GPH-GU2372/3372
Applied Bayesian Analysis in Public Health

Lecture 4: Monte Carlo approximation

Hai Shu, PhD

10/03/2022

Monte Carlo methods: Topics

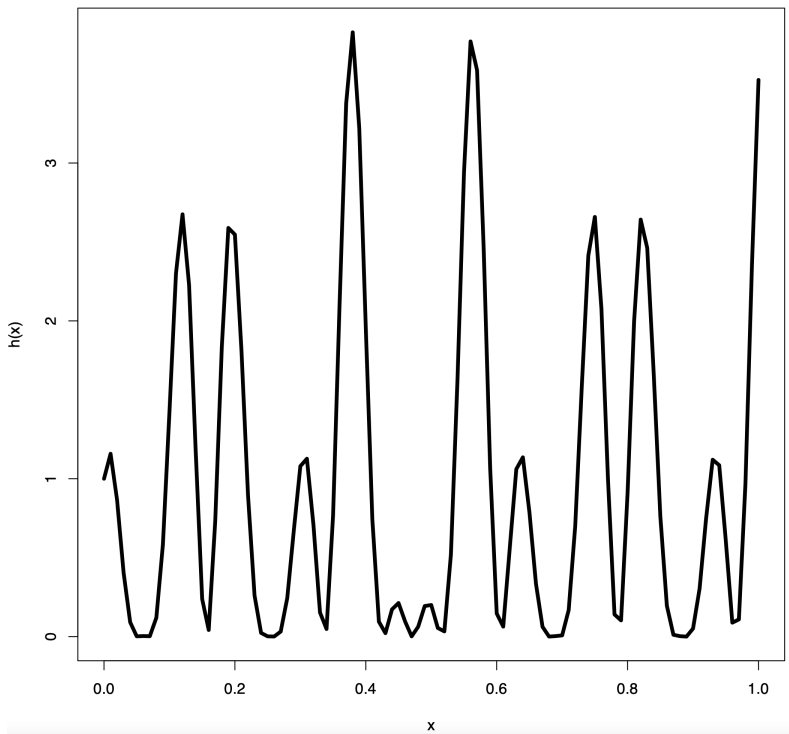
- Introduction to Monte Carlo methods
- Sampling from predictive distribution
- Posterior predictive model checking

Computing an integral

Consider the function

$$h(x) = [\cos(50x) + \sin(20x)]^2$$

and suppose you want to compute the integral



Suppose you are not very good at algebra. How can you compute the integral?

Computing an expectation

Consider the following situation: according to the kinetic theory of gases, the magnitude Y of the velocity of a gas molecule is random and its probability density is given by Maxwell's distribution

$$p(y;\sigma) = \frac{\sqrt{\frac{2}{\pi}}}{\sigma^3} y^2 \exp(-\frac{y^2}{2\sigma^2}), \quad y \geq 0$$

The gas molecule kinetic energy is then $g(Y) = \frac{1}{2}m \cdot Y^2$ where m is the mass of the molecule. Your collaborator is interested in deriving the average kinetic energy of a molecule of mass m. How can you derive it?

You need to compute the integral

$$E(g(Y)) = \int_0^\infty g(y)p(y)dy$$

Suppose you are not very good at algebra. How can you compute the integral?

Monte Carlo methods

- Monte Carlo methods are simulation-based methods to approximately evaluate integrals of the form

$$E(g(Y)) = \int_{\mathcal{Y}} g(y) \cdot p(y) dy$$

  where Y is a random variable with sample space $\mathcal{Y}$ and density $p(y)$ and $g(y)$ is a function defined on $\mathcal{Y}$.

- In the context of Bayesian inference, Monte Carlo methods are used mainly with the following goals:
  - to compute posterior means or posterior standard deviations of functions $g(\theta)$ of a parameter θ.
  - to compute posterior probabilities, e.g. $P(\theta \in A | y_1, ..., y_n)$ where A is some subset of the parameter space Θ.
  - to compute posterior distributions of functions $g(\theta_1, ..., \theta_k)$ of multiple parameters $\theta_1, ..., \theta_k$.

Basic principles

- The main principles upon which Monte Carlo methods are based are:

  1. Strong Law of Large Numbers: If $Y_1, ..., Y_n$ is a sequence of i.i.d. random variables with sample space $\mathcal{Y}$ and density $p(y)$ and $g(y)$ is a function defined on $\mathcal{Y}$, then:

     $$\frac{1}{n}\sum_{i=1}^{n} g(Y_i)$$

     converges almost surely to $E(g(Y))$ as $n \to \infty$.

  2. Central Limit Theorem: If $Y_1, ..., Y_n$ is a sequence of i.i.d. random variables with sample space $\mathcal{Y}$ and density $p(y)$, and $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, then:

     $$\sqrt{n} \cdot \frac{\bar{Y} - E(Y)}{\sqrt{\text{Var}(Y)}}$$

     converges in distribution to a standard normal random variable $Z \sim N(0, 1)$.

Basic principles

- The Central Limit theorem is often restated as:

$$\sqrt{n}\left(\bar{Y} - E(Y)\right) \xrightarrow{d} N(0, \text{Var}(Y))$$

- For both the Strong Law of Large Numbers and Central Limit Theorem, the approximation is better as n gets large.
- If we are able to simulate n values $y_1, ..., y_n$ independently from the density $p(y)$, then the Strong Law of Large Numbers assures us that:
  - If we compute $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, then as n gets large, $\bar{y}$ provides a good approximation to $E(Y)$.
  - If we compute $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, then as n gets large, s^2 provides a good approximation to $\text{Var}(Y)$.
  - If, for any $c \in \mathbb{R}$, we compute $\mathbf{1}[y_i \leq c]$ and we consider $\#\{y_i \leq c\}/n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[y_i \leq c]$, then as n gets large, $\#\{y_i \leq c\}/n$ provides a good approximation to $P(Y \leq c)$.

Basic principles

- Similarly, the Strong Law of Large Numbers assures us that if we are able to simulate n values $y_1, ..., y_n$ independently from the density $p(y)$, then:
  - If we compute $\#\{y_i \leq t\}/n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[y_i \leq t]$ for any t and derive in this way the empirical distribution function $\widehat{F}_n(t)$, then as n gets large, $\widehat{F}_n(t)$ provides a good approximation to the distribution function $F(t)$.
  - If we compute the median of $y_1, ..., y_n$ (via the inverse of the empirical distribution function, or quantile function), then as n gets large, the sample median provides a good approximation to the median $Y_{\frac{1}{2}}$ of the distribution.
  - If we compute the α-percentile of $y_1, ..., y_n$ (via the inverse of the empirical distribution function, or quantile function), then as n gets large, this sample percentile provides a good approximation to the α-percentile Y_α of the distribution.

Monte Carlo methods

- How do we use Monte Carlo methods in our case?
- Suppose we have the following model:

$$\begin{cases} Y_1, ..., Y_n|\theta & \overset{i.i.d.}{\sim} & p(y|\theta) \\ \theta & \sim & p(\theta) \end{cases}$$

  and we are interested in approximating the posterior distribution $p(\theta|y_1, ..., y_n)$ of θ or, more in general, the posterior distribution of functions $g(\theta)$ of θ.

- To achieve this, we will generate B sample values $\theta^{(1)}, ..., \theta^{(B)}$ from the posterior distribution $p(\theta|y_1, ..., y_n)$. We will compute $g(\theta^{(1)}), ..., g(\theta^{(B)})$ and we will use the Strong Law of Large Numbers to approximate the posterior distribution $p(g(\theta)|y_1, ..., y_n)$ of $g(\theta)$.

Monte Carlo methods: example

- Example: Consider the example we looked at before. $Y_1, ..., Y_n$ are random variables indicating the number of children for women who were in their 20s during the 1970s and have a college degree. Then we could set up the following model:

$$\begin{cases} Y_1, ..., Y_n|\theta & \overset{i.i.d.}{\sim} & \text{Poisson}(\theta) \\ \theta & \sim & \text{Gamma}(a, b) \end{cases}$$

- We have seen that in this case,
$p(\theta|y_1, ..., y_n) = \text{Gamma}(\tilde{a} = a + \sum_{i=1}^n y_i, \tilde{b} = b + n)$
- Suppose we take $p(\theta) = \text{Gamma}(2, 1)$ and we have observations on $n = 44$ women for which $\sum_{i=1}^n y_i = 66$. Then:
$p(\theta|y_1, ..., y_n) = \text{Gamma}(2 + 66, 1 + 44) = \text{Gamma}(68, 45)$.
- This implies that:

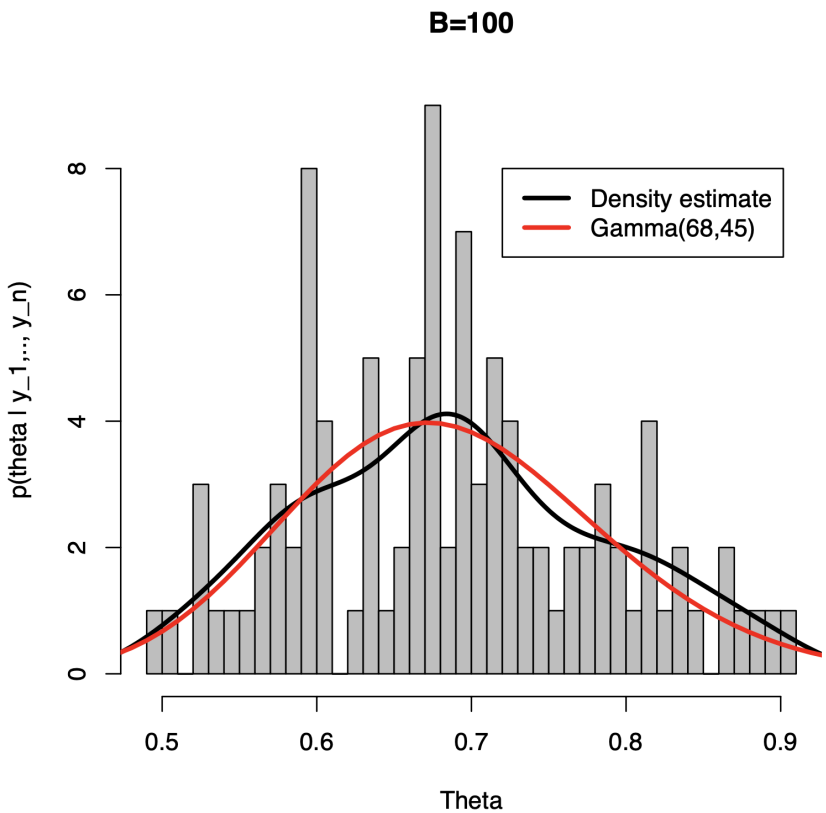
$$\text{E}(\theta|y_1, ..., y_n) = \frac{\tilde{a}}{\tilde{b}} = \frac{68}{45} = 1.51$$
$$\text{Var}(\theta|y_1, ..., y_n) = \frac{\tilde{a}}{\tilde{b}^2} = \frac{68}{45^2} = 0.03$$

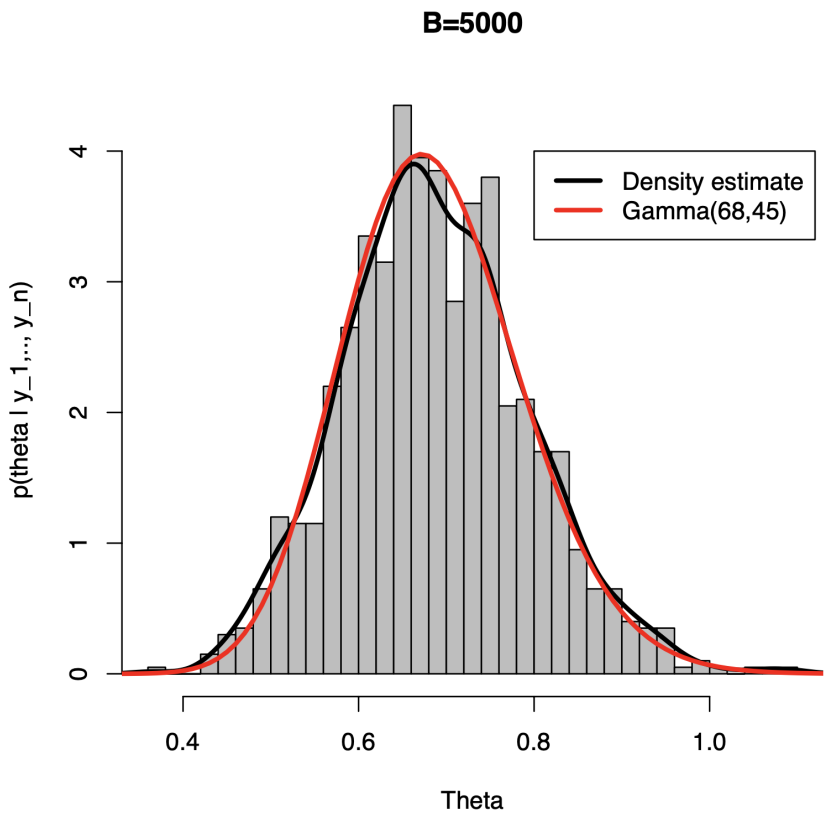
Monte Carlo methods: example

- Let's now simulate B values $\theta^{(1)}, ..., \theta^{(B)}$ from the posterior distribution $p(\theta|y_1, ..., y_n)$ and look at the following quantities:
  1. $\bar{\theta} = \frac{1}{B} \sum_{i=1}^{B} \theta^{(i)}$
  2. $s^2 = \frac{1}{B-1} \sum_{i=1}^{B} (\theta^{(i)} - \bar{\theta})^2$
  3. median of the distribution
  4. a 95% credible interval for θ

Monte Carlo methods: example

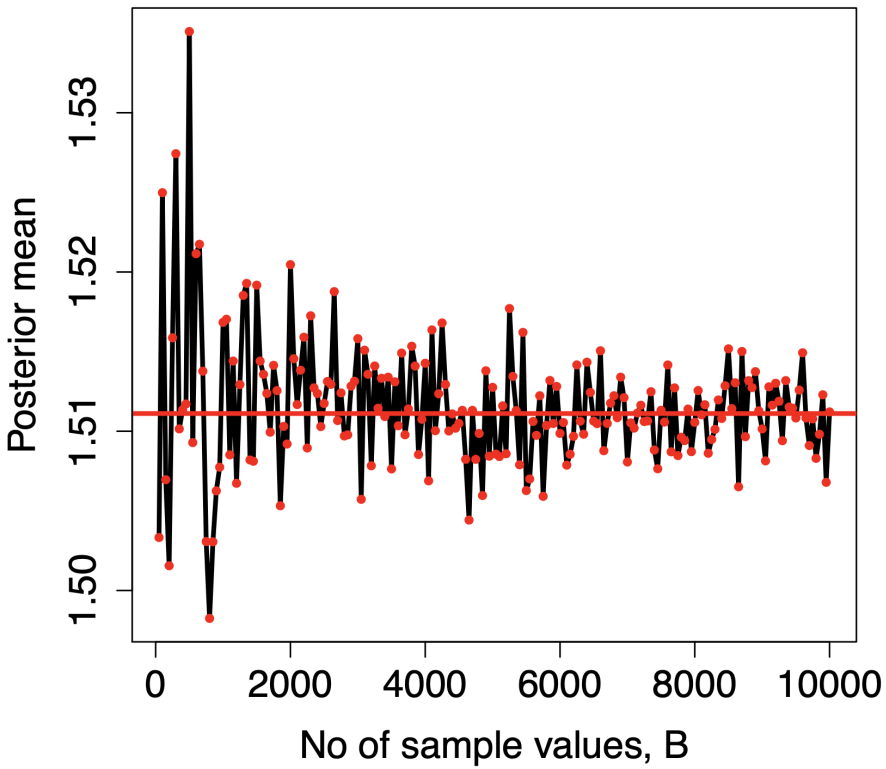
Histograms and density estimates based on samples $\theta^{(1)}, ..., \theta^{(B)}$ from the posterior distribution $p(\theta|y_1, ..., y_n) = \text{Gamma}(68, 45)$ for $B = 100$ and $B = 5,000$ and plot of the $\text{Gamma}(68, 45)$ density function.

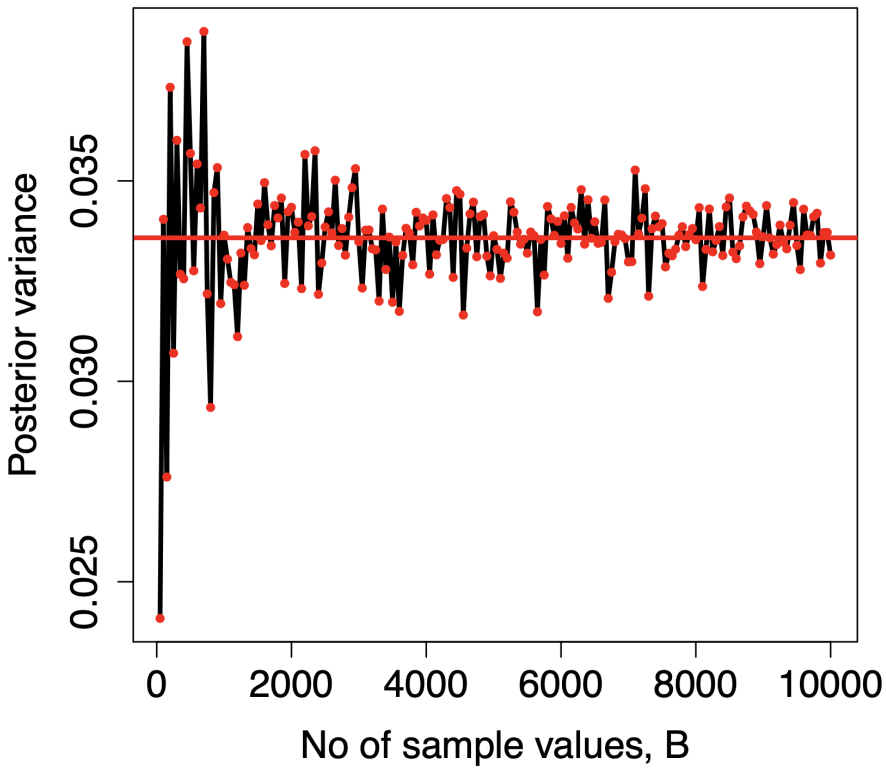




Monte Carlo methods: example

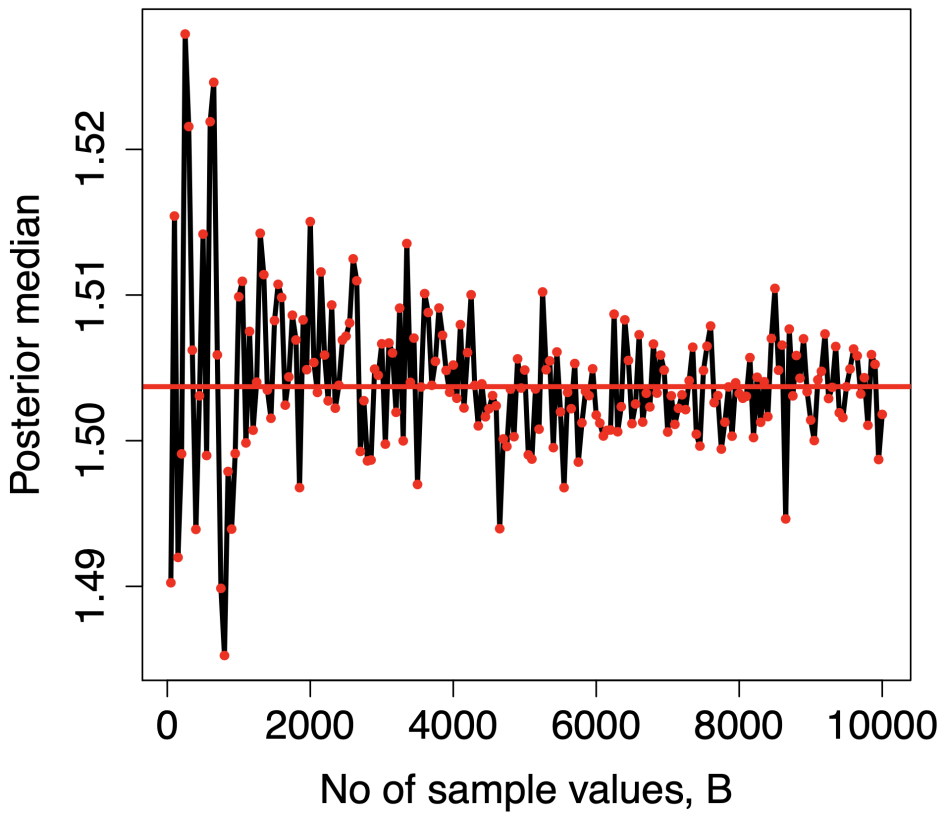
Monte Carlo approximation to the posterior mean and posterior variance based on samples $\theta^{(1)}, ..., \theta^{(B)}$ from the posterior distribution $p(\theta|y_1, ..., y_n) = \text{Gamma}(68, 45)$ for different values of B. The red line indicates the exact value of the posterior mean and posterior variance.





Monte Carlo methods: example

Monte Carlo approximation to the posterior median based on samples $\theta^{(1)}, ..., \theta^{(B)}$ from the posterior distribution $p(\theta|y_1, ..., y_n) = \text{Gamma}(68, 45)$ for different values of B. The red line indicates the exact value of the posterior median.



Monte Carlo methods: example

- Monte Carlo appoximation of a 95% credible interval for θ based on samples $\theta^{(1)}, ..., \theta^{(B)}$ from the posterior distribution $p(\theta|y_1, ..., y_n) = \text{Gamma}(68, 45)$ for different values of B:

B	95% credible interval for θ
50	(1.22,1.82)
100	(1.15,1.89)
500	(1.17,1.83)
1,000	(1.16,1.88)
2,000	(1.16,1.89)
5,000	(1.17,1.90)
10,000	(1.18,1.89)

- The exact 95% credible interval for θ is: $(1.17, 1.89)$.

Monte Carlo standard error

- In each of these plots and in the table with the 95% credible interval, the approximation is more accurate as the number of sampled values, B, gets larger. However, there is always variation around the true value also for large B.
- The Monte Carlo standard error gives a measure of this variability and is derived from the Central Limit Theorem (CLT).
- Applying the CLT to our situation, the CLT states that if $\bar{\theta}$ is given by $\bar{\theta} = \frac{1}{B}\sum_{i=1}^{B}\theta^{(i)}$, then $\bar{\theta}$ is approximately normal with mean equal to $E(\theta|y_1, ..., y_n)$ and standard deviation equal to $\sqrt{\frac{\text{Var}(\theta|y_1,...,y_n)}{B}}$.
- The Monte Carlo standard error is the Monte Carlo approximation to this quantity.
- Since $s^2 = \frac{1}{B-1}\sum_{i=1}^{B}(\theta^{(i)} - \bar{\theta})^2$ is the Monte Carlo approximation to $\text{Var}(\theta|y_1, ..., y_n)$, $\sqrt{\frac{s^2}{B}}$ is the Monte Carlo approximation to $\sqrt{\frac{\text{Var}(\theta|y_1,...,y_n)}{B}}$.

- This result can be used to compute approximate 95% Monte Carlo confidence interval for the posterior mean:

$$\bar{\theta} \pm 1.96\sqrt{\frac{s^2}{B}}$$

- Or it can be used to derive the sample size B needed to obtain a certain accuracy in the Monte Carlo approximation.
- Note that this formula does not apply only to the posterior mean: it could be used for any function $g(\theta)$.

Number of Monte Carlo simulations

- In the previous example, we had the following values of the Monte Carlo standard error $\sqrt{\frac{s^2}{B}}$ for the posterior mean $E(\theta|y_1, ..., y_n)$ (= 1.511):

B	$\bar{\theta}$	s^2	$\sqrt{\frac{s^2}{B}}$
50	1.486	0.034	0.026
100	1.542	0.027	0.017
500	1.526	0.033	0.008
1,000	1.512	0.032	0.006
2,000	1.519	0.034	0.004
5,000	1.511	0.033	0.003
10,000	1.511	0.033	0.002

- So, for example, a 95% confidence interval for $E(\theta|y_1, ..., y_n)$ based on a Monte Carlo approximation with $B = 100$ values is: $1.542 \pm 1.96 \times 0.017$.
- If we want our Monte Carlo estimate to be within 0.01 from the true value, we will need a sample size of at least 500.

Monte Carlo methods for posterior distributions

- Suppose we have observations $y_1, ..., y_n$ of n random variables $Y_1, ..., Y_n$ that we can model as follows:

$$\begin{cases} Y_1, ..., Y_n | \theta & \overset{i.i.d.}{\sim} & p(y|\theta) \\ \theta & \sim & p(\theta) \end{cases}$$

- We want to derive the posterior distribution $p(g(\theta)|y_1, ..., y_n)$ of $g(\theta)$ for some function g.

- We can use Monte Carlo methods to approximate it.

Monte Carlo methods : Algorithm

- The Monte Carlo method in this case will work in the following way:
  - We will generate B values $\theta^{(1)}, ..., \theta^{(B)}$ independently from the posterior distribution $p(\theta|y_1, ..., y_n)$.
  - For each i, we will evaluate $\gamma^{(i)} = g(\theta^{(i)})$, that is:

$$\begin{cases} \text{sample } \theta^{(1)} \text{ from } p(\theta|y_1, ..., y_n) & \longrightarrow & \gamma^{(1)} = g(\theta^{(1)}) \\ \text{sample } \theta^{(2)} \text{ from } p(\theta|y_1, ..., y_n) & \longrightarrow & \gamma^{(2)} = g(\theta^{(2)}) \\ \vdots & & \\ \text{sample } \theta^{(B)} \text{ from } p(\theta|y_1, ..., y_n) & \longrightarrow & \gamma^{(B)} = g(\theta^{(B)}) \end{cases}$$

  - Then, $\gamma^{(1)}, ..., \gamma^{(B)}$ is a sample from the posterior distribution $p(g(\theta)|y_1, ..., y_n)$.
  - We can compute the empirical distribution for $\gamma^{(1)}, ..., \gamma^{(B)}$ (this will be an approximation to $p(g(\theta)|y_1, ..., y_n)$), the mean of $\gamma^{(1)}, ..., \gamma^{(B)}$ (approximation to the posterior mean of $g(\theta)$), the variance of $\gamma^{(1)}, ..., \gamma^{(B)}$ (approximation to the posterior variance of $g(\theta)$), ...

Monte Carlo methods : Example

- Example: The most important factor in early breast cancer is the number of axillary lymph nodes that test positive for breast cancer. There is no standard number of lymph nodes to sample. Sometimes surgeon remove three nodes and sometimes they remove 30. The proportion θ of lymph nodes that are positive has approximately a Beta$(0.1, 5)$ density. So, the probability that a sampled lymph node is positive is $E(\theta) = \frac{a}{a+b} = \frac{0.1}{5+0.1} = 0.02$, approximately 2%.

  A woman with breast cancer had a mastectomy and subsequently, the surgeon removed 20 lymph nodes. None tested positive.

  What is the updated density for the proportion of positive lymph nodes? What about the log-odds?

Monte Carlo methods : Example: Model

- Our data in this situation can be modeled as follows: let Y_i be the negative/positive test result for the i-th removed lymph node. Then, conditional on θ, the probability that a lymph node is positive, $Y_1, ..., Y_{20}$ are conditionally i.i.d. with $p(y|\theta) = \text{Bernoulli}(\theta)$.

  We have the following model:

$$\begin{cases} Y_1, ..., Y_{20}|\theta & \overset{i.i.d.}{\sim} & \text{Bernoulli}(\theta) \\ \theta & \sim & \text{Beta}(0.1, 5) \end{cases}$$

- Our data is: $n = 20$, $\sum_{i=1}^{20} y_i = 0$.
- Thus, the posterior distribution $p(\theta|y_1, ..., y_n)$ for θ is $\text{Beta}(\tilde{a} = a + \sum_{i=1}^{20} y_i = 0.1 + 0 = 0.1, \tilde{b} = b + (n - y) = 5 + (20 - 0) = 25) = \text{Beta}(0.1, 25)$.

Example: Posterior distribution of log-odds

- We want to derive the posterior distribution for the log-odds $g(\theta) = \log(\frac{\theta}{1-\theta})$.
- We can apply the transformation theorem to $p(\theta|y_1, ..., y_n)$ and derive the exact posterior distribution $p(g(\theta)|y_1, ..., y_n)$ for the log-odds.
- We have: $\gamma = g(\theta) = \log(\frac{\theta}{1-\theta})$ which implies $\theta = h(\gamma) = \frac{\exp(\gamma)}{1+\exp(\gamma)}$. Therefore:

$$\frac{d\theta}{d\gamma} = \frac{d\left(\frac{\exp(\gamma)}{1+\exp(\gamma)}\right)}{d\gamma} = \frac{\exp(\gamma)}{(1+\exp(\gamma))^2}$$

$$p(\gamma|y_1, ..., y_n) = p(\theta|y_1, ..., y_n) \cdot |\frac{d\theta}{d\gamma}|$$

$$= \frac{\Gamma(\tilde{a}+\tilde{b})}{\Gamma(\tilde{a})\Gamma(\tilde{b})} \left(\frac{\exp(\gamma)}{1+\exp(\gamma)}\right)^{\tilde{a}-1} \left(1 - \frac{\exp(\gamma)}{1+\exp(\gamma)}\right)^{\tilde{b}-1} \cdot \frac{\exp(\gamma)}{(1+\exp(\gamma))^2}$$

$$= \frac{\Gamma(\tilde{a}+\tilde{b})}{\Gamma(\tilde{a})\Gamma(\tilde{b})} \frac{\exp(\tilde{a}\gamma)}{(1+\exp(\gamma))^{\tilde{a}+\tilde{b}}} = \frac{\Gamma(25.1)}{\Gamma(0.1)\Gamma(25)} \cdot \frac{\exp(0.1\cdot\gamma)}{(1+\exp(\gamma))^{25.1}}$$

Example: Approximation to posterior distribution

- The posterior distribution $p(\gamma|y_1, ..., y_n)$ of the log-odds $\gamma = \log(\frac{\theta}{1-\theta})$ is:

$$p(\gamma|y_1, ..., y_n) = \frac{\Gamma(25.1)}{\Gamma(0.1)\Gamma(25)} \cdot \frac{\exp(0.1 \cdot \gamma)}{(1 + \exp(\gamma))^{25.1}}$$

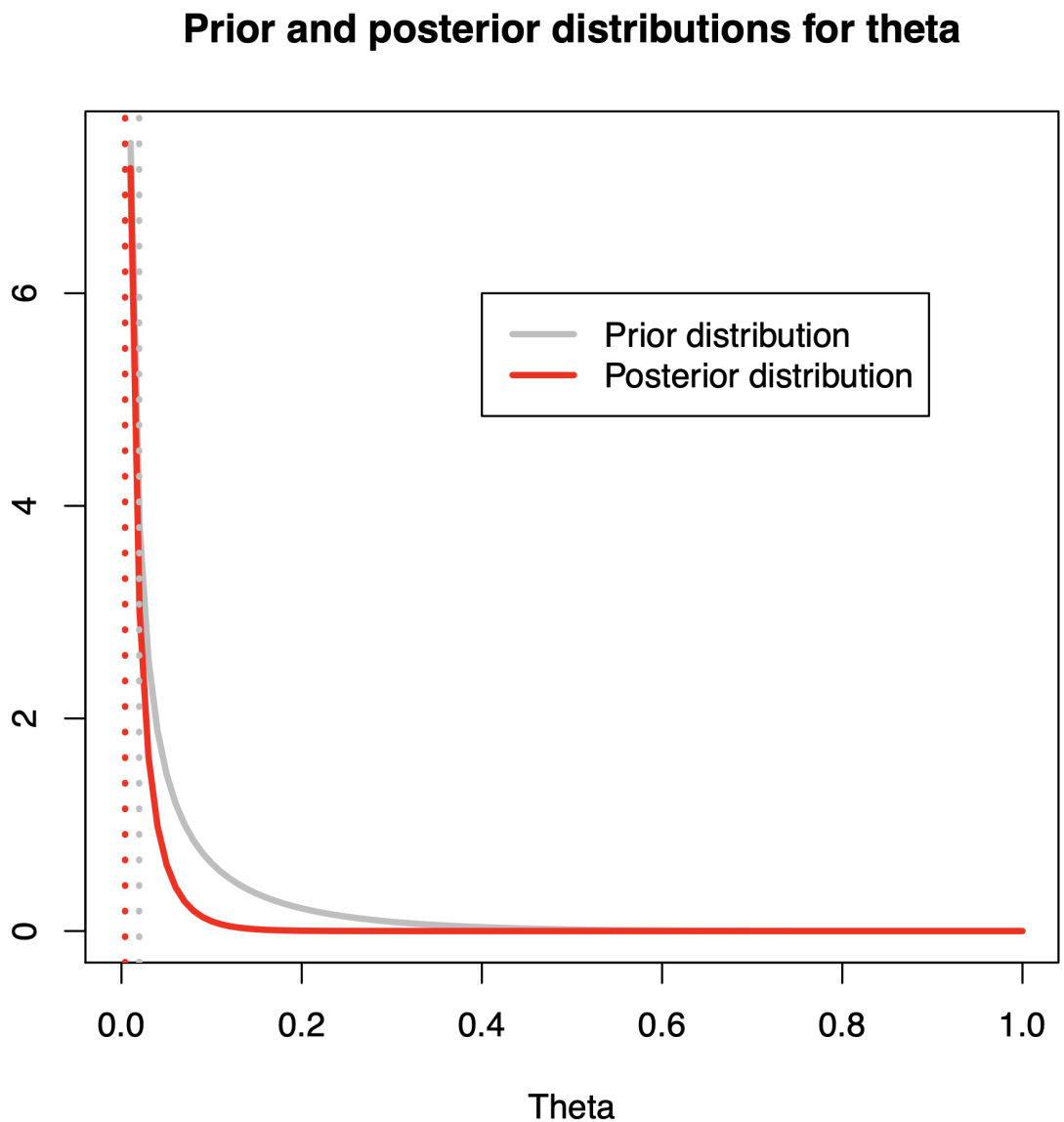
  which does not belong to any of the distributions we know.

- Let's compute an approximation to $p(\gamma|y_1, ..., y_n)$ using Monte Carlo methods.

- We sample $B = 5000$ values $\theta^{(1)}, ..., \theta^{(B)}$ independently from $p(\theta|y_1, ..., y_n) = \text{Beta}(0.1, 25)$ and we compute $\gamma^{(1)}, ..., \gamma^{(B)}$.

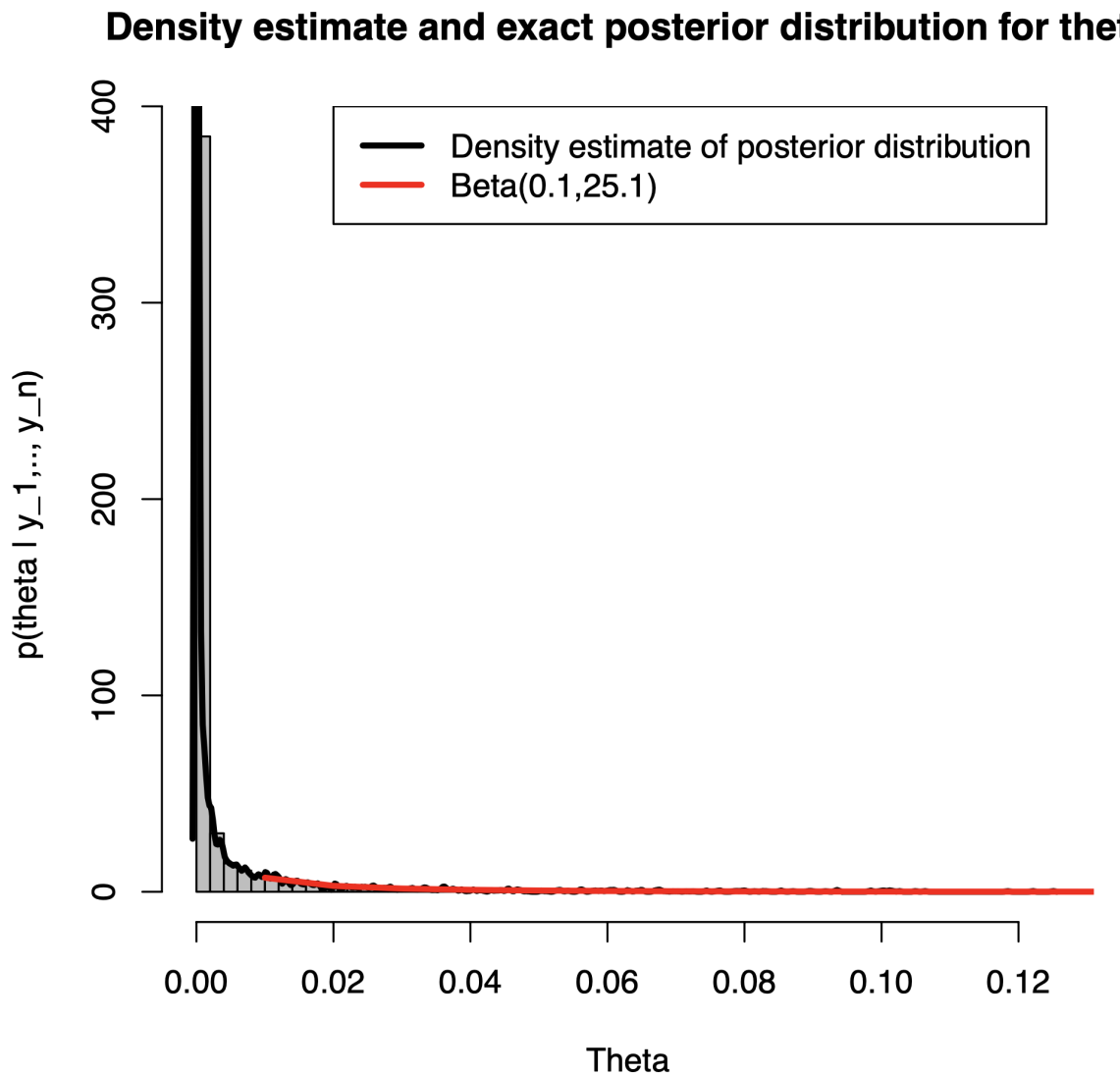
Example: breast cancer

Prior and posterior distributions for θ.



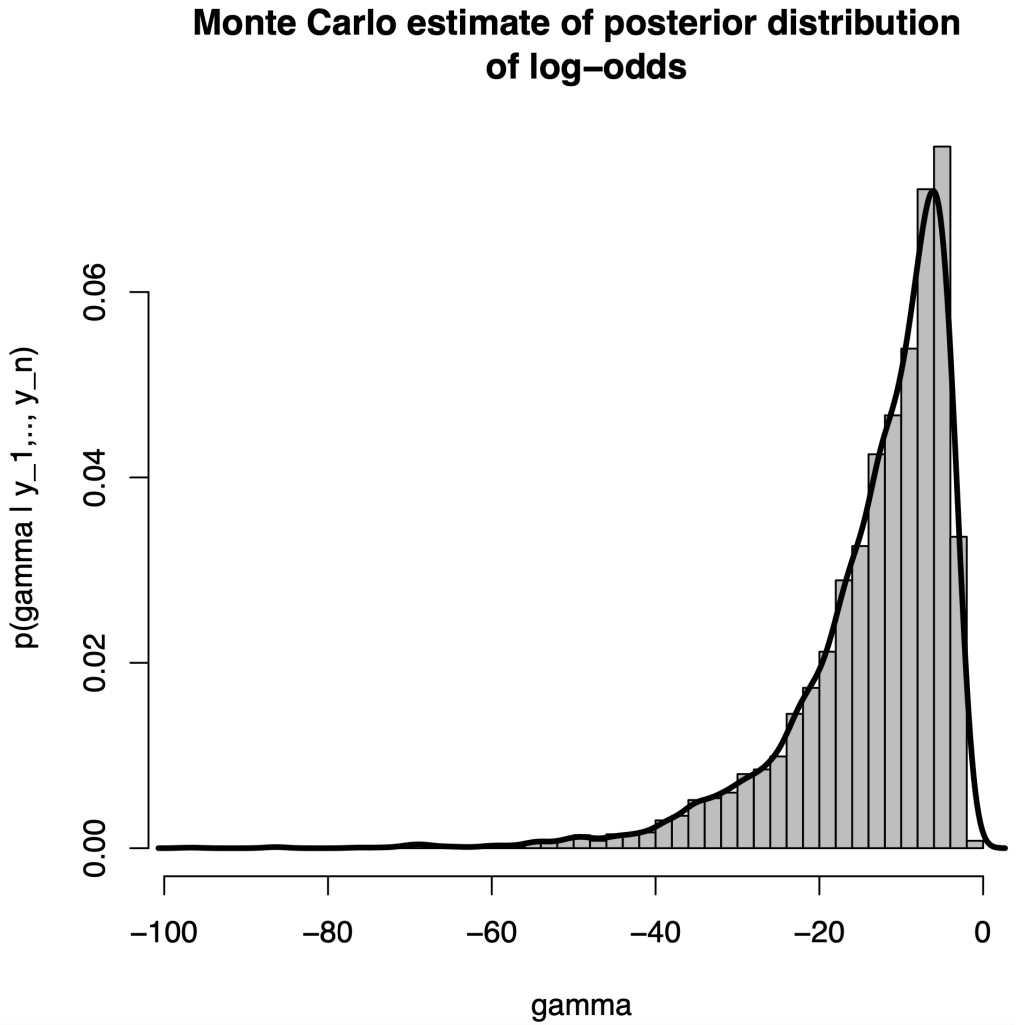
Example: breast cancer

Monte Carlo approximation: Histogram of the $B = 5000$ sampled values $\theta^{(1)}, ..., \theta^{(B)}$ from the posterior distribution $p(\theta|y_1, ..., y_n) = \text{Beta}(0.1, 25)$.



Example: breast cancer

Monte Carlo approximation: Histogram of the $B = 5000$ sampled log-odds values $\gamma^{(1)} = \log(\frac{\theta^{(1)}}{1-\theta^{(1)}}), \ldots, \gamma^{(B)} = \log(\frac{\theta^{(B)}}{1-\theta^{(B)}})$ from the posterior distribution $p(\gamma = g(\theta)|y_1, ..., y_n)$.



Example: breast cancer

- We can also look at how the prior and the posterior distribution of the log-odds change in light of the data.
- In order to do that, we will need to derive the prior density for the log-odds $\gamma = g(\theta) = \log(\frac{\theta}{1-\theta})$ that is induced by the Beta$(0.1, 5)$ prior distribution for θ.
- Since the Beta prior is conjugate for θ and we have derived the posterior distribution for the log-odds that corresponds to a Beta posterior distribution, we know what is the exact form of the prior density for the log-odds induced by the Beta prior on θ. This is:

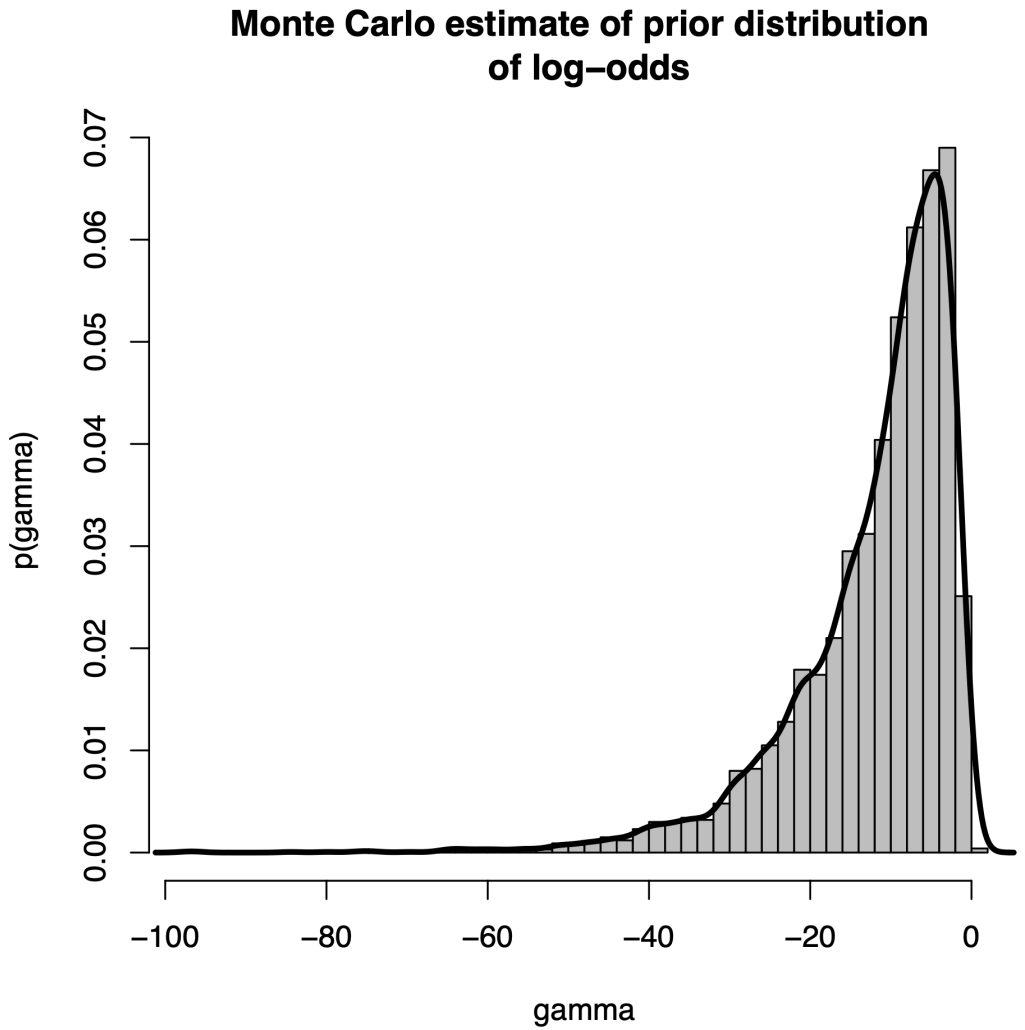
$$p(\gamma) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\exp(a\gamma)}{(1+\exp(\gamma))^{a+b}} = \frac{\Gamma(5.1)}{\Gamma(0.1)\Gamma(5)} \cdot \frac{\exp(0.1\cdot\gamma)}{(1+\exp(\gamma))^{5.1}}$$

- We can obtain a Monte Carlo approximation to it by sampling independently $B = 5000$ values $\theta^{(1)}, ..., \theta^{(B)}$ from the prior distribution $p(\theta) = \text{Beta}(0.1, 5)$, and derive

$$\gamma^{(1)} = \log\left(\frac{\theta^{(1)}}{1-\theta^{(1)}}\right), \ldots, \gamma^{(B)} = \log\left(\frac{\theta^{(B)}}{1-\theta^{(B)}}\right)$$

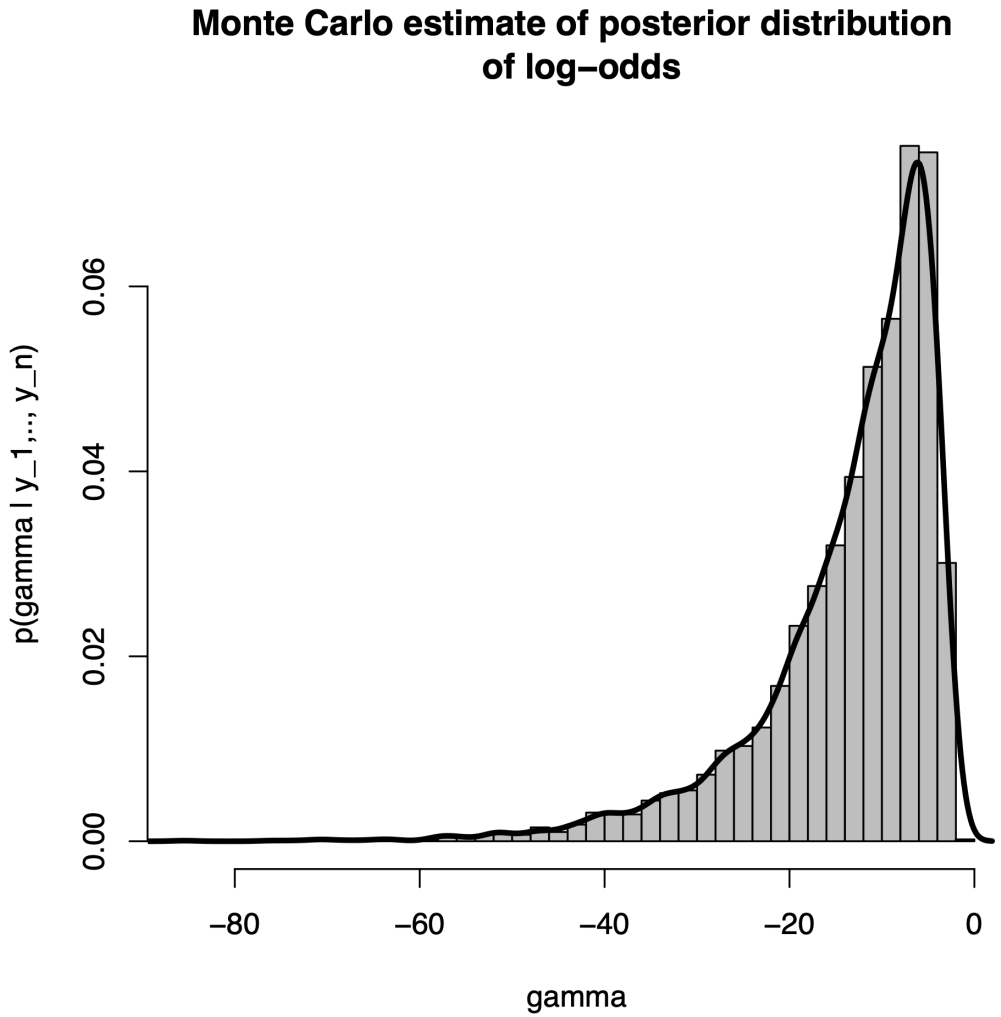
Example: breast cancer

- The empirical distribution of these 5000 values will be an approximation to the prior distribution of the log-odds γ induced by the Beta$(0.1, 5)$ prior on θ.



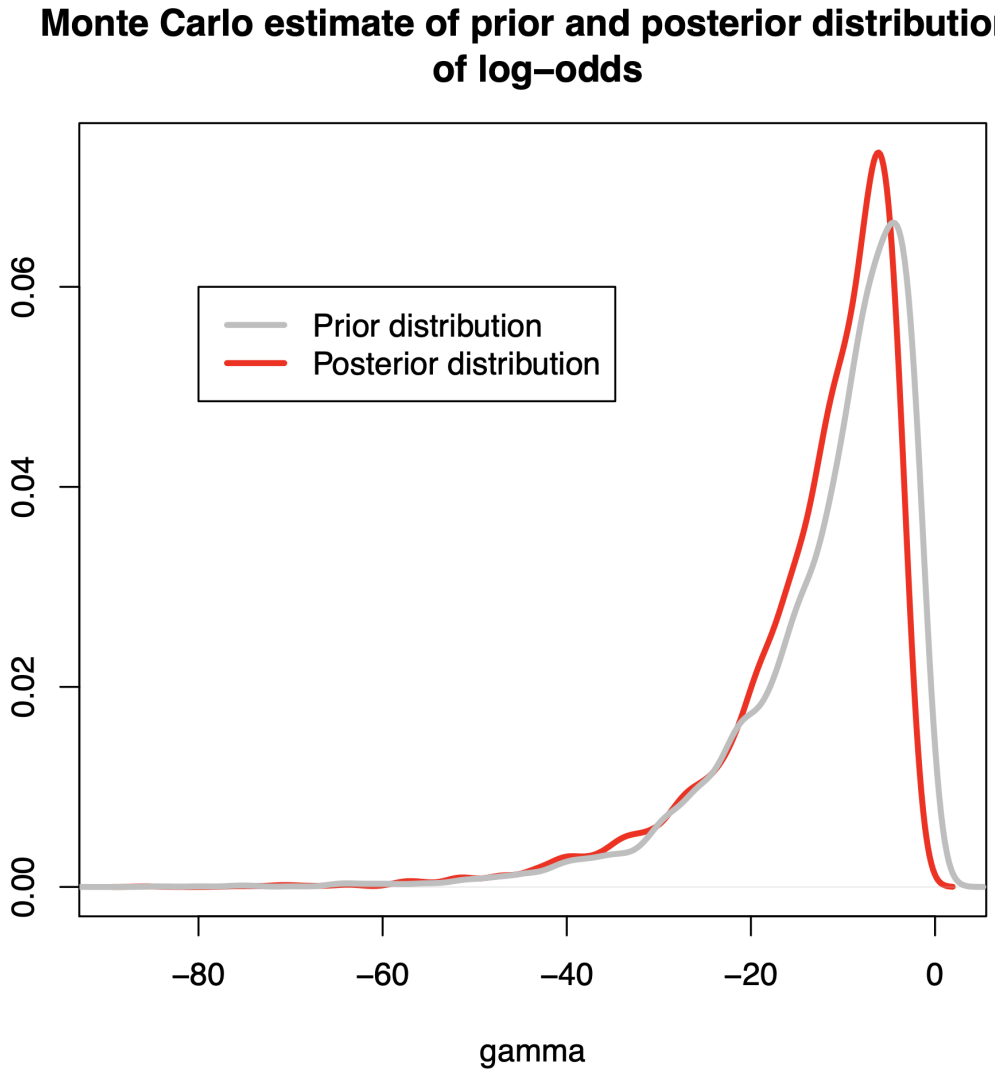
Example: breast cancer

- The empirical distribution of the 5000 values $\gamma^{(1)}, ..., \gamma^{(B)}$ obtained by applying the g function on the 5000 values $\theta^{(1)}, ..., \theta^{(B)}$ sampled from the posterior distribution $p(\theta|y_1, ..., y_n)$ of θ will be an approximation to the posterior distribution of γ.



Example: breast cancer

- Monte Carlo approximation: Approximation of the prior and posterior distributions of the log-odds that a lymph node will test positive based on $B = 5000$ sampled values.



Example: breast cancer

- We can summarize the posterior distribution of the log-odds γ by looking at the posterior mean, the posterior standard deviation, and the credible interval of the log-odds.
- We approximate this by using again Monte Carlo methods computing, respectively, the mean and the standard deviation of the 5000 sampled values $\theta^{(1)}, ..., \theta^{(B)}$ and the 2.5-th and 97.5-th quantiles of the 5000 sampled values.

Parameter	MC est. of post. mean	MC est. of post. variance	MC est. of 95% CI
θ	0.004	0.00013	(4.7e-18,0.04)
γ	-13.3	94.1	(-39.9,-3.3)

- For θ we can compute the exact values for the posterior mean, posterior variance and 95% credible interval:
posterior mean=0.004, posterior variance=0.00015 and
95% credible interval = (2.4e-18, 0.04).

Example: breast cancer

- We can also report our uncertainty regarding the estimate of the posterior mean of the log-odds that we have obtained.

- This is quantified by the Monte Carlo standard error, that is, $\sqrt{\frac{s^2}{B}}$ which in this case is $\sqrt{\frac{s^2}{B}} = \sqrt{\frac{94.1}{5000}} = 0.14$.

Functions of more parameters

- Monte Carlo methods can be used also to evaluate functions of more parameters.
- Example: In the 1990s the General Society Survey gathered data on the educational attainment and number of children of 155 women who were in their 20s in the 1970s. We divide the women participating in the survey in two groups: those without a college degree and those with a college degree.

  Let $Y_{1,1}, ..., Y_{n_1,1}$ denote the number of children for woman $1, ..., n_1$ in the group of women without college degree.

  Let $Y_{1,2}, ..., Y_{n_2,2}$ denote the number of children for woman $1, ..., n_2$ in the group of women with college degree.

  We model $Y_{1,i}, ..., Y_{n_i,i}|\theta_i \overset{iid}{\sim} \text{Poisson}(\theta_i)$ for $i = 1, 2$.

Functions of more parameters

- We assume that θ_1 and θ_2 are independent and we place on each of them a Gamma$(2, 1)$ prior.
  We collect the following data:

$$\begin{aligned} n_1 &= 111, \quad \sum_{i=1}^{n_1} y_{i,1} = 217 \\ n_2 &= 44, \quad \sum_{i=1}^{n_2} y_{i,2} = 66 \end{aligned}$$

  In light of the observed data, we want to compute the posterior probability that $\theta_1 > \theta_2$.

- We want to compute $P(\theta_1 > \theta_2 | y_{1,1}, ..., y_{n_1,1}, y_{1,2}, ..., y_{n_2,2})$.

- To compute this, we need to evaluate the following integral

$$\int_0^{\infty} \int_0^{\theta_1} p(\theta_1, \theta_2 | y_{1,1}, ..., y_{n_1,1}, y_{1,2}, ..., y_{n_2,2}) d\theta_1 d\theta_2$$

Functions of more parameters

- Given that θ_1 and θ_2 are independent, we have the following:

$$\begin{cases} Y_{1,1}, ..., Y_{n_1,1}|\theta_1 & \overset{i.i.d.}{\sim} & \text{Poisson}(\theta_1) \\ \theta_1 & \sim & \text{Gamma}(2,1) \end{cases}$$

$$\begin{aligned} \implies p(\theta_1|y_{1,1}, ..., y_{n_1,1}) &= \text{Gamma}(\tilde{a} = a + \textstyle\sum_{i=1}^{n} y_{i,1}, \tilde{b} = b + n_1) \\ &= \text{Gamma}(2 + 217, 1 + 111) \\ &= \text{Gamma}(219, 112) \end{aligned}$$

$$\begin{cases} Y_{1,2}, ..., Y_{n_2,2}|\theta_1 & \overset{i.i.d.}{\sim} & \text{Poisson}(\theta_2) \\ \theta_2 & \sim & \text{Gamma}(2,1) \end{cases}$$

$$\begin{aligned} \implies p(\theta_2|y_{1,2}, ..., y_{n_2,2}) &= \text{Gamma}(\tilde{a} = a + \textstyle\sum_{i=1}^{n} y_{i,2}, \tilde{b} = b + n_2) \\ &= \text{Gamma}(2 + 66, 1 + 44) \\ &= \text{Gamma}(68, 45) \end{aligned}$$

Functions of more parameters

- $p(\theta_1|y_{1,1}, ..., y_{n_1,1}) = \text{Gamma}(219, 112)$ and
  $p(\theta_2|y_{1,2}, ..., y_{n_2,2}) = \text{Gamma}(68, 45)$
- Additionally, $\theta_1|y_{1,1}, ..., y_{n_1,1}$ and $\theta_2|y_{1,2}, ..., y_{n_2,2}$ are independent.
- Therefore:

$$\begin{aligned} &p(\theta_1, \theta_2|y_{1,1}, ..., y_{n_1,1}, y_{1,2}, ..., y_{n_2,2}) \\ &= p(\theta_1|y_{1,1}, ..., y_{n_1,1}) \cdot p(\theta_2|y_{1,2}, ..., y_{n_2,2}) \end{aligned}$$

which implies

$$\begin{aligned} &P(\theta_1 > \theta_2|y_{1,1}, ..., y_{n_1,1}, y_{1,2}, ..., y_{n_2,2}) \\ &= \int_0^\infty \int_0^{\theta_1} p(\theta_1, \theta_2|y_{1,1}, ..., y_{n_1,1}, y_{1,2}, ..., y_{n_2,2}) d\theta_1 d\theta_2 \\ &= \int_0^\infty \int_0^{\theta_1} \text{Gamma}(219, 112) \cdot \text{Gamma}(68, 45) d\theta_1 d\theta_2 \end{aligned}$$

Functions of more parameters

- The integral that we need to compute for $P(\theta_1 > \theta_2 | y_{1,1}, ..., y_{n_1,1}, y_{1,2}, ..., y_{n_2,2})$ has been simplified but it is still not immediate to calculate.
- We can approximate it using Monte Carlo methods by doing the following:
  1. Choose a number B of values to sample.
  2. Repeat the following: for i that goes from 1 to B
     - Sample $\theta_1^{(i)}$ from the posterior distribution $p(\theta_1 | y_{1,1}, ..., y_{n_1,1}) = \text{Gamma}(219, 112)$.
     - Sample $\theta_2^{(i)}$ from the posterior distribution $p(\theta_2 | y_{1,2}, ..., y_{n_2,2}) = \text{Gamma}(68, 45)$.

     (This will be a sample of size B from the joint posterior distribution of $\theta_1, \theta_2 | y_{1,1}, ..., y_{n_1,1}, y_{1,2}, ..., y_{n_2,2}$)
  3. Compute the proportion of times $\theta_1^{(i)}$ is greater than $\theta_2^{(i)}$. This will be the Monte Carlo estimate of the integral.

Functions of more parameters

- We chose $B = 5000$ and simulated independently 5000 values for θ_1 and 5000 values for θ_2 from the respective posterior distributions.

- Of these 5000 pairs, 4840 had θ_1 greater than θ_2.
  Thus, we estimate

$$P(\theta_1 > \theta_2 | y_{1,1}, ..., y_{n_1,1}, y_{1,2}, ..., y_{n_2,2}) = \frac{4840}{5000} = 0.968$$

Predictive distributions

- In Bayesian theory, prediction of future observations are based on predictive distributions, distributions of the data **averaged** over all possible parameter values.
- If the data $y_1, ..., y_n$ has not been observed, predictions are based on the prior predictive distribution

$$p(y) = \int_\Theta p(y|\theta) \cdot p(\theta) d\theta$$

$p(y)$ is also called marginal likelihood, since it is the likelihood averaged over all the possible values of θ supported by our prior beliefs $p(\theta)$.

- After the data $y_1, ..., y_n$ has been observed, predictions are based on the posterior predictive distribution

$$p(y_{n+1}|y_1, ..., y_n) = \int_\Theta p(y_{n+1}|\theta) \cdot p(\theta|y_1, ..., y_n) d\theta$$

this is is the likelihood of future data averaged over the posterior distribution $p(\theta|y_1, ..., y_n)$ of θ.

Prior predictive distribution

- The predictive distributions quantify our knowledge about the future, but they also provide the probability of re-observing each y_i, $i = 1, ..., n$ assuming that the adopted model is true.
- For example, the prior predictive distribution can be used to check whether our assumptions regarding the prior distribution of θ lead to reasonable prior beliefs for the data.
  - Choose a size n for the data.
  - Sample B independent values, $\theta^{(1)}, ..., \theta^{(B)}$ from the prior distribution $p(\theta)$.
  - For each $i = 1, ..., B$ and each value $\theta^{(i)}$, sample n independent values $y_1^{(i)}, ..., y_n^{(i)}$ from the sampling model $p(y|\theta^{(i)})$.
  - Then, $\left\{(y_1^{(1)}, ..., y_n^{(1)}, \theta^{(1)}), ..., (y_1^{(B)}, ..., y_n^{(B)}, \theta^{(B)})\right\}$ are B independent samples from the joint distribution $p(y_1, ..., y_n, \theta)$.
  - Look at the marginal distribution of $p(y_1, ..., y_n)$ (e.g. look at histogram, kernel density plots) and check if this distribution matches your prior beliefs about the data.

Posterior predictive distribution

- Similarly, we can use the posterior predictive distribution for model checking.
- By definition, the posterior predictive distribution is given by:

$$p(y_{n+1}|y_1, ..., y_n) = \int_\Theta p(y_{n+1}|\theta) \cdot p(\theta|y_1, ..., y_n) d\theta$$

- For some models, it is possible to compute the posterior predictive distribution in closed form (e.g. Bernoulli/Binomial sampling model and Beta prior or Poisson sampling model and Gamma prior or ...). In general, however, this is not possible.
- We will employ Monte Carlo methods to determine the posterior predictive distribution.

  We will sample B times from the posterior distribution $p(\theta|y_1, ..., y_n)$ of θ and from the sampling model $p(y|\theta)$.

Sampling from the posterior predictive distribution

- We will employ Monte Carlo methods to determine the posterior predictive distribution.
- Precisely we will:
  - sample $\theta^{(1)}$ from $p(\theta|y_1, ..., y_n)$ THEN sample $y_{n+1}^{(1)}$ from $p(y|\theta^{(1)})$
  - sample $\theta^{(2)}$ from $p(\theta|y_1, ..., y_n)$ THEN sample $y_{n+1}^{(2)}$ from $p(y|\theta^{(2)})$
  - ...
  - sample $\theta^{(B)}$ from $p(\theta|y_1, ..., y_n)$ THEN sample $y_{n+1}^{(B)}$ from $p(y|\theta^{(B)})$
- Then, $\left\{(y_{n+1}^{(1)}, \theta^{(1)}), (y_{n+1}^{(2)}, \theta^{(2)}), ..., (y_{n+1}^{(B)}, \theta^{(B)})\right\}$ are B independent samples from the joint posterior predictive distribution $p(y_{n+1}, \theta|y_1, ..., y_n)$.
- If we look at the marginal distribution $p(y_{n+1}|y_1, ..., y_n)$ (via histogram or kernel density plots), this is the posterior predictive distribution!

Sampling from the posterior predictive distribution: example

- The method used above to simulate from the joint posterior predictive distribution is the method of composition sampling.
- We will use it extensively in Markov Chain Monte Carlo (MCMC) algorithms.
- Example: Number of children for women that were in their 20s in the 1970s.
  $Y_{1,1}, Y_{2,1}, ..., Y_{n_1,1}$: number of children for n_1 women without college degree.
  $Y_{1,2}, Y_{2,2}, ..., Y_{n_2,2}$: number of children for n_2 women with college degree.
  We assume that:

$$\begin{cases} Y_{1,i}, ..., Y_{n_i,i}|\theta_i & \overset{i.i.d.}{\sim} & p(y_i|\theta_i) = \text{Poisson}(\theta_i) \\ \theta_i & \sim & p(\theta_i) = \text{Gamma}(2,1) \end{cases}$$

  for $i = 1, 2$ and $p(\theta_1) \perp p(\theta_2)$, i.e. θ_1 and θ_2 are assumed to be independent.

Sampling from the posterior predictive distribution: example

- We observed: $\begin{cases} n_1 &= 111, & \sum_{i=1}^{n_1} y_{i,1} = 217 \\ n_2 &= 44, & \sum_{i=1}^{n_2} y_{i,2} = 66 \end{cases}$

$$\Longrightarrow \begin{cases} p(\theta_1 | y_{1,1}, \ldots, y_{n_1,1}) &= \text{Gamma}(\tilde{a} = 219, \tilde{b} = 112) \\ p(\theta_2 | y_{1,2}, \ldots, y_{n_2,2}) &= \text{Gamma}(\tilde{a} = 68, \tilde{b} = 45) \end{cases}$$

- Given the data observed, the probability that a woman without college degrees has $\tilde{y}_1$ children is:

$$P(\tilde{Y}_1 = \tilde{y}_1 | y_{1,1}, \ldots, y_{n_1,1}) = \text{NegBin}(\tilde{y}_1; \tilde{a} = 219, \tfrac{1}{\tilde{b}+1} = \tfrac{1}{113})$$

for $\tilde{y}_1 = 0, 1, 2, \ldots$.

Similarly: given the data observed, the probability that a woman with college degree has $\tilde{y}_2$ children is:

$$P(\tilde{Y}_2 = \tilde{y}_2 | y_{1,2}, \ldots, y_{n_2,2}) = \text{NegBin}(\tilde{y}_2; \tilde{a} = 68, \tfrac{1}{\tilde{b}+1} = \tfrac{1}{46})$$

for $\tilde{y}_2 = 0, 1, 2, \ldots$.

Sampling from the posterior predictive distribution: example

- We want to compute the predictive probability that $\tilde{Y}_1 > \tilde{Y}_2$.
- How do we compute this? By definition, this is:

$$
\begin{aligned}
&P(\tilde{Y}_1 > \tilde{Y}_2|\text{data}) \\
&= \sum_{(\tilde{y}_1,\tilde{y}_2):\tilde{y}_1>\tilde{y}_2} p(\tilde{y}_1, \tilde{y}_2|\text{data}) \\
&= \sum_{\tilde{y}_2=0}^{\infty} \sum_{\tilde{y}_1=\tilde{y}_2+1}^{\infty} p(\tilde{y}_1|y_{1,1}, \ldots, y_{n_1,1}) \cdot p(\tilde{y}_2|y_{1,2}, \ldots, y_{n_2,2}) \\
&= \sum_{\tilde{y}_2=0}^{\infty} \sum_{\tilde{y}_1=\tilde{y}_2+1}^{\infty} \text{NegBin}(\tilde{y}_1; 219, \tfrac{1}{113}) \cdot \text{NegBin}(\tilde{y}_2; 68, \tfrac{1}{46})
\end{aligned}
$$

(Note that I am using a different parametrization of the Negative Binomial distribution than the book!)

- We can approximate this probability using Monte Carlo approximation and the definition of posterior predictive distribution:

$$p(\tilde{y}_i|y_{1,i}, ..., y_{n_i,i}) = \int_{\Theta_i} p(\tilde{y}_i|\theta_i) \cdot p(\theta_i|y_{1,i}, ..., y_{n_i,i}) d\theta_i \quad \text{for} \quad i = 1, 2$$

Sampling from the posterior predictive distribution: example

- We can proceed as follows: iterate $i = 1, \ldots, B$
  - sample $\theta_1^{(i)}$ from $p(\theta_1|y_{1,1}, ..., y_{n_1,1})$ THEN sample $\tilde{y}_1^{(i)}$ from $p(y_1|\theta_1^{(i)})$;
    sample $\theta_2^{(i)}$ from $p(\theta_2|y_{1,2}, ..., y_{n_2,2})$ THEN sample $\tilde{y}_2^{(i)}$ from $p(y_2|\theta_2^{(i)})$.

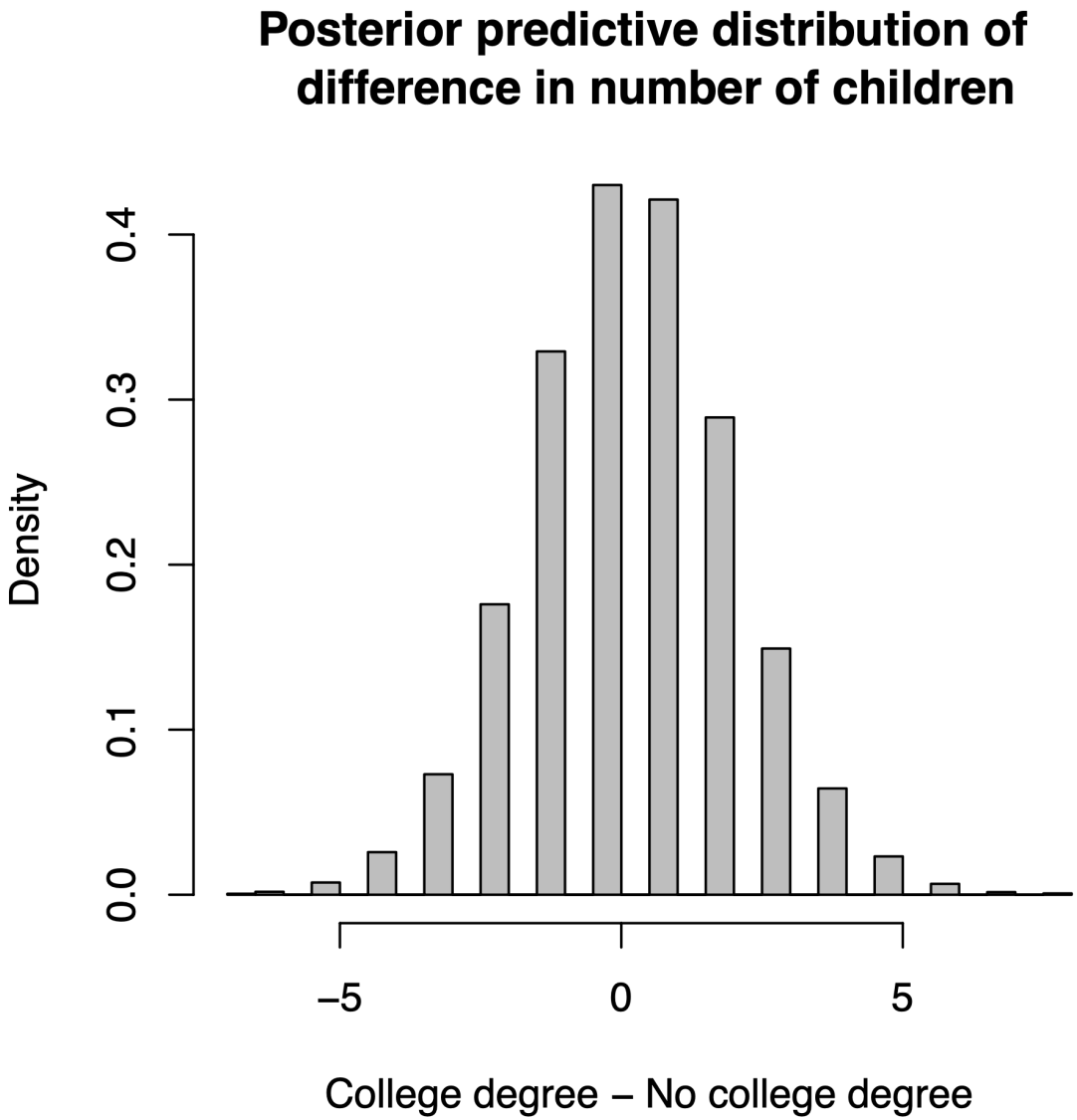
- The samples $\left\{(\tilde{y}_1^{(1)}, \tilde{y}_2^{(1)}), \ldots, (\tilde{y}_1^{(B)}, \tilde{y}_2^{(B)})\right\}$ are B independent samples from the joint posterior predictive distribution $p(\tilde{y}_1, \tilde{y}_2|\text{data})$.

Sampling from the posterior predictive distribution: example

- We can use those samples to compute $P(\tilde{Y}_1 > \tilde{Y}_2|\text{data})$: we simply have to count how many of those B samples have $\tilde{y}_1^{(i)} > \tilde{y}_2^{(i)}$.
- Alternatively, we can create a new variable $\tilde{D} = \tilde{Y}_1 - \tilde{Y}_2$ and see how many of the B values $\tilde{d}^{(1)}, ..., \tilde{d}^{(B)}$ of $\tilde{D}$ that we can compute from $(\tilde{y}_1^{(i)}, \tilde{y}_2^{(i)})$, $i = 1, ..., B$ are positive.

Sampling from the posterior predictive distribution: example

- We sampled $B = 10{,}000$ times from the joint posterior predictive distribution and obtained: $P(\tilde{Y}_1 > \tilde{Y}_2 | \text{data}) \approx 0.48$.



Posterior predictive model checking

- Posterior predictive distributions can be used for checking the assumptions of a model and its goodness of fit.
- Suppose that we have set up the following model:

$$\begin{cases} Y_1, ..., Y_n|\theta & \overset{i.i.d.}{\sim} & p(y|\theta) \\ \theta & \sim & p(\theta) \end{cases}$$

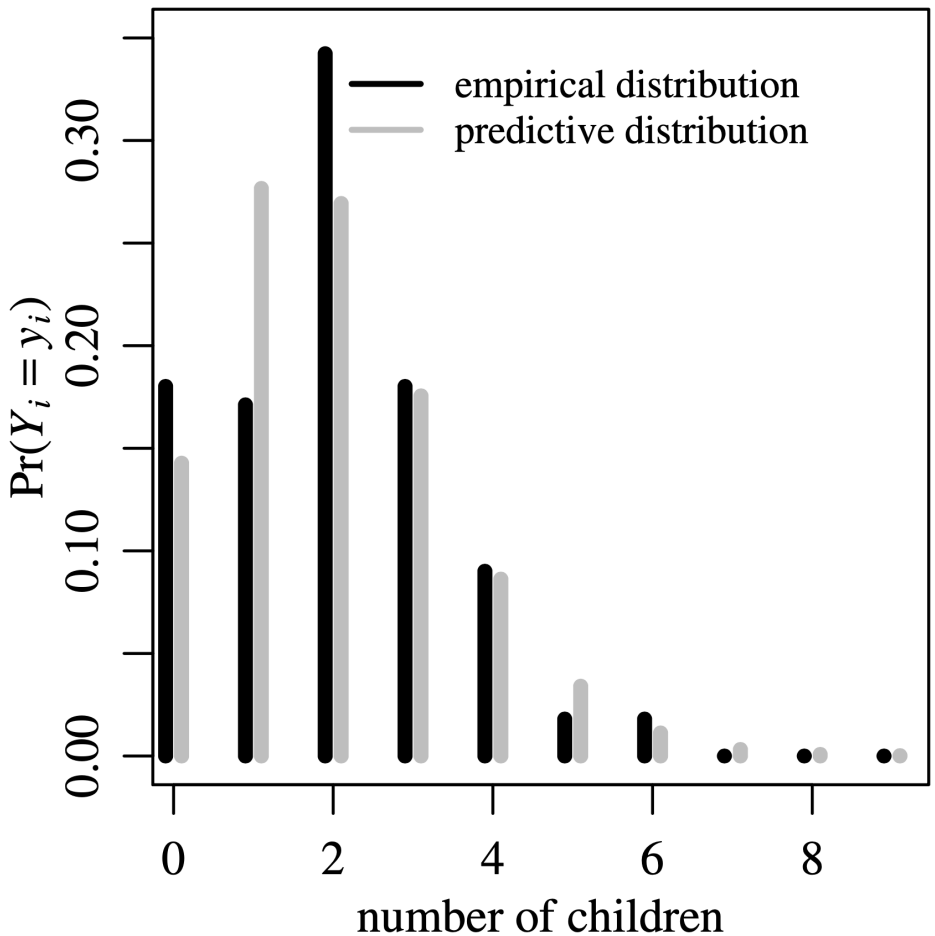
- Suppose that we observe data $y_1, ..., y_n$ and we derive the posterior distribution $p(\theta|y_1, ..., y_n)$.
- How can we check if our model assumptions are correct and if our model is fitting the data well?
- One way is to plot the empirical distribution of the data and the posterior predictive distribution and check if they have similar shapes. Significant discrepancies between the two might indicate that the model assumptions are not correct.

Posterior predictive model checking

- Example: Consider again the example on the number $Y_{1,1}, ..., Y_{n_1,1}$ of children of women without college degree who were in their 20s in the 1970s.
- We modeled $Y_{1,1}, ..., Y_{n_1,1}|\theta_1 \overset{iid}{\sim} \text{Poisson}(\theta_1)$ and $\theta_1 \sim \text{Gamma}(2, 1)$.
- With the data observed, we obtained the posterior distribution $p(\theta_1|y_{1,1}, ..., y_{n_1,1}) = \text{Gamma}(219, 112)$ and posterior predictive distribution $p(\tilde{y}_1|y_{1,1}, ..., y_{n_1,1}) = \text{NegBin}(219, \frac{1}{113})$.
- We can plot the empirical distribution of the data and the posterior predictive distribution.

Posterior predictive model checking

- We sampled $B = 10{,}000$ times from the posterior predictive distribution $p(\tilde{y}_1 | y_{1,1}, ..., y_{n_1,1})$.
- Posterior predictive distribution for the number of children of women without college degree and empirical distribution of the data



Posterior predictive model checking

- These two distributions seem to be in conflict. If the observed data have twice as many women with two children than one, why should we be predicting otherwise?

Posterior predictive model checking: posterior p-value

- Another way to check model assumptions is by simulating B replicates of the data under the model assumptions and compute what are called posterior p-values.
- Suppose we have n observations $y_1, ..., y_n$ of n conditionally i.i.d. random variables for which we assume the following model: $Y_1, ..., Y_n|\theta \overset{iid}{\sim} p(y|\theta)$ and $\theta \sim p(\theta)$.
- Let $t(y_1, ..., y_n)$ be a test statistic and let t_{obs} be the value of the test statistic on the observed data.
- How do we do compute posterior p-values?

  Repeat the following for $i = 1, \ldots, B$:
  - We sample $\theta^{(i)}$ from the posterior distribution $p(\theta|y_1, ..., y_n)$, THEN we sample n independent values $\tilde{y}_1^{(i)}, ..., \tilde{y}_n^{(i)}$ from $p(\tilde{y}|\theta^{(i)})$ and we compute $\tilde{t}^{(i)} = t(\tilde{y}_1^{(i)}, ..., \tilde{y}_n^{(i)})$.

Posterior predictive model checking: posterior p-value

- The posterior p-value is defined as:

$$\text{Posterior p-value} = P\left(\tilde{t}(\tilde{y}_1, ..., \tilde{y}_n) > t_{\text{obs}} | y_1, ..., y_n\right)$$

- The posterior p-value is a posterior probability and can be interpreted as the probability of observing in the future samples where the test statistic is as high or higher than the one already observed.

- Values around 0.5 indicate that the distributions of the replicated data (i.e., the posterior predictive distribution) and the actual data are close.
  Values close to 0 or 1 indicate differences between them.

- Note that posterior p-values must not be used or interpreted as the "probability that the model is true"!

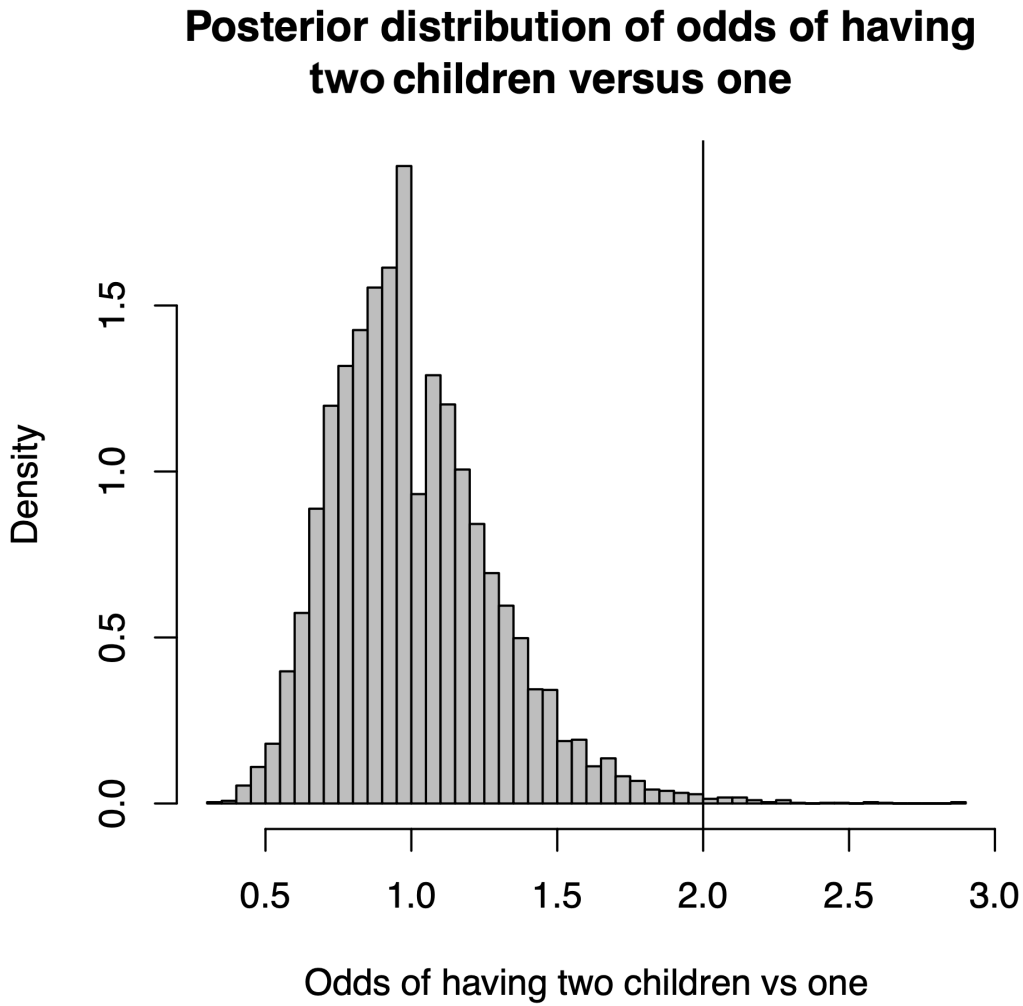
- In practice we compute the posterior p-values by looking at how many of the B values $\tilde{t}^{(1)}, ..., \tilde{t}^{(B)}$ are greater than t_{obs}.

Posterior predictive model checking: posterior p-value

- Example: Consider again the example on the number $Y_{1,1}, ..., Y_{n_1,1}$ of children of women without college degree.
- We have already seen that the empirical distribution of the data and the posterior predictive distribution differ in the probability of having 1 and 2 children. We want to compute now a posterior p-value.
- We choose as test statistic $t(y_1, ..., y_n)$ the odds of having 2 children versus one child.
- For the data observed: $t_{\text{obs}} = \frac{38}{19} = 2$.
- We simulate $B = 10{,}000$ datasets of size $n = 111$ from the posterior predictive distribution by simulating for each $i = 1, ..., 10{,}000$ first a $\theta_1^{(i)}$ from a Gamma$(129, 112)$, THEN $y_{1,1}^{(i)}, ..., y_{111,1}^{(i)}$ from a Poisson$(\theta_1^{(i)})$ and computing $\tilde{t}^{(i)}$.
- We obtain: posterior p-value = 0.05.

Posterior predictive model checking: posterior p-value

- Plot of the empirical distribution of the test statistics $\tilde{t}^{(i)}$ given the data $y_{1,1}, ..., y_{n_1,1}$. The vertical line indicates the observed value t_{obs}.
- This seems to indicate that the Poisson model is flawed.



Posterior predictive model checking: words of caution

- If the posterior predictive model checking indicates that the model is not adequate for the data, it does not necessarily mean that we need to consider a different model.
- It depends on what aspects of the model we are interested in.
- If the interest is in deriving the true $p(y_1)$, then the Poisson model is not a good choice for the example considered.
- However, the sample mean and the sample variance for the data collected are respectively 1.95 and 1.90, thus a Poisson model catpure these aspects of the distribution of the data.
- In practice, it is important that the datasets generated from the posterior predictive distribution match the observed data in terms of aspects that are of interest.
- To compute the posterior p-values, the data are used twice: for estimation of the posterior predictive density and for comparison between the predictive density and the data.

Posterior predictive model checking: PPO

- Another posterior predictive model checking statistic is the posterior predictive ordinate (PPO).
- Suppose we have n observations $y_1, ..., y_n$ of n conditionally iid random variables with: $Y_1, ..., Y_n|\theta \overset{iid}{\sim} p(y|\theta)$ and $\theta \sim p(\theta)$.
- The PPO is:

$$\text{PPO}_i = p(\tilde{Y} = y_i|y_1, ..., y_n) = \int_\Theta p(y_i|\theta) \cdot p(\theta|y_1, ..., y_n)d\theta$$

  and it provides the probability of observing again y_i after having observed $y_1, ..., y_n$.
- Low PPO_i values indicate observations originating from the tails of the distribution and extreme low PPO_i values indicate outliers.
- A large amount of observations y_i with small PPO_i may indicate a poorly fitted model.
- Note: The scaling of the PPOs depends on the structure of the assumed sampling model.

Posterior predictive model checking: CPO

- To avoid using the data twice, a predictive model checking statistic that can be calculated is the conditional predictive ordinate (CPO).
- The CPO for observation y_i is given by:

$$\begin{aligned}\text{CPO}_i &= p(\tilde{Y} = y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n) = P(\tilde{Y} = y_i | \boldsymbol{y}_{-i}) \\ &= \int_\Theta p(y_i|\theta) \cdot p(\theta | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n) d\theta \\ &= \int_\Theta p(y_i|\theta) \cdot p(\theta | \boldsymbol{y}_{-i}) d\theta\end{aligned}$$

  and it provides a method for tracing outliers.
- Small CPO values indicate observations that are not expected under the posterior predictive distribution of the current model.