Bayesian Statistics

Chapter 4. Monte Carlo Approximation

Hojin Yang

Department of Statistics
Pusan National University

Introduction

- We will often want to summarize other aspects of a posterior distribution
- For example, we may want to calculate $P(\theta \in A | y_1, \ldots, y_n)$ for arbitrary sets A
- Alternatively, the function of parameters, $|\theta_1 - \theta_2|$, θ_1/θ_2, or $\max\{\theta_1, \theta_2\}$ as well as means and standard deviations for these
- If obtaining exact values for these posterior quantities is difficult or impossible, all of these posterior quantities of interest can be approximated by using the Monte Carlo method

4.1. Monte Carlo Method

- In the last chapter we obtained the posterior distributions

$$\theta_1|\{n_1 = 111, \sum Y_{i,1} = 217\} \sim \text{gamma}(2 + 217, 1 + 111) = \text{gamma}(219, 112)$$

$$\theta_2|\{n_2 = 44, \sum Y_{i,2} = 66\} \sim \text{gamma}(2 + 66, 1 + 44) = \text{gamma}(68, 45)$$

- It was claimed that

$$P(\theta_1 > \theta_2 | \sum_{i=1}^{n_1} Y_{i,1} = 217, \sum_{i=1}^{n_2} Y_{i,2} = 66) = 0.97$$

- How was this probability calculated?

- We modeled θ_1 and θ_2 as conditionally independent given the data
- Hence,

$$\begin{aligned}
&\Pr(\theta_1 > \theta_2 | y_{1,1}, \ldots, y_{n_2,2}) \\
&= \int_0^\infty \int_0^{\theta_1} p(\theta_1, \theta_2 | y_{1,1}, \ldots, y_{n_2,2}) \, d\theta_2 d\theta_1 \\
&= \int_0^\infty \int_0^{\theta_1} \text{dgamma}(\theta_1, 219, 112) \times \text{dgamma}(\theta_2, 68, 45) \, d\theta_2 d\theta_1 \\
&= \frac{112^{219} 45^{68}}{\Gamma(219)\Gamma(68)} \int_0^\infty \int_0^{\theta_1} \theta_1^{218} \theta_2^{67} e^{-112\theta_1 - 45\theta_2} \, d\theta_2 d\theta_1.
\end{aligned}$$

- There are a variety of ways to calculate this integral
  - can be done with pencil using results from calculus
  - can be done with mathematical software packages
- We will use an integration method as the general principles

- The method, known as Monte Carlo approximation, is based on random sampling and its implementation does not require a deep knowledge of calculus or numerical analysis
- Let θ be a parameter of interest
- $y_1, \ldots, y_n$ be the numerical values from $p(y_1, \ldots, y_n|\theta)$
- Suppose we could sample some number S of independent, random θ-values from the posterior distribution, i.e.,

$$\theta^{(1)}, \ldots, \theta^{(S)} \sim p(\theta|y_1, \ldots, y_n)$$

- The empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ would approximate $p(\theta|y_1, \ldots, y_n)$ with increasing S

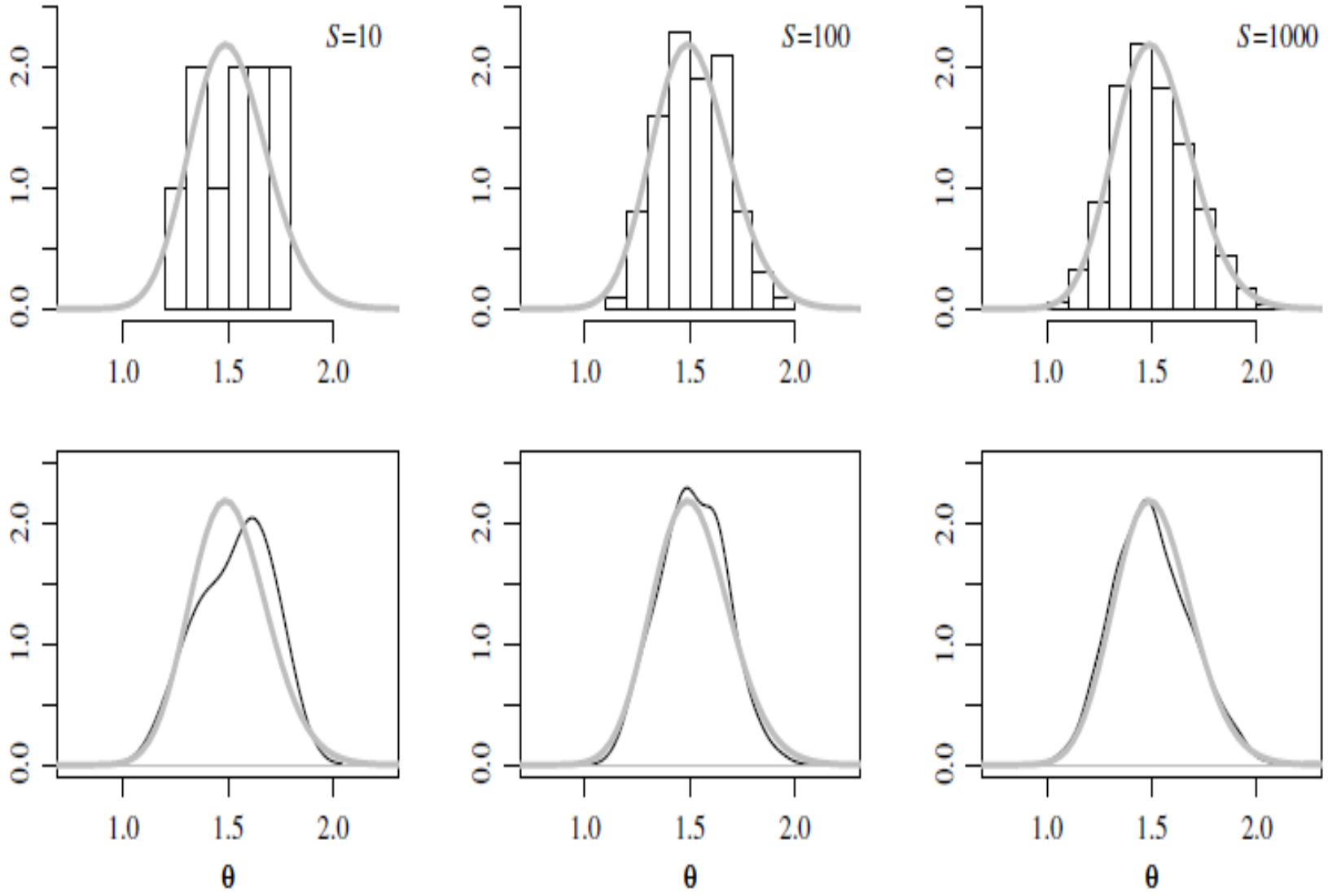
- The empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ is called as a Monte Carlo approximation to $p(\theta|y_1, \ldots, y_n)$
- Many computer languages and computing environments have procedures for simulating this sampling process
- Additionally, let $g(\theta)$ be any function
- The law of large numbers says that if $\theta^{(1)}, \ldots, \theta^{(S)}$ are i.i.d. samples from $p(\theta|\mathbf{y})$, then

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \rightarrow \mathrm{E}[g(\theta)|y_1, \ldots, y_n] = \int g(\theta)p(\theta|y_1, \ldots, y_n)\, d\theta \text{ as } S \rightarrow \infty$$

- This implies that as $S \to \infty$

$$\bar{\theta} = \sum_{s=1}^{S} \theta^{(s)}/S \to \mathrm{E}[\theta|y_1, \ldots, y_n];$$
$$\sum_{s=1}^{S} (\theta^{(s)} - \bar{\theta})^2/(S-1) \to \mathrm{Var}[\theta|y_1, \ldots, y_n];$$
$$\#(\theta^{(s)} \leq c)/S \to \Pr(\theta \leq c|y_1, \ldots, y_n);$$
$$\text{the empirical distribution of } \{\theta^{(1)}, \ldots, \theta^{(S)}\} \to p(\theta|y_1, \ldots, y_n);$$
$$\text{the median of } \{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_{1/2};$$
$$\text{the } \alpha\text{-percentile of } \{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_{\alpha}.$$

- Any aspect of a posterior distribution we may be interested in can be approximated arbitrarily exactly with a large enough Monte Carlo sample



- Figure shows the empirical distribution of the Monte Carlo samples provides close approximation to the true density (gamma(68,45))

Numerical Evaluation

- Suppose $Y_1, \ldots, Y_n | \theta \sim \text{Poi}(\theta)$ and $\theta \sim G(a, b)$
- Observing $Y_1 = y_1, \ldots, Y_n = y_n$,

$$p(\theta|\mathbf{y}) \propto G(a + \sum y_i, b + n)$$

- For the college-educated population in the birthrate example, recall $(a = 2, b = 1)$ and $(\sum y_i = 66, n = 44)$

- Expectation: The posterior mean is $(a + \sum y_i)/(b + n) = 1.51$
- Monte Carlo approximations to this for $S \in \{10, 100, 1000\}$ can be obtained in R

```
a<-2    ;  b<-1
sy<-66  ;  n<-44

theta.mc10<-rgamma(10,a+sy,b+n)
theta.mc100<-rgamma(100,a+sy,b+n)
theta.mc1000<-rgamma(1000,a+sy,b+n)
```

- Results will vary depending on the seed of the random number generator

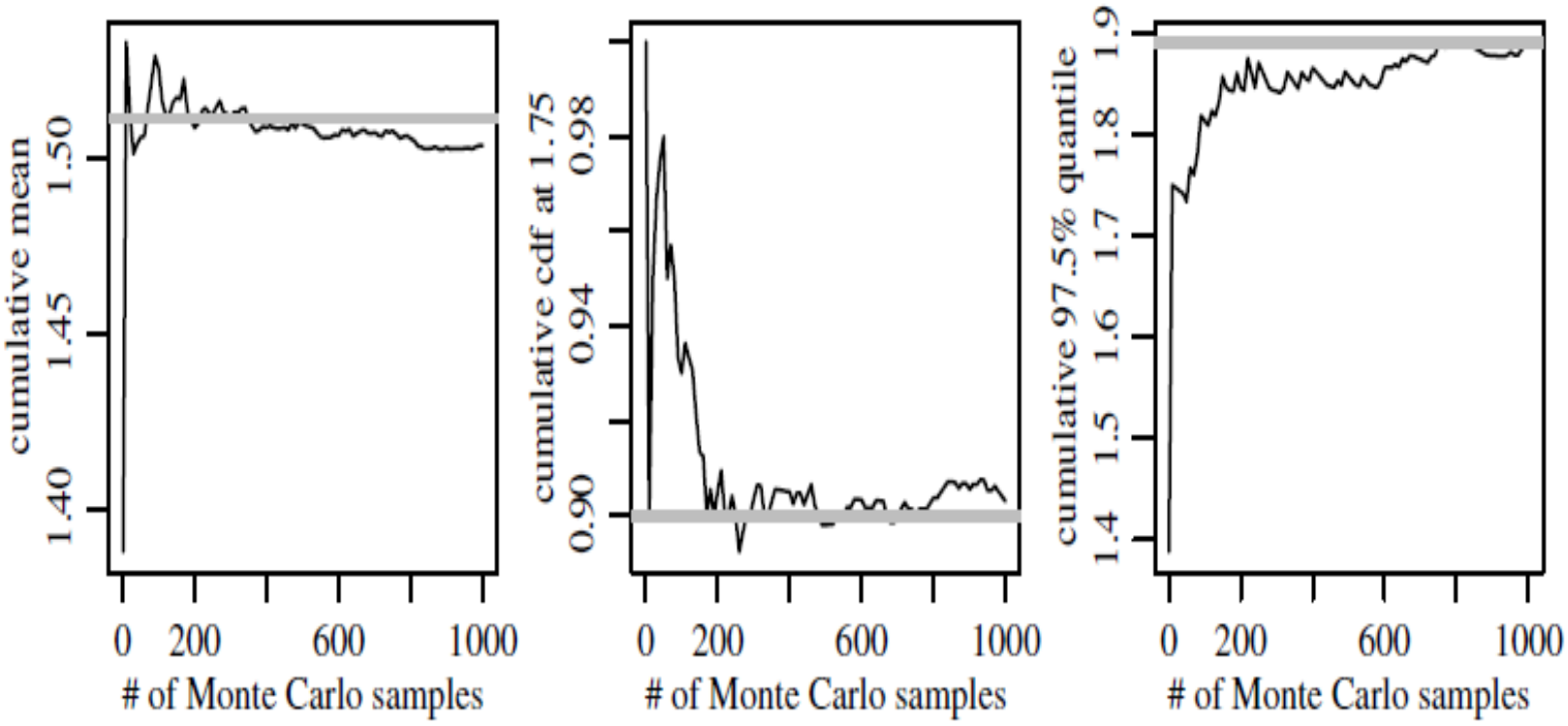
```
> mean(theta.mc10)
[1] 1.532794
> mean(theta.mc100)
[1] 1.513947
> mean(theta.mc1000)
[1] 1.501015
```

- Probabilities: Posterior probability that $\theta < 1.75$ can be obtained from $pgamma(1.75, a + sy, b + n)$, which yields 0.8998

```
> mean(theta.mc10<1.75)
[1] 0.9
> mean(theta.mc100<1.75)
[1] 0.94
> mean(theta.mc1000<1.75)
[1] 0.899
```

- Quantiles: A 95% credible ci can be obtained with $qgamma(c(.025, .975), a + sy, b + n)$, which yields $(1.173, 1.891)$

```
> quantile(theta.mc10, c(.025,.975))
    2.5%    97.5%
1.260291 1.750068
> quantile(theta.mc100, c(.025,.975))
    2.5%    97.5%
1.231646 1.813752
> quantile(theta.mc1000, c(.025,.975))
    2.5%    97.5%
1.180194 1.892473
```



- Figure shows the convergence of the Monte Carlo (MC) estimates, based on cumulative estimates from a sequence of $S = 1000$ samples from the $G(68, 45)$

4.2 Posterior Inference for Arbitrary Functions

- We are interested in the posterior distribution of $g(\theta)$
- For example, we are sometimes interested in the log odds:

$$\text{log odds}(\theta) = \log \frac{\theta}{1-\theta} = \gamma$$

- If we generate a sequence $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ from the posterior distribution of θ,

$$\frac{1}{S}\sum_{s=1}^{S} \log \frac{\theta^{(s)}}{1-\theta^{(s)}} \xrightarrow{\mathcal{P}} E[\log \frac{\theta}{1-\theta} | y_1, \ldots, y_n]$$

- These too can be computed using a Monte Carlo approach

$$\left.\begin{array}{ll}
\text{sample } \theta^{(1)} \sim p(\theta|y_1,\ldots,y_n), & \text{compute } \gamma^{(1)} = g(\theta^{(1)}) \\
\text{sample } \theta^{(2)} \sim p(\theta|y_1,\ldots,y_n), & \text{compute } \gamma^{(2)} = g(\theta^{(2)}) \\
\quad\vdots & \\
\text{sample } \theta^{(S)} \sim p(\theta|y_1,\ldots,y_n), & \text{compute } \gamma^{(S)} = g(\theta^{(S)})
\end{array}\right\} \text{independently}$$

- The sequence $\{\gamma^{(1)},\ldots,\gamma^{(S)}\}$ constitutes S independent samples from $p(\gamma|y_1,\ldots,y_n)$ and as $S \to \infty$

$$\begin{array}{l}
\bar{\gamma} = \sum_{s=1}^{S} \gamma^{(s)}/S \to \mathrm{E}[\gamma|y_1,\ldots,y_n], \\
\sum_{s=1}^{S} (\gamma^{(s)} - \bar{\gamma})^2/(S-1) \to \mathrm{Var}[\gamma|y_1,\ldots,y_n], \\
\text{the empirical distribution of } \{\gamma^{(1)},\ldots,\gamma^{(S)}\} \to p(\gamma|y_1,\ldots,y_n),
\end{array}$$

Example: Log-odds

	$Y = 1$	$Y = 0$	n_{j+}
$X = 1$	441	419	860
$X = 0$	335	676	1011
n_{+j}	776	1095	1871

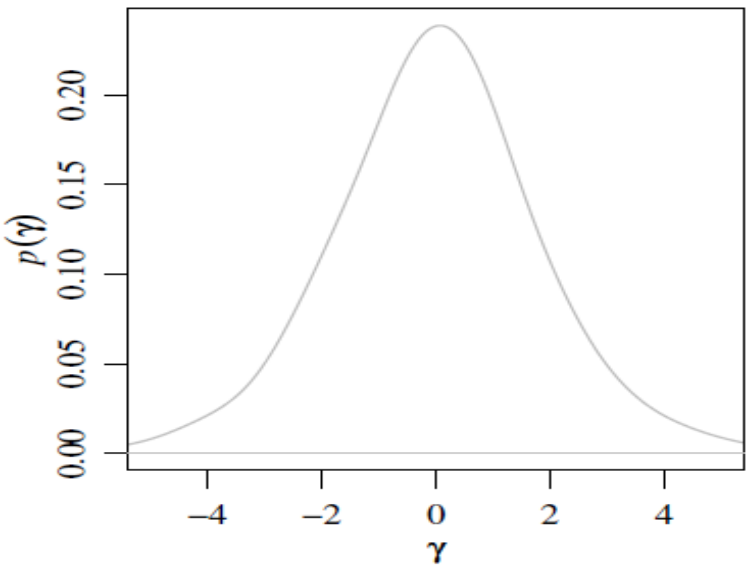
- Of the n = 860 individuals in the religious minority (non-Protestant), y = 441 (51%) said they agreed with the Supreme Court ruling, whereas 353 of the 1011 Protestants (35%) agreed with the ruling.
- Let θ be the population proportion agreeing with the ruling in the minority population ($x = 1$).
- Using a binomial sampling model and a uniform prior distribution, the posterior distribution of θ is beta(442, 420)

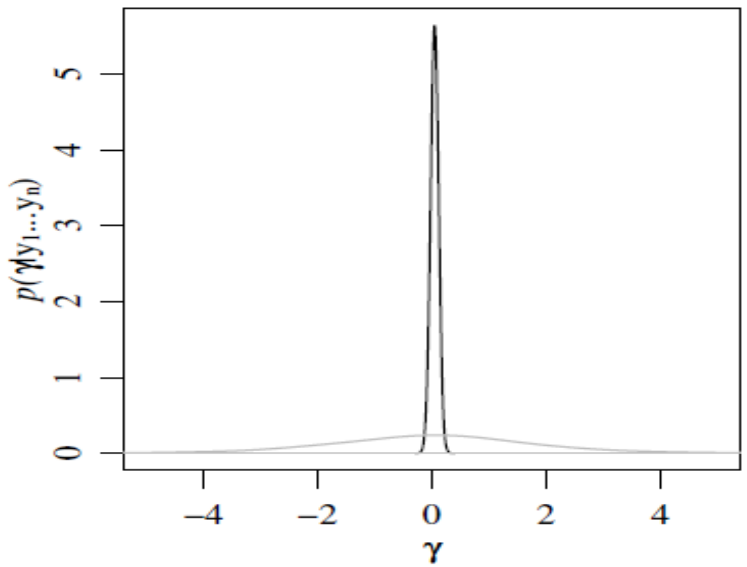
- Using the Monte Carlo algorithm described above, we can obtain samples of the log-odds $\gamma = \log \frac{\theta}{1-\theta}$ from both the prior distribution and the posterior distribution of γ .

```
a<-1 ; b<-1
theta.prior.mc<-rbeta(10000,a,b)
gamma.prior.mc<- log( theta.prior.mc/(1-theta.prior.mc) )

n0<-860-441 ; n1<-441
theta.post.mc<-rbeta(10000,a+n1,b+n0)
gamma.post.mc<- log( theta.post.mc/(1-theta.post.mc) )
```

- Using the density() function in R , we can plot smooth kernel density approximations to these distributions .





Example: Functions of two parameters

- Based on the prior distributions and the data in the birthrate example, the posterior distributions for the two educational groups are

$$\{\theta_1|y_{1,1},\ldots,y_{n_1,1}\} \sim \text{gamma}(219,112) \text{ (women without bachelor's degrees)}$$
$$\{\theta_2|y_{1,2},\ldots,y_{n_2,2}\} \sim \text{gamma}(68,45) \quad \text{(women with bachelor's degrees).}$$

- There are a variety of ways to describe our knowledge about the difference between θ_1 and θ_2

$$P(\theta_1 > \theta_2|\sum_{i=1}^{n_1} Y_{i,1} = 217, \sum_{i=1}^{n_2} Y_{i,2} = 66)$$

- Or in the posterior distribution of θ_1/θ_2

- Both of these quantities can be obtained with Monte Carlo sampling

$$\begin{array}{ll}
\text{sample } \theta_1^{(1)} \sim p(\theta_1 | \sum_{i=1}^{111} Y_{i,1} = 217), & \text{sample } \theta_2^{(1)} \sim p(\theta_2 | \sum_{i=1}^{44} Y_{i,2} = 66) \\
\text{sample } \theta_1^{(2)} \sim p(\theta_1 | \sum_{i=1}^{111} Y_{i,1} = 217), & \text{sample } \theta_2^{(2)} \sim p(\theta_2 | \sum_{i=1}^{44} Y_{i,2} = 66) \\
\vdots & \vdots \\
\text{sample } \theta_1^{(S)} \sim p(\theta_1 | \sum_{i=1}^{111} Y_{i,1} = 217), & \text{sample } \theta_2^{(S)} \sim p(\theta_2 | \sum_{i=1}^{44} Y_{i,2} = 66)
\end{array}$$

- $\{(\theta_1^{(1)}, \theta_2^{(1)}), \ldots, (\theta_1^{(S)}, \theta_2^{(S)})\}$ consists of S independent samples from the joint posterior distribution of θ_1 and θ_2
- This can be used to make Monte Carlo approximations to posterior quantities of interest
- We use

$$\frac{1}{S} \sum_{s=1}^{S} I(\theta_1^{(s)} > \theta_2^{(s)}) \xrightarrow{\mathcal{P}} P(\theta_1 > \theta_2 | \sum_{i=1}^{n_1} Y_{i,1} = 217, \sum_{i=1}^{n_2} Y_{i,2} = 66)$$

- The approximation can be calculated in R .

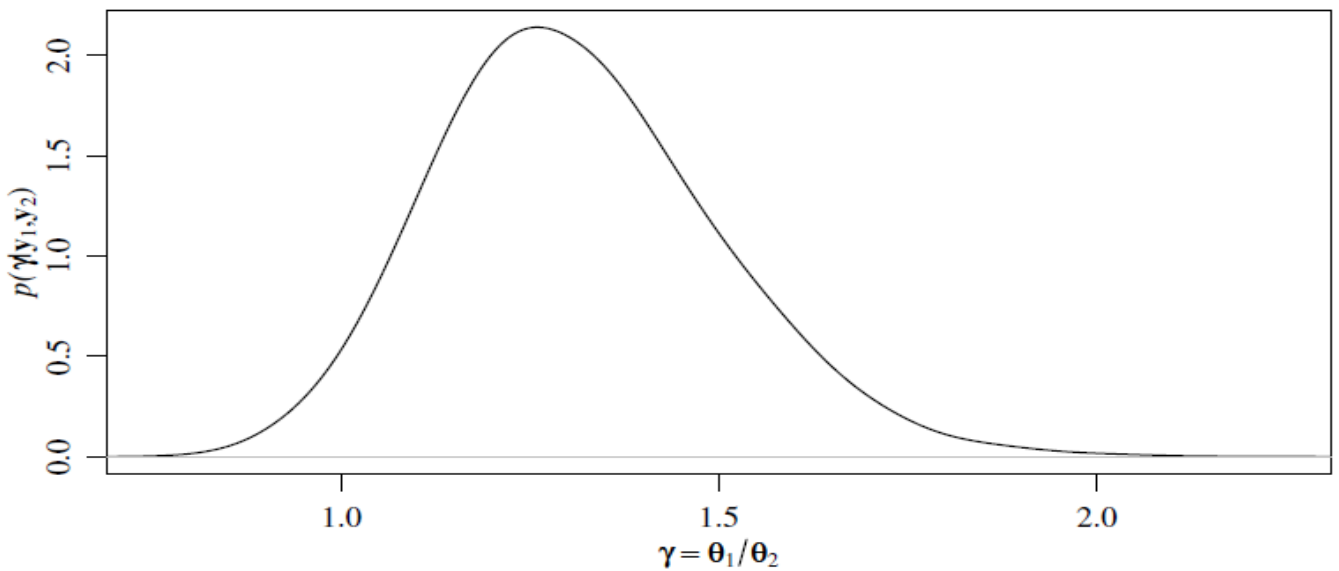
```
> a<-2 ; b<-1
> sy1<-217 ;  n1<-111
> sy2<-66  ;  n2<-44

> theta1.mc<-rgamma(10000,a+sy1, b+n1)
> theta2.mc<-rgamma(10000,a+sy2, b+n2)

> mean(theta1.mc>theta2.mc)

[1] 0.9708
```

- Empirical distribution of $\{\theta_1^{(1)}/\theta_2^{(1)}, \ldots, \theta_1^{(S)}/\theta_2^{(S)}\}$ approximate the posterior distribution of θ_1/θ_2



4.3 Sampling from predictive distributions

- Predictive distribution of a random variable $\tilde{Y}$ is a probability distribution s.t.
  - known quantities have been conditioned on
  - unknown quantities have been integrated out
- Let $\tilde{Y}$ be the number of children of a person who is sampled from the population of women aged 40 with a college degree
  - Sampling model: $p(\tilde{Y} = \tilde{y}|\theta) = p(\tilde{y}|\theta) = \theta^{\tilde{y}} e^{-\theta}/\tilde{y}\,!$
  - Predictive model: $p(\tilde{Y} = \tilde{y}) = \int p(\tilde{y}|\theta)p(\theta)d\theta$
- In the case $\theta \sim G(a, b)$, the predictive distribution is $NB(a, b)$

- A predictive distribution that integrates over unknown parameters but is not conditional on observed data is called a prior predictive distribution
- After we have observed a sample $Y_1, \ldots, Y_n$ from the population, the relevant predictive distribution for a new observation becomes .

$$\begin{aligned}\Pr(\tilde{Y} = \tilde{y} | Y_1 = y_1, \ldots, Y_n = y_n) &= \int p(\tilde{y}|\theta, y_1, \ldots, y_n) p(\theta|y_1, \ldots, y_n)\, d\theta \\ &= \int p(\tilde{y}|\theta) p(\theta|y_1, \ldots, y_n)\, d\theta.\end{aligned}$$

- This is called a posterior predictive distribution, because it conditions on an observed dataset
- In the case of a Poisson model with a gamma prior distribution, the posterior predictive distribution is $NB(a + \sum y_i, b + n)$

- We will be able to sample from $p(\theta|\mathbf{y})$ and $p(\mathbf{y}|\theta)$ when $p(\tilde{y}|\mathbf{y})$ is complicated
- Since $p(\tilde{y}|y_1,\ldots,y_n) = \int p(\tilde{y}|\theta)p(\theta|y_1,\ldots,y_n)d\theta$, we see $p(\tilde{y}|y_1,\ldots,y_n)$ is the posterior expectation of $p(\tilde{y}|\theta)$
- Sample $\{\theta^{(1)},\ldots,\theta^{(S)}\} \sim p(\theta|y_1,\ldots,y_n)$ then, the Monte Carlo allows

$$\frac{1}{S}\sum_{s=1}^{S} p(\tilde{y}|\theta^{(s)}) \xrightarrow{\mathcal{P}} p(\tilde{y}|y_1,\ldots,y_n)$$

- This procedure will work well if $p(y|\theta)$ is discrete and we are interested in quantities that are easily computed from $p(y|\theta)$

- Obtaining these samples can be done quite easily as follows

$$\begin{aligned}
&\text{sample } \theta^{(1)} \sim p(\theta|y_1,\ldots,y_n), \quad \text{sample } \tilde{y}^{(1)} \sim p(\tilde{y}|\theta^{(1)}) \\
&\text{sample } \theta^{(2)} \sim p(\theta|y_1,\ldots,y_n), \quad \text{sample } \tilde{y}^{(2)} \sim p(\tilde{y}|\theta^{(2)}) \\
&\qquad\qquad\qquad\qquad \vdots \\
&\text{sample } \theta^{(S)} \sim p(\theta|y_1,\ldots,y_n), \quad \text{sample } \tilde{y}^{(S)} \sim p(\tilde{y}|\theta^{(S)}) \,.
\end{aligned}$$

- $\{(\theta^{(1)}, \tilde{y}^{(1)}), \ldots, (\theta^{(S)}, \tilde{y}^{(S)})\}$ constitutes S independent samples from the joint posterior distribution of $(\theta, \tilde{y})$
- $\{\tilde{y}^{(1)}, \ldots, \tilde{y}^{(S)}\}$ constitutes S independent samples from the marginal posterior distribution of $\tilde{y}$, which is the posterior predictive distribution

Example: Poisson model

- At the end of Chapter 3 it was reported that the predictive probability that an age-40 woman without a college degree would have more children than an age-40 woman with a degree was 0.48.
- To arrive at this answer exactly we would have to do .

$$\Pr(\tilde{Y}_1 > \tilde{Y}_2 | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66) =$$

$$\sum_{\tilde{y}_2=0}^{\infty} \sum_{\tilde{y}_1=\tilde{y}_2+1}^{\infty} \text{dnbinom}(\tilde{y}_1, 219, 112) \times \text{dnbinom}(\tilde{y}_2, 68, 45)$$

- Alternatively, this sum can be approximated with Monte Carlo sampling
- Since $\tilde{Y}_1$ and $\tilde{Y}_2$ are a posteriori independent, posterior predictive samples from the conjugate Poisson model can be generated .

$$
\begin{aligned}
&\text{sample } \theta^{(1)} \sim \text{gamma}\left(a+\sum y_i, b+n\right), \quad \text{sample } \tilde{y}^{(1)} \sim \text{Poisson}\left(\theta^{(1)}\right) \\
&\text{sample } \theta^{(2)} \sim \text{gamma}\left(a+\sum y_i, b+n\right), \quad \text{sample } \tilde{y}^{(2)} \sim \text{Poisson}\left(\theta^{(2)}\right) \\
&\qquad \vdots \\
&\text{sample } \theta^{(S)} \sim \text{gamma}\left(a+\sum y_i, b+n\right) \quad \text{sample } \tilde{y}^{(S)} \sim \text{Poisson}\left(\theta^{(S)}\right).
\end{aligned}
$$

- Monte Carlo samples from the posterior predictive dists of our two educational groups can be obtained with

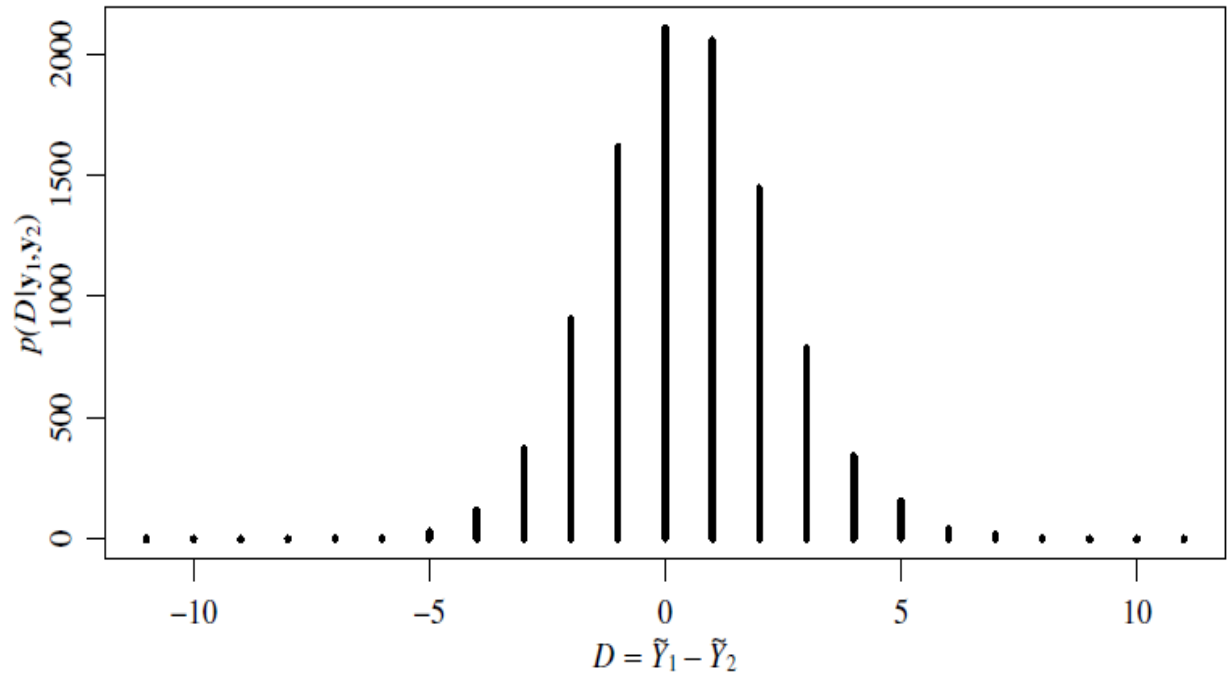
```
> a<-2 ; b<-1
> sy1<-217 ;   n1<-111
> sy2<-66  ;   n2<-44

> theta1.mc<-rgamma(10000,a+sy1, b+n1)
> theta2.mc<-rgamma(10000,a+sy2, b+n2)
> y1.mc<-rpois(10000,theta1.mc)
> y2.mc<-rpois(10000,theta2.mc)

> mean(y1.mc>y2.mc)
[1] 0.4823
```

- Once we have generated these Monte Carlo samples from the posterior predictive distribution, we can use them again to calculate other posterior quantities of interest

- Monte Carlo approximation to the posterior distribution of $\delta = (\tilde{Y}_1 - \tilde{Y}_2)$ would be possible



- The difference in number of children between two individuals, one sampled from each of the two groups