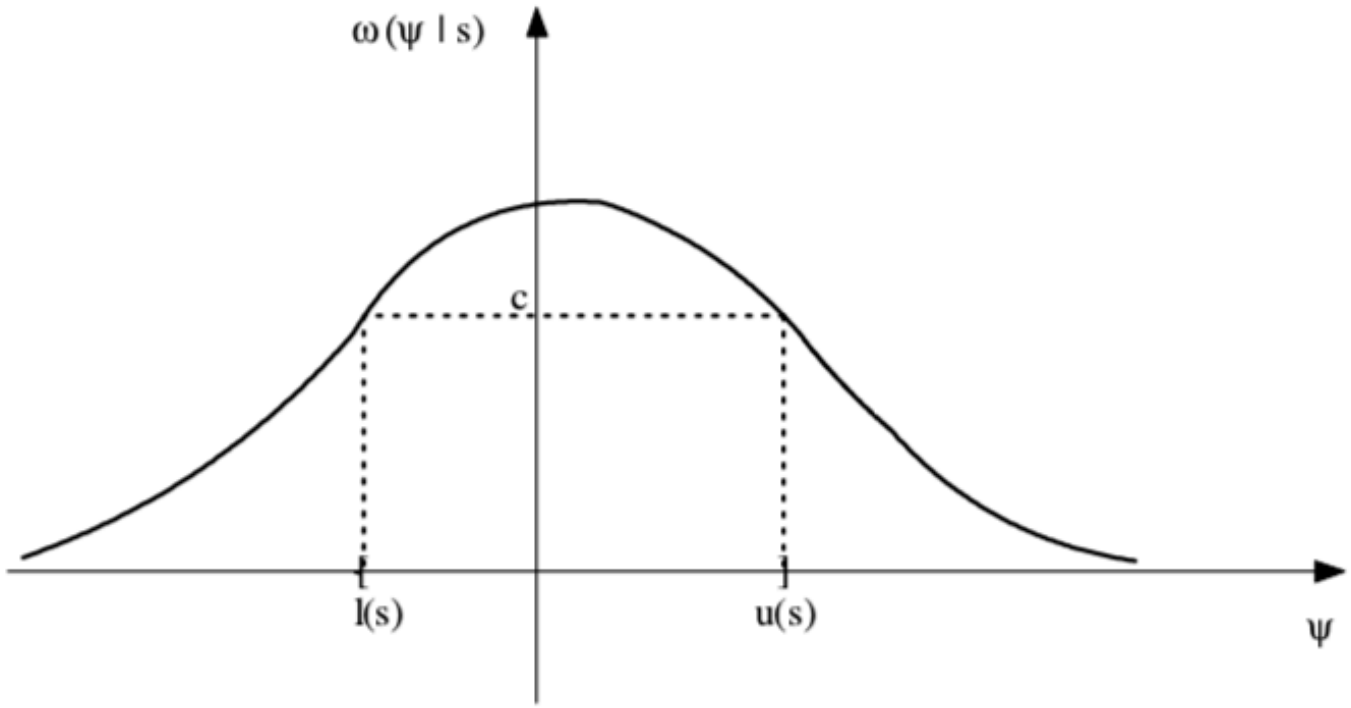


Figure 7.2.1: An HPD interval $C(s) = [l(s), u(s)] = \{\psi : \omega(\psi \mid s) \geq c\}$.

Example 1: Normal Model

- Since Normal distribution is symmetric about its mode (also mean) μ_*, a shortest γ-HPD interval is of the form

$$\mu_* \pm \sigma_* c$$

where c is such that

$$\begin{aligned}\gamma &= \Pi\left(\mu \in [\mu_* \pm \sigma_* c] \,|y_1, ..., y_n\right) \\ &= \Pi\left(-c \leq \frac{(\mu - \mu_*)}{\sigma_*} \leq c | y_1, ..., y_n\right)\end{aligned}$$

Since

$$\frac{(\mu - \mu_*)}{\sigma_*} | y_1, ..., y_n \sim N(0, 1)$$

we have $\gamma = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1$, where Φ is the CDF of the standard normal distribution.

- This immediately implies that $c = z_{(1+\gamma)/2}$ and the γ-HPD interval is given by

$$\mu_* \pm \sigma_* z_{(1+\gamma)/2}$$

Equivalently,

$$[(1-k)\mu_0 + k\bar{y}] \pm \sqrt{k\frac{\sigma_0^2}{n}} z_{(1+\gamma)/2}$$

Example 2: Normal Model (σ known)

- Note that as $\tau_0^2 \to \infty$, namely, as the prior becomes increasingly diffuse, this interval converges to the interval

$$\bar{y} \pm \frac{\sigma_0}{\sqrt{n}} z_{(1+\gamma)/2}$$

  which is also the γ-confidence interval derived in the previous chapter for this problem.

- Therefore, under a diffuse normal prior, the Bayesian and frequentist approaches agree.

Bayesian predictive inference

- Prediction problems arise when we have an unobserved response value t in a sample space T and observed response $s \in S$. Furthermore, we have the statistical model $\{P_\theta : \theta \in \Omega\}$ for s and the conditional statistical model $\{Q_\theta(\cdot|s) : \theta \in \Omega\}$ for t given s.

- We assume that both models have the same true value of $\theta \in \Omega$. The objective is to construct a prediction $\tilde{t}(s) \in T$ unobserved value t, based on the observed data s.

- The conditional density of (t, θ), given s, is

$$q(t, \theta|s) = \frac{q(t|s, \theta) f(s|\theta)\pi(\theta)}{\int_\Omega f(s|\theta)\pi(\theta)d\theta} = \frac{q(t|s, \theta) f(s|\theta)\pi(\theta)}{m(s)}$$

- Then the marginal posterior distribution of t, known as the posterior predictive of t, is

$$q(t|s) = \int_\Omega q(t, \theta|s)d\theta = \int_\Omega \frac{q(t|s, \theta) f(s|\theta)\pi(\theta)}{m(s)} d\theta = \int_\Omega q(t|s, \theta)\pi(\theta|s)d\theta$$

Bayesian predictive inference

- Notice that the posterior predictive of t is obtained by averaging the conditional density of t, given (θ, s), with respect to the posterior distribution of θ.
- Now that we have obtained the posterior predictive distribution of t, we can use it to select an estimate of the unobserved value.
- We could again choose the posterior mode $\hat{t}$ or the posterior expectation $E(t|x) = \int_T q(t|s)dt$ as our prediction, whichever is deemed most relevant.
- We can also construct a γ-prediction region $C(s)$ for a future value t from the model $\{q_\theta(\cdot|s) : \theta \in \Omega\}$.
- A γ-prediction region for t satisfies $Q(C(s)|s) \geq \gamma$, where $Q(\cdot|s)$ is the posterior predictive measure for t. One approach to constructing $C(s)$ is to apply the HPD concept to $q(t|s)$.

Example 1: Bernoulli Model

- Suppose we want to predict the next independent outcome Y_{n+1}, having observed a sample $y_1, ..., y_n$ from $Bernoulli(p)$ and $p \sim Beta(\alpha, \beta)$. Here, the future observation is independent of the observed data.

- The posterior predictive probability function of Y_{n+1} at t is then given by

$$q(t|y_1, ..., y_n) = \begin{cases} \frac{n\bar{y}+\alpha}{n+\alpha+\beta} & t = 1 \\ \frac{n(1-\bar{y})+\beta}{n+\alpha+\beta} & t = 0 \end{cases}$$

  which is the probability function of $Bernoulli((n\bar{y} + \alpha)/(n + \alpha + \beta)$.

- Using the posterior mode as the predictor, i.e., maximizing $q(t|y_1, ..., y_n)$ for t, leads to the prediction.

$$\hat{t} = \begin{cases} 1 & \frac{n\bar{y}+\alpha}{n+\alpha+\beta} > \frac{n(1-\bar{y})+\beta}{n+\alpha+\beta} \\ 0 & \text{Otherwise} \end{cases}$$

- The posterior expectation predictor is given by $E(t|y_1, ...y_n) = \frac{n\bar{y}+\alpha}{n+\alpha+\beta}$

- Note that the posterior mode takes a value in $\{0, 1\}$, and the future Y_{n+1} will be in this set, too. The posterior mean can be any value in $[0, 1]$.

Example 2: Normal Model (σ known)

- Suppose $y_1, \ldots, y_n$ is a sample from a $N(\mu, \sigma_0^2)$ distribution, where $\mu \in \mathbb{R}$ is unknown and σ_0^2 is known and we use the prior $N(\mu_0, \tau_0^2)$.
- Suppose we want to predict a future observation Y_{n+1} which is sampled from $N(\bar{y}, \sigma_*^2)$ where recall that $\sigma_*^2 = k\frac{\sigma_0^2}{n}$.
- Thus, the future observation is not independent of the observed data, but it is independent of the parameter. We can show that $N(\bar{y}, \sigma_*^2)$ is the posterior predictive distribution of Y_{n+1} at t. Therefore, we would predict t by $\bar{y}$, as this is both the posterior mode and mean.
- Suppose we want a γ-prediction interval for a future observation Y_{n+1} from $N(\bar{y}, \sigma_*^2)$. Since this is also the posterior predictive distribution of Y_{n+1} and is symmetric about $\bar{y}$, a γ-prediction interval for Y_{n+1}, derived via the HPD concept, is given by

$$\bar{y} \pm \sigma_* z_{(1+\gamma)/2}$$

Example 3: Predicting m future values - Normal model (σ known)

- Suppose that there is interest in m future random sample values $(Y_1, ..., Y_m)|(y, \mu)$ from $N(\mu, \sigma_0^2)$. Find the posterior predictive distribution of $\bar{Y}_m$
- Solution: The poserior distribution μ is given by $N(\mu_*, \sigma_*^2)$. Now $(\bar{Y}_m|y, \mu) \sim N(\mu, \sigma_0^2/m)$. Therefore

$$q(\bar{y}_m|y) = \int q(\bar{y}_m|y, \mu)\pi(\mu|y)d\mu \propto \int_{-\infty}^{\infty} e^{-\frac{(\bar{y}_m-\mu)^2}{2\sigma_0^2/m}} e^{-\frac{(\mu-\mu_*)^2}{2\sigma_*^2}} d\mu$$

- This is the integral of the exponent of a quadratic in both $\bar{y}_m$ and μ and so must equal the exponent of a quadratic in $\bar{y}_m$.

Example 3: Predicting m future values - Normal model (σ known)

- It follows that $(\bar{Y}_m|y) \sim N(\eta, \delta^2)$ where

$$\eta = E(\bar{Y}_m|y) = E(E(\bar{Y}_m|y,\mu)|y) = E(\mu|y) = \mu_*$$

and

$$\delta^2 = V(\bar{Y}_m|y) = E(V(\bar{Y}_m|y,\mu)|y) + V(E(\bar{Y}_m|y,\mu)|y) = \frac{\sigma_0^2}{m} + \sigma_*^2$$

- Thus generally we have that $(\bar{Y}_m|y) \sim N(\mu_*, \sigma_*^2 + \frac{\sigma_0^2}{m})$
- A special case is where there is no prior information regarding the normal mean μ. In this case, assuming it is appropriate to set $\tau_0 = \infty$ (i.e, $f(\mu) \propto 1$, $\mu \in \mathbb{R}$), we have that $k = 1$ and hence

$$(\bar{Y}_m|y) \sim N(\bar{y}_n, \frac{\sigma_0^2}{m} + \frac{\sigma_0^2}{n})$$

Bayesian Computations

Bayesian Computations

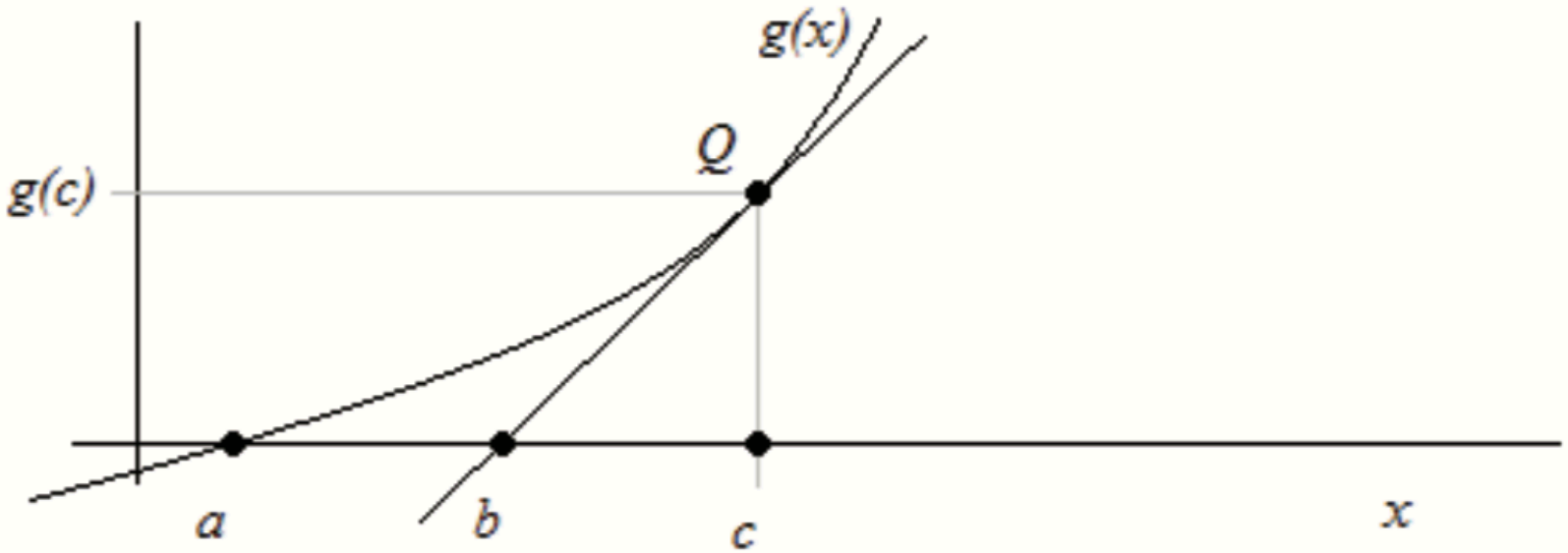
- In all the examples in this chapter so far, we have been able to work out the exact form of the posterior distributions and carry out a number of important computations using these
- It often occurs, however, that we cannot derive any convenient form for the posterior distribution. Furthermore, even when we can derive the posterior distribution, there computations might arise that cannot be carried out exactly.
- when we apply Bayesian inference in a practical example, we need to have available methods for approximating these quantities.

The Newton-Raphson Algorithm

- The Newton-Raphson (N-R) algorithm is a useful technique for solving equations of the form $g(x) = 0$.
- This algorithm involves choosing a suitable starting value $x^{(0)}$ and iteratively applying the following equation until convergence had been achieved to a desired degree of precision.

$$x^{(j+1)} = x^{(j)} - g'(x^{(j)})^{-1} g(x^{(j)})$$

- How does the N-Ralgorithm work?



Example 1: The N-R Algorithm

Suppose that the posterior CDF of a parameter is $\Pi(\theta|y) = \theta^6$. Find the posterior median of the posterior distribution.

```
NR <- function(th,J=5){
thvec <- th; for(j in 1:J){
num <- th^6-1/2;den <- 6*th^5
th <- th - num/den; thvec <- c(thvec,th) }; thvec }
# theta's posterior cdf minus 1/2 (numerator)
# theta's posterior pdf (denominator)
options(digits=4)
NR(th=1,J=6)
```

```
## [1] 1.0000 0.9167 0.8926 0.8909 0.8909 0.8909 0.8909
```

```
NR(th=0.8,J=6)
```

```
## [1] 0.8000 0.9210 0.8933 0.8909 0.8909 0.8909 0.8909
```

```
0.8909-(0.8909^6-0.5)/(6*0.8909^5) # 0.8909
```

```
## [1] 0.8909
```

Example 2 : The N-R Algorithm

Use the N-R algorithm to solve the equation $t^2 = e^t$

```
options(digits=6); t=0; tv=t; for(j in 1:7){ t=t-(t^2-exp(t))/(2*t-exp(t))
tv=c(tv,t) }; tv; t^2-exp(t)
```

```
## [1]  0.000000 -1.000000 -0.733044 -0.703808 -0.703467 -0.703467 -0.703467
## [8] -0.703467
```

```
## [1] 0
```

```
(-0.703467)^2-exp(-0.703467) # -8.03508e-07
```

```
## [1] -8.03508e-07
```

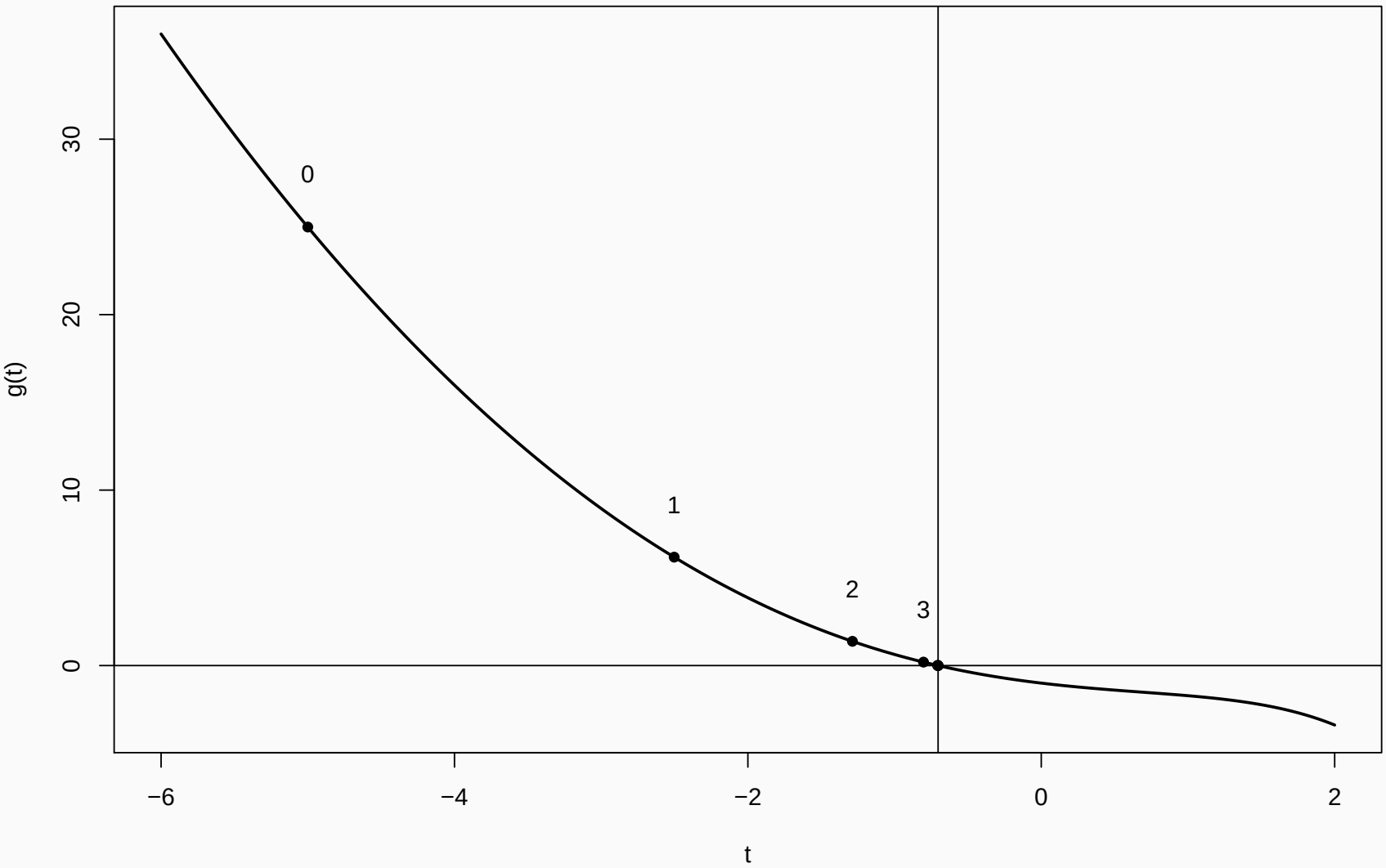
```
t=1; tv=t; for(j in 1:7){ t=t-(t^2-exp(t))/(2*t-exp(t)); tv=c(tv,t) }; tv
```

```
## [1]  1.000000 -1.392211 -0.835088 -0.709834 -0.703483 -0.703467 -0.703467
## [8] -0.703467
```

```
t=-5; tv=t; for(j in 1:7){ t=t-(t^2-exp(t))/(2*t-exp(t)); tv=c(tv,t) }; tv
```

```
## [1] -5.000000 -2.502357 -1.287421 -0.802834 -0.707162 -0.703473 -0.703467
## [8] -0.703467
```

Example 2 : The N-R Algorithm



Example 3: The N-R Algorithm

- Consider the Bayesian model: $(y|p) \sim Binom(n = 3, p)$ and $p \sim U(0, 1)$ and suppose the observed value of y is 2. Find the posterior median of p.

- The posterior distribution of p is given by $p|y = 2 \sim Beta(3, 2)$ with the density function $\pi(p|y) = 12p^2(1 - p)$ for $0 < p < 1$. So, the posterior CDF is

$$\Pi(p|y) = \int_0^p 12r^2(1 - r)dr = 4p^3 - 3p^4, \qquad 0 < p < 1$$

- To find the posterior median of p we need to solve $\Pi(p|y) = 0.5$, it requires solving $4p^3 - 3p^4 = 0.5$; that is $g(p) = 0$ where

$$g(p) = \Pi(p|y) - 0.5 = 4p^3 - 3p^4 - 0.5$$

- Thus, the N-R algorithm is defined by iterating

$$p_{j+1} = p_j - \frac{g(p_j)}{g'(p_j)} = p_j - \frac{4p_j^3 - 3p_j^4 - 0.5}{12p_j^2 - 12p_j^3}$$

Example 3 : The N-R Algorithm

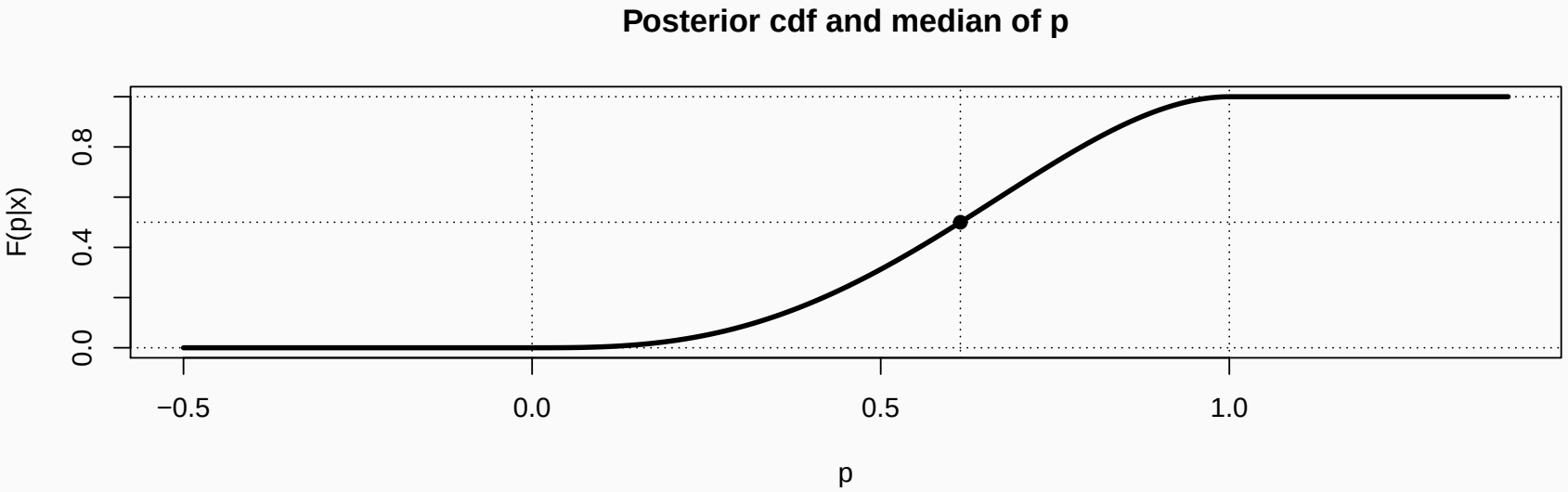
```
##  Intial values= {2/3, 0.5, 0.9, 0.1, 0.614272}; Repeat= {7,20,1}

## [1] 0.667 0.615 0.614 0.614 0.614 0.614 0.614 0.614

## [1] 0.500 0.625 0.614 0.614 0.614 0.614 0.614 0.614

## [1] 0.900 0.439 0.649 0.615 0.614 0.614 0.614 0.614

## [1] 0.100 4.695 3.627 2.834 2.251 1.832 1.543 1.362

##  [1] 0.100 4.695 3.627 2.834 2.251 1.832 1.543 1.362 1.273 1.249 1.247
## [12] 1.247 1.247 1.247 1.247 1.247 1.247 1.247 1.247 1.247 1.247 1.247

## [1] 0.499999
```

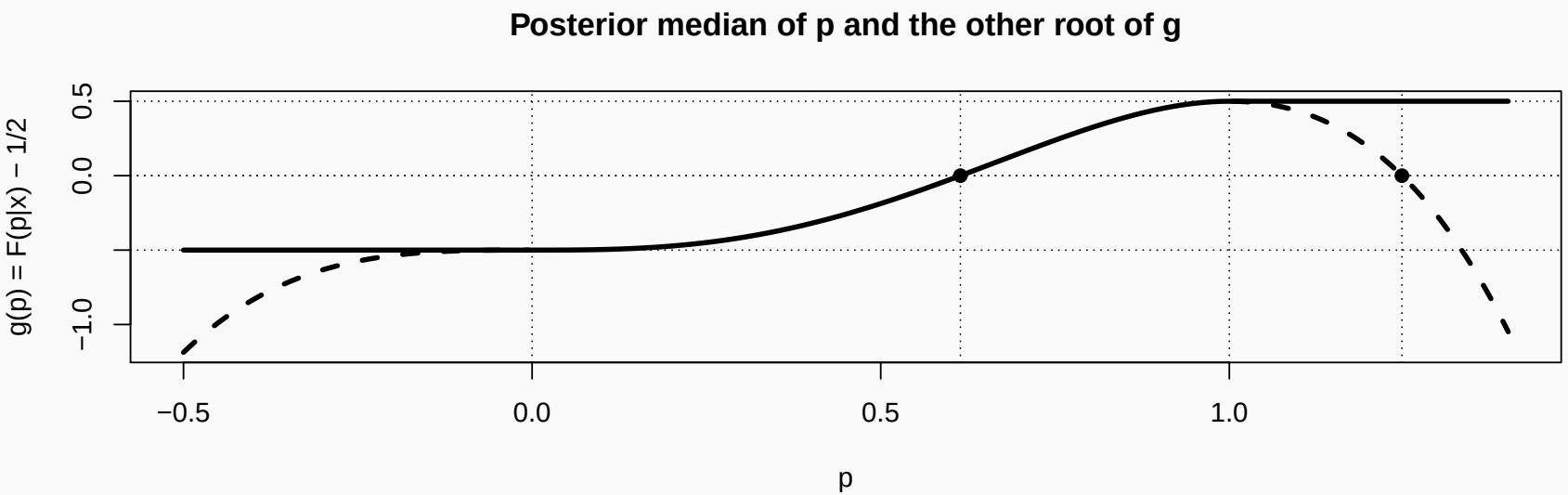
Remarks

- What is a good starting value here? Try the MLE, $p_0 = 2/3$.
- Starting at other values $(0.5, 0.9, 0.1)$, we obtain different sequences.
- The last sequence does not seem to have converged even run this for a bit longer ($n = 20$).
- Thus if we start at 0.1, the algorithm converges to an impossible value of p, namely 1.24748.
- It appears that the required posterior median is 0.61427. As a check, we may calculate

$$\Pi(0.61427|y) = 4(0.61427)^3 - 3(0.61427)^4 = 0.49999 \approx 0.5$$

Example 3 : The N-R Algorithm





The multivariate N-R algorithm

- The N-Ralgorithm can also be used to solve several equations simultaneously, say

$$g_k(x_1, ..., x_K) = 0 \qquad k = 1, ..., K$$

- Let $\mathbf{x} = (x_1, ..., x_K)^T$, $g(\mathbf{x}) = (g_1(\mathbf{x}), ..., g_K(\mathbf{x}))^T$, and $\mathbf{0} = (0, ..., 0)^T$.
- The system of K equations may be expressed as $g(\mathbf{x}) = \mathbf{0}$ and the N-R algorithm involves iterating according to

$$\mathbf{x}^{(j+1)} = \mathbf{x}^{(j)} - g'(\mathbf{x}^{(j)})^{-1} g(\mathbf{x}^{(j)})$$

where $\mathbf{x}^{(j)} = (x_1^{(j)}, ..., x_K^{(j)})^T$, $g(\mathbf{x}^{(j)}) = (g_1(\mathbf{x}^{(j)}), ..., g_K(\mathbf{x}^{(j)}))^T$ and $g'(\mathbf{x}^{(j)})$ is

$$\begin{pmatrix} \partial g_1(\mathbf{x})/\partial x_1 & ... & \partial g_1(\mathbf{x})/\partial x_K \\ \vdots & \ddots & \vdots \\ \partial g_K(\mathbf{x})/\partial x_1 & \cdots & \partial g_K(\mathbf{x})/\partial x_K \end{pmatrix}$$

evaluated at $\mathbf{x}^{(j)}$.

Example : Finding a HPD via the multivariate N-R algorithm

- Consider the Bayesian model: $y|\lambda \sim Poisson(\lambda)$ and $\pi(\lambda) = 1$, $\lambda > 0$ and suppose we observe $y = 1$. Find the 80% HPD for λ.
- First, since $y = 1$

$$\pi(\lambda|y) \propto \pi(\lambda) f(y|\lambda) = e^{-\lambda} \frac{\lambda^y}{y!} = e^{-\lambda} \lambda$$

  Thus $(\lambda|y) \sim Gamma(2, 1)$ with $\pi(\lambda|y) = \lambda e^{-\lambda}$, $\lambda > 0$

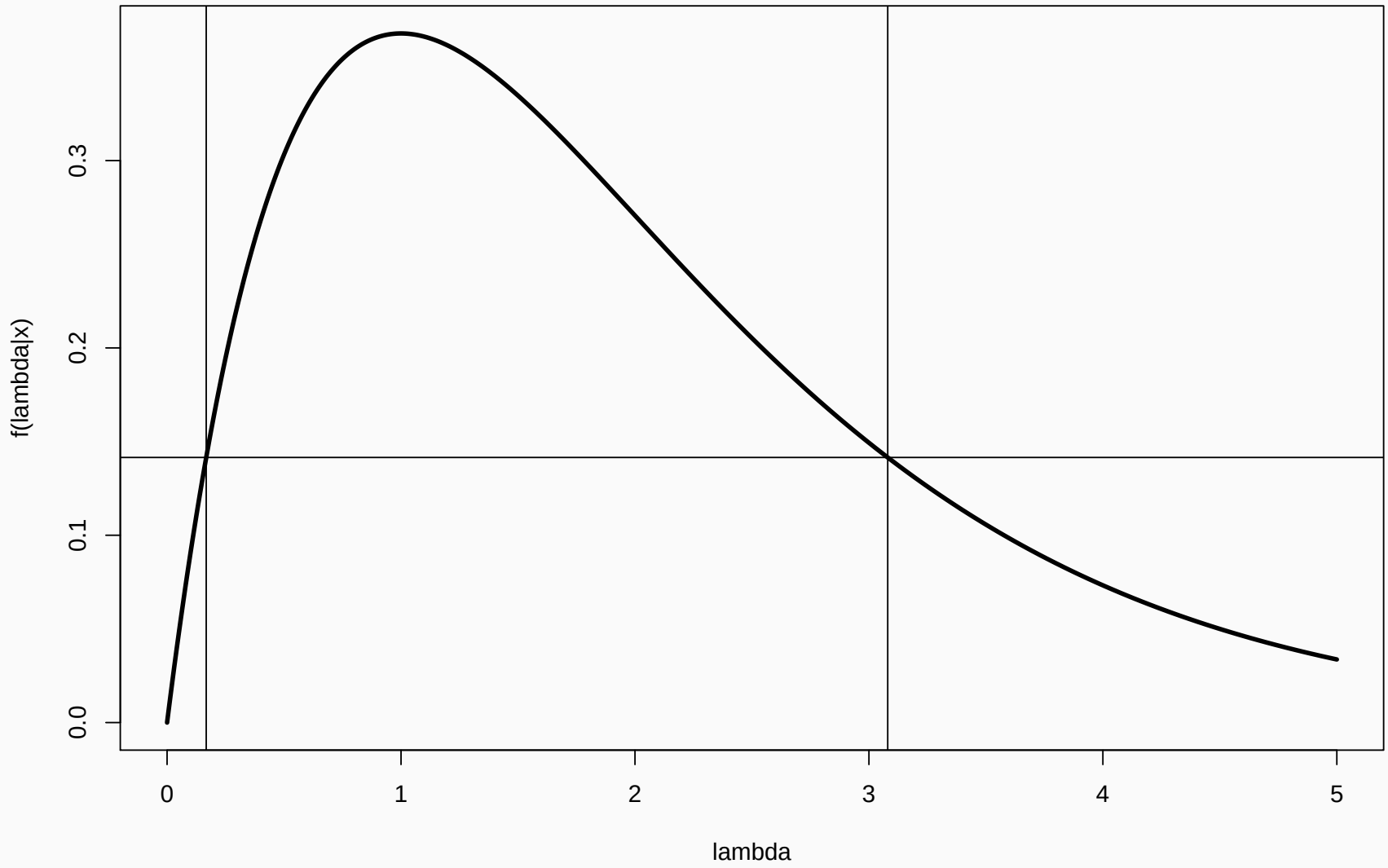
- The 80% HPD interval for λ is (a, b), where a and b satisfy the two equations:

$$\Pi(b|y) - \Pi(a|y) = 0.8$$
$$\pi(b|y) = \pi(a|y)$$

Example : Finding a HPD via the multivariate N-R algorithm

```
##  Intial Values: 0.5,0.3 and Repeat=7

##      gmat
## [1,]  0.5 0.07765 0.1632 0.1673 0.1673 0.1673 0.1673 0.1673
## [2,]  3.0 2.74069 3.0256 3.0793 3.0803 3.0803 3.0803 3.0803

##  Values of a and b and the Corresponding value on the Gamma curves (c)

## [1] 0.1673 3.0803 0.1415 0.1415

##  Areas under the gamma curve before and between  a and  b

## [1] 0.01253 0.81253 0.80000
```

Example : Finding a HPD via the multivariate N-R algorithm



Integration techniques

- Bayesian inference typically involves a great deal of integration (and/or summation). For example, consider the posterior density $\pi(\theta|y) = 6\theta^5$ for $0 < \theta < 1$.
- Suppose that we wish to find the posterior mean estimate of $\lambda = \theta^2$. This estimate is

$$\hat{\lambda} = E(\theta^2|y) = \int_0^1 \theta^2(6\theta^5)d\theta = 0.75$$

- But, what if this integral did not have a simple analytical solution?
- We may then consider using a numerical integration technique.

Example : Integration techniques

- Suppose that $X \sim N(\mu, \sigma^2)$ and $Y = (X|X > c)$ where $\mu = 8$, $\sigma = 3$ and $c = 10$. Find $E(Y)$ using numerical techniques and compare your answer with the exact value,

$$\mu + \sigma \frac{\phi(\frac{c-\mu}{\sigma})}{1 - \Phi(\frac{c-\mu}{\sigma})}$$

- Note that $E(Y) = \int_c^\infty g(x)dx$ where $g(x) = \frac{xf(x)}{P(X>c)}$
- Applying the $integrate()$ function directly to $g(x)$, we obtain $E(Y) = 11.7955$.

```
8+3*(dnorm((10-8)/3))/(1-pnorm((10-8)/3))
```

```
## [1] 11.7955
```

Example : Integration techniques

```
INTEG <- function(xvec, yvec, a = min(xvec), b = max(xvec)){
fit <- smooth.spline(xvec, yvec)
spline.f <- function(x){predict(fit, x)$y }
integrate(spline.f, a, b)$value }
mu=8; sig=3; c = 10; options(digits=6)
PXpos = (1-pnorm((c-mu)/sig))
gfun=function(x){ x * dnorm(x,mu,sig) / PXpos }
integrate(gfun,c,20)$value
```

```
## [1] 11.7929
```

```
integrate(gfun,c,30)$value
```

```
## [1] 11.7955
```

```
xvec <- seq(c,20,0.1); gvec <- gfun(xvec); INTEG(xvec,gvec,c,20)
```

```
## [1] 11.7929
```

```
xvec <- seq(c,30,0.1); gvec <- gfun(xvec); INTEG(xvec,gvec,c,30)
```

```
## [1] 11.7955
```

The optim() function

- The function $optim()$ in R is a very useful and versatile tool for maximising or minimising functions, both of one and of several variables.
- This R function can also be adapted for solving single or simultaneous equations and provides an alternative to other techniques such as trial and error, the Newton-Raphson algorithm, etc.
- The next exercise shows how the $optim()$ function can be used to specify a prior distribution.

Specification of parameters using the optim() function

- Consider the normal-gamma model given by the random sample $(y_1, ..., y_n|\lambda)$ from $N(\mu, 1/\lambda)$ and $\lambda \sim Gamma(\eta, \tau)$.
- Use the $optim()$ function in R to find the values of η and τ which correspond to a prior belief that the population standard deviation $\sigma = 1/\sqrt{\lambda}$ lies between 0.5 and 1 with 95% probability, and that σ is equally likely to be below 0.5 as it is to be above 1.
- We wish to find the values of η and τ which satisfy the two equations:

$$\Pi(\sigma < a|y) = \alpha/2 \qquad \text{and} \qquad \Pi(\sigma < b|y) = 1 - \alpha/2$$

where $a = 0.5$, $b = 1$, and $\alpha = 0.05$

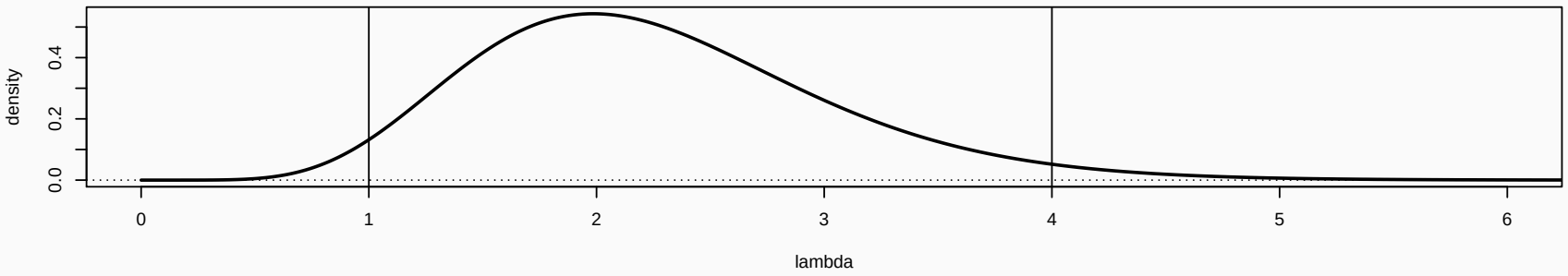
Example : Specification of Parameters using optim() function

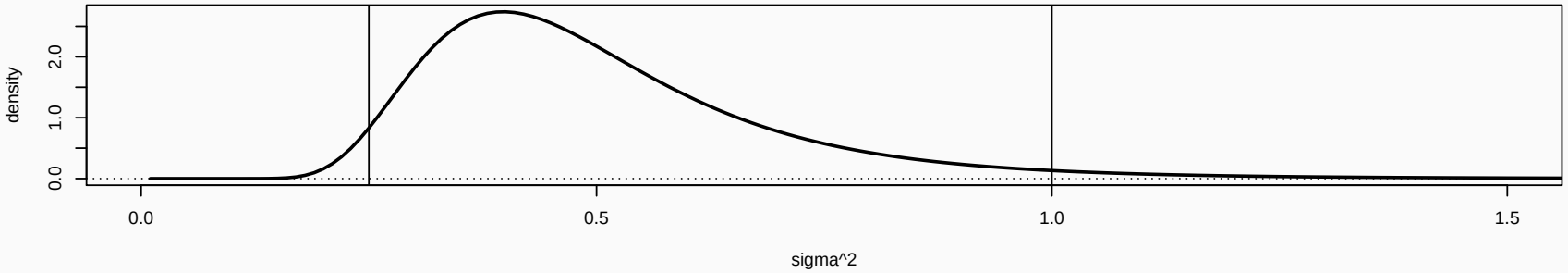
```
## [1] 8.4764 3.7679

## [1] 0.025048 0.975104

## [1] 8.4748 3.7654

## [1] 0.025 0.975

## [1] 8.4753 3.7655

## [1] 0.024992 0.974996

## [1] 8.4748 3.7654

## [1] 0.025 0.975

## [1] 1

## [1] 0.025

## [1] 0.025
```

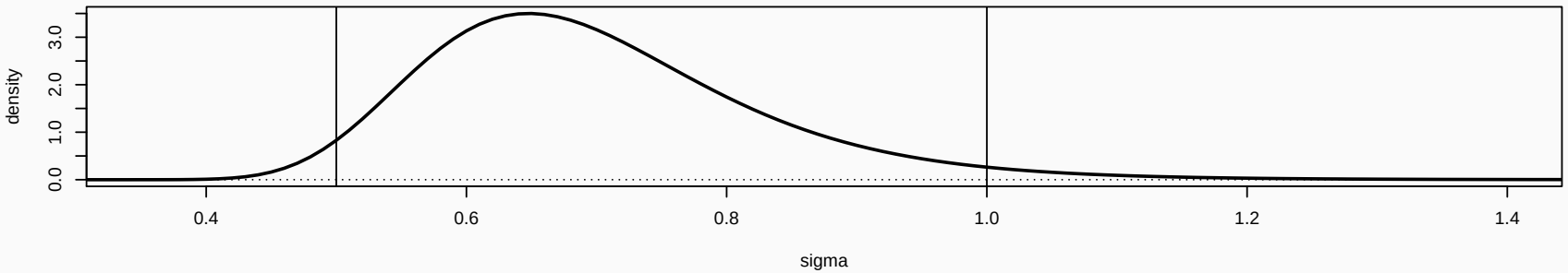
Remark

- With the default settings and starting at $\eta = 0.2$ and $\tau = 6$, $optim()$ produced some warning messages (which we ignored) and provided the solution, $\eta = 8.4764$ and $\tau = 3.7679$.
- Now, this solution is not exactly correct, because the probabilities of a $Gamma(8.4764, 3.7679)$ random variable lying below $1/b^2 = 1$ and below $1/a^2 = 4$, respectively, are 0.025048 and 0.975104 (i.e. not exactly 0.025 and 0.975 as desired).
- However, applying the $optim()$ function again but starting at the previous solution, namely $\eta = 8.4764$ and $\tau = 3.7679$, yielded a 'refined' solution, $\eta = 8.4748$ and $\eta = 3.7654$.
- This solution may be considered correct, because the probabilities of a $Gamma(8.4748, 3.7654)$ random variable being less than $1/b^2 = 1$ and less than $1/a^2 = 4$, respectively, are exactly 0.025 and 0.975.
- As a check, the next three densities show that the area under that density is exactly 1, and that the areas underneath it to the left of 0.5 and to the right of 1 are both exactly 0.025.

Example : Specification of Parameters using optim() function







Monte Carlo Methods

Monte Carlo methods in Bayesian inference

- The term Monte Carlo (M.C) methods refers to a broad collection of tools that are useful for approximating quantities based on artificially generated random samples, such as M.C integration for estimating means, importance sampling, the rejection algorithm, etc.
- Most of the ideas above in this chapter are directly applicable to Bayesian inference. Suppose we have derived a posterior distribution or density $\pi(\theta|x)$ but it is complicated and difficult to work with directly.
- We can then try to generate a random sample from that posterior with a view to estimating all the required inferential quantities (e.g. point and interval estimates) via the method of M.C.

Monte Carlo methods in Bayesian inference

- First, denote the Monte Carlo sample as $\theta_1, ..., \theta_J \sim \pi(\theta|y)$. Then, the M.C estimate of the posterior mean of θ

$$\hat{\theta} = E(\theta|y) = \int \theta \pi(\theta|y) d\theta$$

is $\bar{\theta} = \frac{1}{J} \sum_{j=1}^{J} \theta_j$ (M.C sample mean) and a $100(1-\alpha)\%$ C.I for $\hat{\theta}$ is

$$\bar{\theta} \pm z_{\alpha/2} \frac{s_\theta}{\sqrt{J}}$$

where, $s_\theta = \sqrt{\frac{1}{J-1} \sum_{j=1}^{J} (\theta_j - \bar{\theta})^2}$

- A M.C $100(1-\alpha)\%$ Credible Interval for θ is $(\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2})$, where $\hat{q}_p$ is the empirical p-quantile of $\theta_1, ..., \theta_J$ and the M.C estimate of the posterior median is $\hat{q}_{0.5}$.

- When the posterior density $\pi(\theta|y)$ does not have a closed form expression (as it is often the case), it can be estimated by smoothing a probability histogram of $\theta_1, ..., \theta_J$.

Monte Carlo methods in Bayesian inference

- Once an estimate of the posterior density has been obtained, the mode of that estimate defines the M.C estimate of the posterior mode.
- Suppose we are interested in some posterior probability $p = \Pi(\theta \in A|y)$ where A is a subset of the parameter space. Then, the M.C estimate of p is

$$\hat{p} = \frac{1}{J}\sum_{j=1}^{J} I(\theta_j \in A)$$

- That is, the proportion of the θ_j values which lie in A, and a $100(1-\alpha)\%$ C.I for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/J}$$

Monte Carlo methods in Bayesian inference

- Suppose we are interested in a function of the parameter, $\psi = g(\theta)$. Then regardless of how complicated g is, we can perform M.C inference on ψ easily.
- Simply calculate $\psi_j = g(\theta_j)$ for each $j = 1, ..., J$. This results in a random sample from the posterior distribution of ψ, namely the values $\psi_1, ..., \psi_J$ random sample from $\pi(\psi|x)$.
- For example, the posterior mean of ψ, namely

$$\hat{\psi} = E(\psi|y) = \int \psi\omega(\psi|y)d\psi = \int g(\theta)\pi(\theta|y)d\theta$$

- It can be estimated by its M.C estimate, $\bar{\psi} = \frac{1}{J}\sum_{j=1}^{J}\psi_j$, and a $100(1-\alpha)\%$ C.I for $\hat{\psi}$ is

$$\bar{\psi} \pm z_{\alpha/2}\frac{s_\psi}{\sqrt{J}}$$

Example : M.C inference - Normal model (σ known)

- Suppose we observe the data vector $y = \{2.1, 3.2, 5.2, 1.7\}$
- Generate $J = 1000$ values from the posterior distribution of μ. Use this sample to perform M.C inference on μ. Illustrate your inferences with a suitable graph.
- Solution: Recall that the posterior distribution of $\frac{\mu - \mu_*}{\sigma_*} \sim N(0,1)$. Hence, we generate random sample from $z_1, ..., z_J \sim N(0,1)$ and then calculate $\mu_j = \mu_* + \sigma_* z_j$ for $j = 1, ..., J = 1000$.
- We use the posterior sample $\mu_1, ..., \mu_J \sim \pi(\mu|y)$ for M.C inference on μ.
- The posterior mean $\hat{\mu} = E(\mu|y)$ is estimated by $\bar{\mu} = \frac{1}{J}\sum_{j=1}^{J} \mu_j = 3.0076$.
- A M.C 95% C.I for $\hat{\mu}$ is $\bar{\mu} \pm z_{\alpha/2}\frac{s_\mu}{\sqrt{J}}$ where $s_\mu = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}(\mu_j - \bar{\mu})^2}$ which yields (2.979573, 3.035642).
- Further, A M.C 95% HPD for μ is $(\hat{q}_{0.025}, \hat{q}_{0.975})$ which is

$$(2.174511, 3.863680)$$

Example : M.C inference - Normal model (σ known)

```
mu0=2.75; tau0=0.5; sig0=1;y=c(2.1, 3.2, 5.2, 1.7);n=length(y);
ybar=mean(y);k=n*tau0^2/(n*tau0^2+sig0^2/n); k ;ybar=mean(y); ybar
```

```
## [1] 0.8
```

```
## [1] 3.05
```

```
mus = (1-k)*mu0 + k*ybar; sigs2=k*sig0^2/n;c(mus,sigs2)
```

```
## [1] 2.99 0.20
```

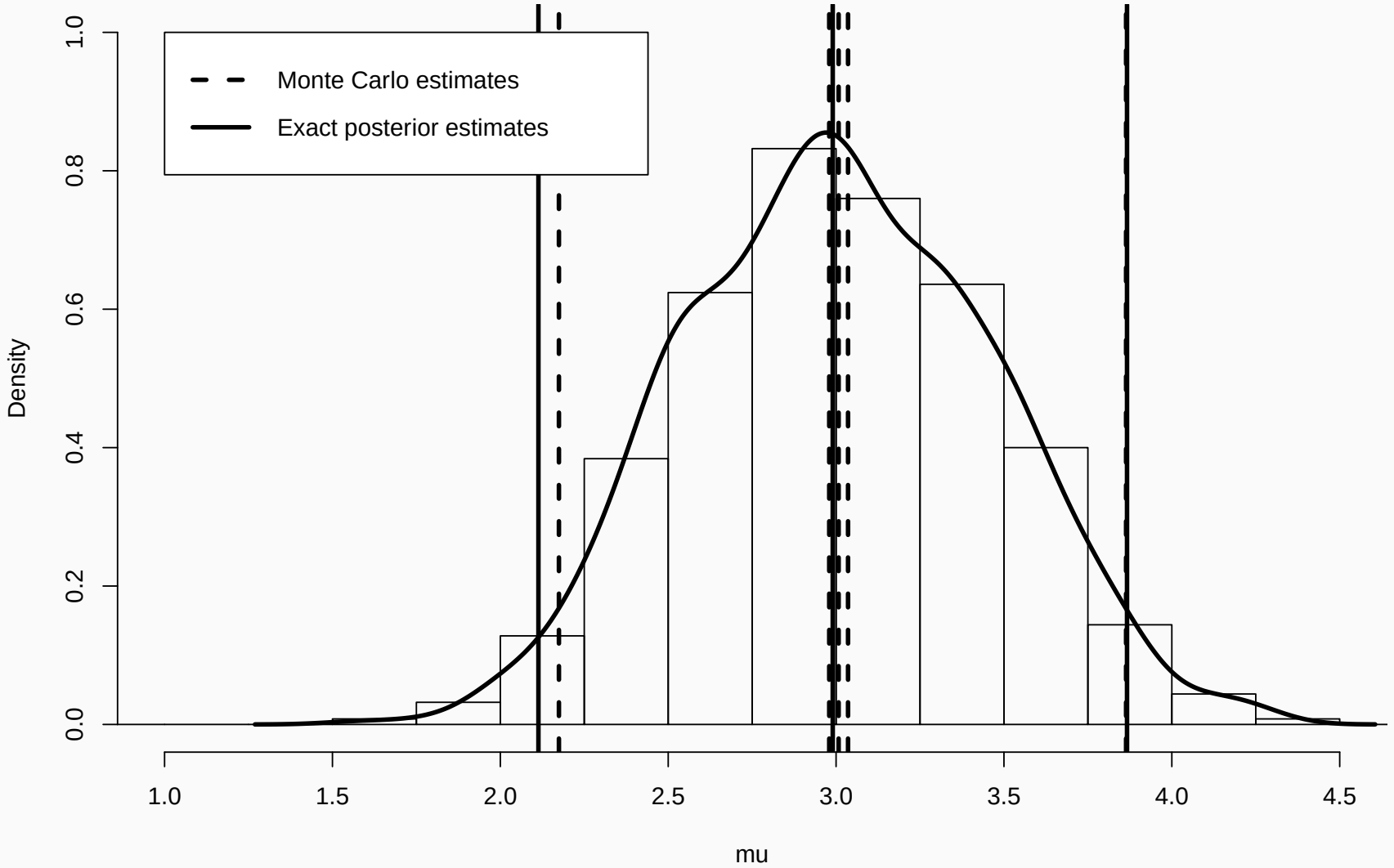
```
J=1000; set.seed(144); zj<-rnorm(J);muj=ybar+sqrt(sigs2)*zj
mubar=mean(muj); muCI=mubar + c(-1,1)*qnorm(0.975)*sd(muj)/sqrt(J)
muHPD=quantile(muj,c(0.025,0.975));c(mubar,muCI,muHPD)
```

```
##                          2.5%  97.5%
## 3.0076 2.9796 3.0356 2.1745 3.8637
```

```
muhat=mus; muHPDtrue= mus+ sqrt(sigs2)*qnorm(c(0.025,0.975))
c(muhat,muHPDtrue)
```

```
## [1] 2.9900 2.1135 3.8665
```

Example : M.C inference - Normal model (σ known)



Monte Carlo prediction

- Suppose that in the context of a Bayesian model defined by $q(y|\theta)$ and $\pi(\theta)$, we wish to predict a value x whose distribution is specified by $q(x|y,\theta)$. Recall that the posterior predictive density is

$$q(x|y) = \int_{\Omega} q(x|y,\theta)\pi(\theta|y)d\theta$$

- If this density is complicated, we may choose to perform M.C predictive inference on x using a sample $x_1, ..., x_J \sim q(x|y)$. The question then arises as to how such a sample may be obtained.

- One answer is to sample from $q(x|y)$ directly. But that may be difficult since $q(x|y)$ is complicated. Another answer is to apply the method of composition through the equation $q(x,\theta|y) = q(x|y,\theta)\pi(\theta|y)$.

- This means that we should first sample $\theta' \sim \pi(\theta|y)$ and then sample $x' \sim q(x|y,\theta')$. The result is $(x',\theta') \sim q(x,\theta|y)$. If we then discard θ', the result is the required $x' \sim q(x|y)$. Implementing this process a total of J times results in the required sample, $x_1, ..., x_J \sim q(x|y)$.

Example : Monte Carlo prediction - binomial-beta model

The probability of heads coming up on a bent coin follows a standard uniform distribution a priori. We toss the coin 50 times and get 28 heads. Estimate using M.C the probability that heads will come up on at least six of the next 10 tosses of the same bent coin.

- The binomial-beta model is $f(y|\theta) \sim Binom(n, \theta)$ when $\pi(\theta) \sim Beta(\alpha, \beta)$. Then, the posterior distribution is given by $\pi(\theta|y) \sim Beta(\alpha + n, \beta + n - y)$.
- We can show that if the future data X has distribution defined by $q(x|y, \theta) \sim Binom(m, \theta)$, then the posterior predictive distribution has the following probability function (beta binomial distribution)

$$q(x|y) = \left(\begin{array}{c} m \\ x \end{array} \right) \frac{B(y + x + \alpha, n - y + m - x + \beta)}{B(y + \alpha, n - y + \beta)} \quad \text{for} \quad x = 0, .., m$$

Example : Monte Carlo prediction - binomial-beta model

- Rather than trying to sample from this distribution directly, we may do the following:
  - Sample $\theta' \sim Beta(\alpha + n, \beta + n - y)$.
  - Sample $x'|y, \theta' \sim Binom(m, \theta')$
  - Discarding θ', we obtain the required sample value, $x' \sim q(x|y)$.

- In the situation here: $\alpha = \beta = 1$, $n = 50$, $y = 32$, $m = 10$.

- Implementing the above sampling strategy $J = 10000$ times with these specifications, we obtain a large M.C sample, $x_1, ..., x_J \sim q(x|y)$.

- It is found that 7.084 of the sample values are at least 6. Hence, we estimate $p = P(X \geq 6|y)$ by $\hat{p} = 0.7084$. A 95% CI for p is then

$$(\hat{p} \pm z_{0.025}\sqrt{\hat{p}(1-\hat{p})/J}) = (0.69949, 0.71731)$$

- Note that the exact probability here is $p = \sum_{x=6}^{10} q(x|y) = 0.70296$ which lies in the 95% C.I obtained using the M.C method.

Example : Monte Carlo prediction - binomial-beta model

```
options(digits=5)
n=50; y=32; alp=1;bet=1; a=alp+y; b=bet+n-y; m=10; J=10000
set.seed(443); tv=rbeta(J,a,b); xv=rbinom(J,m,tv)
phat=length(xv[xv>=6])/J; CI=phat+c(-1,1)*qnorm(0.975)*sqrt(phat*(1-phat)/J)
c(phat,CI)
```

```
## [1] 0.70840 0.69949 0.71731
```

```
xvec=0:m;
fxgiveny= choose(m,xvec)*beta(y+xvec+alp,n-y+m-xvec+bet)/beta(y+alp,n-y+bet)
sum(fxgiveny)  # Check it is a proper probability function
```

```
## [1] 1
```

```
sum(fxgiveny[xvec>=6]) # Check the exact value of P(X >= 6)
```

```
## [1] 0.70296
```

Introduction to MCMC methods

- Monte Carlo methods include basic techniques for generating a random sample and methods for using such a sample to estimate quantities such as difficult integrals.
- An advanced techniques for generating a random sample, in particular the class of techniques known as Markov Chain Monte Carlo (MCMC) methods.
- Applying an MCMC method involves designing a suitable Markov chain, generating a large sample from that chain for a burn-in period until stochastic convergence, and making appropriate use of the values following that burn-in period.

Introduction to MCMC

- Like other iterative techniques such as the Newton-Raphson algorithm, MCMC methods require an arbitrary starting values and then iterating repeatedly until convergence.
- But, MCMC methods are distinguished from these other methods by the fact that the update at each iteration is not deterministic but stochastic, with the probability distributions involved dependent on results from the previous iteration.
- Typically, MCMC methods are used to sample from multivariate probability distributions rather than univariate ones. This is because a univariate distribution can usually be sampled from using simpler methods.

Choosing Priors

Choosing Priors

- The issue of selecting a prior for a problem is an important one. Because this will typically vary from statistician to statistician, this is often criticized as being too subjective for scientific studies.
- What then justifies one choice of a statistical model or prior over another?
- The only way to assess whether or not an appropriate choice was made is to check whether the observed s is reasonable given this choice.
- If s is surprising, when compared to the distribution prescribed by the model and prior, then we have evidence against the statistician's choices.
- Methods designed to assess this are called model-checking procedures.

Conjugate Priors

Definition 2

The family of priors $\{\pi_\lambda : \lambda \in \Lambda\}$ for the parameter θ of the model $\{f(.|\theta) : \theta \in \Omega\}$ is conjugate, if for all data $s \in S$ and all $\lambda \in \Lambda$, the posterior distribution $\pi_\lambda(.|s) \in \{\pi_\lambda : \lambda \in \Lambda\}$.

- Many conjugate families offer sufficient variety to allow for the expression of a wide spectrum of prior beliefs.
- Examples: we have effectively shown that the family of all Beta distributions is conjugate for sampling from the Bernoulli model. We showed that the family of normal priors is conjugate for sampling from the location normal model.

Elicitation

- Elicitation involves explicitly using the statistician's beliefs about the true value of θ to select a prior in $\{\pi_\lambda : \lambda \in \Lambda\}$ that reflects these beliefs.
- Typically, these involve the statistician asking questions of himself, or of experts in the application area, in such a way that the answers specify a prior from the family.

Example : Elicitation - Normal model (σ known)

- Suppose we are sampling from $N(\mu, \sigma_0^2)$ distribution with unknown μ and known σ_0^2, and we restrict attention to the family $\{N(\mu_0, \tau_0^2) : \mu_0 \in \mathbb{R}, \tau_0^2 > 0\}$ of priors for μ. So, here, $\lambda = (\mu_0, \tau_0^2)$ and there are two degrees of freedom in this family. Thus, specifying two independent characteristics specifies a prior.
- We might ask an expert to specify a number μ_0 such that the true value of μ was as likely to be greater than as less than μ_0, so that μ_0 is the median of the prior.
- We might also ask the expert to specify a value ν_0 such that there is 99% certainty that the true value of μ is less than ν_0. This of course is the 0.99-quantile of their prior.
- We could ask the expert to specify the center μ_0 of their prior distribution and for a constant τ_0 such that $\mu_0 \pm 3\tau_0$ contains the true value of μ with virtual certainty.

Empirical Bayes

- When the choice of λ_0 is based on the data s, the methods are referred to as empirical Bayesian methods.
- Logically, such methods would seem to violate a basic principle of inference, namely, the principle of conditional probability. For when we compute the posterior distribution of θ using a prior based on s, in general this is no longer the conditional distribution of θ given the data.
- While this is certainly an important concern, in many problems the application of empirical Bayes leads to inferences with satisfying properties.
- For example, one empirical Bayesian method is to compute the prior predictive $m_\lambda(s)$ for the data s, and then base the choice of λ on these values.

Example : Empirical Bayes - Bernoulli Model

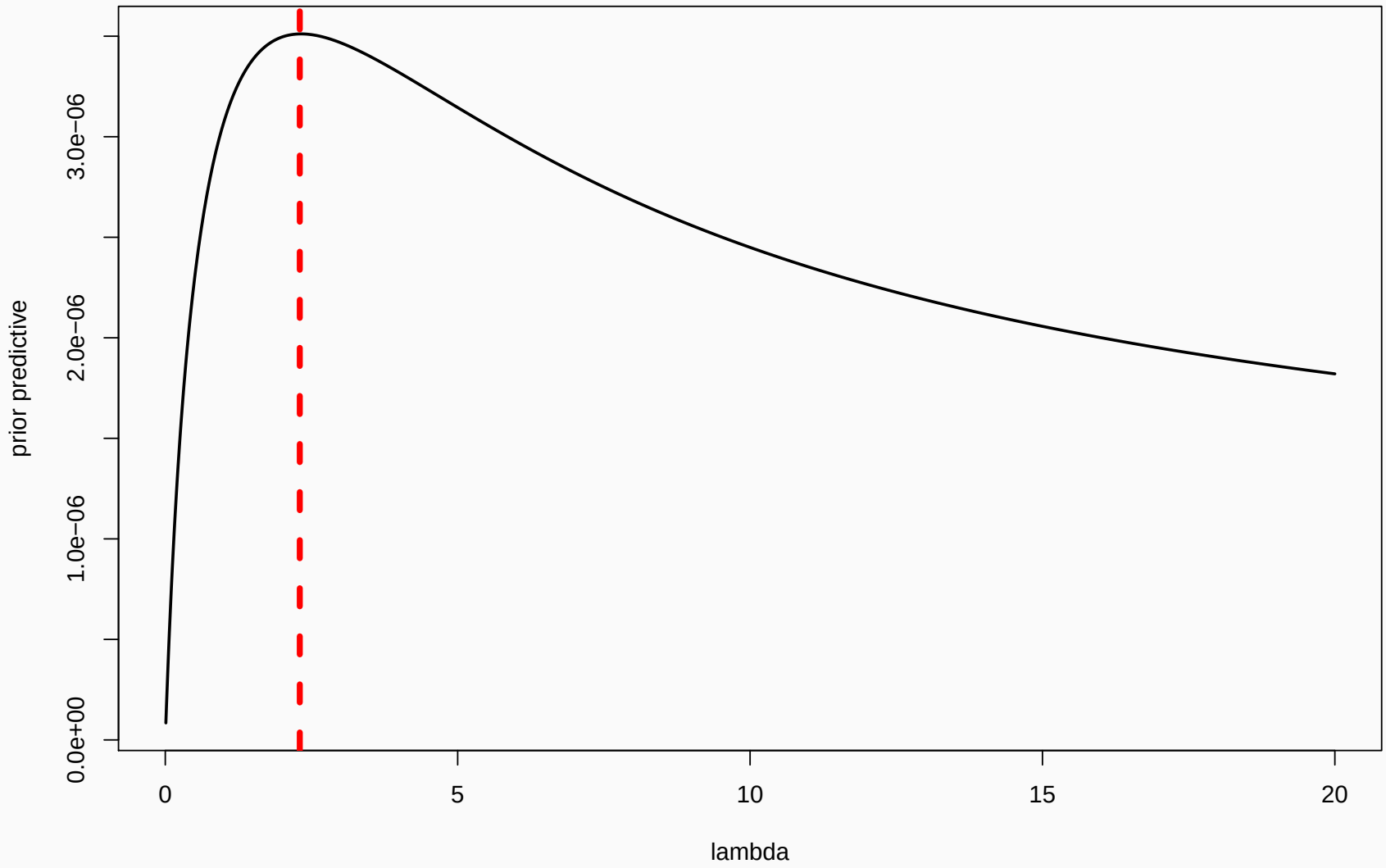
- Suppose we have a sample $y_1, ..., y_n$ from a $Bernoulli(p)$ distribution and we contemplate putting a $Beta(\lambda, \lambda)$ prior on p for some $\lambda > 0$. So, the prior is symmetric about $1/2$ and the spread in this distribution is controlled by λ. Since the prior mean is $1/2$ and the prior variance is $\lambda^2/[(2\lambda+1)(2\lambda)^2] = 1/4(2\lambda+1) \to 0$ as $\lambda \to \infty$. It follows that choosing λ large leads to a very precise prior. Then we have that

$$m_\lambda(y_1, ..., y_n) = \frac{\Gamma(2\lambda)\Gamma(n\bar{y}+\lambda)\Gamma(n(1-\bar{y})+\lambda)}{\Gamma^2(\lambda)\Gamma(n+2\lambda)}$$

- It is difficult to find the value of λ that maximizes this. But, for real data we can tabulate and plot $m_\lambda(y_1, ..., y_n)$ to obtain this value.

- More advanced computational methods can also be used. For example, suppose that $n = 20$ and we obtained $n\bar{y} = 5$ as the number of 1's observed. In the next Figure, we have plotted the graph of $m_\lambda(y_1, ..., y_n)$ as a function of λ. We can show that the maximum occurs at $\lambda = 2.3$. Accordingly, we use the $Beta(n\bar{y}+\lambda, n(1-\bar{y})+\lambda) = Beta(7.3, 17.3)$ distribution for inferences about p.

Example : Empirical Bayes - Bernoulli Model



Hierarchical Bayes

- An alternative to choosing a prior for θ in $\{\pi_\lambda : \lambda \in \Lambda\}$ consists of putting yet another prior distribution ω called a hyperprior, on λ.
- This approach is commonly called hierarchical Bayes. The prior for θ basically becomes $\pi(\theta) = \int_\Lambda \pi_\lambda(\theta)\omega(\lambda)d\lambda$, so we have in effect integrated out the hyperparameter.
- The problem then is how to choose the prior ω. In essence, we have simply replaced the problem of choosing the prior on θ with choosing the hyperprior on λ. So, in this situation, the posterior density of θ is equal to

$$\pi(\theta|s) = \frac{f(s|\theta)\int_\Lambda \pi_\lambda(\theta)\omega(\lambda)d\lambda}{m(\lambda)} = \int_\Lambda \frac{f(s|\theta)\pi_\lambda(\theta)}{m_\lambda(s)}\frac{m_\lambda(s)\omega(\lambda)}{m(s)}d\lambda$$

  where $m(s) = \int_\Lambda m_\lambda(s)\omega(\lambda)d\lambda$ and $m_\lambda(s) = \int f(s|\theta)\pi_\lambda(\theta)d\theta$.
- Therefore, we can use $\pi(\theta|s)$ for inferences about the model parameter θ. (e.g., estimation, credible regions, and hypothesis assessment) and $m_\lambda(s)\omega(\lambda)/m(s)$ for inferences about λ.

Improper Priors and Non-informativity

- One approach to choosing a prior, and to stop the chain of priors in a hierarchical Bayes approach, is to prescribe a non-informative prior. Such a prior is also referred to as a default prior or reference prior.
- The idea here is to give a rule such that, if a statistician has no prior beliefs about the value of a parameter or hyperparameter, then a prior is prescribed that reflects this. Surprisingly, in many contexts, statisticians have been led to choose non-informative priors that are improper, i.e., $\int_\Omega \pi(\theta) d\theta = \infty$, so they do not correspond to probability distributions.
- The interpretation of an improper prior is not at all clear, and their use is somewhat controversial. Of course, (s, θ) no longer has a joint probability distribution when we are using improper priors, and we cannot use the principle of conditional probability to justify basing our inferences on the posterior.

Example : Normal model with an Improper Prior

- Suppose $y_1, ..., y_n$ is a random sample from $N(\mu, \sigma_0^2)$ distribution, where $\mu \in \mathbb{R}$ is unknown and σ_0^2 is known. Many arguments for default priors in this context lead to the choice $\pi(\mu) = 1$, which is clearly improper.
- Pretending that this π is a proper probability density, we get that the posterior density of μ is proportional to $e^{-\frac{n}{2\sigma_0^2}(\mu-\bar{y})^2}$ which immediately implies that the posterior distribution of μ is $N(\bar{y}, \sigma_0^2/n)$.
- Note that this is the same as the limiting posterior as $\tau_0 \to \infty$, although the point of view is quite different.

Improper Priors and Non-informativity

- Jeffreys proposed defining a non-informative prior for θ as $\pi(\theta) \propto I(\theta)^{1/2}$ where $I(\theta)$ is the Fisher information for θ

$$I(\theta) = -E\left[\frac{\partial^2 \log f(y|\theta)}{\partial \theta^2}\right] = E\left[\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right)^2\right]$$

- This is referred to as Jeffreys' prior which is dependent on the model.

- Jeffreys' prior has an important invariance property. That is, Jeffreys' prior is invariant to reparameterisation because if $\psi = g(\theta)$ then

$$I(\psi) = I(\theta)^{1/2}\left|\frac{d\theta}{d\psi}\right|$$

Example 1: Jeffreys' priors

- Example 1: Consider the Normal case with unknown mean μ, known variance σ_0^2. The random sample is $y_1, ..., y_n$ from $N(\mu, \sigma_0^2)$. Thus,

$$\log f(y_1, ..., y_n|\mu) \propto -\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma_0^2}$$

  which leads to $I(\mu) = n/\sigma_0^2$. Hence, the Jeffreys' prior for μ is proportional to 1, i.e. the Uniform distribution.

Example 2: Jeffreys' priors

- Example 2: Consider the Normal case with known mean μ_0, unknown variance σ^2. Thus,

$$\log f(y_1, ..., y_n|\sigma^2) \propto -\frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(y_i - \mu_0)^2}{2\sigma^2}$$

  which leads to $I(\sigma^2) = n/2\sigma^4$. Hence, Jeffreys' prior for σ^2 is proportional to $1/\sigma^2$

- This improper distribution is approximated by a $Gamma(\varepsilon, \varepsilon)$ distribution with $\varepsilon \to 0$.

- Note that $\pi(\sigma^2) \propto 1/\sigma^2$ is equivalent to a uniform prior on $\log\sigma^2$.

What We Have Just Learned

- The Prior and Posterior Distributions
- Inferences Based on the Posterior
- Bayesian Computations
- Choosing Priors