Introduction to Bayesian Statistics
(Chapter 2)

Michael Tsiang

Stats 102C: Introduction to Monte Carlo Methods

UCLA

Acknowledgements: Qing Zhou

Outline

1. The Frequentist Perspective

2. The Bayesian Perspective

The Frequentist and Bayesian Perspectives

- The underlying difference between the frequentist and Bayesian perspectives is what probability represents.
- The frequentist perspective:
  - Probability represents the long-run relative frequency of random events.

    fixed number
  - Parameters are considered (often unknown) fixed constants.
- The Bayesian perspective:
  - Probability represents one's subjective belief about random events.
  - Parameters are considered random variables.

Likelihood

- Consider the scenario of flipping a coin n times, where the probability of heads on any given flip is θ. p(head)

sum("1")

- If Y is the number of heads in n flips, then $Y \sim \text{Bin}(n, \theta)$, with PMF/density

n times

give p(head) = θ, p(Y=y)

$$P_\theta(Y = y) = f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}. = f(y|\theta)$$

p(head) if Y times head

$= L(\theta|y)$

- The density of Y, considered as a function of θ, is called the **likelihood function** (or just **likelihood**): $L(\theta|y) = f(y|\theta)$.

- Suppose we observed $Y = y$ heads from n flips. Based on this data, we want to estimate θ.

Maximum Likelihood

- In the frequentist perspective, θ is a fixed constant: If we could repeat the scenario (flipping the coin n times) infinitely many times, the relative frequency of times that the coin lands on heads would be θ.

$\theta \to \frac{\text{sum}(\text{"1"})}{n}$

- A standard (frequentist) way to estimate θ would be the maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{MLE}} = \underset{\theta}{\operatorname{argmax}}\, L(\theta|y) = \underset{\theta}{\operatorname{argmax}}\, f(y|\theta).$$

- What is $\hat{\theta}_{\mathrm{MLE}}$ for $Y \sim \mathrm{Bin}(n, \theta)$?

Maximum Likelihood

To maximize the likelihood, we differentiate the log-likelihood $\log f(y|\theta)$ with respect to θ:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(y|\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta} \log \left[\binom{n}{y} \theta^y (1-\theta)^{n-y} \right] \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta} \left[\log \binom{n}{y} + y \log \theta + (n-y) \log(1-\theta) \right] \\
&= \frac{y}{\theta} - \frac{n-y}{1-\theta} \quad = 0
\end{aligned}
$$

Maximum Likelihood

Setting the derivative to 0 and solving for θ:

$$\begin{aligned} \frac{n-y}{1-\theta} &= \frac{y}{\theta} \\ (n-y)\theta &= y(1-\theta) \\ n\theta - y\theta &= y - y\theta \\ \theta &= \frac{y}{n} \end{aligned}$$

when $\theta = \frac{y}{n}$, the biggest value to make y times head.

So the maximum likelihood estimator for θ is

$$\hat{\theta}_{\text{MLE}} = \frac{y}{n}.$$

$\frac{\text{sum}(\text{"1"})\ (\text{success})}{n \text{ times}}$

That is, if we observe y heads in n coin flips, we would estimate the probability of heads to be $\frac{y}{n}$.

Confidence Intervals

- A $\mathbf{100(1-\alpha)\%}$ **confidence interval for** $\boldsymbol{\theta}$ is a random interval $[\ell(Y), u(Y)]$ such that, *before the data is gathered*,

$$P[\ell(Y) < \theta < u(Y)|\theta] = 1 - \alpha.$$

- Once we observe $Y = y$, then the interval $[\ell(y), u(y)]$ is no longer random, so

$$P[\ell(y) < \theta < u(y)|\theta] = \begin{cases} 0 & \text{if } \theta \notin [\ell(y), u(y)] \\ 1 & \text{if } \theta \in [\ell(y), u(y)]. \end{cases}$$

- If we were to take many random samples and form a $100(1-\alpha)\%$ confidence interval from each one, about $100(1-\alpha)\%$ of these intervals would contain θ.

- The probability $1 - \alpha$ is called the (frequentist) **coverage probability**.

Outline

1. The Frequentist Perspective

2. The Bayesian Perspective

The Prior

$0 \leq \theta \leq 1$

- In the Bayesian perspective, we are able to take our prior beliefs into account. We represent our beliefs about θ prior to observing data by a **prior distribution** $\pi(\theta)$.
- Suppose, before observing any data, we believe the coin should be fair, but we are not 100% sure.
- For example, we can model our prior beliefs by a beta distribution $\mathrm{Beta}(\alpha, \beta)$, so

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{for } \theta \in [0,1].$$

- Parameters of the prior distribution (α and β in this example) are called **hyperparameters**.

The Prior

before

- For example, for hyperparameters $\alpha = 4, \beta = 4$, the prior mean (what we expect θ to be prior to observing data) is

$$E(\theta) = \frac{\alpha}{\alpha + \beta} = \frac{4}{4+4} = 0.5.$$

The prior mode of θ is

$$\text{mode}(\theta) = \frac{\alpha - 1}{\alpha + \beta - 2} = \frac{4-1}{4+4-2} = 0.5.$$

- How do we incorporate the observed data $Y = y$ to update our prior beliefs?

The Posterior

- The **posterior** after **distribution** $\pi(\theta|y)$ represents our updated beliefs about θ *after* observing the data.

- To find the posterior distribution, we apply Bayes Theorem:

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{f(y)} = \frac{\pi(\theta)f(y|\theta)}{\int \pi(\theta)f(y|\theta)\,\mathrm{d}\theta}$$

nothing about θ

- $f(y) = \int \pi(\theta)f(y|\theta)\,\mathrm{d}\theta$ is called the **marginal likelihood**.

fixed Z

- Notice that the marginal likelihood does not depend on θ, so we have the key result:

$$\pi(\theta|y) \quad \propto \quad \pi(\theta)f(y|\theta)$$

$$\textbf{posterior} \quad \propto \quad \textbf{prior} \times \textbf{likelihood}$$

$\pi_0(\theta)$

For our coin flipping example (the **Beta-Binomial Model**):

- If $\underbrace{\pi(\theta) \sim \text{Beta}(\alpha, \beta)}_{}$ and $f(y|\theta) \sim \text{Bin}(n, \theta)$, then the posterior distribution of θ is

  piror   likelihood   likelihood

$$\underbrace{\pi(\theta|y)}_{\text{posterior}} \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \binom{n}{y} \theta^{y}(1-\theta)^{n-y}$$

$$\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1},$$

$\theta^{x-1}(1-\theta)^{y-1} \sim \text{Beta}(x, y)$

  which we recognize as a beta distribution with parameters $\alpha' = y + \alpha$ and $\beta' = n - y + \beta$.

- So $\pi(\theta|y) \sim \text{Beta}(\underset{\alpha}{y + \alpha}, \underset{\beta}{n - y + \beta})$.

  beta bin model

  $E(\pi(\theta|y)) = \dfrac{y+\alpha}{y+\alpha+n-y+\beta}$

- If the posterior is in the same parametric family as the prior, the prior and posterior are called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood.

The Posterior Mean

- A standard Bayesian estimator for θ is the **posterior mean**

$$E(\theta|y) = \int \theta\pi(\theta|y)\,\mathrm{d}\theta, \quad = \frac{\alpha'}{\alpha'+\beta'} = \frac{\alpha+y}{\alpha+\beta+n}$$

is about α β and n and y

which represents our updated beliefs about what we expect θ to be after observing the data.

- For conjugate distributions, the posterior distribution and posterior mean can usually be computed analytically.

- For distributions that are not conjugate, the posterior distribution and posterior mean can be difficult or impossible to compute in closed form, so Markov Chain Monte Carlo methods are applied.

The Posterior Mean

For our coin flipping example:

- The posterior distribution is

$$\pi(\theta|y) \sim \text{Beta}(\alpha' = y + \alpha, \beta' = n - y + \beta).$$

- The posterior mean is then

$$E(\theta|y) = \frac{\alpha'}{\alpha' + \beta'} = \frac{y + \alpha}{n + \alpha + \beta}.$$ (α, β, y, n)

$E(\theta) = \dfrac{\alpha}{\alpha + \beta}$

α : y : head times

$\alpha+\beta$: n : total times.

The Posterior Mean

For our coin flipping example:

- The posterior mean can be written as

$$
\begin{aligned}
E(\theta|y) &= \frac{y+\alpha}{n+\alpha+\beta} \\
&= w\frac{\alpha}{\alpha+\beta} + (1-w)\frac{y}{n},
\end{aligned}
$$

where $w = \dfrac{\alpha+\beta}{n+\alpha+\beta}$.

$$\underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{(w)} \times \frac{\alpha}{\alpha+\beta} + \underbrace{\frac{n}{\alpha+\beta+n}}_{(1-w)} \times \frac{y}{n}$$

- The posterior mean is thus a weighted average of the prior mean and the data mean.

- In this example, $\alpha+\beta$ is the **prior effective sample size**, and α is the prior number of heads. Large values of α and β represent strongly held prior beliefs.

- As $n \to \infty$, the data outweighs the prior, and $E(\theta|y) \to \dfrac{y}{n}$.

For our coin flipping example:

- Suppose our prior is

$$\pi(\theta) \sim \text{Beta}(\alpha = 4, \beta = 4).$$ $E(\pi(\theta)) = 0.5$ (prior)

- If we observe $Y = 3$ out of $n = 10$ coin flips ($\hat{\theta}_{\text{MLE}} = 0.3$), then the posterior distribution of θ is

$$\pi(\theta|y) \sim \text{Beta}(\alpha' = 7, \beta' = 11),$$

$\frac{\alpha + y}{\alpha + \beta + n} = \frac{4+3}{4+3+11}$

with posterior mean

$$E(\theta|y) = \frac{\alpha'}{\alpha' + \beta'} = \frac{7}{7+11} = \frac{7}{18} \approx 0.3888889.$$

$E(\pi(\theta)) = 0.5$ prior

$E(\pi_1(\theta)) = 0.3889$ posterior

Example: Beta-Binomial Model

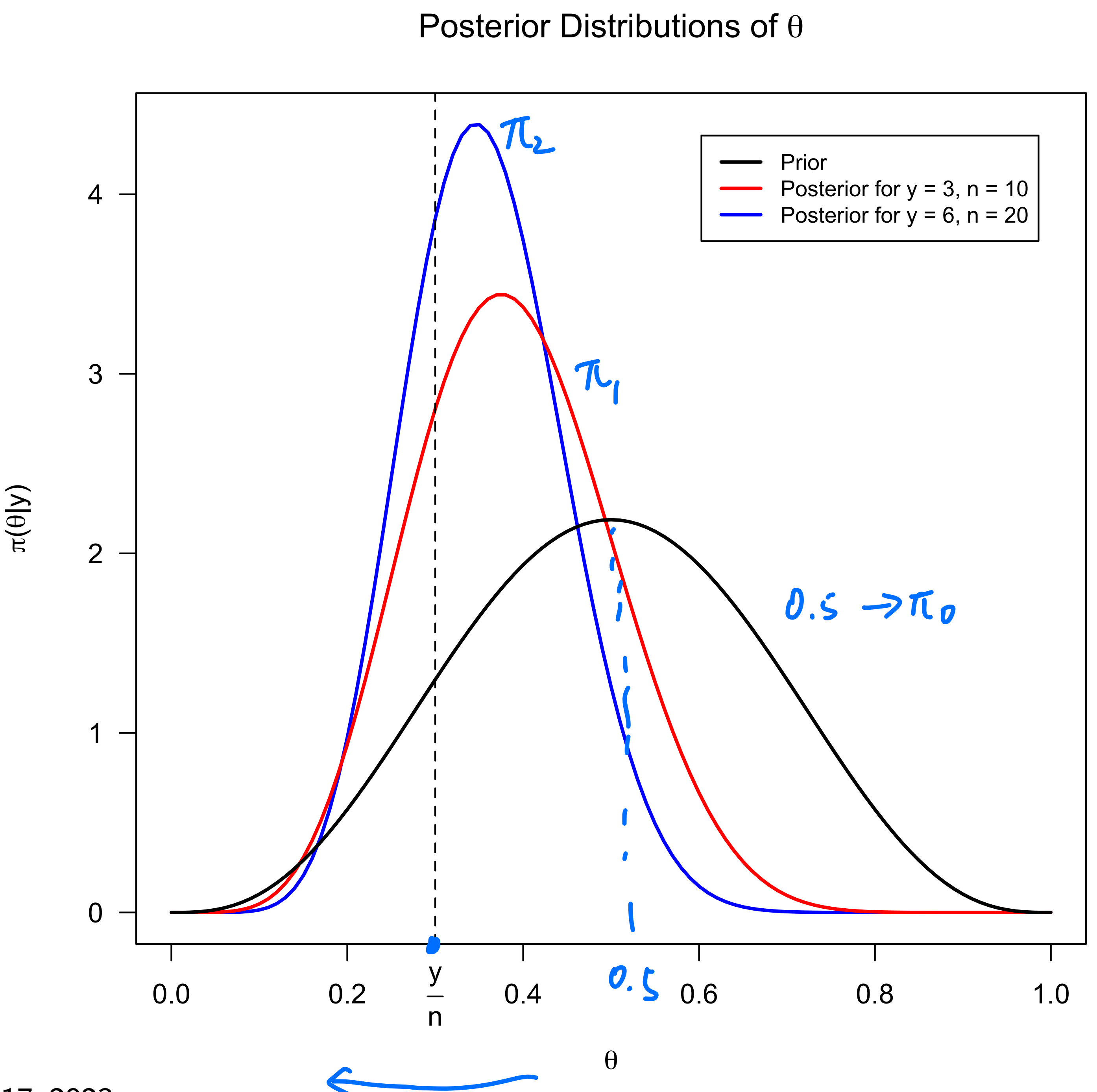
If we instead observe $Y = 6$ out of $n = 20$ coin flips ($\hat{\theta}_{\text{MLE}} = 0.3$), then the posterior distribution of θ is

$$\pi(\theta|y) \sim \text{Beta}(\alpha' = 10, \beta' = 18),$$

with posterior mean

$$E(\theta|y) = \frac{\alpha'}{\alpha' + \beta'} = \frac{10}{10 + 18} = \frac{10}{28} \approx 0.3571429.$$

Example: Beta-Binomial Model



The Uninformative Prior

- The prior distribution can represent past information, such as past experiments or literature, or subjective beliefs from a knowledgeable person.

- If no prior information is available (or we do not want to take it into account), we can use an **uninformative** (or **flat**) **prior**, which assigns equal density to all possibilities of the parameter.

- When using an uninformative prior, Bayesian estimators tends to be similar (sometimes identical) to frequentist estimators: The data easily outweighs a prior with no information.

The Uninformative Prior

For the coin flipping example: previous : $\theta \sim$ Beta(α, β)

- An uninformative prior would be $\theta \sim \text{Unif}(0,1)$, so $\frac{1}{1-0} = 1$

$$\pi(\theta) = 1, \quad \text{for } \theta \in [0,1].$$

- The posterior distribution would then be

$$\begin{aligned} \pi(\theta|y) &\propto \pi(\theta) f(y|\theta) \\ &= 1 \cancel{\binom{n}{y}} \theta^y (1-\theta)^{n-y} \\ &\propto \theta^y (1-\theta)^{n-y}, \end{aligned}$$

which we recognize as a beta distribution with parameters $\alpha' = y + 1$ and $\beta' = n - y + 1$.

- So $\pi(\theta|y) \sim \text{Beta}(y+1, n-y+1)$.

$\theta^{\underbrace{y+1}-1} \quad (1-\theta)^{\underbrace{n-y}+1}$

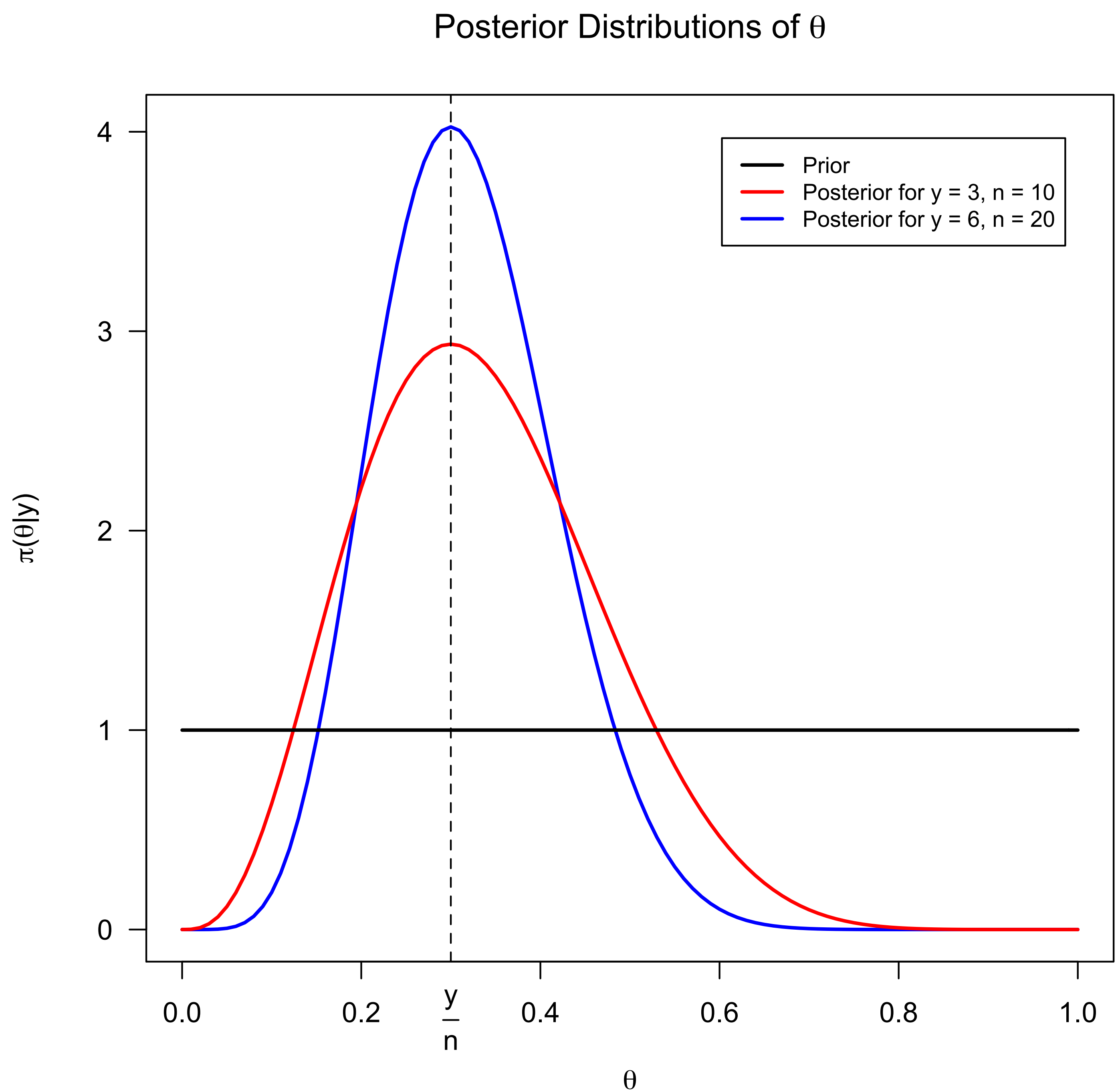
The Uninformative Prior

- Another common Bayesian estimator is the **posterior mode**, also called the **maximum a posteriori (MAP) estimator**.
- In the coin flipping example with uninformative prior:

$$\hat{\theta}_{\mathrm{MAP}} = \mathrm{mode}(\theta|y) = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{y}{n}. = MLE$$

- In other words: When we do not account for prior information, the MAP estimator of θ coincides with the MLE.

Example: Beta-Binomial Model



Credible Intervals

- An interval $[\ell(y), u(y)]$, based on the observed data $Y = y$, is a **$100(1-\alpha)\%$ credible interval for θ** if

$$P[\ell(y) < \theta < u(y) | Y = y] = 1 - \alpha.$$

  The probability $1-\alpha$ is called the **(Bayesian) coverage probability**.

- The interpretation of a credible interval is that it describes the information about the location of the true value of θ *after* you have observed $Y = y$.

- This is different from the frequentist interpretation of coverage probability, which describes the probability that the interval will cover the true value *before* the data is observed.

Quantile-based Credible Intervals

- The method for constructing a credible interval from a posterior distribution is not unique.

- A Bayesian analogue to a frequentist confidence interval is to use posterior quantiles.

- If $\theta_{\alpha/2}$ and $\theta_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ posterior quantiles of θ, then

$$P(\theta_{\alpha/2} < \theta < \theta_{1-\alpha/2} | Y = y) = 1 - \alpha,$$

  so $[\theta_{\alpha/2}, \theta_{1-\alpha/2}]$ is a **$100(1-\alpha)\%$ quantile-based credible interval for θ**.

- The quantile function for the Beta distribution in R is `qbeta()`.

High Posterior Density Regions

- A common alternative to a quantile-based interval is a **high posterior density (HPD) region (or interval)**.
- The HPD region chooses the narrowest region with $1-\alpha$ coverage probability. All points in an HPD region have higher posterior density than points outside the region.
- The basic construction:

  Starting from the high point of the posterior density, gradually move a horizontal line down across the density until the posterior probability of θ-values in the region reaches $1-\alpha$.
- For symmetric and unimodal distributions, the HPD interval will be the same as the quantile-based interval. For multimodal distributions, the HPD region may not be a single interval.

High Posterior Density Regions

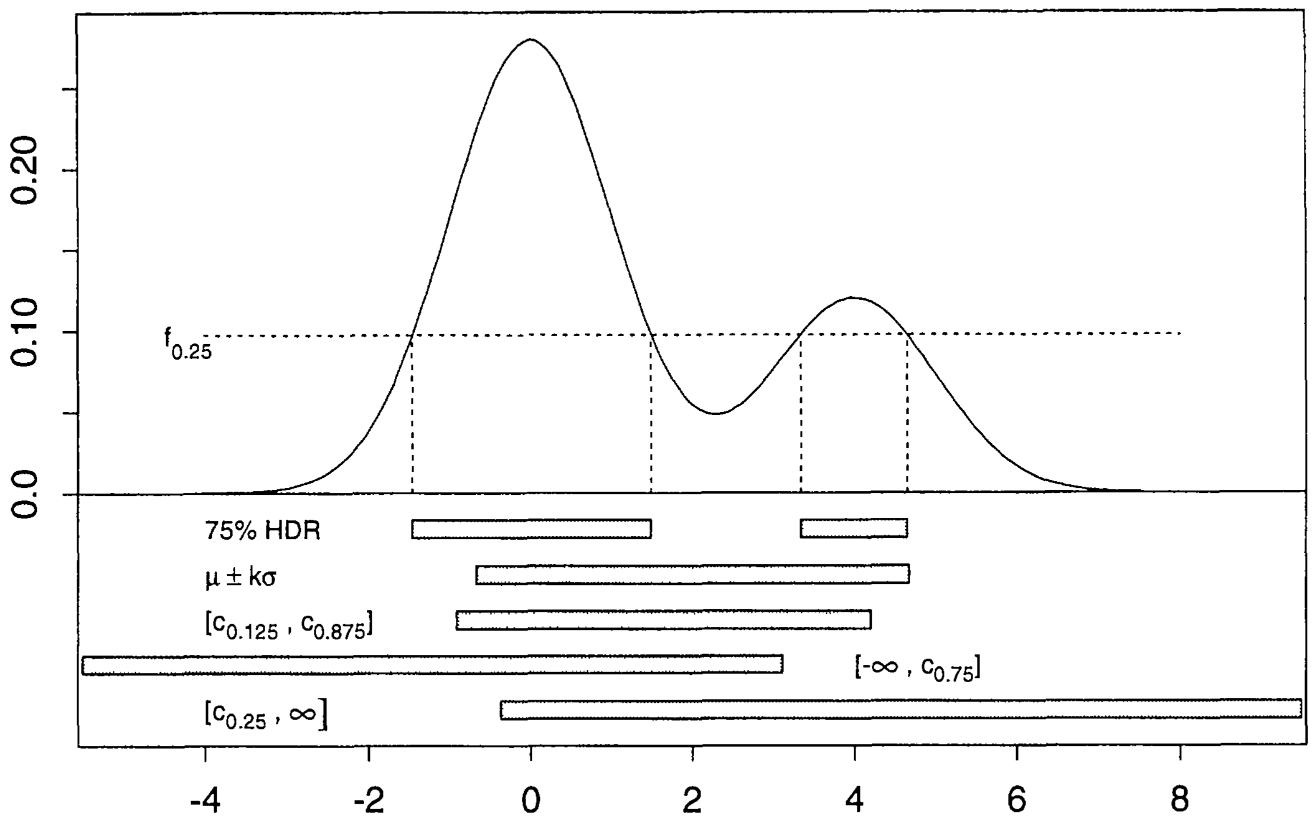


Figure 1. Five Different 75% Probability Regions From a Normal Mixture Density. Here, c_q denotes the qth quantile, μ denotes the mean, and σ denotes the standard deviation of the density.

Source: Hyndman, R. J., *Computing and Graphing Highest Density Regions*, The American Statistician, Vol. 50, No. 2, 1996.

High-Dimensional Bayesian Inference

- Since parameters are considered random in the Bayesian framework, scenarios with even a few parameters can involve high-dimensional multivariate distributions.
- Hyperparameters of the prior can themselves have prior distributions (called **hyperpriors**). Models which have hyperpriors are called **(Bayesian) hierarchical models**.
- Classical methods are often inadequate to deal with high-dimensional problems.
- Markov Chain Monte Carlo methods make much of Bayesian inference possible.