COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Monte Carlo Approximation

- When we know the exact posterior distribution of θ (for example, when we use conjugate priors), then it is relatively straightforward to compute the posterior mean and variance.
- However, often we will want to summarize other aspects of a posterior distribution.
- For example, we might be interested in a contrast $|\theta_1 - \theta_2|$ or a ratio θ_1/θ_2.
- Obtaining exact theoretical formulas for these posterior quantities can be difficult and, often impossible.
- If we can generate random sample values of the parameters from their posterior distributions, then all these posterior quantities of interest can be approximated to an arbitrary degree of precision using the *Monte Carlo method*.

Recall Example 1: Birth rates (p. 48)

Over the course of the 1990s the General Social Survey gathered data on the educational attainment and number of children of 155 women who were 40 years of age at the time of their participation in the survey. In this example we will compare the women with college degrees to those without in terms of their numbers of children. We have the model:

$$X_{1,1}, \dots, X_{n_1,1} | \theta_1 \sim Poisson(\theta_1)$$
$$X_{1,2}, \dots, X_{n_2,2} | \theta_2 \sim Poisson(\theta_2)$$

where Group 1 is women without college degree, and Group 2 is women with degrees. Consider independent priors

$$\theta_1, \theta_2 \sim \Gamma(2, 1)$$

Then the posterior distributions are independent with:

$$\theta_1 | x_{1,1}, \dots, x_{n_{1,1}} \sim \Gamma(219, 112)$$
$$\theta_2 | x_{1,2}, \dots, x_{n_{2,2}} \sim \Gamma(68, 45)$$

We were interested in P($\theta_1 > \theta_2$).

On the side

General question: Suppose that we have independent variables

$$X \sim \Gamma(a_1, b_1)$$
$$Y \sim \Gamma(a_2, b_2)$$

Find P($X > Y$).

We will use the following fact from probability theory:

If $U \sim \Gamma(a_1, 1)$ & $V \sim \Gamma(a_2, 1)$ are independent, then

$$\frac{U}{U+V} \sim Beta(a_1, a_2)$$

In our case, $b_1 X \sim \Gamma(a_1, 1), b_2 Y \sim \Gamma(a_2, 1)$. Therefore,

$$P(X > Y) = P(b_1 X > (b_1 + b_2)Y - b_2 Y)$$

$$= P\left(\frac{Y}{b_1 X + b_2 Y} < \frac{1}{b_1 + b_2}\right) = P\left(\frac{b_2 Y}{b_1 X + b_2 Y} < \frac{b_2}{b_1 + b_2}\right)$$

That is, $P(X > Y) = F_{\text{Beta}(a_2, a_1)}\left(\frac{b_2}{b_1 + b_2}\right)$

Example 1 Solution

In our example the posterior distributions are independent with:

$$\theta_1 | x_{1,1}, \dots, x_{n_{1,1}} \sim \Gamma(219, 112)$$
$$\theta_2 | x_{1,2}, \dots, x_{n_{2,2}} \sim \Gamma(68, 45)$$

Then, $P(\theta_1 > \theta_2) = F_{\text{Beta}(68,219)}\left(\frac{45}{112+45}\right)$

In R we use `pbeta(0.2866, 68, 219) = 0.9725039`

All of this can (and should be!) avoided with a Monte Carlo approximation.

Monte Carlo Approximation

Suppose we could sample some number S of *independent*, random θ-values from the posterior distribution $f(\theta|x_1,..., x_n)$:

$$\theta^{(1)}, \ldots, \theta^{(S)} \sim f(\theta|x_1,..., x_n)$$

The empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ is known as a *Monte Carlo approximation* to $f(\theta|x_1,..., x_n)$. From this sample and empirical distribution, we can compute pretty much any numerical characteristic of the posterior distribution or even approximate the whole posterior pdf. If the posterior distribution is one of the "named" distributions, then the task of sampling from it is particularly straightforward.

The `R` system for generating random numbers

- For a given well-known distribution you have four choices
- Here the normal distribution is given as an example
- Give me 100 observations from standard normal: `rnorm(100)`
- The density of standard normal at given point x: `dnorm(x)`
- The cumulative distribution at given x: `pnorm(x)`
- The quantile at given probability p (that is, inverse cdf): `qnorm(p)`

The `R` system for generating random numbers

The following “famous” distributions are available in R base:

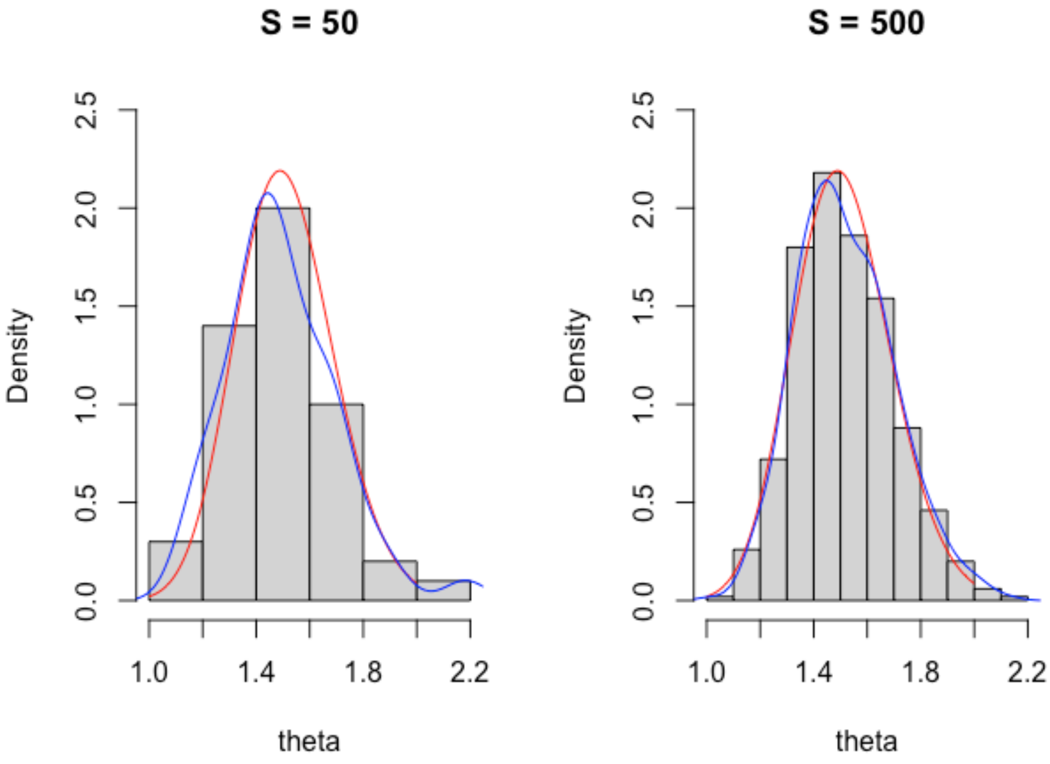
- Beta: `beta`
- Binomial: `binom`
- Cauchy: `cauchy`
- Chi-squared: `chisq`
- Exponential: `exp`
- Fisher F: `f`
- Gamma: `gamma`
- Geometric: `geom`
- Hypergeometric: `hyper`
- Negative binomial `nbinom`
- Normal: `norm`
- Poisson: `pois`
- Student t: `t`
- Uniform: `unif`

Example 1 Continued

Consider the posterior of θ_2

$$\theta_2 | x_{1,2}, \ldots, x_{n_{2,2}} \sim \Gamma(68, 45)$$

Estimated posterior from simulated samples:



Exercise 1

Simulate the posterior of

$$\theta_2 | x_{1,2}, \ldots, x_{n_{2,2}} \sim \Gamma(68, 45)$$

and compare the histogram to the true pdf for various values of *S*.

Estimating Numerical Characteristics

Let g(θ) be a function of the parameter. By the LLN we have that if $\theta^{(1)}, \ldots, \theta^{(S)}$ are iid from $f(\theta|x_1, \ldots, x_n)$, then as $S \to \infty$

$$\frac{1}{S}\sum_{i=1}^{S} g\left(\theta^{(i)}\right) \to E(g(\theta)|x_1, \ldots, x_n) = \int g(\theta) f(\theta|x_1, \ldots, x_n) d\theta$$

Notice this includes very broad collection of results, like

$$\bar{\theta} \to E(\theta|x_1, \ldots, x_n)$$

$$\frac{\#\left(\theta^{(i)} < c\right)}{S} \to P(\theta < c|x_1, \ldots, x_n)$$

The sample can also be used to estimate variance, median, quantiles, …

Example 1 Continued

Consider the posterior of θ_2

$$\theta_2 | x_{1,2}, \ldots, x_{n_{2,2}} \sim \Gamma(68, 45)$$

- Then the posterior mean is known to be

$$E(\theta_2 | x_1, \ldots, x_n) = \frac{68}{45} = 1.51$$

- $\mathrm{P}(\theta_2 < 1.46 | x_1, \ldots, x_n)$ can be computed "exactly" with

```
pgamma(1.46, 68, 45) = 0.4
```

- However, the median of the Gamma distribution does not have any closed form formula!
- We will estimate all of these in R with Monte Carlo simulation.

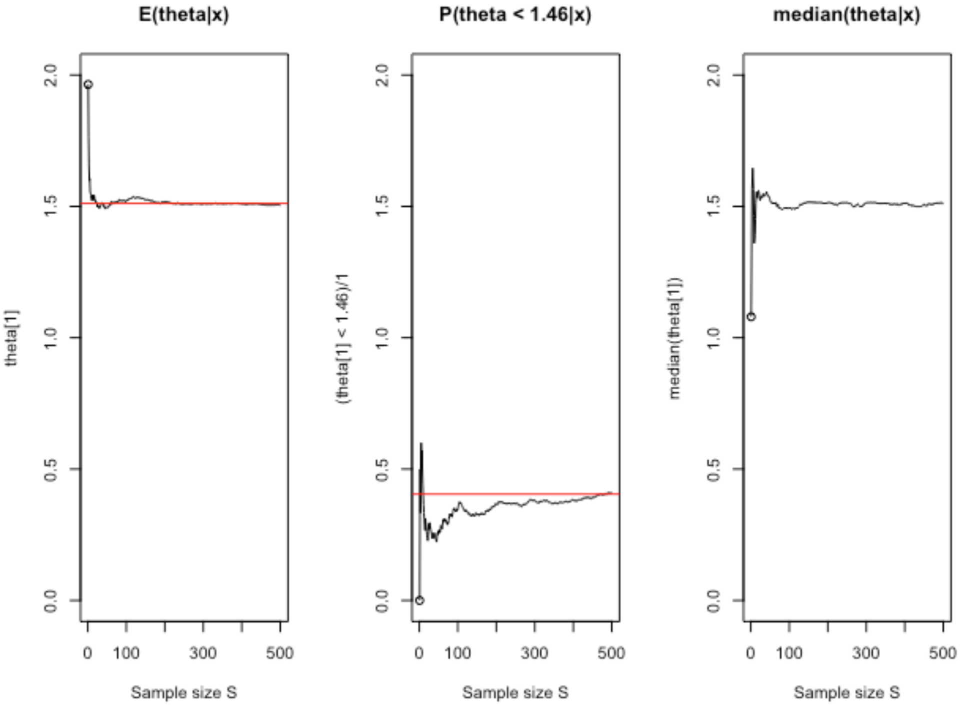
Example 1 Continued

Consider the posterior of θ_2

$$\theta_2 | x_{1,2}, \ldots, x_{n_{2,2}} \sim \Gamma(68, 45)$$

Then the posterior mean is known to be $E(\theta_2 | x_1, \ldots, x_n) = \frac{68}{45} = 1.51$

$\mathrm{P}(\theta_2 < 1.46 | x_1, \ldots, x_n)$ can be computed “exactly” `pgamma(1.46, 68, 45) = 0.4`



Monte Carlo Standard Error

Monte Carlo standard errors can be obtained to assess the accuracy of approximations to posterior means. Let

$$\bar{\theta} = \frac{\sum_{i=1}^{S} \theta^{(i)}}{S}$$

Then the CLT guarantees that $\bar{\theta} \approx N\left(\theta, \sqrt{\frac{Var(\theta|x_1, \dots, x_n)}{S}}\right)$

The Monte Carlo standard error is the approximation to this standard deviation. Let

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{S}\left(\theta^{(i)} - \bar{\theta}\right)^2}{S-1}$$

Then the Monte Carlo standard error is:

$$\sqrt{\frac{\hat{\sigma}^2}{S}}$$

Exercise 1 Continued

Calculate Monte Carlo SEs of the quantities in Example 1 and obtain 95% CI for them.

Posterior Inference for Functions of θ

Suppose we are interested in the posterior distribution of some computable function $g(\theta)$ of θ. In the binomial model, for example, we are sometimes interested in the logit:

$$logit(\theta|x_1, \ldots x_n) = \log \frac{\theta}{1-\theta} = \gamma$$

The LLN and the continuous mapping theorem guarantees that if $\theta^{(1)}, \ldots, \theta^{(S)}$ are iid from $f(\theta|x_1, \ldots, x_n)$ then the sample average of $\log \frac{\theta^{(i)}}{1-\theta^{(i)}}$ converges to $E\left(\log \frac{\theta}{1-\theta} \middle| x_1, \ldots, x_n\right)$

That is, simulate the sequence $\gamma^{(1)}, \ldots, \gamma^{(S)}$ and use that sequence in the same way we used $\theta^{(1)}, \ldots, \theta^{(S)}$

Example 2: Logit

Fifty-four percent of the respondents in the 1998 General Social Survey reported their religious preference as Protestant, leaving non-Protestants in the minority. Respondents were also asked if they agreed with a Supreme Court ruling that prohibited state or local governments from requiring the reading of religious texts in public schools. 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. Using a binomial model and a uniform prior distribution, the posterior distribution of θ is Beta(442, 420).

We will obtain Monte Carlo approximation of $\log \frac{\theta}{1-\theta}$

Exercise 2

Use the data from Example 1 and obtain a Monte Carlo approximation of the posterior distribution of $\frac{\theta_1}{\theta_2}$

Predictive Distributions

If we knew the true value of the parameter, then the predictive distribution of a new data point is $f(x|\theta)$. However, we will never know θ, so we must integrate over all possible θ values.

If we did not have any sample data from the population, our predictive distribution would be obtained by integrating out θ using the *prior*:

$$f(x_{new}) = \int f(x_{new}|\theta)\pi(\theta)d\theta$$

A predictive distribution that integrates over unknown parameters but is *not* conditional on observed data is called a *prior predictive distribution*.

Such a distribution can be useful in evaluating if a prior distribution for θ actually translates into reasonable prior beliefs for observable data x_{new}

Predictive Distributions

After we have observed a sample $X_1, \ldots, X_n$ from the population, the relevant predictive distribution for a new observation becomes

$$f(x_{new}|x_1, \ldots, x_n) = \int f(x_{new}|\theta, x_1, \ldots, x_n)\, f(\theta|x_1, \ldots, x_n) d\theta$$

$$= \int f(x_{new}|\theta)\, f(\theta|x_1, \ldots, x_n) d\theta$$

This is called a *posterior predictive distribution*.

Very often $f(x_{new}|x_1, \ldots, x_n)$ will be too complicated or even impossible to sample from directly.

What do we do?

Sampling from Predictive Distributions

General result: If we need a sample from $f(x)$ which is too complicated, but we can write $f(\mathrm{x}) = f(x|y)f(\mathrm{y})$ for some other variable y, then we can obtain the sample sequentially:

1. Draw y from $f(y)$
2. Draw x from $f(x|\mathrm{y})$
3. Ignore the y and keep only the x.
4. Start over at step 1.

In our case this means:

sample $\theta^{(1)} \sim f(\theta|x_1, \ldots, x_n)$, then sample $x_{new}^{(1)} \sim f(x_{new}|\theta)$

…

sample $\theta^{(S)} \sim f(\theta|x_1, \ldots, x_n)$, then sample $x_{new}^{(S)} \sim f(x_{new}|\theta)$

From the resulting sequence $\{(\theta, x_{\mathrm{new}})^{(1)}, \ldots (\theta, \mathrm{x}_{\mathrm{new}})^{(S)}\}$ keep only the subsequence $\{x_{\mathrm{new}}{}^{(1)}, \ldots \mathrm{x}_{\mathrm{new}}{}^{(S)}\}$.

Example 3

In the Poisson example suppose we want to calculate the predictive probability that an age-40 woman without a college degree would have more children than an age-40 woman with a degree:

$$P\left(X_1^{(new)} > X_2^{(new)} \middle| data\right)$$

$$= \sum_{x_2=0}^{\infty} \sum_{x_1=x_2+1}^{\infty} dbnbinom(x_1, 219, 212) \times dnbinom(x_2, 68, 45)$$

The answer is 0.48 and we will compute it in R using 3 different methods.

Exercise 3

In the above example, obtain the graph of the distribution of

$$X_1^{(new)} - X_2^{(new)}$$

Monte Carlo Integration

- One of the factors that restricted the early development of Bayesian methods was the complexity of the integrals involved in determining posterior distributions and the associated estimators.
- *Monte Carlo integration* is an algorithm which allow us to construct probabilistic estimates of these integrals.
- This method is useful not only in Bayesian statistics, but even for classic methods, like the MSE of an estimator:

$$MSE\left(\hat{\theta}\right) = E\left[\left(\hat{\theta}(x_1, \dots, x_n) - \theta\right)^2\right] = \int \left(\hat{\theta}(x_1, \dots, x_n) - \theta\right)^2 f_{\hat{\theta}}(x_1, \dots, x_n)\, dx_1 \dots dx_n$$

- For many estimators and models that are of practical value, it is often the case that this integral is intractable.
- In such cases the only option is to approximate the integral numerically.

Monte Carlo Integration

- Suppose that we need to evaluate any integral where $g(x)$ is not a pdf

$$J = \int_a^b g(x)dx$$

- Suppose also that $f(x)$ is any pdf with support $[a, b]$. Then

$$J = \int_a^b g(x)dx = \int_a^b \frac{g(x)}{f(x)} f(x)dx = E\left[\frac{g(X)}{f(X)}\right]$$

- Consider we have a random sample $X_1, \ldots, X_n$ from $f(x)$. Then by LLN and the continuous mapping theorem we have that

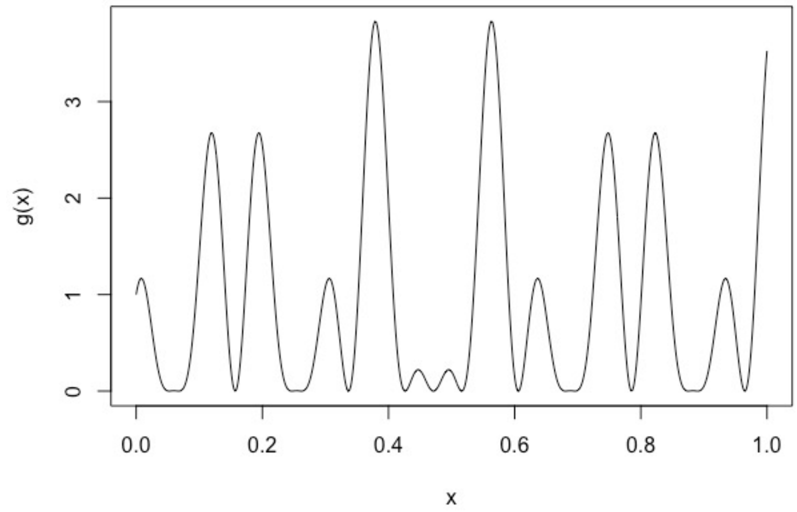
$$\hat{J} = \frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)} \xrightarrow{P} E\left[\frac{g(X)}{f(X)}\right] = J$$

Example 4

Compute the integral

$$\int_0^1 (\cos 50x \ + \sin 20x)^2 dx$$

For this we can use $f(x) = 1$ for U(0, 1) distribution.



Exercise 4

Approximate the integral

$$\int_0^4 \sqrt{x + \sqrt{x + \sqrt{x + \sqrt{x}}}}\, dx$$

Answer: 7.6766

How to generate random numbers from "unknown" distributions

- The uniform distribution is always the starting point
- Apply a function: `-log(runif(100))` gives exponential distribution
- Inverse transform: compute quantiles of desired distribution (if possible)
- Acceptance-rejection sampling
- Combinations of distributions
- Mixtures

Simple transformations

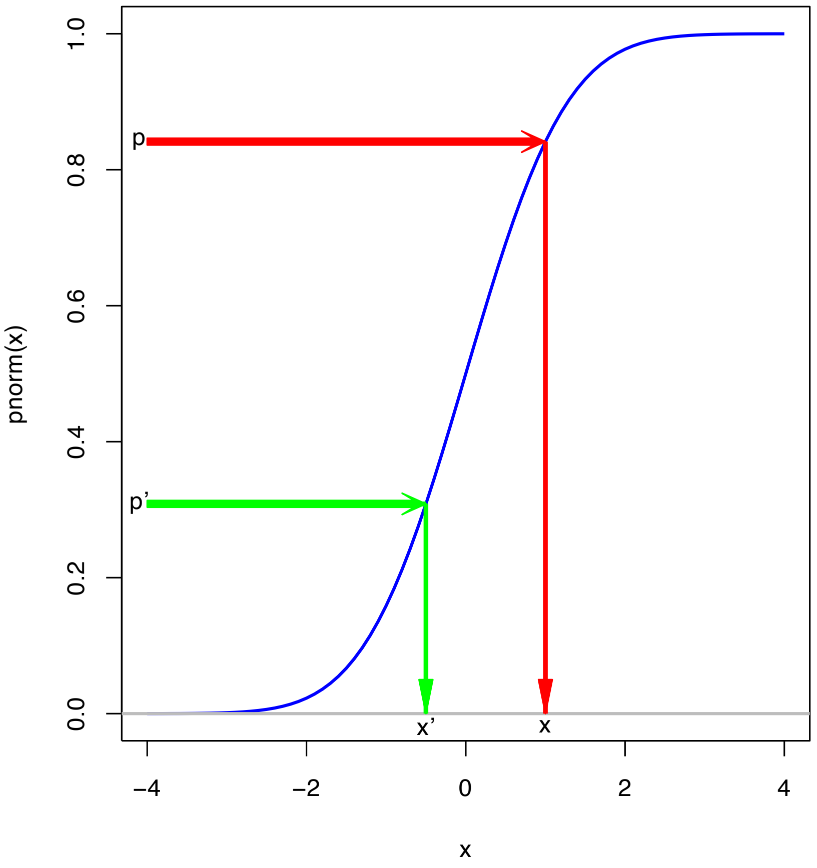
- Apply a function to uniformly distributed random numbers
- Example: `y = -log(runif(n))` gives exponential distribution
- Sums of these leads to the gamma distribution
- Many of such “tricks” exist
- See examples in Rizzo’s book on computational statistics

Inverse transform or quantiles

- Remember the cumulative distribution: $p = F(x)$

- Where $F(x) = \int_{-\infty}^{x} f(t)dt$
- And $f(x)$ is the probability density (pdf)
- We have that $0 < p < 1$
- Turn things around: $x = F^{-1}(p)$
- Generate uniformly distributed p
- Compute $x = F^{-1}(p)$
- Then x will have density f
- Of course, you need a procedure to invert F
- Example: qnorm(runif(n)) is equivalent to rnorm(n)

Illustration of the quantile approach



A special case: the exponential distribution

- Density: $f(x) = e^{-x}$ (positive x)
- Cumulative distribution: $F(x) = 1 - e^{-x}$
- You can check this by integrating it yourself
- From $p = 1 - e^{-x}$ follows $x = -\log(1 - p)$
- But if p from uniform distribution, so is $1 - p$
- Hence $x = -\log(p)$ has exponential distribution

Inverting the cumulative distribution

- You need a formula for $F^{-1}(p)$
- In many cases an exact formula is not available
- But over the years very good approximations were found
- Try ?qnorm to see pointers to literature
- Luckily, R has many built-in procedures