STAT3016/4116/7016:
Introduction to Bayesian Data Analysis

RSFAS, College of Business and Economics, ANU

Monte Carlo Approximation and Sampling Methods

Introduction

The use of conjugate priors leads to posterior distributions with nice formulae for the posterior mean and variance.

However, we are often interested in posterior quantities that are functions of the parameter(s) of interest for which obtaining exact values may be difficult. For example $Pr(\theta \in A|y_1, ..., y_n)$ or $SD|(\theta_1 - \theta_2)|$.

Solution??

analytic solution might be very tedious

but a simulation is easy to conduct, and

with high precision.

Example

In a previous example we looked at birthrates of women with and without bachelor degrees. From the assumed priors and data collected, we found:

$$(\theta_1 | \sum_i Y_{i,1} = 217) \sim \text{Gamma}(219, 112)$$

$$(\theta_2 | \sum_i Y_{i,2} = 66) \sim \text{Gamma}(68, 45)$$

It was claimed that
$Pr(\theta_1 > \theta_2 | \sum_i Y_{i,1} = 217, \sum_i Y_{i,2} = 66) = 0.97$

How was this probability calculated??

We could try and solve the integral
$\int_0^\infty \int_0^{\theta_1} p(\theta_1, \theta_2 | y_{1,1}, ..., y_{n_2,2}) d\theta_2 d\theta_1$
- requires calculus or numerical analysis.

The Monte Carlo Method

Suppose we could sample S independent random θ values from the posterior distribution $p(\theta|y)$:

$$\theta^{(1)}, ..., \theta^{(S)} \overset{\text{iid}}{\sim} p(\theta|y)$$

Then the empirical distribution of the samples $\{\theta^{(1)}, ..., \theta^{(S)}\}$ would approximate $p(\theta|y)$, with the approximation improving with increasing S. ↑ precision ↑

The empirical distribution of $\{\theta^{(1)}, ..., \theta^{(S)}\}$ is known as a *Monte Carlo* approximation to $p(\theta|y)$.

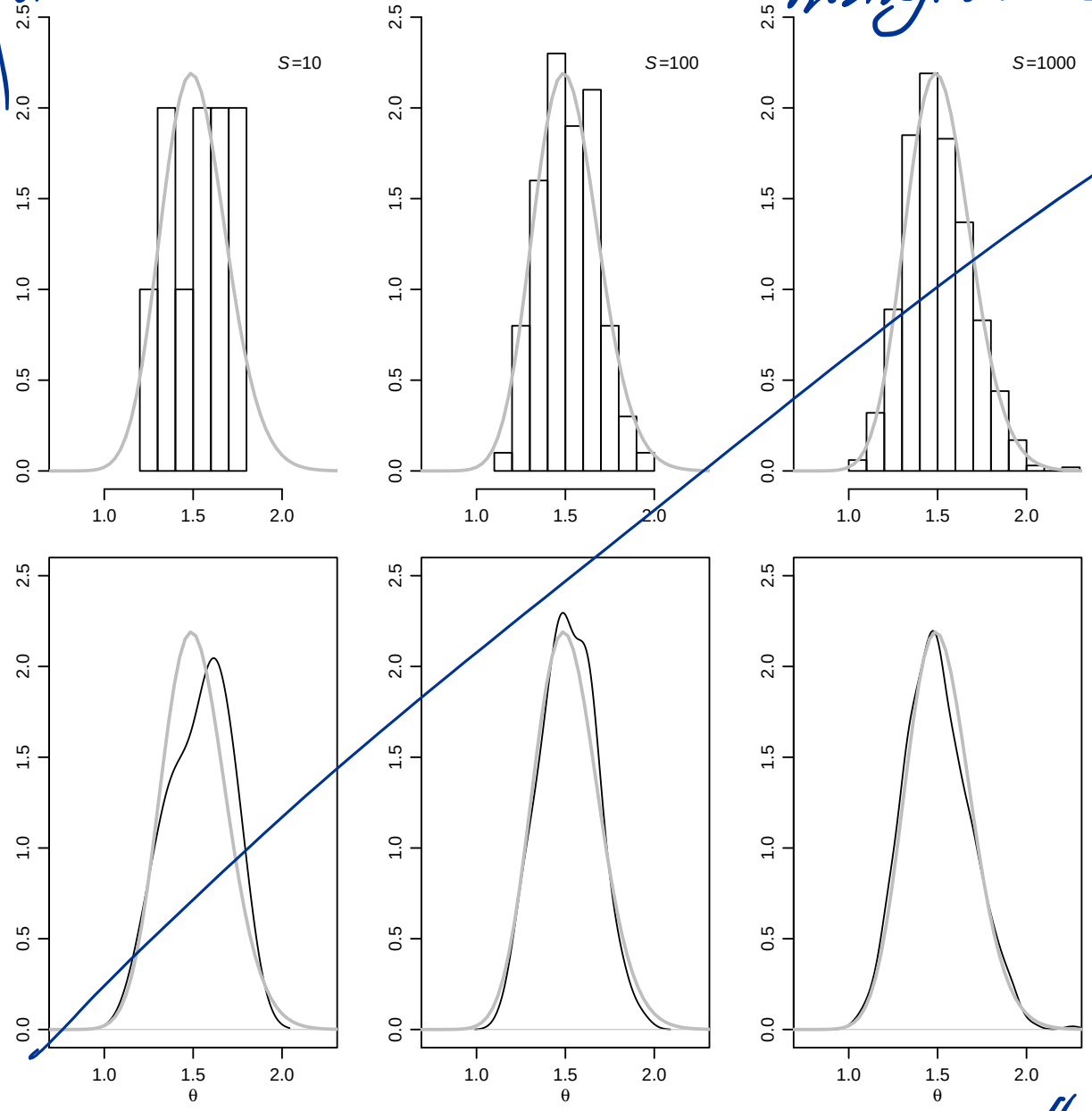
- sometimes $p(\theta|y)$ itself won't be known analytically but MC would still work (by simulation)
  - ↳ pretty cool !

The Monte Carlo Method

light grey line: true distribution

histogram: empirical distribution

PRECISION



# of simulations

The Monte Carlo Method

By the law of large numbers, if $\{\theta^{(1)}, ..., \theta^{(S)}\}$ are i.i.d samples from $p(\theta|y)$: theoretical foundation of MC

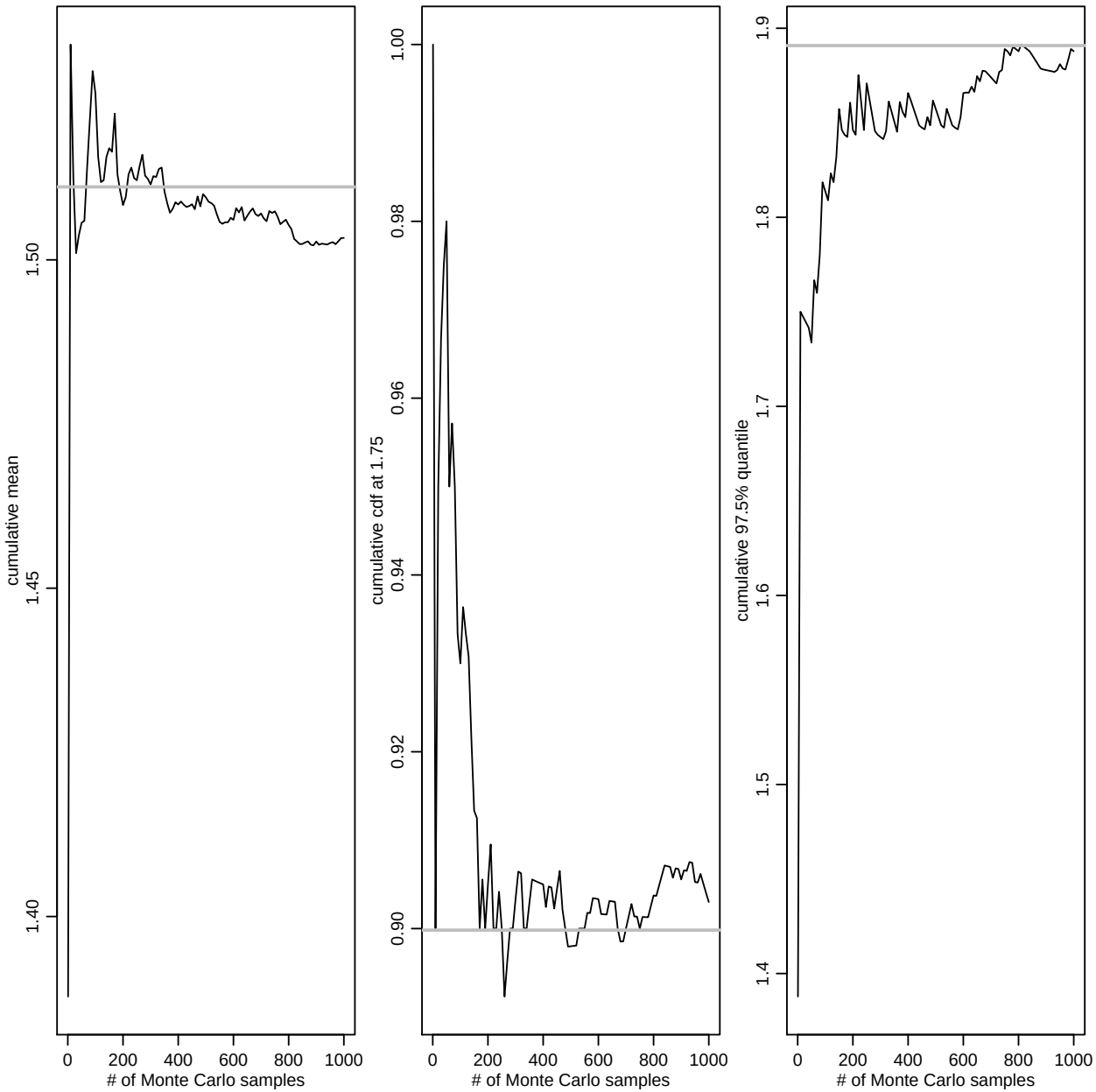
$$\frac{1}{S}\sum_{s=1}^{S}\theta^{(s)} \to E[\theta|y] = \int \theta p(\theta|y)d\theta \text{ as } S \to \infty$$

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

- $\bar{\theta} = \sum_{s=1}^{S}\theta^{(s)}/S \to$?? θ
- $\sum_{s=1}^{S}(\theta^{(s)} - \bar{\theta})^2/(S-1) \to$?? $var(\theta)$
- $\#\ (\theta^{(s)} \leq c)/S \to$?? $P(\theta \leq c)$
- the empirical distribution of $\{\theta^{(1)}, ..., \theta^{(S)}\} \to$?? distribution of θ
- the median of $\{\theta^{(1)}, ..., \theta^{(S)}\} \to$?? median of θ
- the α-percentile of $\{\theta^{(1)}, ..., \theta^{(S)}\} \to$?? the α-percentile of target distribution

Note how you can change the number of MC samples (S) to achieve a desired precision in your interval estimate of the posterior mean of θ.

The Monte Carlo Method



The Monte Carlo Method - for arbitrary functions of θ

Suppose we are interested in a function of θ, say $g(\theta)$.

For example, in the binomial model, we are sometimes interested in the log-odds $= \log \frac{\theta}{1-\theta}$.

Suppose we generate a sequence $\left\{\theta^{(1)}, ..., \theta^{(S)}\right\}$ from the posterior distribution of θ, and compute $\log \frac{\theta^{(s)}}{1-\theta^{(s)}}$ for each draw s=1,...,S.

By the law of large numbers, the average value of the sequence $\log \frac{\theta^{(s)}}{1-\theta^{(s)}}$ converges to $E[\log \frac{\theta}{1-\theta}|y]$.

What about other aspects of the posterior distribution of $\log \frac{\theta}{1-\theta}$?

The Monte Carlo Method - for arbitrary functions of θ

For each value in the sequence $\{\theta^{(1)}, ..., \theta^{(S)}\}$, compute $\gamma^{(s)} = g(\theta^{(s)})$ (s=1,...,S). The sequence $\{\gamma^{(1)}, ..., \gamma^{(S)}\}$ constitutes S independent samples from $p(\gamma|y)$, and so as $S \to \infty$:

- $\bar{\gamma} = \sum_{s=1}^{S} \gamma^{(s)}/S \to$?? $E[\gamma|y]$
- $\sum_{s=1}^{S} (\gamma^{(s)} - \bar{\gamma})^2/(S-1) \to$?? $Var[\gamma|y]$
- the empirical distribution of $\{\gamma^{(1)}, ..., \gamma^{(S)}\} \to$?? the target distribution of $g(\theta|y)$

The Monte Carlo Method - for arbitrary functions of θ

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

n<-860
y<-441
theta.post.mc<-rbeta(10000,a+y,b+n-y)
gamma.post.mc<-log(theta.post.mc/(1-theta.post.mc))

plot(density(gamma.post.mc),xlim=c(-5,5),
   xlab="",lwd=3,
   main="MC approx. to prior and post. log odds")
lines(density(gamma.prior.mc),xlim=c(-5,5),
   xlab="",col="grey",lwd=3)
```

Beta(1,1)

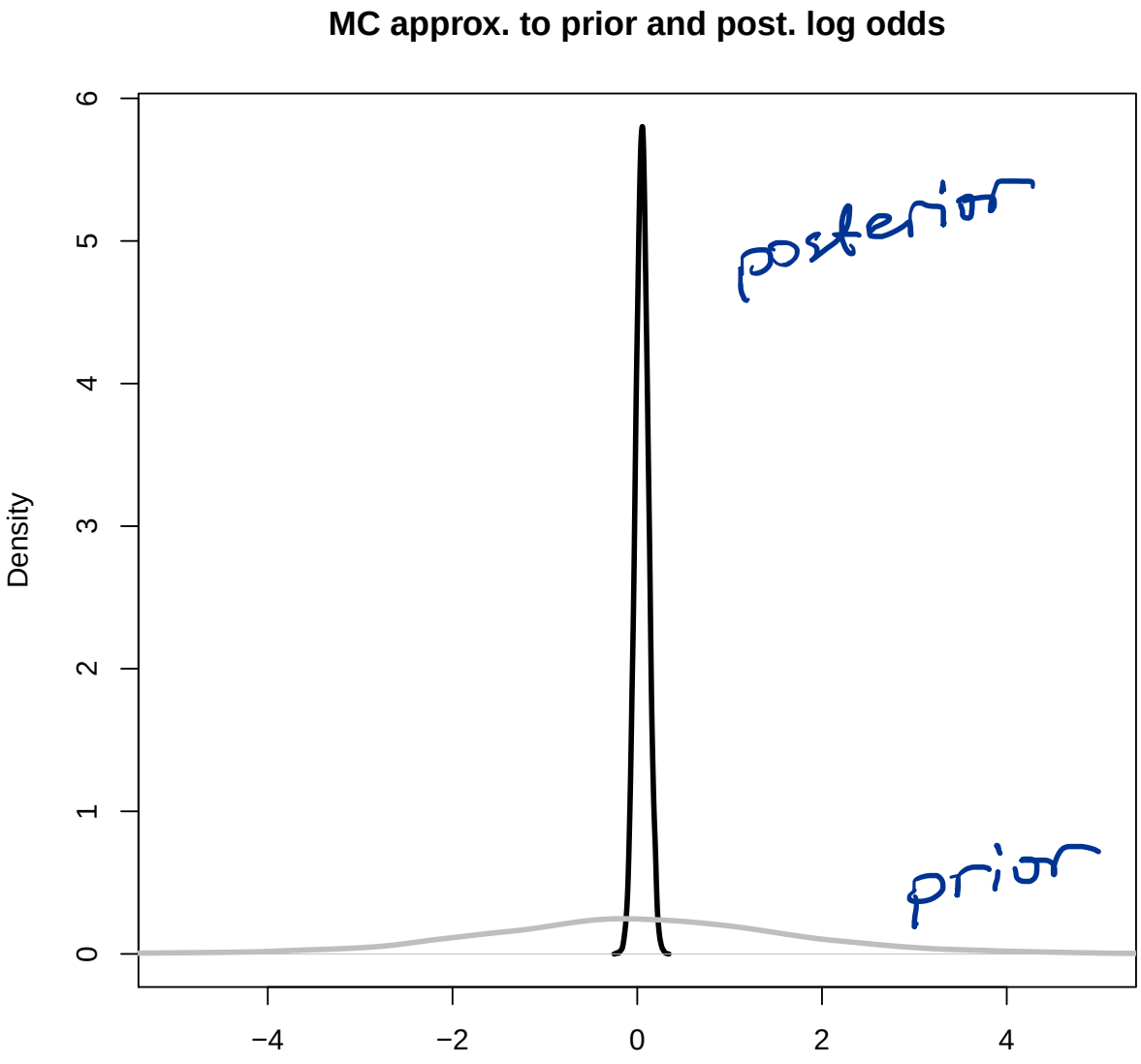
uniform flat prior

logit-transformation

posterior

Beta-Binomial model

The Monte Carlo Method - for arbitrary functions of θ



The Monte Carlo Method - for arbitrary functions of θ

Exercise 1: Let θ_1 be the birthrate for women without bachelor degrees. Let θ_2 be the birthrate for women with bachelor degrees. Previously, we derived the following posterior disitributions based on independent priors:

Poisson-Gamma model.

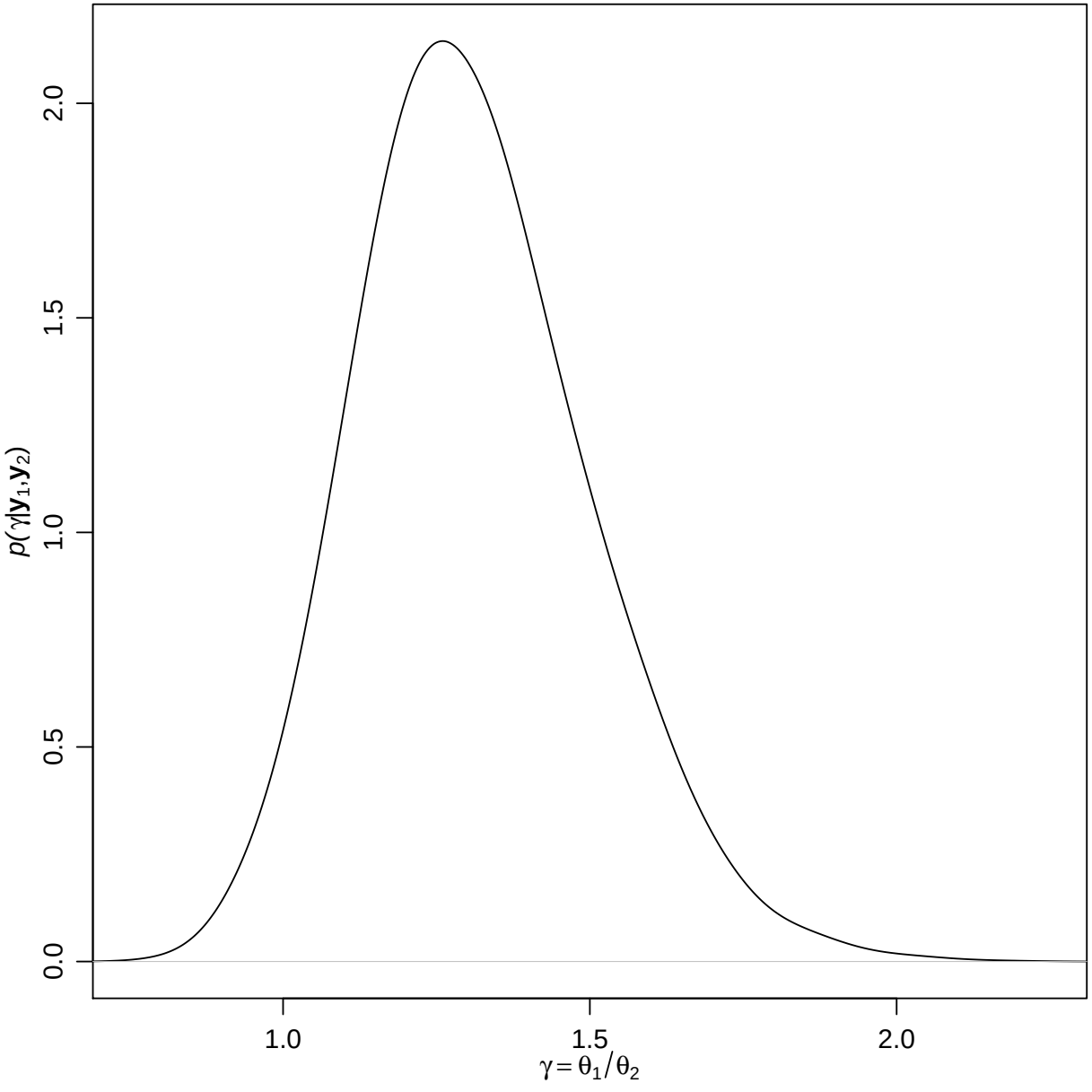
$$\theta_1|\mathbf{y_1} \sim \text{Gamma}(219, 112)$$

$$\theta_2|\mathbf{y_2} \sim \text{Gamma}(68, 45)$$

Find $Pr(\theta_1 > \theta_2|y_1, y_2) \approx \frac{1}{S}\sum_{s=1}^{S} 1(\theta_1^{(s)} > \theta_2^{(s)})$

see Lecture notes from Chapter 3

The Monte Carlo Method - for arbitrary functions of θ



Sampling from predictive distributions

Recall the predictive distribution is a probability distribution for $\tilde{Y}$ such that

- known quantities are conditioned on;
- unknown quantities have been integrated out

Distinguish between the *prior predictive distribution* and the *posterior predictive distribution*.

We use the *prior predictive distribution* to evaluate if we have a reasonable prior. We use the *posterior predictive distribution* to check the sampling model.

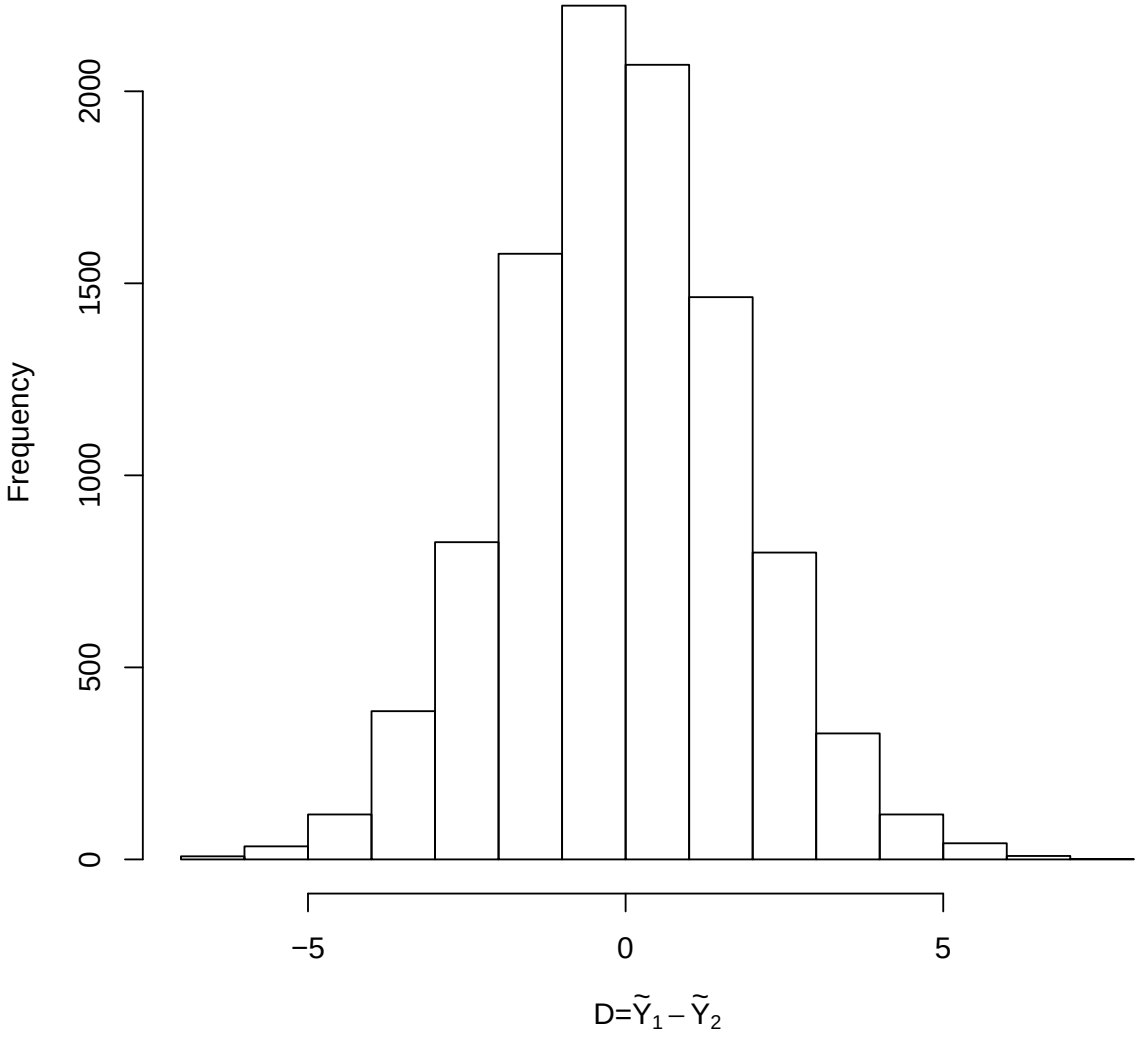
Sampling from predictive distributions

How would you implement the Monte Carlo method to obtain draws from the posterior predictive distribution $p(\tilde{y}|y)$??

Exercise 2: For the birthrate example, find the posterior predictive probability that an age-40 woman without a college degree would have more children than an age-40 woman with a college degree. What would be the analytic solution - - is there a closed form solution?

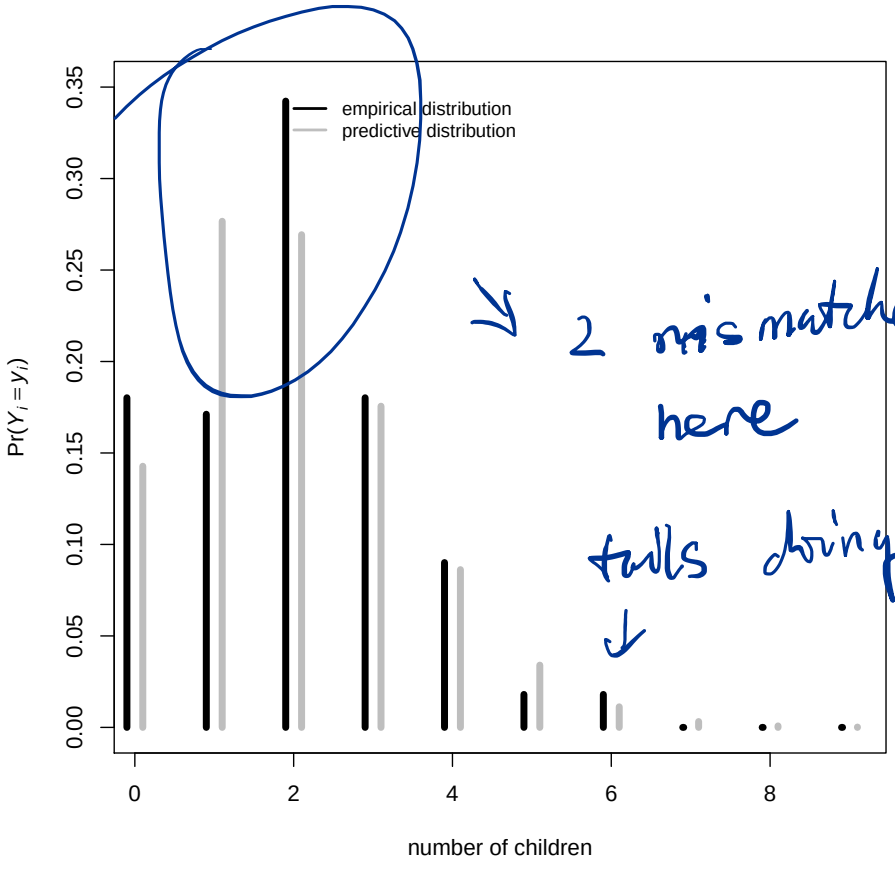
Approximate the probability with Monte Carlo simulations.

Sampling from predictive distributions



Posterior predictive model checking

Assess discrepancies between empirical and predictive distributions. Comment on the plot below (comparison between empirical and posterior predictive distribution of number of children of women without bachelor's degree)



"PPP"
posterior predictive p-value

2 mismatches here

obvious discrepancy.

tails doing all right.

Posterior predictive model checking

If the model fits, the replicated data under the model should look similar to the observed data. In other words, the observed data should look plausible under the posterior predictive distribution. If there is a discrepancy, is it due to sampling variability or sampling model misfit??

that's ok

remodeling?

Technique: draw Monte Carlo samples from the posterior predictive distribution and compare these samples to the observed data. Any systematic differences between the simulations and the observed data indicate potential failings of the model.

test statistics of your choice

- How to measure discrepancy? $\rightarrow$ define a test quantity, eg the ratio of number of women with two children to the number of women with one child

could be sample mean, ratio, etc.

- What size of discrepancy should we be concerned about? $\rightarrow$ posterior predictive p-value (tail area probability)

Posterior predictive model checking

For each $s \in \{1, ..., S\}$:

1. sample $\theta^{(s)} \sim p(\theta | \mathbf{Y} = \mathbf{y_{obs}})$
2. sample $\tilde{\mathbf{Y}}^{(\mathbf{s})} = (\tilde{\mathbf{y}}_{\mathbf{1}}^{(\mathbf{s})}, ..., \mathbf{y}_{n}^{(s)}) \sim p(y | \theta^{(s)})$
3. compute $t(s) = t(\tilde{\mathbf{Y}}^{(\mathbf{s})})$

Posterior predictive model checking

Let y^{rep} denote the replicated data that *could have been* observed. Note the different defintion of $\tilde{y}$ and y^{rep}. We will work with the distribution of y^{rep} given the current state of knowledge, that is, $p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta$.

Classical p-value for the test statistic $T(y)$ is:

$$p_C = Pr(T(y^{rep}) \geq (T(y))|\theta)$$

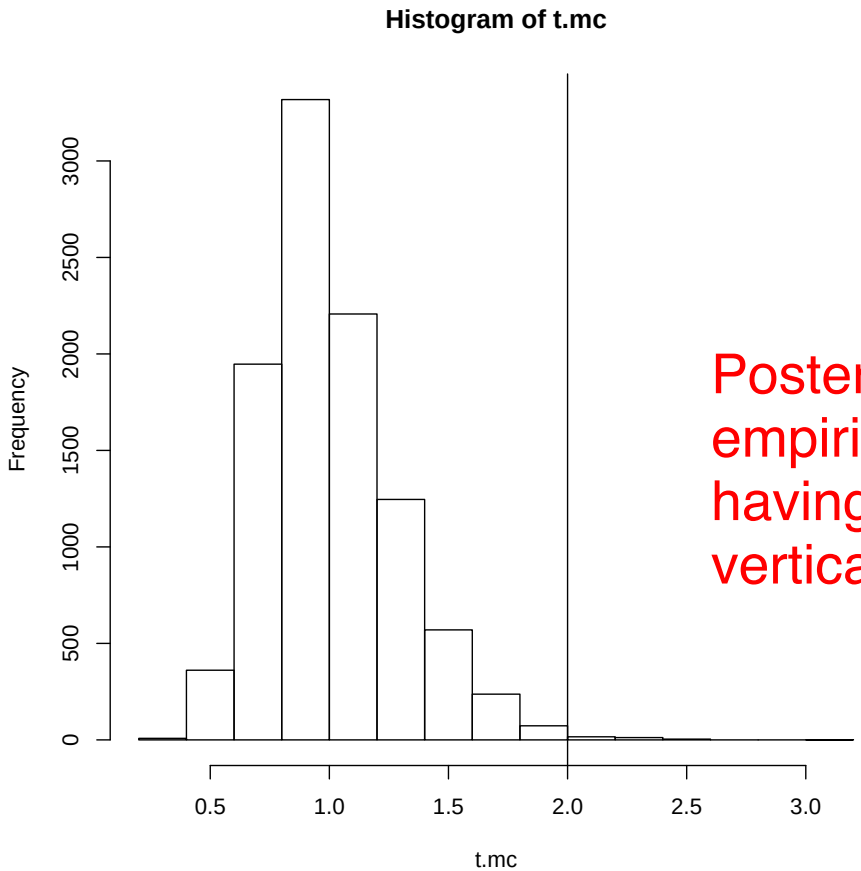
"Bayesian posterior predictive p-value :"

$$p_B = Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y)$$

That is, the probability that the replicated data could be more extreme than the observed data. Note

$$p_B = \int \int I_{T(y^{rep}),\theta \geq T(y,\theta)} p(y^{rep}|\theta)p(\theta|y)dy^{rep}d\theta$$

Posterior predicitive model checking



Posterior predictive distribution of the empirical odds of having 2 children versus having 1 child. Observed odds given by vertical line.

It's a model-checking procedure.
But not a judgement to decide if the model is right or wrong. (More like to seek for space for improvement / fundamental flaw!)

Posterior predictive model checking — exercise

Assume we observe the following sequence of binary outcomes:
1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0.
Assume a uniform prior on the common probability of success $ \theta$.
Perform a posterior predictive check to assess whether the binomial sampling model assumption is appropriate. What test quantity would you use?

Monte Carlo methods to produce independent samples from non-standard distributions

$g(\theta)$: proposal density (some density I can sample from directly)

1. Rejection sampling

Let $p(\theta|y)$ be our target density. (Note: we could also work with the unnormalized density $q(\theta|y)$). Let there be a positive function $g(\theta)$ defined for all θ for which $p(\theta|y) > 0$ that has the following properties:

- We can draw from the probability density proportional to g.
- The importance ratio $\frac{p(\theta|y)}{g(\theta)}$ must have a known bound, that is $\frac{p(\theta|y)}{g(\theta)} \leq M$, for a known constant M.

important ratio

$p(\theta|y) \leq M g(\theta)$

↓ upper bound

Monte Carlo methods to produce independent samples from non-standard distributions

Rejection sampling algorithm:

1. Sample a candidate θ_c at random from the probability density proportional to $g(\theta)$.
2. Calculate acceptance probability α for θ_c.

$$\alpha = \frac{p(\theta_c|y)}{Mg(\theta_c)}$$

3. Draw a value u from the U(0,1) distribution
4. *Accept* θ_c as a draw from $p(\theta|y)$ if $\alpha \geq u$. Otherwise reject θ_c and go back to step 1.

Rejection sampling

Used as a fast method to sample from univariate distributions and truncated multivariate distributions.

A version of rejection sampling forms the basis for the Metropolis-Hastings algorithm (Hoff Ch 10) that we will use later to sample from posteriors without knowing the normalizing constant.

Rejection sampling - Example 1

Suppose our target density is the triangle density

$$f(x) = \begin{cases} 8x & \text{if } 0 \leq x \leq 0.25 \\ \frac{8}{3} - \frac{8}{3}x & \text{if } 0.25 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

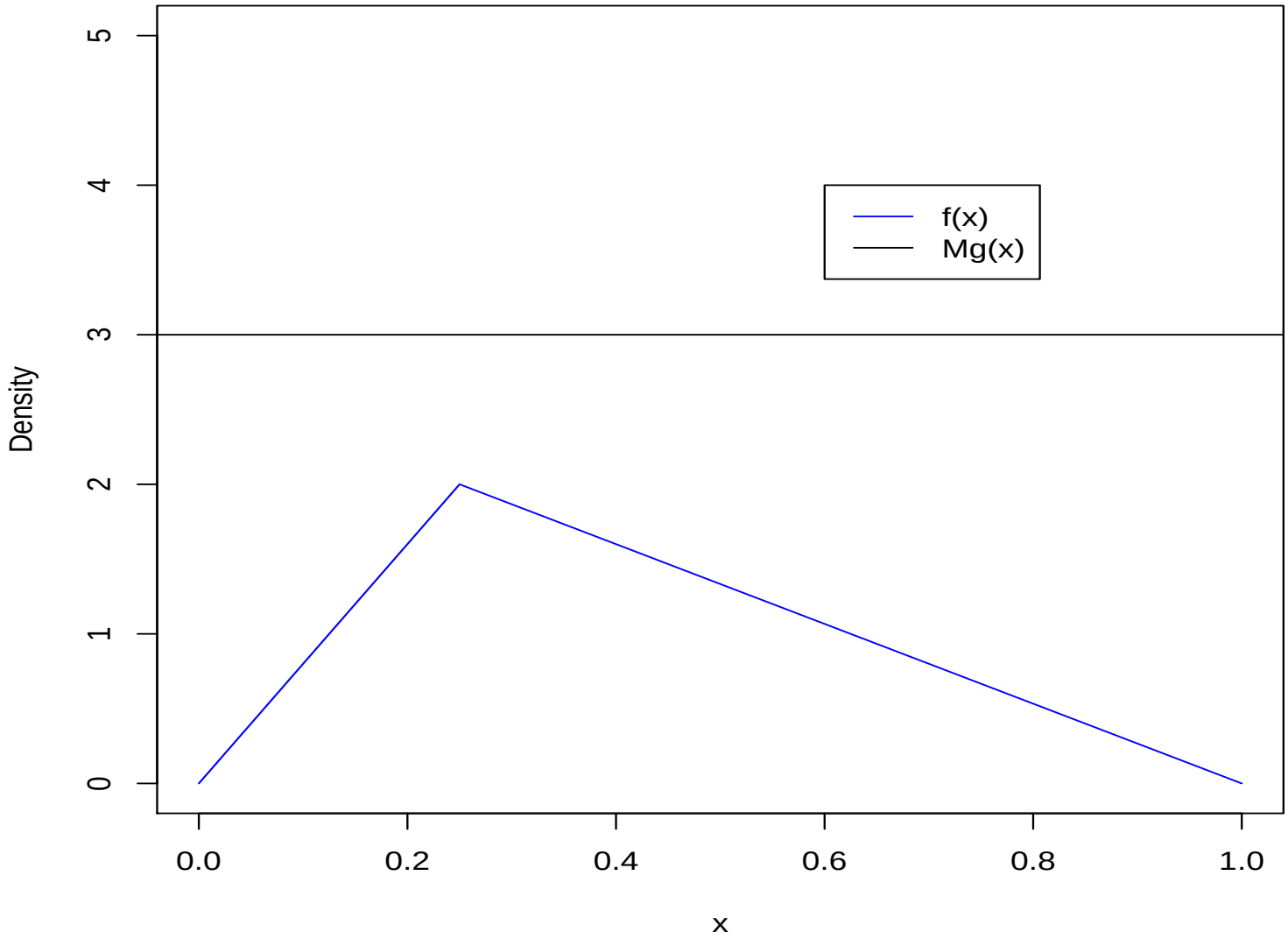
Rejection sampling - Example 1

```
#target density
t.x<-function(x){
if(x>=0&&x<0.25)
8*x
else if(x>=0.25&&x<=1)
8/3-8/3*x
}
#candidate density
g.x<-function(x){
if(x>=0&&x<=1)
     1
     else 0

     }

M<-3
```

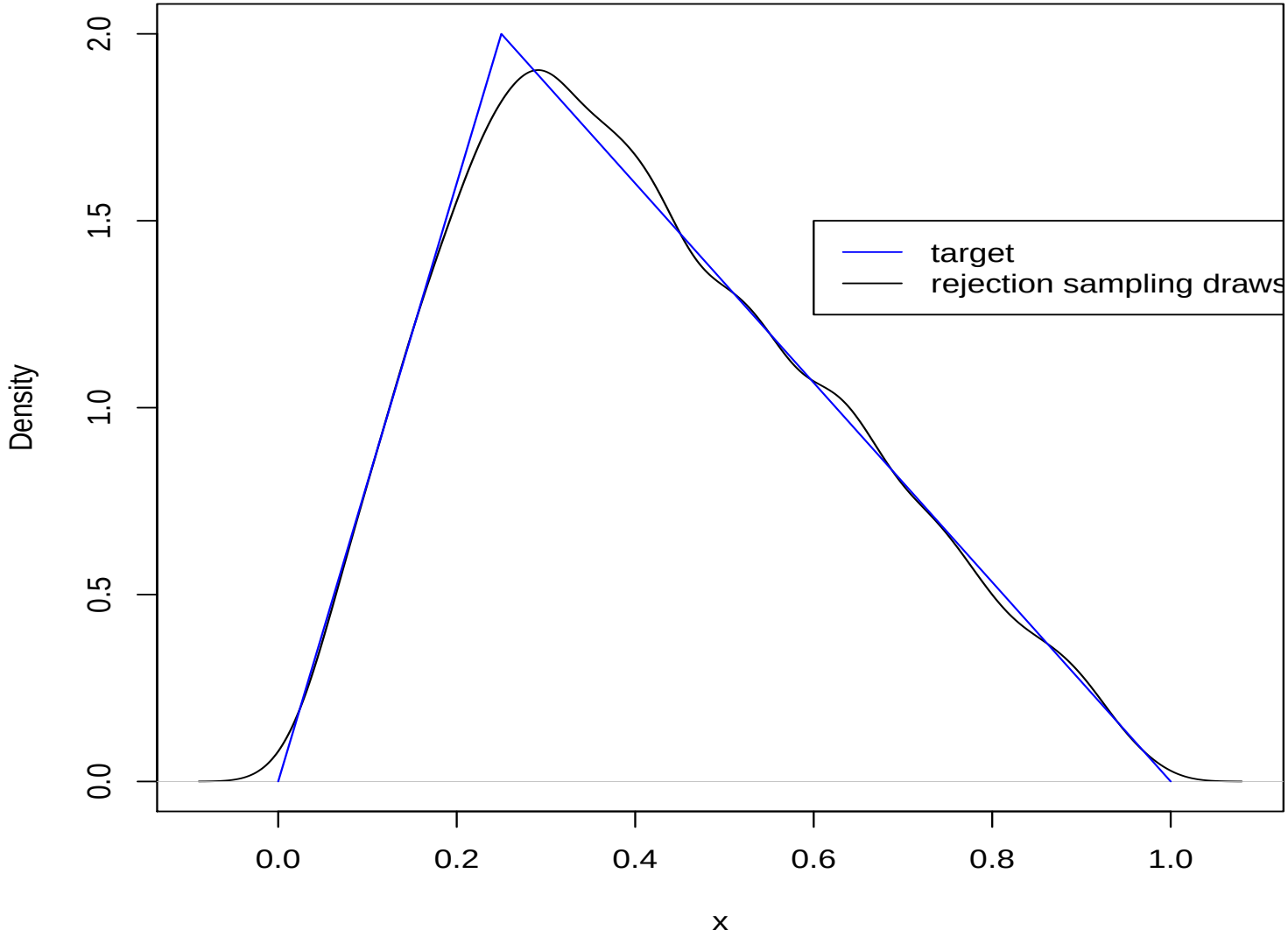
Rejection sampling - Example 1



Rejection sampling - Example 1

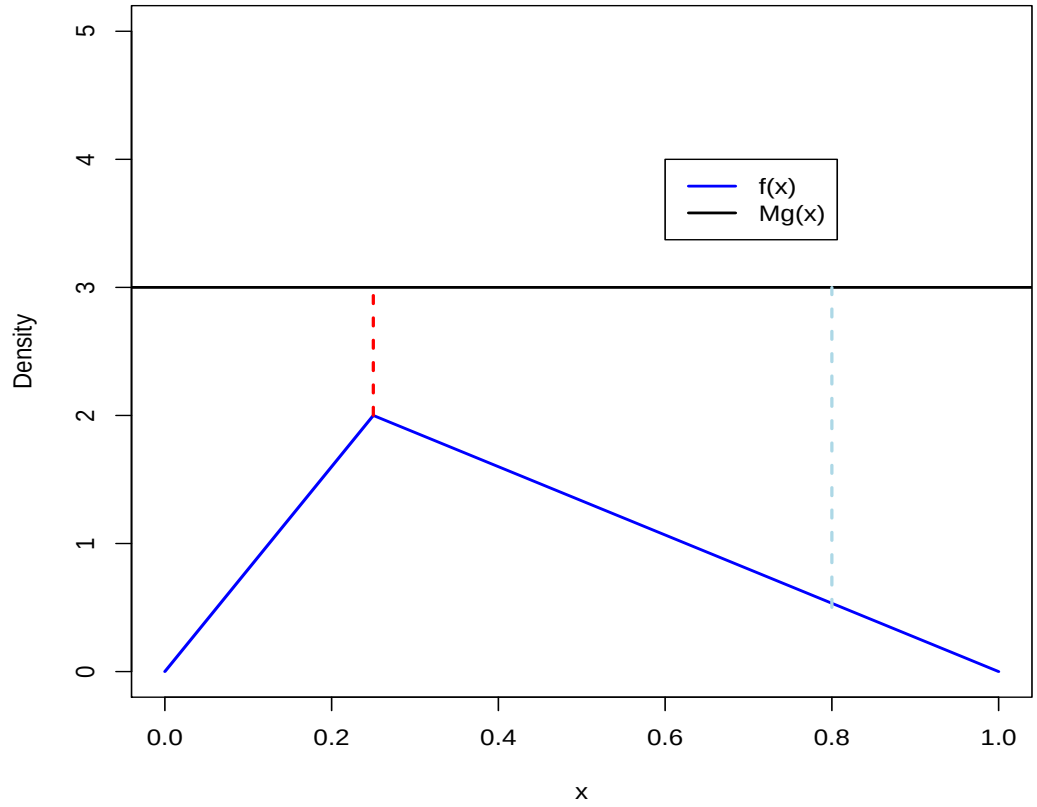
```
m<-10000
n.draws<-0
draws<-c()
x.grid<-seq(0,1,by=0.01)
while(n.draws<m){
  x.c<-runif(1,0,1)
  accept.prob<-t.x(x.c)/(M*g.x(x.c))
  u<-runif(1,0,1)
  if(accept.prob>=u){
   draws<-c(draws,x.c)
   n.draws<-n.draws+1
  }
}
```

Rejection sampling - Example 1



Rejection sampling - Example 1

Why does it work?



The difference between $f(x)$ and $Mg(x)$ at places with higher density (ie at around x=0.25) is smaller than at places with lower density (ie around x=0.8), so the acceptance probability at x=0.25 is higher and more draws of x=0.25 are accepted.

Rejection sampling

$$\text{inverse logit}(\alpha+\beta x_j) = \frac{e^{\alpha+\beta x_j}}{1+e^{\alpha+\beta x_j}}$$

$$p(\alpha,\beta|y_1,\dots,y_n) \propto p(\alpha)p(\beta)\prod_{j=1}^{n} p(y_j|\alpha,\beta)$$

$$= t_4(0,2^2)\,t_4(0,1^2)\prod_{j=1}^{n} \text{inverse logit}(\dots)^{y_j}(1-\text{inverse}\dots)$$

There are an infinite number of candidate densities $g(\theta)$ and constants M that we can use.

The only difference between them is computation time.

If $g(\theta)$ is significantly different in shape than $p(\theta|y)$ or if $Mg(\theta)$ is significantly greater than $f(x)$, then more of our candidate draws will be rejected.

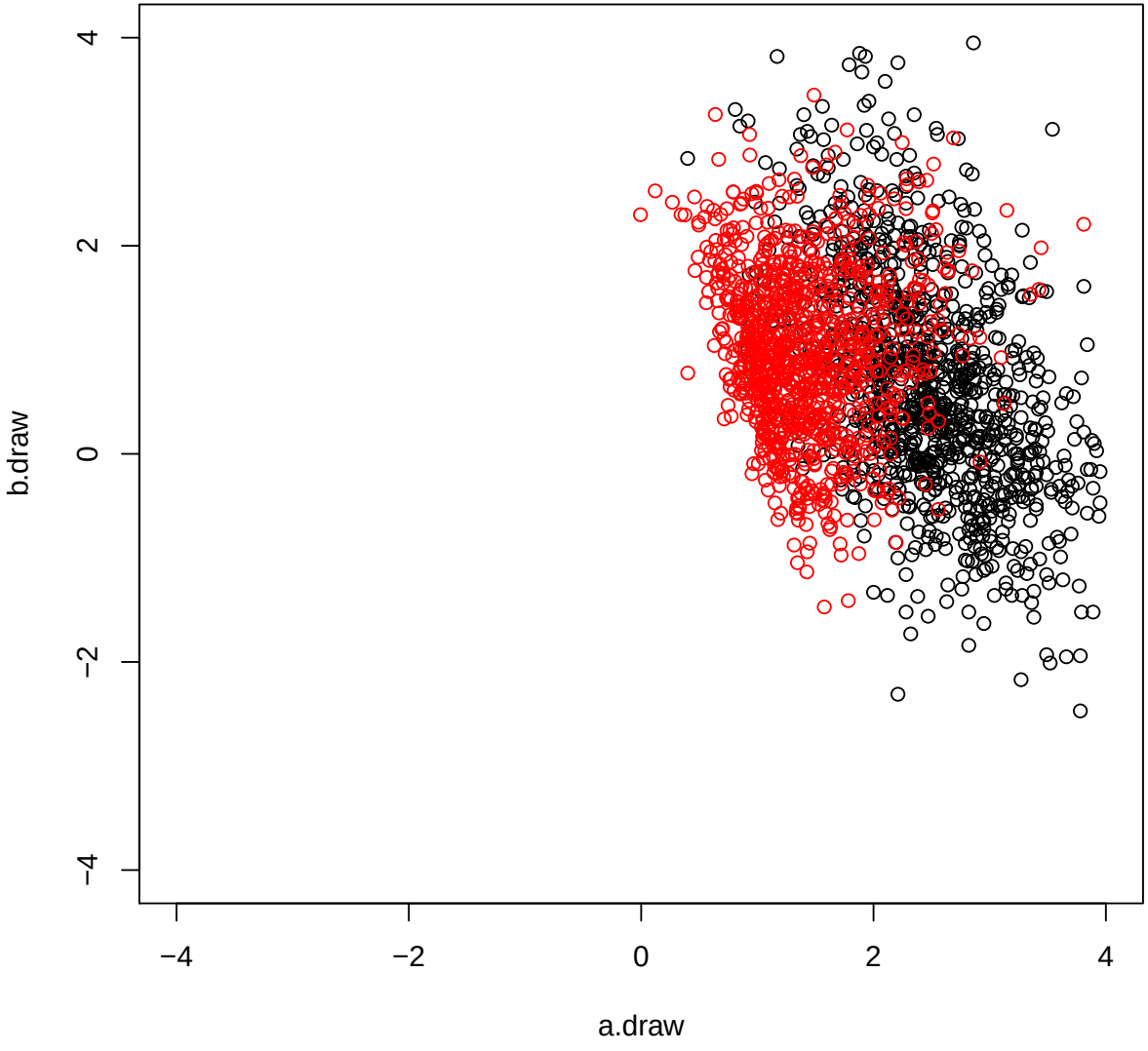
If $p(\theta|y) = Mg(\theta)$, then all our draws will be accepted.

Rejection sampling - Example 2

Exercise 3: Consider the model $y_j \sim \mathrm{Bin}(n_j, \theta_j)$ where $\theta_j = \mathrm{logit}^{-1}(\alpha + \beta x_j)$, for (j =1....J) and with independent prior distributions, $\alpha \sim t_4(0, 2^2)$ and $\beta \sim t_4(0, 1)$. Suppose J=10, the x_j values are randomly drawn from a U(0,1) distribution, and $n_j \sim \mathrm{Pois}^+(5)$ where Poisson is the Poisson distribution restricted to positive values.

- Sample a dataset at random from the model
- Use rejection sampling to get 1000 independent posterior draws from (α, β).

Rejection sampling - Example 2



Monte Carlo methods to produce independent samples from non-standard distributions

2. Importance sampling

(Related to rejection sampling and a precursor to the Metropolis algorithm (an MCMC method)).

We are interested in $E[h(\theta)|y]$ but we cannot generate random draws of θ from $p(\theta|y)$. Let $g(\theta)$ be a probability density from which we can generate random draws, and let $q(\theta|y)$ be the unnormalized density of $p(\theta|y)$. We can write:

both divide & multiply by $g(\theta)$

$$E[h(\theta)|y] = \frac{\int h(\theta)q(\theta|y)d\theta}{\int q(\theta|y)d\theta} = \frac{\int [h(\theta)q(\theta|y)/g(\theta)]g(\theta)d\theta}{\int [q(\theta|y)/g(\theta)]g(\theta)d\theta}$$

(Note:Unlike rejection sampling, importance sampling is not a method to sample from $p(\theta|y)$, but a method to compute $E[h(\theta)|y]$ (a posterior quantity).

Draw $\theta^{(1)} \dots \theta^{(S)}$ from $g(\theta)$

$\downarrow$

$h(\theta^{(1)}) \dots h(\theta^{(S)})$

$$\frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}) = E_g[h(\theta)]$$

expectation under function $g(\cdot)$

extra step

$E_{p(\theta|y)}[h(\theta)|y]$ → goal

Monte Carlo methods to produce independent samples from non-standard distributions

2. Importance sampling

which can be estimated using S draws $\theta^1, ..., \theta^S$ from $g(\theta)$ by the expression:

$$\frac{\frac{1}{S}\sum_{s=1}^{S} h(\theta^s) w(\theta^s)}{\frac{1}{S}\sum_{s=1}^{S} w(\theta^s)}$$

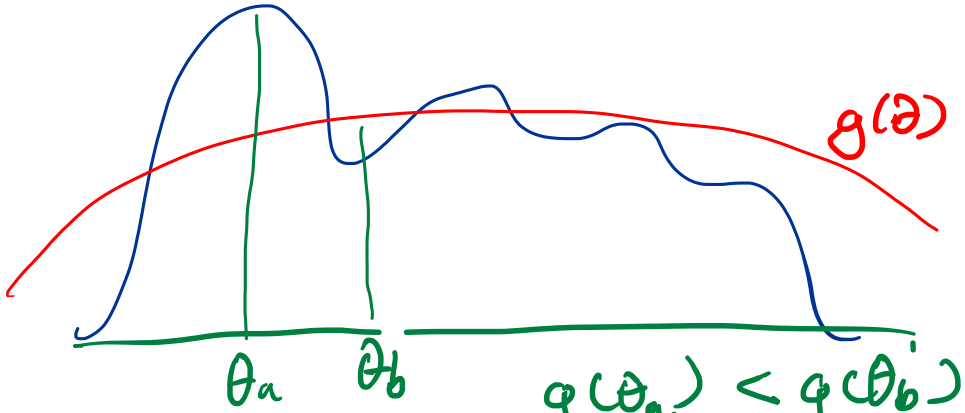
where the factors

$$w(\theta^s) = \frac{q(\theta^s|y)}{g(\theta^s)}$$

are called the *importance weights* or *importance ratios*.

Ideally choose $g(\theta)$ so that $\frac{hq}{g}$ is roughly constant.

Sps $p(\theta|y)$ / $g(\theta|y)$



$g(\theta_a) < g(\theta_b)$

The intuitive $p(\theta_a|y) > p(\theta_b|y)$

without importance ratio
θ_b will be over represented than relative to θ_a.

explanation of weights (importance ratio) is to ~~weight the posterior~~ (?) balance the over representiness & underrepresentiness of target quantity (?)

Monte Carlo methods to produce independent samples from non-standard distributions

2. Importance sampling

Examine the distribution of sampled importance weights to assess the precision of your estimate using importance sampling.

Importance sampling is sometimes used to obtain a starting point for an MCMC algorithm and when considering mild changes to the posterior distribution (eg replacing a normal distribution with a t-distribution).

Importance Sampling - Example

$p(\theta|y) \sim t_3$
$g(\theta) \sim N(0,1)$

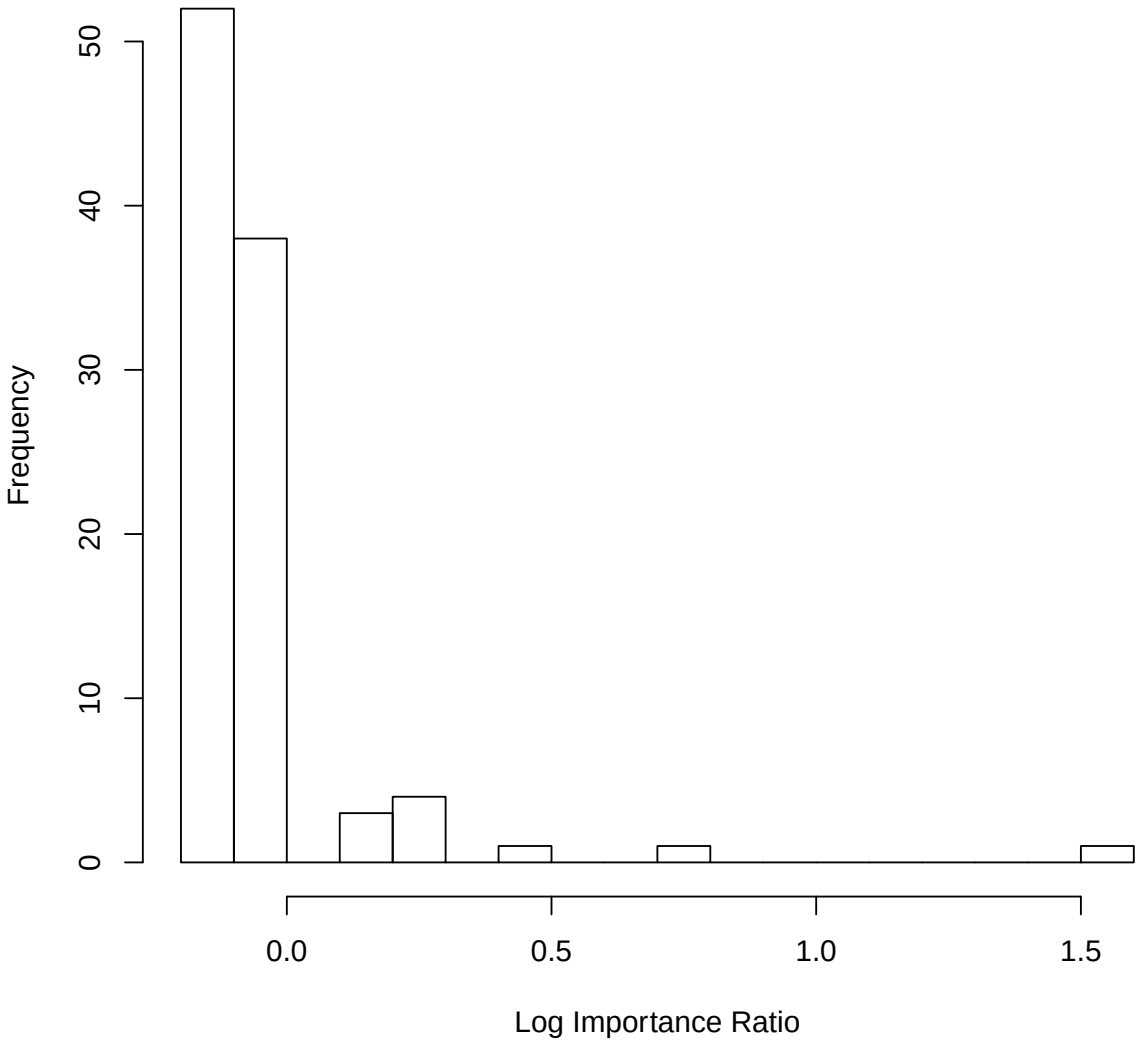
Exercise 4: Consider a univariate posterior distribution $p(\theta|y)$ which we wish to approximate and then calculate moments of using importance sampling from an unnormalised density $g(\theta)$. Suppose the posterior distribution is a t_3 distribution, and the approximation is N(0,1).

(a) Draw a sample of size S=100 from the approximate density and compute the importance ratios. Plot a histogram of the log importance ratios.

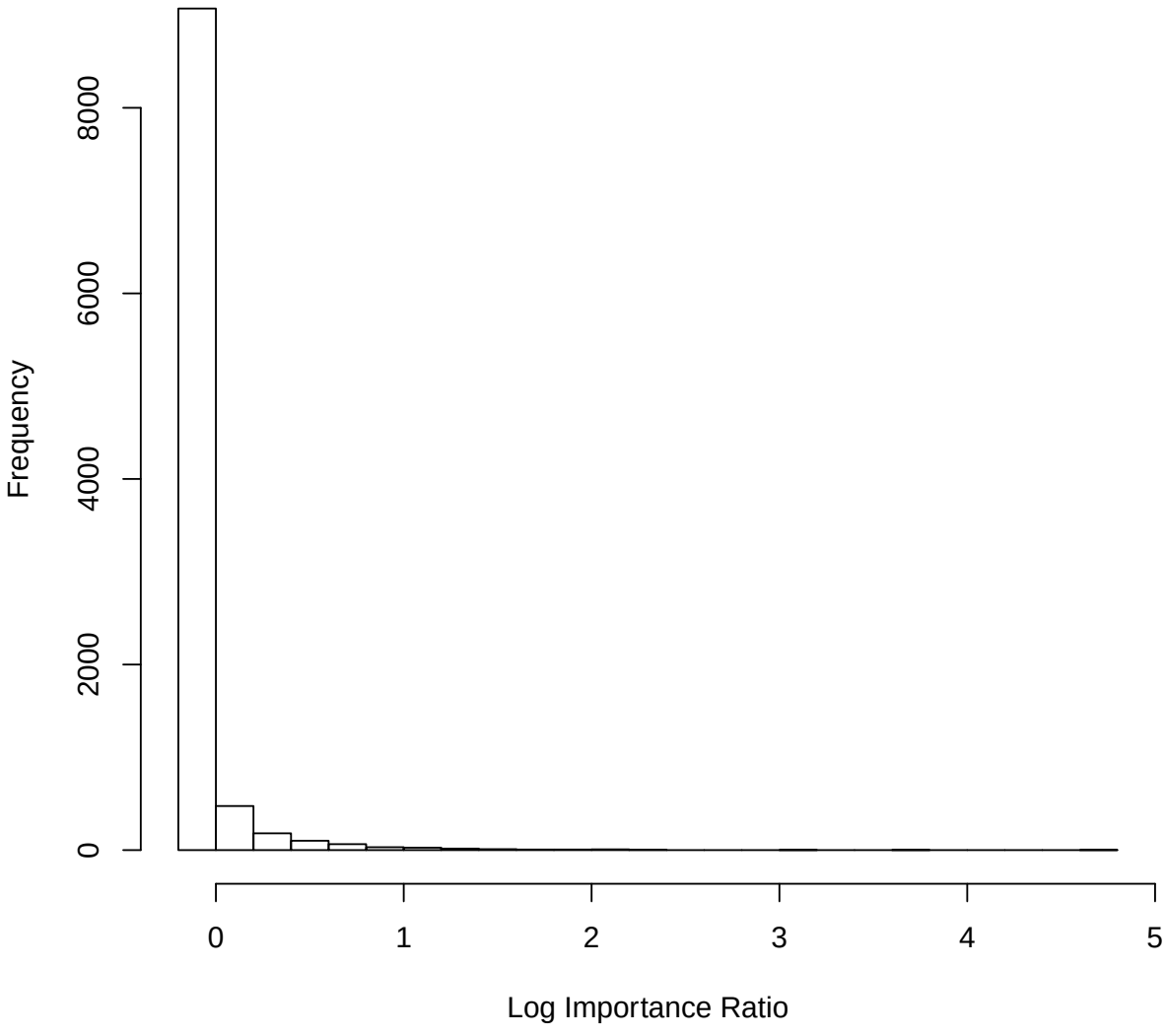
(b) Estimate $E[\theta|y]$ and $Var[\theta|y]$ using importance sampling. Compare to the true values.

(c) Repeat (a) and (b) for $S = 10,000$. Explain why the estimates of $Var[\theta|y]$ are systematically low.

Importance sampling - Example



Importance sampling - Example



Coming up.....

Rejection sampling and importance sampling are Monte Carlo methods.

Markov Chain Monte Carlo (MCMC) methods are better adapted to high-dimensional complex distributions but produce dependent samples. We will begin to look at MCMC methods in Hoff Chapter 6.