## Monte Carlo Approximation

What quantities associated with the posterior do we often need?

- Probabilities
- Summary measures, such as mean, mode, standard deviation and percentiles
- Distributions of *functions* of $\boldsymbol{\theta}$

There are cases where some or all of these quantities can be difficult or impossible to compute exactly. In such cases Monte Carlo methods can be very helpful.

*Example 5 revisited* Suppose the company decides they want to infer $1/\theta$, which is the mean lifetime of their electrical components.

We know that the posterior of $\theta$ for their data is $\mathrm{gamma}(50, 37815)$.

It follows that the posterior of $1/\theta$ is an *inverse-gamma* distribution. We could use this fact to determine probabilities, moments, etc.

Let's pretend we don't know much math stat, and can't figure out the posterior of $1/\theta$. However, *we do know how to generate values from a gamma distribution*.

We can use this knowledge to approximate the posterior and related quantities.

A cool thing about simulating from a posterior is that *we can generate as many values as we wish, and thereby minimize errors in approximating quantities of interest*.

So, I'll generate 100,000 i.i.d. values from the gamma$(50, 37815)$ distribution, and then take the reciprocal of each one.

In so doing *we have a random sample of size 100,000 from the posterior of* $1/\theta$*.*

I'll use the R function `rgamma` to generate the observations.

I used the following R command to obtain a hundred thousand draws from a gamma distribution with $a = 50$ and $b = 37,815$:

```
theta=rgamma(100000,50,rate=37815)
```

Then I simply did

```
beta=1/theta
```

to get the necessary draws from the posterior of $1/\theta$.

Posterior mean=771.73

Sample mean of 100,000 values=771.99

Posterior SD=111.39

Sample SD of 100,000 values=111.57

A nice device for estimating a density from observations $X_1, \ldots, X_n$ is the *kernel density estimate*

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

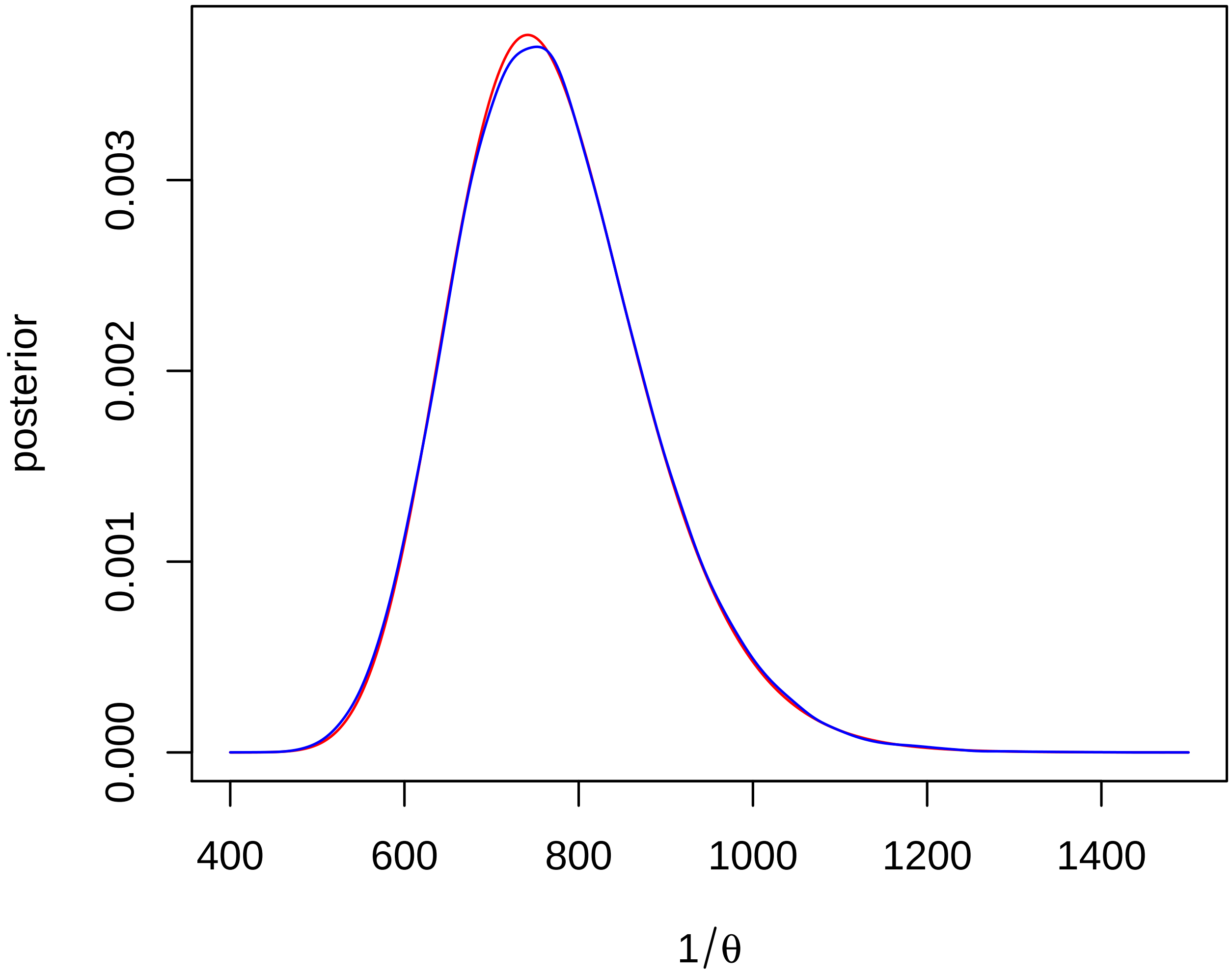
The function $K$ is a density called the *kernel*. It is often taken to be a standard normal density.

The *bandwidth* is $h$, a positive constant that controls the smoothness of the estimate.

In R the following command was used to obtain a plot of a kernel estimate of the posterior of $1/\theta$:

```
plot(density(beta,from=400,to=1500,
        n=1000,bw=15))
```

*Actual posterior of* $1/\theta$ *and kernel estimate based on 100,000 draws from posterior*



One could use the kernel estimate to approximate an HPD region for $1/\theta$.

A "lazy," but often adequate, method of approximating a credible interval is to use empirical quantiles of the Monte Carlo data.

For a 95% credible interval, I can order our 100,000 draws, and then select the 2500th and 97,500th ordered observations.

This yields an approximate credible interval of

$$(583.1, 1019.5).$$

The actual 95% credible interval extending from the 2.5th to the 97.5th percentiles of the posterior is

$$(583.7, 1019.0).$$

## Sampling from predictive distributions

A *prior* predictive distribution provides a means of predicting a data value before any data have been observed.

A *posterior* predictive distribution provides an enhanced means of predicting a future datum given a set of data that has already been observed.

Given a probability model $p(\boldsymbol{y}|\boldsymbol{\theta})$ and a prior $p(\boldsymbol{\theta})$, the prior predictive distribution for a future data vector $\boldsymbol{y}$ is

$$m(\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Given a vector of observations $\boldsymbol{y}$, the posterior predictive distribution of a future vector $\boldsymbol{y}_f$ is

$$m(\boldsymbol{y}_f|\boldsymbol{y}) = \frac{1}{m(\boldsymbol{y})} \int_{\Theta} p(\boldsymbol{y}_f, \boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

If $\boldsymbol{Y}_f$ and $\boldsymbol{Y}$ are independent conditional on $\boldsymbol{\theta}$, then

$$\begin{aligned} m(\boldsymbol{y}_f|\boldsymbol{y}) &= \int_\Theta p(\boldsymbol{y}_f|\boldsymbol{\theta}) \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{m(\boldsymbol{y})}\, d\boldsymbol{\theta} \\ &= \int_\Theta p(\boldsymbol{y}_f|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}. \end{aligned}$$

Even if the distributions in the last integrand are "familiar," the integral may not be, and Monte Carlo can be a useful way to approximate the predictive distribution.

Anytime I want to approximate a distribution, I can in principle do so by generating a large random sample from the distribution, and then computing the empirical distribution or kernel density estimate from all the observations.

*How can I generate from $m(\,\cdot\,|\boldsymbol{y})$ when I don't know its form?*

If I know how to generate $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\boldsymbol{y})$, and how to generate $Y_f$ from $p(\boldsymbol{y}_f|\boldsymbol{\theta})$, then I know how to generate draws $(\boldsymbol{Y}_f, \boldsymbol{\theta})$ from the conditional of $(\boldsymbol{Y}_f, \boldsymbol{\theta})$ given $\boldsymbol{y}$.

Then I just use all the values of $\boldsymbol{Y}_f$ as my sample from the marginal of $\boldsymbol{Y}_f$ given $\boldsymbol{y}$.

## Digression

There are a couple of interesting ideas at work in the method described at the top of the page.

1. First of all we have the simple observation that if $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a bivariate distribution, then $X_1, \ldots, X_n$ is a random sample from the marginal distribution of $X_i$.

2. We may know how to generate a bivariate random sample even when we don't know how to generate from one of the two marginals. Here we use the idea of a *hierarchical model*, which is a prevalent idea in Bayesian methods.

Suppose we know that $X$ and $Y$ have joint density $f(x, y)$. We want to be able to generate values of $X$ but we don't know how to.

Suppose that we know how to generate

- values of $Y$, and
- values from the conditional distribution of $X$ given $Y = y$.

This means we can generate a bivariate random sample, and hence (using idea **1**), we can get a random sample $X_1, \ldots, X_n$.

We have

$$f(x,y) = f(x|y)f_Y(y).$$

The right hand side is an example of a *hierarchical distribution*. At the first level we have the distribution of $Y$, and at the second level the distribution of $X$ given $Y$.

This notion is used in Gibbs sampling (a type of MCMC) and the construction of hierarchical models.

---

---

*Example 8* Suppose, as in Example 5, that $Y_1, \ldots, Y_n$ are a random sample from the exponential density $f(y|\theta) = \theta e^{-\theta y} I_{(0,\infty)}(y)$.

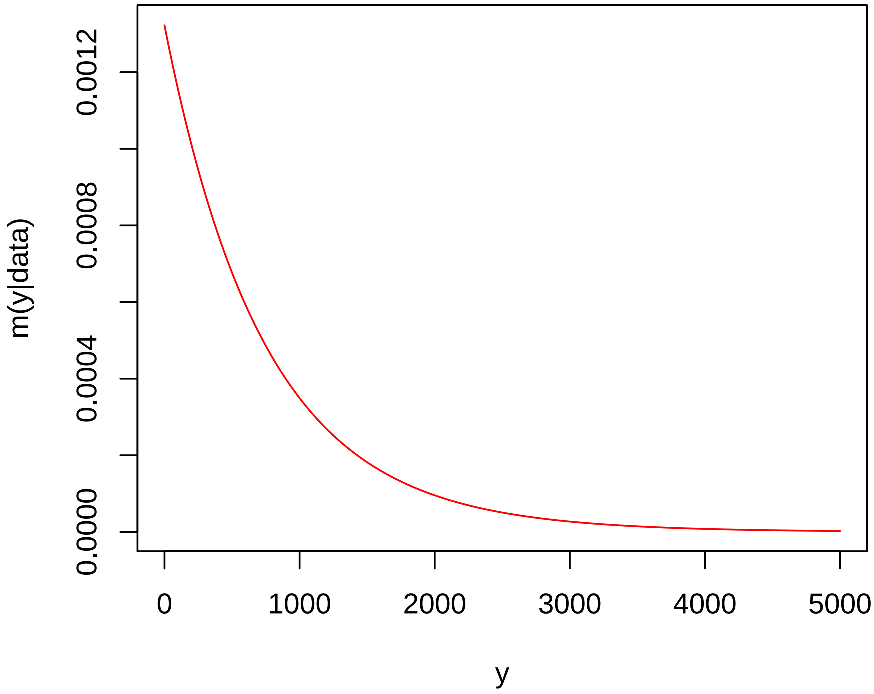
A gamma$(a,b)$ density will be used as prior. Having observed $y_1, \ldots, y_n$ with corresponding sample mean $\bar{y}$, suppose we want to predict a future observation $Y$ that is independent of $Y_1, \ldots, Y_n$ given $\theta$.

The posterior predictive distribution of $Y$ is

$$m(y|\bar{y}) = \int_0^\infty \theta e^{-\theta y} \cdot \frac{(b+n\bar{y})^{n+a}}{\Gamma(n+a)} \theta^{n+a-1} \times e^{-(b+n\bar{y})\theta}\, d\theta$$

$$= \frac{(n+a)(b+n\bar{y})^{n+a}}{(y+b+n\bar{y})^{n+a+1}}.$$

*Posterior predictive distribution for the setting of Example 5:* $a = 0$, $b = 0$, $n = 50$, $\bar{y} = 756.3$



Let's pretend we don't know what the posterior predictive distribution is. We'll generate values of $Y$ as described on p. 84N.

## Algorithm

- Generate $\theta$ from gamma$(50, 37815)$.

- Generate $Y$ from the exponential distribution with rate $\theta$, where $\theta$ is the value generated in the previous step.

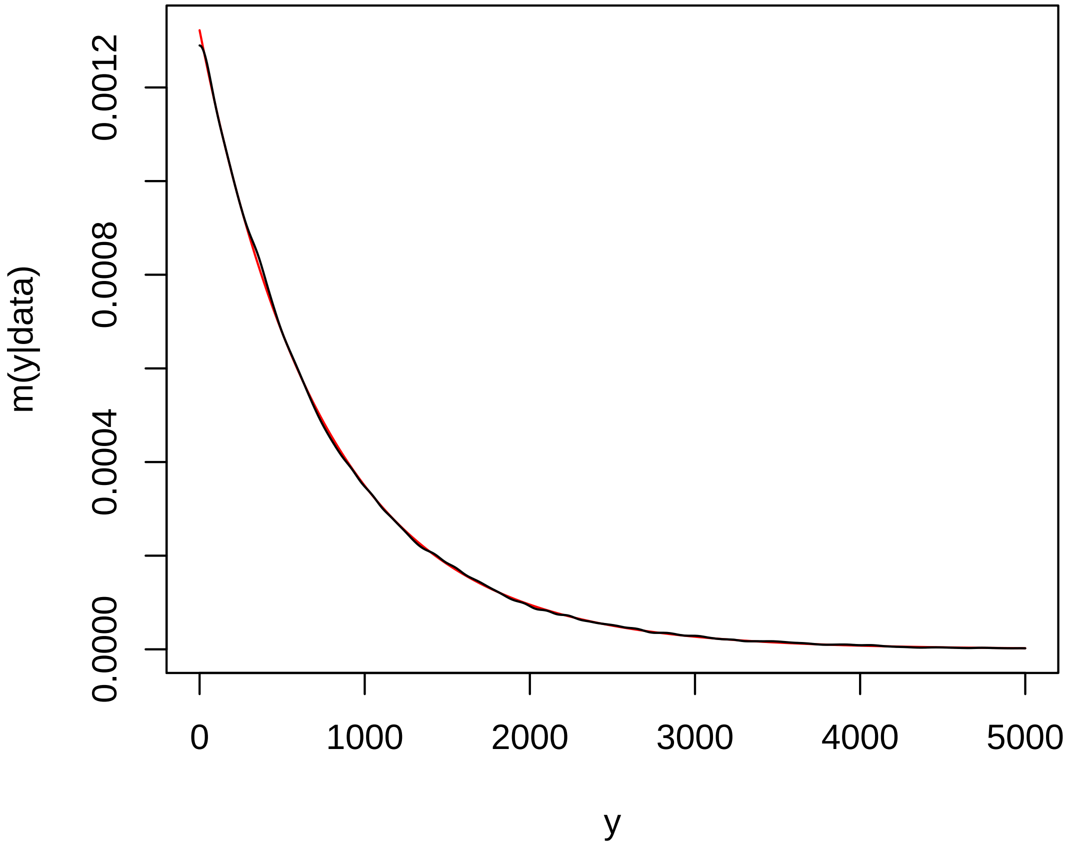
- Repeat the above steps 100,000 times, thus obtaining a random sample $Y_1, \ldots, Y_{100{,}000}$ from the posterior predictive density.

You can check using calculus that the mean and standard deviation of the posterior predictive distribution are 771.7 and 787.6, respectively.

The sample mean and standard deviation of the 100,000 values I generated using the algorithm on the previous page were, respectively,

770.0 and 787.5.

Red line: Actual posterior predictive density
Black line: Kernel estimate



## Model checking via the posterior predictive distribution

Whenever a particular probability model $p(\boldsymbol{y}|\boldsymbol{\theta})$ is used, it is advisable to check whether it fits the data.

Suppose $Y_1, \ldots, Y_n$ are a random sample from $f(y|\boldsymbol{\theta})$. A common way for a frequentist to check model fit is to compare the empirical distribution of $Y_1, \ldots, Y_n$ with $f(y|\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is the MLE or some other estimate of $\boldsymbol{\theta}$.

A formal test, such as Kolmogorov-Smirnov (KS) or Cramér-von Mises, can be done. For example, the KS test statistic is

$$\sup_{-\infty<y<\infty} \left| F_n(y) - \int_{-\infty}^{y} f(t|\widehat{\boldsymbol{\theta}})\, dt \right|,$$

where $F_n$ is the empirical cdf.

There are formal Bayesian methods for checking the fit of a model, but an informal way of doing so is to *compare the empirical distribution with the posterior predictive distribution.*

If these two distributions are not relatively close to each other, then there is evidence that the probability model does not fit.

It turns out that this is similar to the frequentist method described on the previous page.

When $n$ is large, the posterior distribution is highly concentrated at the MLE $\widehat{\boldsymbol{\theta}}$ and hence

$$
\begin{aligned}
m(y|y_1,\ldots,y_n) &= \int_\Theta f(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|y_1,\ldots,y_n)\,d\boldsymbol{\theta} \\
&\approx f(y|\widehat{\boldsymbol{\theta}})\int_\Theta p(\boldsymbol{\theta}|y_1,\ldots,y_n)\,d\boldsymbol{\theta} \\
&= f(y|\widehat{\boldsymbol{\theta}}).
\end{aligned}
$$