# Monte Carlo

Statistics 110

Summer 2006

# Monte Carlo Example

**Example:** Bayesian Analysis of the ED50 of an Anti-pneumococcus serum

An experiment was performed to study an anti-pneumococcus serum. The serum was given to 200 mice (40 at each of 5 doses), the mice were exposed to pneumococcus (the bacterium associated with pneumonia) and observed for a week. Survival after the week was recorded for each mouse.

| Serum Dose (cc) | # Survived out of 40 |
|---|---|
| 0.0028 | 5 |
| 0.0056 | 19 |
| 0.0112 | 31 |
| 0.0225 | 34 |
| 0.0450 | 39 |

One factor of interest was the dose that would lead to a 50% survival rate (known as the $ED50$) based on the following Bayesian logistic model

$$\begin{aligned} X_i|\alpha,\beta &\overset{ind}{\sim} Bern(p_i) \\ \alpha &\overset{ind}{\sim} N(10,100) \\ \beta &\overset{ind}{\sim} N(1,100) \end{aligned}$$

where

$$\begin{aligned} p_i &= \frac{e^{\alpha+\beta\log(\text{dose}_i)}}{1+e^{\alpha+\beta\log(\text{dose}_i)}} \\ \log\frac{p_i}{1-p_i} &= \alpha+\beta\log(\text{dose}_i) \end{aligned}$$

The prior distributions on $\alpha$ and $\beta$ used here were chosen to make the problem concrete, but in general could be based on studies of similar sera.

To get a 50% survival rate, we need to find the dose satisfying

$$\frac{e^{\alpha+\beta\log(\text{dose})}}{1+e^{\alpha+\beta\log(\text{dose}_i)}} = 0.5 \Longleftrightarrow \alpha+\beta\log(\text{dose}) = 0$$

yielding

$$ED50 = e^{-\alpha/\beta}$$

Since $\alpha$ and $\beta$ are random variables in this setting, so $ED50$. Thus we are interested in the posterior distribution of $ED50$ given the survival data. Lets focus on the quantities

$$E[ED50|\mathbf{X}] = E\left[\frac{-\alpha}{\beta}|\mathbf{X}\right]$$

$$\text{Med}(ED50|\mathbf{X}) = \text{Med}\left(\frac{-\alpha}{\beta}|\mathbf{X}\right)$$

$$\text{SD}(ED50|\mathbf{X}) = \text{SD}\left(\frac{-\alpha}{\beta}|\mathbf{X}\right)$$

The posterior density of $\alpha$ and $\beta$, given the data $\mathbf{X} = \{X_1, \ldots, X_{200}\}$ satisfies

$$f(\alpha, \beta|\mathbf{X}) \propto \frac{1}{10\sqrt{2\pi}} \exp\left(\frac{-(\alpha-10)^2}{200}\right) \frac{1}{10\sqrt{2\pi}} \exp\left(\frac{-(\beta-1)^2}{200}\right) \times \prod_{i=1}^{200} \left(\frac{e^{\alpha+\beta\log(\text{dose}_i)}}{1+e^{\alpha+\beta\log(\text{dose}_i)}}\right)^{x_i} \left(\frac{1}{1+e^{\alpha+\beta\log(\text{dose}_i)}}\right)^{1-x_i}$$

This is not proportional to a known density thus its properties are not known. In addition, trying to calculate expected values based on this exactly are likely to be impossible.

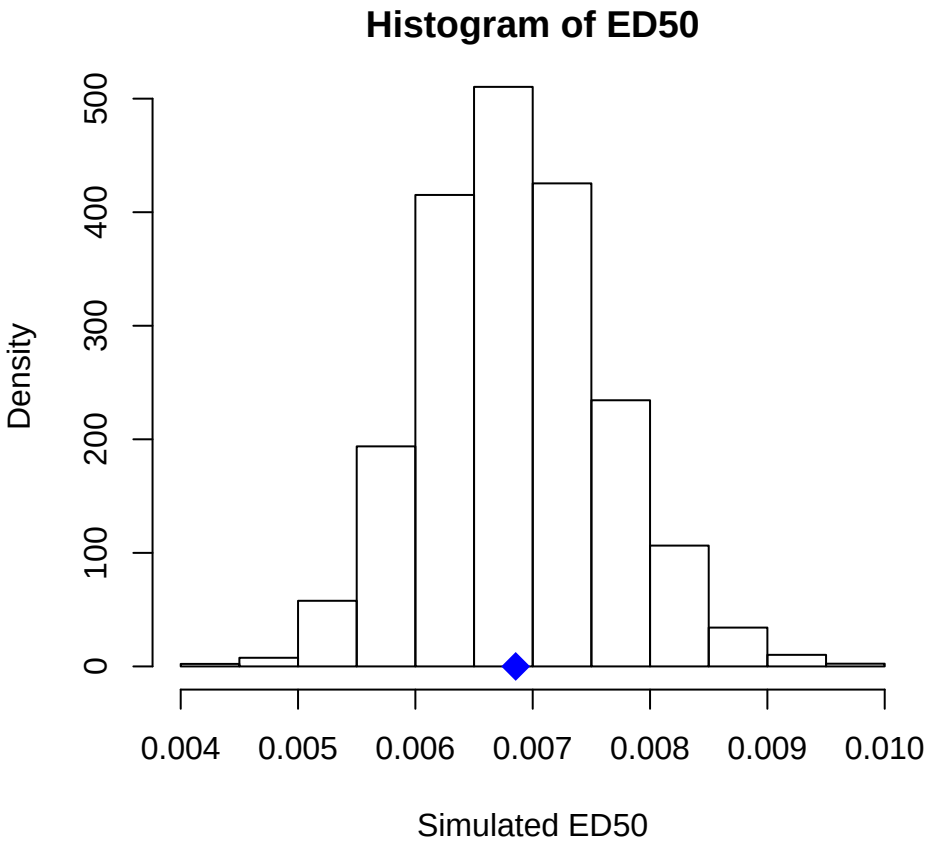
However it is possible to generate samples from $f(\alpha, \beta|\mathbf{X})$ and thus investigate properties of this distribution via Monte Carlo.

To study the posterior distribution of $ED50$, 10000 samples $(\alpha_i, \beta_i)$ were simulated from $f(\alpha, \beta|\mathbf{X})$ and

$$ED50_i = e^{-\alpha_i/\beta_i}$$

was calculated, giving 10000 samples from $f(ED50|\mathbf{X})$. Based on these samples

$$\hat{E}[ED50] = 0.00686$$
$$\text{Med}(ED50) = 0.00683$$
$$\widehat{\text{SD}}(ED50) = 0.00078$$




The sampling scheme used in this case was Metropolis-Hastings, a form Markov Chain Monte Carlo.

# Monte Carlo

Many computations involving distributions can not be done exactly as we have seen during to term. A popular approximation approach is Monte Carlo Simulation.

Suppose we are interested in a calculation of the form

$$E[g(X)] = \int g(x)f(x)dx \qquad X \text{ is continuous}$$

or

$$= \sum g(x_i)p(x_i) \qquad X \text{ is discrete}$$

Calculations that fit into this framework are

- Moments: $E[X]$, $\mathrm{Var}(X)$, etc

  These were of interest, for example, in the SST project

- Probabilities: $P[a \leq X \leq b]$

$$P[a \leq X \leq b] = E[I(a \leq X \leq b)]$$

We have seen this calculation done with the histograms of sampling distributions.

- True confidence level of a confidence interval on a population mean

  The common confidence interval for $\mu$, the population mean, is

$$\left(\bar{X} - t^* \frac{s}{\sqrt{n}}, \bar{X} + t^* \frac{s}{\sqrt{n}}\right)$$

  where $t^* = t_{1-\alpha/2}$.

  If $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, the true coverage probability (confidence level) will be $1 - \alpha$. However if the data is sampled from a different distribution, the true confidence level depends on how well the normal approximation to the distribution $\bar{X}$ works.

  Want to know the real probability that a randomly chosen sample leads to having the truth in the interval, i.e.

$$P\left[\bar{X} - t^* \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t^* \frac{s}{\sqrt{n}}\right]$$

- Width of confidence intervals

  In many situations, confidence intervals for parameters are of the form

  $$\text{Est} \pm t^* S_{\text{Est}}$$

  (such as the previous case). So the width of the interval (Width $= 2t^* S_{\text{Est}}$) is a random quantity. Often of interest are

  $$E[\text{Width}] \quad \text{or} \quad \text{Var}(\text{Width})$$

Sample $X_1, X_2, \ldots, X_n$ from CDF $F_X(x)$ and approximate $\mu_g = E[g(X)]$ by

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

Under certain regularity conditions

$$\frac{1}{n} \sum_{i=1}^{n} g(x_i) \to E[g(X)]$$

by the Law of Large Numbers.

In addition, if $\mathrm{Var}(g(X)) < \infty$ (and maybe other regularity conditions), by the Central Limit Theorem

$$\frac{\bar{g} - \mu_g}{\mathrm{SD}(\bar{g})} \xrightarrow{\mathcal{D}} N(0, 1)$$

Thus, we can get as accurate an approximation to $E[g(X)]$ by Monte Carlo as desired.

## Issues

While the basic scheme appears to be easy, there are a number of issues to consider. We want to be able to get a precise answer quickly. The basic approach may be inefficient.

- Distribution of $X$
  - Univariate versus multivariate
  - Form of distribution. Do we know the density exactly or the basic form of it.
- Function being integrated
- Sampling Scheme
  - Independent and identically distributed draws
  - Simple Random Sample vs Stratfied vs ???
  - Importance Sampling (sample from a different distribution)
  - Dependent samples (Markov Chain Monte Carlo)

# Simulation of Random Variables

While the basic distribution are built into many packages that can do simulation, not all that are necessary are. We need additional techniques to be able to sample from these other distributions.

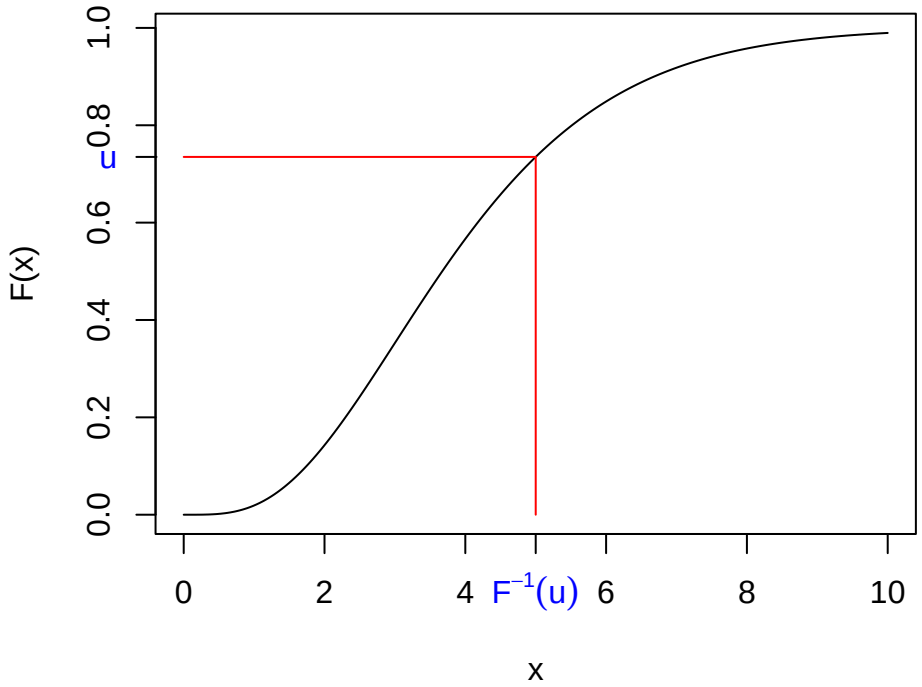
- Inverse CDF
- Relationships with other distributions
- Acceptance - Rejection Sampling
- Plus many more

# Inverse CDF Method

Let $F(x) = P[X \leq x]$ be the CDF of the random variable $X$.

Then the inverse CDF (or quantile function) is defined by

$$F^{-1}(u) = \inf\{x : F(x) \leq u\}$$




For continuous RVs

$$P[F(X) \leq u] = P[X \leq F^{-1}(u)] = F(F^{-1}(u)) = u$$

i.e. $F(X) \sim U(0, 1)$

Thus given and iid $U(0,1)$ sample $u_1, \ldots, u_m$, an iid sample $x_1, \ldots, x_m$ from $F$ can be obtained by

$$x_i = F^{-1}(u_i)$$

Examples:

1. $Cauchy(\mu, \sigma)$

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-\mu}{\sigma}\right)$$
$$F^{-1}(u) = \mu + \sigma \tan(\pi(u - 0.5))$$

2. $Exp(\mu)$

$$F(x) = 1 - e^{-x\mu}$$
$$F^{-1}(u) = \frac{-1}{\mu} \log(1-u)$$

3. Discrete distributions

Suppose that the random variable $X$ has possible values $s_1, s_2, \ldots, s_k$ ($k$ possibly infinite) and let the values of the CDF be

$$P_j = \sum_{i=1}^{j} P[X = s_i] = P[X \leq s_j]$$

Then independent observations $x_i$ can be generated by setting $x_i = s_j$ if

$$P_{j-1} < u_i \leq P_j$$

where $P_0 = 0$ and $u_i \sim U(0, 1)$

Note that

$$P[X = x_j] = P_j - P_{j-1} = P[X \leq x_j] - P[X \leq x_{j-1}]$$

so the draws do have the correct distribution.

**Advantages:**

- Will give draws from the correct distribution

**Disadvantages:**

- While the density is often of a nice form, the CDF and its inverse often aren't (e.g. Normal, Gamma, Beta, etc).

- Though there are often good approximations for the quantile function (e.g. `R` and `Matlab` often use rational function approximations), these are often slow and poor for simulation purposes (particularly in the tails of the distribution).

- For a discrete distribution with many classes, they may be many comparisons made to determine $x_j$. For example, `R` doesn't use this approach for Binomial draws if $np > 30$

# Relationships with Other Distributions

Examples:

- $X \sim N(\mu, \sigma^2)$ then $Y = e^X \sim LogN(\mu, \sigma^2)$
- $X \sim N(0, 1)$ then $Y = X^2 \sim \chi^2_1$
- $X_\alpha \sim Gamma(1, \alpha), X_\beta \sim Gamma(1, \beta)$ then

$$Y = \frac{X_\alpha}{X_\alpha + X_\beta} \sim Beta(\alpha, \beta)$$

- $X \sim U(0, 1)$ then $Y = -\log X \sim Exp(1)$

The inverse CDF method can be thought of as a special case of this.

**Advantages:**

- Will give draws from the correct distribution

**Disadvantages:**

- Many distributions don't have useful relationships
- Can be inefficient as functions like $\log, \sin, \cos$ can be somewhat expensive to calculate

# Acceptance-Rejection

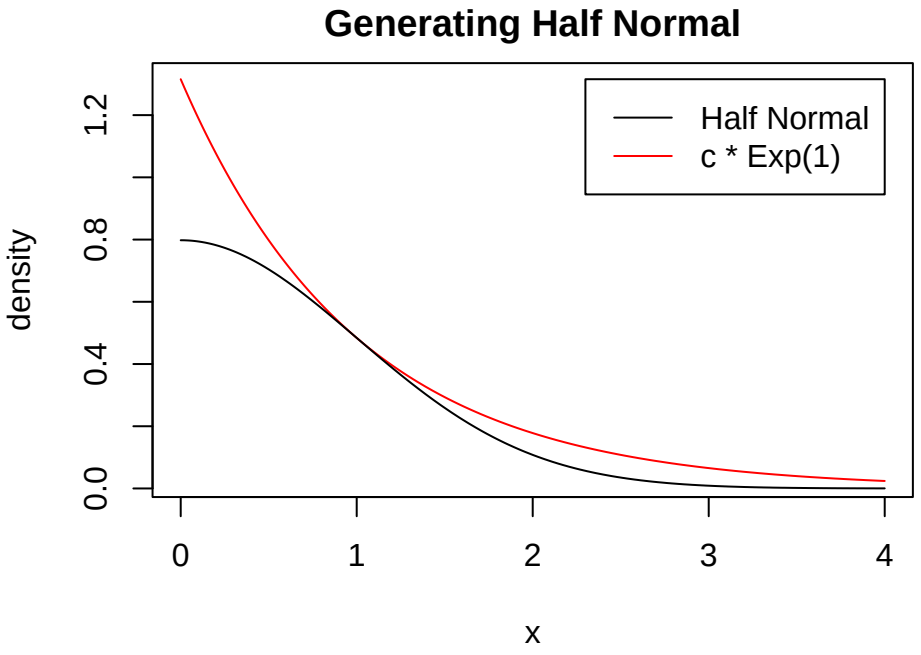
Due to von Neumann (1951)

Want to simulate from a distribution with density $f(x)$.

Need to find a "dominating" or majorizing distribution $g(x)$ where $g$ is easy to sample from and

$$f(x) \leq cg(x) = h(x)$$

for all $x$ and some constant $c > 1$.



Sampling scheme

1. Sample $x$ from $g(x)$ and compute the acceptance ratio

$$r(x) = \frac{f(x)}{cg(x)} = \frac{f(x)}{h(x)} \leq 1$$

2. Sample $u \sim U(0, 1)$

   If $u \leq r(x)$ accept and return $x$

   If $u > r(x)$ reject and go back to 1)

   Note that this step is equivalent to flipping a biased coin with success probability $r(x)$

Then the resultant sample is a draw from the density $f(x)$.

**Proof.** Let $I$ be the indicator of whether a sample $x$ is accepted. Then

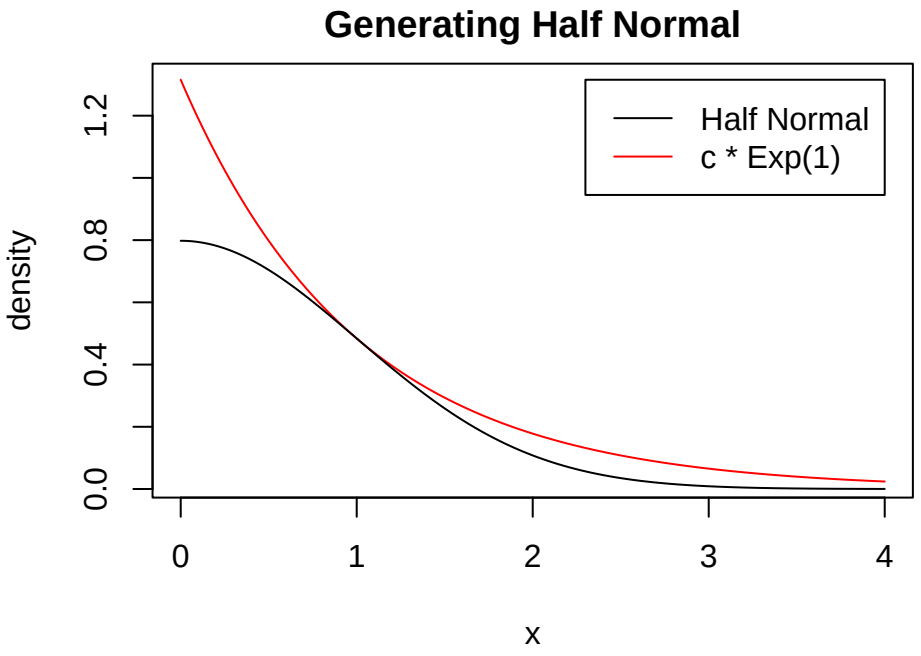
$$\begin{aligned} P[I=1] &= \int P[I=1|X=x]g(x)dx \\ &= \int r(x)g(x)dx \\ &= \int \frac{f(x)}{cg(x)}g(x)dx = \frac{1}{c} \end{aligned}$$

Next

$$\begin{aligned} p(x|I=1) &= \frac{f(x)}{cg(x)}g(x) \Big/ P[I=1] \\ &= \frac{f(x)}{c}c = f(x) \end{aligned}$$

□

For a more geometrical proof see Flury (1990). Its based on the idea of drawing uniform points $(x, y)$ under the curve $h(x)$ and only accepting the points that also lie under the curve $f(x)$.




The number of draws needed until an acceptance occurs is $Geometric(\frac{1}{c})$ and thus the expected number of draws until a sample is accepted is $c$.

The acceptance probability satisfies

$$\frac{1}{c} = \frac{\int f(x)dx}{\int cg(x)dx} = \frac{\text{Area under } f(x)}{\text{Area under } h(x)}$$

One consequence of this is that $c$ should be made as small as possible to minimize the number of rejections.

The optimal $c$ is given by

$$c = \sup \frac{f(x)}{g(x)}$$

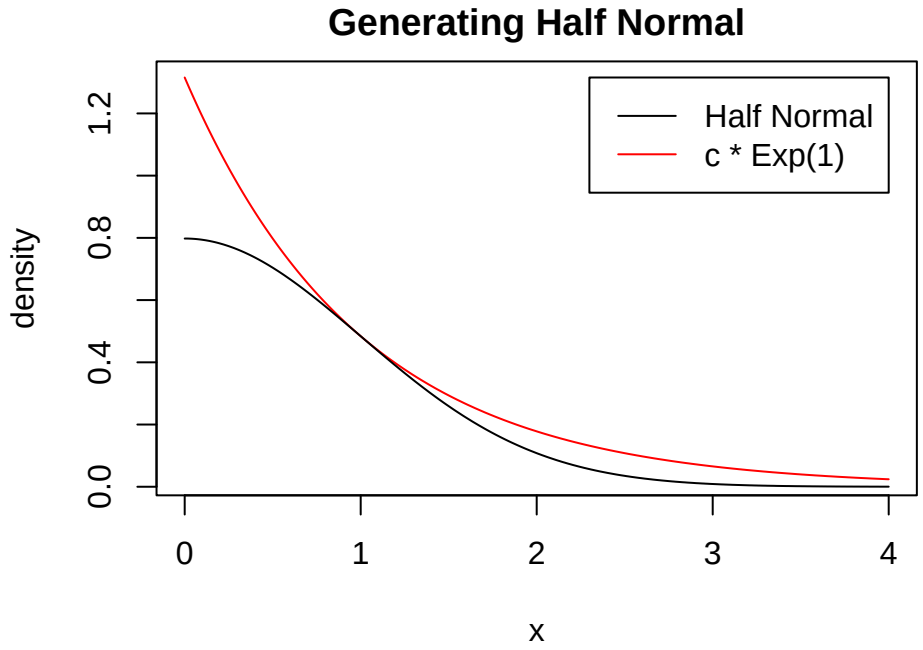
Note that the best $c$ need not be determined, just one that satisfies

$$f(x) \leq cg(x) = h(x)$$

for all $x$.

Example: Generating from the half normal distribution

$$f(x) = 2\phi(x)I(x \geq 0)$$
$$= \sqrt{\frac{2}{\pi}}\exp(-0.5x^2)I(x \geq 0)$$




Lets use an $Exp(1)$ as the dominating density

$$g(x) = e^{-x}I(x \geq 0)$$

The optimal $c$ for this example is

$$c = \sqrt{\frac{2}{\pi}}\exp(0.5) \approx 1.315$$

so the acceptance rate is approximately 76%

This the acceptance-rejection scheme is

1. Draw $x \sim Exp(1)$

$$r(x) = \exp(-0.5(x-1)^2)$$

2. Draw $u \sim U(0,1)$

   If $u \leq r(x)$ accept and return $x$

   If $u > r(x)$ reject and go back to 1)

Note that this approach can be used to simulate $N(0,1)$ random variables. Draw $X$ by this scheme. Then draw $S$ from the distribution

$$P[S=1] = P[S=-1] = 0.5$$

Then $Z = SX \sim N(0,1)$ by the symmetry of the normal density.

**Advantages:**

- Will give draws from the correct distribution.
- Extremely flexible.
- Approach will work for a wide range of problems.
- For many problems there are good choices for the majorizing distribution (i.e. log concave densities).
- Will work for multivariate distributions.

**Disadvantages:**

- Maybe inefficient (large $c$).
- How to pick majorizing distribution not always clear.

# Simulation - Joint and Marginal Distributions

**Joint Distribution:**

Want to simulate $X, Y$ from $f(x, y)$

- Sample $x_i$ from $f_X(x); i = 1, \ldots, n$
- Sample $y_i$ from $f_{Y|X}(y|x_i); i = 1, \ldots, n$

Justification that this scheme actually draws from the joint distribution:

The joint empirical CDF of $(x_i, y_i); i = 1, \ldots, n$ is

$$\hat{F}(x, y) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x, y_i \leq y) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) I(y_i \leq y)$$

The expected value of the ECDF is

$$E[\hat{F}(x,y)] = E[I(X \leq x)I(Y \leq y)] = P[X \leq x, Y \leq y] = F(x,y)$$

since

$$\begin{aligned} E[I(x_i \leq x)I(y_i \leq y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x_i \leq x)I(y_i \leq y) f_X(x_i) f_{Y|X}(y_i|x_i) dy_i dx_i \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x_i, y_i) dy_i dx_i \\ &= P[X \leq x, Y \leq y] \end{aligned}$$

The ECDF is an unbiased estimate of the CDF.

In addition

$$\mathrm{Var}(\hat{F}(x,y)) = \frac{F_{X,Y}(x,y)(1-F_{X,Y}(x,y))}{n} \longrightarrow 0$$

as $n \to \infty$, which implies $\hat{F}(x,y) \xrightarrow{P} F_{X,Y}(x,y)$.

This result can also be seen by noting that $n\hat{F}(x,y) \sim Bin(n, F_{X,Y}(x,y))$ and applying standard binomial convergence results.

Of course this scheme can be extended to an arbitrary number of random variables based on

$$F_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = F_{X_1}(x_1)F_{X_2|X_1}(x_2|x_1)\ldots \\ \times F_{X_k|X_1,\ldots,X_{k-1}}(x_k|x_1,\ldots,x_{k-1})$$

**Marginal Distribution:**

Want to simulate $Y$ based on $f(x, y)$ (assume that $f_Y(y)$ isn't nice)

- Sample $x_i$ from $f_X(x); i = 1, \ldots, n$
- Sample $y_i$ from $f_{Y|X}(y|x_i); i = 1, \ldots, n$
- Keep only $y_i; i = 1, \ldots, n$

Justification that this scheme actually draws from the marginal distribution $F_Y(y)$:

The empirical CDF of $y_i; i = 1, \ldots, n$ is

$$\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq y)$$

The expected value of the ECDF is

$$E[\hat{F}(y)] = E[I(Y \leq y)] = P[Y \leq y] = F_Y(y)$$

since

$$\begin{aligned} E[I(y_i \leq y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(y_i \leq y) f_X(x_i) f_{Y|X}(y_i|x_i) dy_i dx_i \\ &= \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{X,Y}(x_i, y_i) dx_i dy_i \\ &= \int_{-\infty}^{y} f_Y(y_i) dy_i \\ &= P[Y \leq y] \end{aligned}$$

Similarly to before

$$\mathrm{Var}(\hat{F}(y)) = \frac{F_Y(y)(1 - F_Y(y))}{n} \longrightarrow 0$$

as $n \to \infty$, which implies $\hat{F}(y) \xrightarrow{P} F_Y(y)$.

# Monte Carlo Examples

- Confidence Interval Properties

  Let $X_1, X_2, \ldots, X_n$ be iid draws from $N(\mu, \sigma^2)$. The a $100(1-\alpha)$ % confidence interval for $\mu$ is

  $$\bar{x} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

  where $t_{1-\alpha/2} = t^*$ is the $100(1-\frac{\alpha}{2})$ percentile from a $t$ distribution with $n-1$ degrees of freedom.

  This interval gets used in many situations when the data isn't normal. This interval is considered to be fairly robust when the data isn't normal, but lets check it by examining the properties of this procedure when the data isn't normal.

Of interest are

1. True confidence level

$$C = P\left[\bar{X} - t^* \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t^* \frac{S}{\sqrt{n}}\right]$$

   If $X_i \sim N(\mu, \sigma^2)$, this probability is $1 - \alpha$.

2. Expected width of the confidence level

$$E[\text{Width}] = E\left[2t^* \frac{S}{\sqrt{n}}\right] = 2\frac{t^*}{\sqrt{n}} E[S]$$

   Note that for $X_i \sim N(\mu, \sigma^2)$,

$$E[\text{Width}] = 2t^* \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \frac{\sqrt{2}}{\sqrt{n(n-1)}} \sigma$$

   If $n = 10$ and $\sigma = 1$, the mean width is 1.392.

For simulation $i, i = 1, \ldots, 1000$, let $\bar{x}_i$ and $s_i$ be the sample average and standard deviation of the simulated data and $w_i$ be the width of the corresponding confidence interval. Then the estimates for the confidence level and expected width are

1. Confidence level

$$\hat{C} = \frac{1}{1000} \sum_{i=1}^{1000} I\left(\bar{x}_i - t\frac{s_i}{\sqrt{10}} \leq \mu \leq \bar{x}_i + t\frac{s_i}{\sqrt{10}}\right)$$

This is just the sample proportion of intervals containing the truth.

The standard error of this estimate is given by

$$SE_{\hat{C}} = \sqrt{\frac{\hat{C}(1-\hat{C})}{1000}}$$

2. $E[\text{Width}]$

$$\begin{aligned}\bar{w} &= \frac{1}{1000}\sum_{i=1}^{1000} w_i = \frac{1}{1000}\sum_{i=1}^{1000}\frac{2t^*}{\sqrt{10}}s_i \\ &= \frac{2t^*}{\sqrt{10}}\bar{s}\end{aligned}$$

The standard error of this estimate is given by

$$SE_{\bar{w}} = \frac{s_w}{\sqrt{1000}} = \frac{2t^*}{\sqrt{10}}\frac{s_s}{\sqrt{1000}}$$

where $s_w$ is the sample standard deviation of the $w_i$ and $s_s$ is the sample standard deviation of the $s_i$.

Lets examine these in the following 4 situations

1. $N(0,1)$
2. $t_3$ $(\mu = 0, \sigma = \sqrt{3} = 1.732)$
3. $Exp(0.2)$ $(\mu = \sigma = 5)$
4. $U(-1,1)$ $(\mu = 0, \sigma = 0.577)$

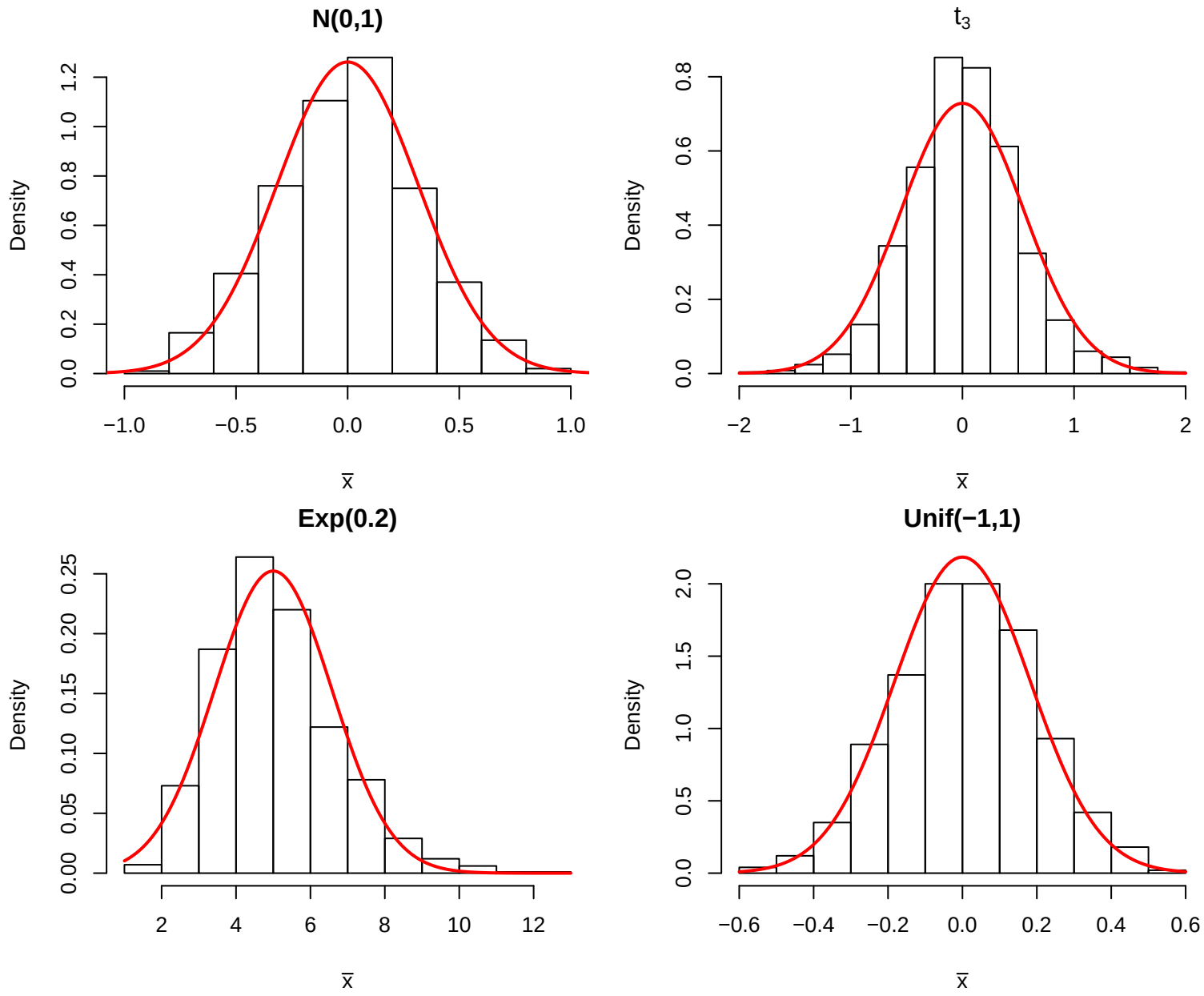
where the confidence level $C = 0.95$ and sample size $n = 10$ with $m = 1000$ simulated data sets.

For simulated data set $i$, $X_1, X_2, \ldots, X_{10}$ are generated from the desired distribution and $\bar{x}_i$, $s_i$, and $w_i$ are calculated and then $I\left(\bar{x}_i - t\frac{s_i}{\sqrt{10}} \leq \mu \leq \bar{x}_i + t\frac{s_i}{\sqrt{10}}\right)$ is determined.

1. Confidence level (nominal = 0.95)

| Distribution | $\hat{C}$ | $SE_{\hat{C}}$ | 95% CI |
|---|---|---|---|
| $N(0,1)$ | 0.954 | 0.0066 | (0.9410, 0.9670) |
| $t_3$ | 0.965 | 0.0058 | (0.9536, 0.9764) |
| $Exp(0.2)$ | 0.904 | 0.0093 | (0.8857, 0.9223) |
| $Unif(-1,1)$ | 0.932 | 0.0080 | (0.9164, 0.9476) |

So the simulations suggest that the true confidence level for the normal case is correct, it a bit higher than the nominal level for the $t_3$ case, and a bit lower than the nominal level for the exponential and uniform cases.

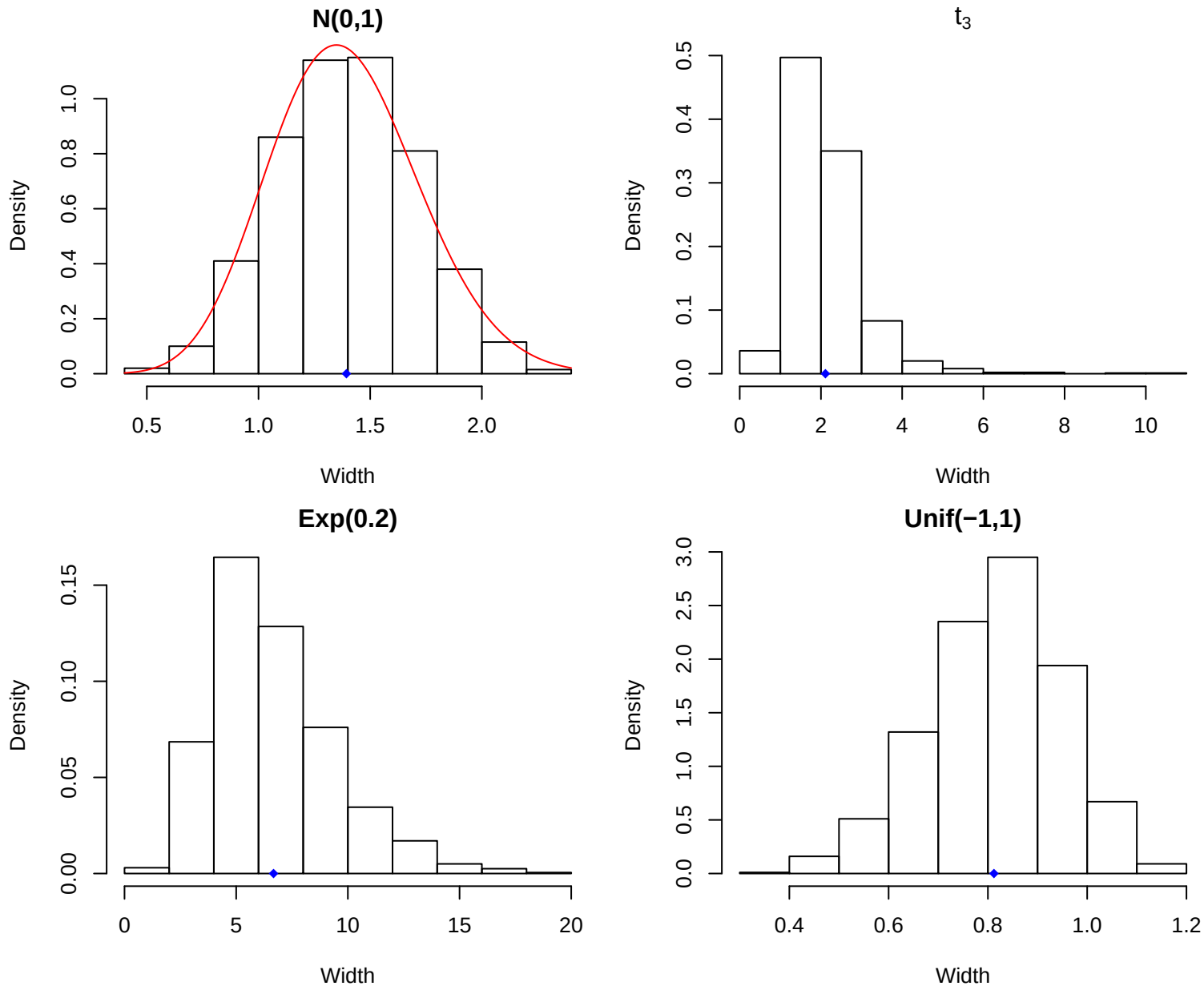

N(0,1)
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−1.0
−0.5
0.0
0.5
1.0
x̄
t3
Density
0.0
0.2
0.4
0.6
0.8
−2
−1
0
1
2
x̄
Exp(0.2)
Density
0.00
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
12
x̄
Unif(−1,1)
Density
0.0
0.5
1.0
1.5
2.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
x̄

2. $E$[Width]

| Distribution | $\bar{w}$ | $SE_{\bar{w}}$ | 95% CI |
|---|---|---|---|
| $N(0,1)$ | 1.394 | 0.0099 | (1.374, 1.413) |
| $t_3$ | 2.110 | 0.0292 | (2.053, 2.167) |
| $Exp(0.2)$ | 6.673 | 0.0865 | (6.503, 6.842) |
| $Unif(-1,1)$ | 0.812 | 0.0043 | (0.804, 0.821) |

Cauchy Example: The assumption that $E[|g(X)|] < \infty$ underlying the Law of Large Numbers is important. Since the Cauchy distribution has no finite moments, $E[S] = \infty$, and thus the mean interval width is $\infty$. Thus the reported sample average and standard error are not meaningful.

In the simulations, there are 15 intervals with widths $> 200$, with the largest being 10640.


N(0,1)
Density
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
Width
$t_3$
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
Width
Exp(0.2)
Density
0.00
0.05
0.10
0.15
0
5
10
15
20
Width
Unif(−1,1)
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.4
0.6
0.8
1.0
1.2
Width

- Bayesian Analysis

  Air Conditioning Failures in a Boeing 720 (Proschan, 1963)

  For plane 7910 (there are 13 planes in the complete dataset), the times between failures (in hours) are 74, 57, 48, 29, 502, 12, 70, 21, 29, 386, 59, 27, 153, 26, 326 (n = 15) ($\bar{x}$= 121.27)

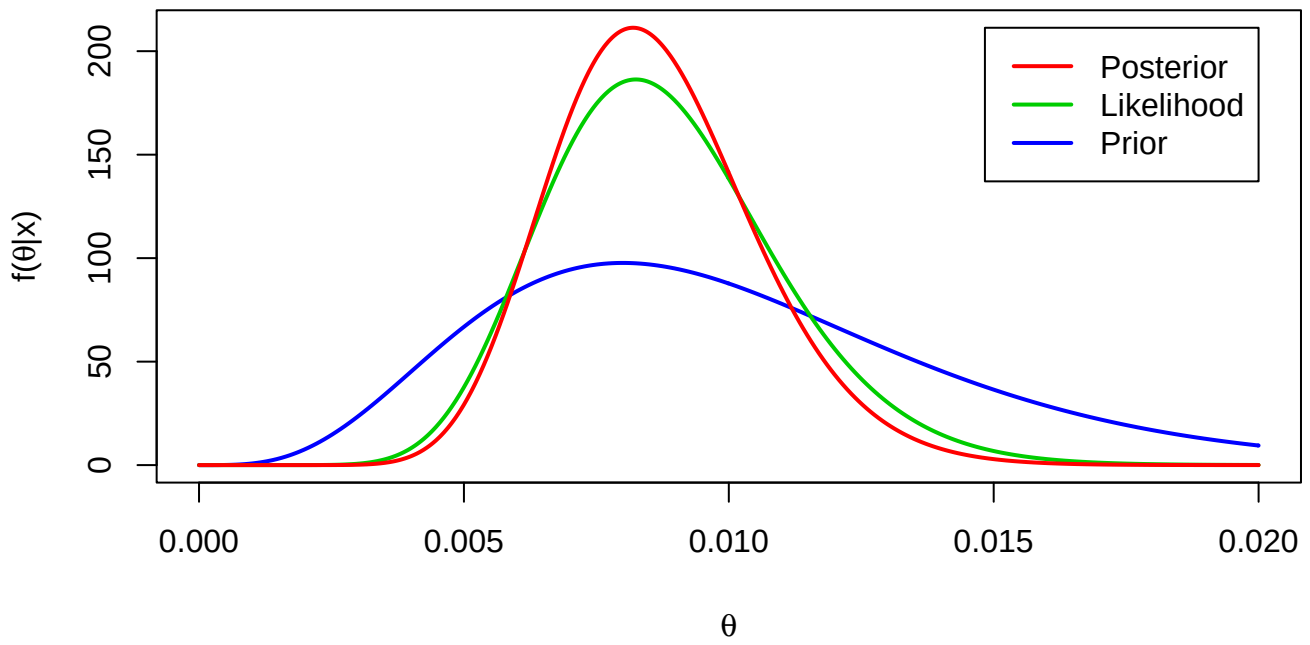
  Assume that $X_i \overset{iid}{\sim} Exp(\theta)$. However the value for $\theta$ is unknown. Suppose that based on information from other planes, we can put a distribution on it

$$\theta \sim Gamma(5, 500)$$

  which has mean 0.01 and standard deviation 0.00447.

  How does the data collected change our belief on $\theta$. This is described by the conditional distribution of $\theta$ given the observations $x_1, x_2, \ldots, x_{15}$. It can be shown that

$$\theta | x_1, \ldots, x_{15} \sim Gamma(5 + 15, 500 + 15\bar{x}) = Gamma(20, 2319)$$




$$E[\theta] = 0.01 \qquad E[\theta|x_1, \ldots, x_{15}] = 0.00862$$

$$\mathrm{Var}(\theta) = 0.0000200 \qquad \mathrm{Var}(\theta|x_1, \ldots, x_{15}) = 0.0000037$$

$$\mathrm{SD}(\theta) = 0.00447 \qquad \mathrm{SD}(\theta|x_1, \ldots, x_{15}) = 0.00193$$

Note: The conditional distribution can be shown by using the facts

1. $\bar{X}|\theta \sim Gamma(n, \theta/n)$
2. The distribution $\theta|\bar{X}$ is the same as the distribution of $\theta|x_1, \ldots, x_{15}$

Suppose that we are interested in

1. Mean time between failures

$$\mu_F = E\left[\frac{1}{\theta}\right]$$

   I think the truth is 122.05 hours.

2. For future observations, what is the chance of going over 150 hours before a failure?

$$p_{150} = P[X > 150] = E\left[e^{-150\theta}\right]$$

   I believe this probability is 0.2855.

Lets use simulation to check my calculus.

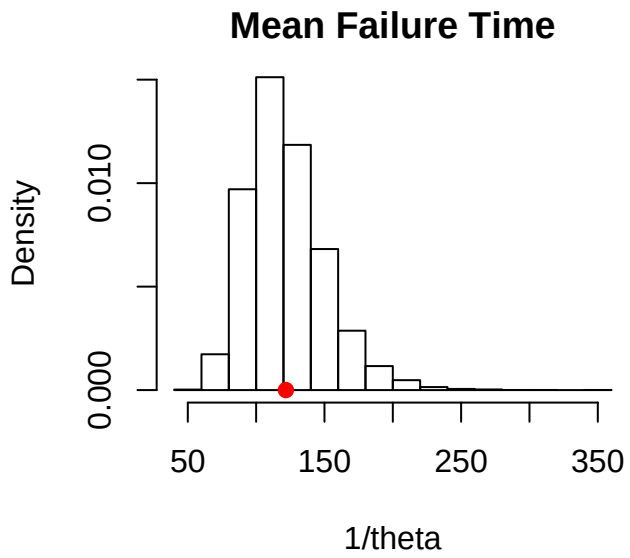
We can examine both calculations with a single simulation. Generate $\theta_1, \theta_2, \ldots, \theta_m$ from $Gamma(20, 2319)$. Then estimate the two quantities by

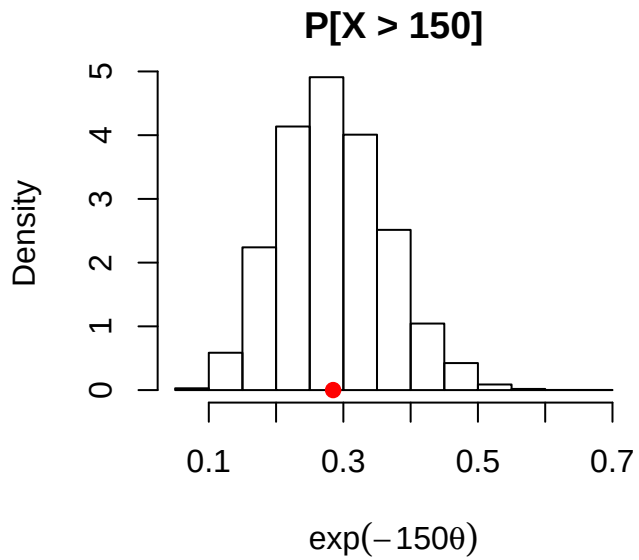
$$\hat{\mu}_F = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\theta_i}$$

$$\hat{p}_{150} = \frac{1}{m} \sum_{i=1}^{m} e^{-150\theta_i}$$

Aside: One advantage of simulation procedures is that a single sample can be used to do more than one calculation. In this example, one sample is being used to calculate 2 different expectations (and could be used to do many more).

The following simulation results are based on $m = 10,000$ imputations.




$$\hat{\mu}_F = 121.781 \qquad \hat{p}_{150} = 0.2848$$

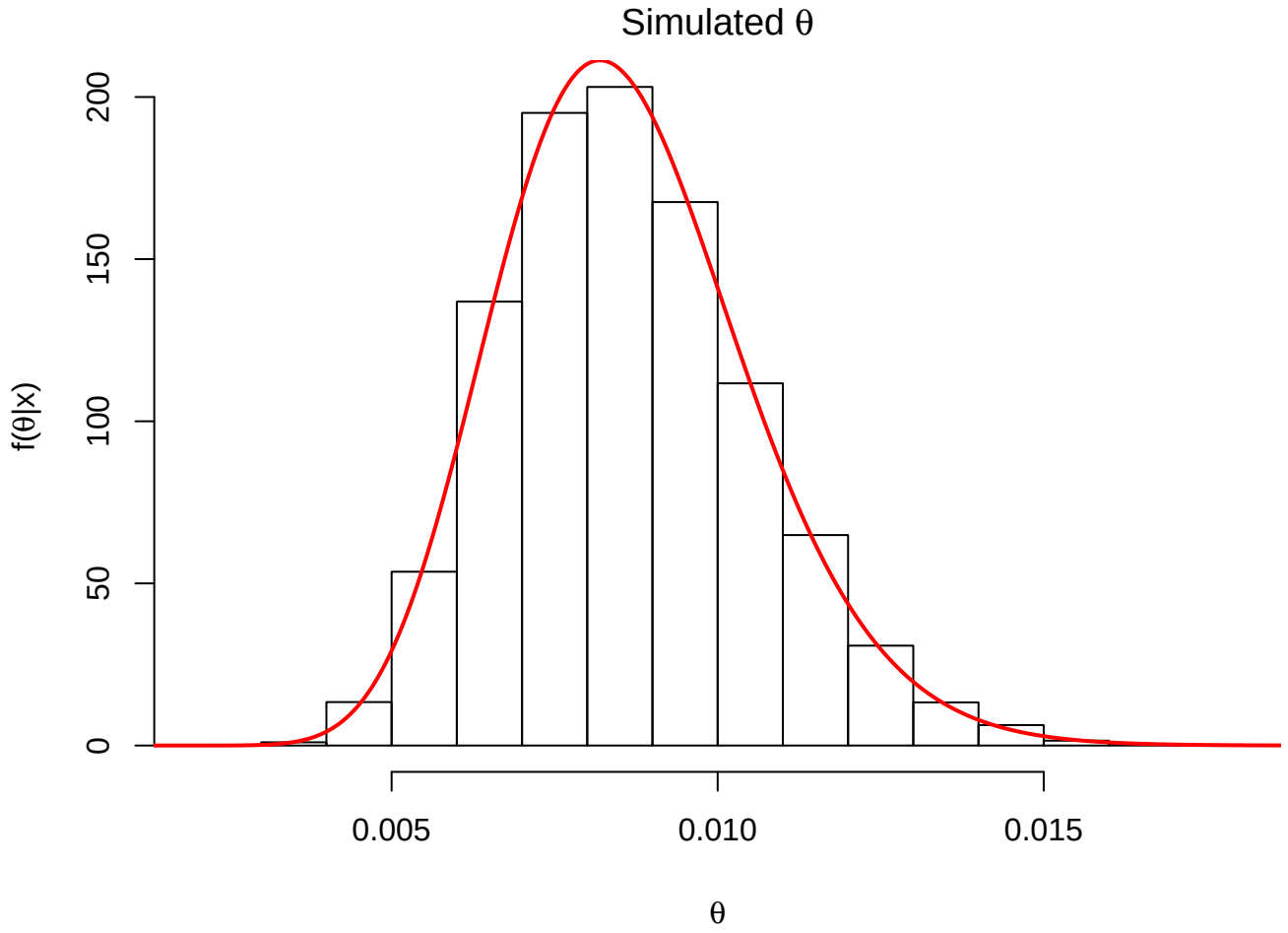
$$SE(\hat{\mu}_F) = 0.285 \qquad SE(\hat{p}_{150}) = 0.00079$$

$$\mu_F = 122.05 \qquad p_{150} = 0.2855$$

Both estimated values are about 1 standard error from the proposed

values, suggesting I did the calculus correctly.

For completeness, lets check to see how well the simulated $\theta$s agree with the conditional distribution of $\theta$.



# How many imputations?

When designing a Monte Carlo study, the sample size $m$ need to be determined.

The usual approach is by bounding the standard error of the estimate.

If the scheme for generating the simulated values is to have iid samples, we want

$$SE \leq \frac{\sigma_g}{\sqrt{m}}$$

which gives

$$m \geq \frac{\sigma_g^2}{SE^2}$$

where $SE$ is the desired standard error and $\sigma_g^2 = \mathrm{Var}(g(X))$.

There is the same potential problem in doing this as in a sample survey: $\sigma_g^2$ is usually unknown. Sometimes you can guess. However there is a usually simple solution in this situation. Do a small simulation (say 100 samples) to estimate $\sigma_g^2$ and use this in the above formula.

Note when choosing the desired $SE$, often people think about bounding the relative standard error, i.e.

$$\frac{SE}{E[g(X)]}$$

People will often think about trying to get an answer within 1% or 0.1% say. This often relates to the number of significant digits you have confidence in. This approach fits into the original scheme by working with a rough guess of $E[g(X)]$. If you don't have a rough guess of this, use the small simulation idea here to get one.

# Sampling Schemes

For the examples shown so far, they have either used Simple Random Sampling or IID sampling from the desired distribution.

These may not be the best approaches as they be slow, inefficient (large standard errors) or difficult to implement (distribution is hard to sample from).

Instead we can modify the sampling scheme to avoid these problems. Two possible approaches are

- Importance Sampling
- Markov Chain Monte Carlo (MCMC)

# Importance Sampling

Used for a number of purposes

- Variance reduction (smaller standard errors)
- Allows for difficult distributions to be sampled from
- Sensitivity analysis
- Reusing samples to reduce computational burden

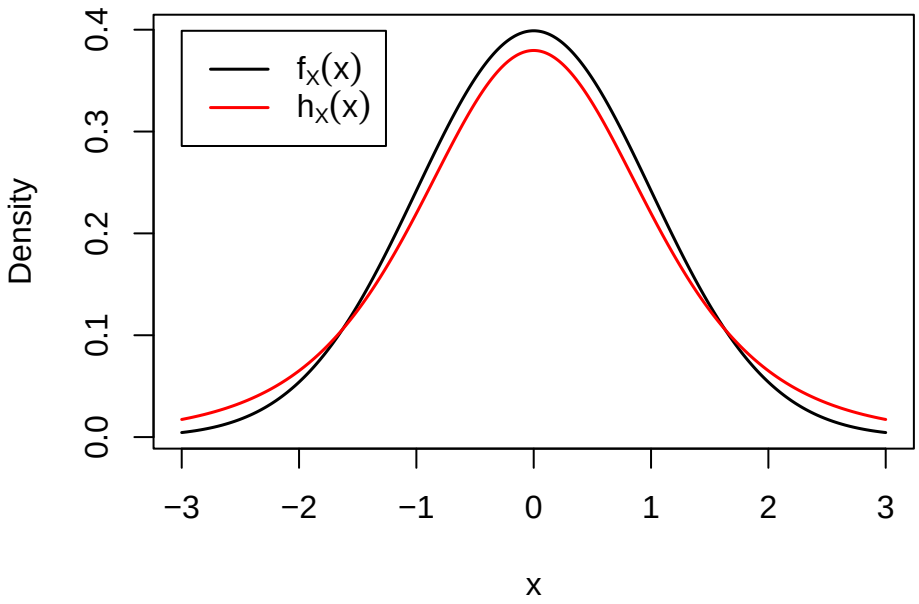
Idea is to sample from a different distribution that picks points in "important" regions of the sample space and is easy to sample from.

As before, want to estimate

$$E[g(X)] = \int g(x) f_X(x) dx$$

Instead of sampling from density (or pmf) $f_X(x)$, sample from a different distribution with density (or pmf) $h_X(x)$.

Since we are sampling from the "wrong" distribution, we have to make adjustments in out estimator. In the example plot to the right, this setup will sample too many values around 3 and not enough values around 0.




$$
\begin{aligned}
E_f[g(X)] &= \int g(x) f_X(x) dx \\
&= \int g(x) \frac{f_X(x)}{h_X(x)} h_X(x) dx \qquad (*) \\
&= E_h \left[ g(X) \frac{f_X(X)}{h_X(X)} \right]
\end{aligned}
$$

This suggests the following sampling scheme

1. Sample $x_1, x_2, \ldots, x_m$ from $h_X(x)$

2. Calculate weights

$$w_i = \frac{f_X(x_i)}{h_X(x_i)}$$

3. Estimate $E[g(X)]$ by

$$\mu_{IS} = \frac{1}{m}\sum_{i=1}^{m} w_i g(x_i) = \frac{1}{m}\sum_{i=1}^{m} \frac{f_X(x_i)}{h_X(x_i)} g(x_i)$$

So instead of a regular average, this estimator is a weighted average.

Points that occur more often under $h_X(x)$ than $f_X(x)$ get downweighted and those that occur less often get upweighted.

Based on (*), $\mu_{IS}$ is an unbiased estimate of $E[g(X)]$ regardless of which proposal distribution $h_X(x)$ is chosen, as long as $h_X(x)$ as the same support as $f_X(x)$, i.e.,

$$f_X(x) > 0 \text{ implies that } h_X(x) > 0$$

This implies any value that you want to sample under $f_X(x)$ you must be able to sample under $h_X(x)$

Note that $h_X(x) > 0$ can be allowed to occur when $f_X(x) = 0$, though doing this tends to be inefficient, as you can generate samples you don't really need ($w_i = 0$). However there can be times when you want to do this.

Example: Simulating to Calculate Small Probabilities

Suppose $Z \sim N(0,1)$ and we are interested in

$$P[Z \geq a] = E[I(Z \geq a)] = \Phi(-a) = p_a$$

for large values of $a > 0$ (say 5 which has $P[Z \geq 5] = 2.87 \times 10^{-07}$)

Motivation: This was based on probability calculations involved in Genetic Linkage analysis. The distribution of the quantities of interest converge in distribution to a $N(0,1)$ as the size of the families goes to $\infty$. However for the family sizes of interest, the normal approximation is extremely poor so it was decided to use simulation to calculate the necessary probabilities. The following argument suggests how the pedigrees need to be simulated in an efficient sampler.

**Naive approach:**

Sample $z_1, z_2, \ldots, z_m$ from a $N(0,1)$ and calculate

$$\hat{p}_a = \frac{1}{m} \sum_{i=1}^{m} I(z_i \geq a)$$

The variance of this estimator is

$$\mathrm{Var}(\hat{p}_a) = \frac{p_a(1-p_a)}{m}$$

Note that virtually all samples will have $z_i < a$. In fact you will only expect to see a $z_i \geq a$ about every 3.5 million draws.

**Importance sampling approach:**

The lack of extreme $z_i$s in the naive approach makes it inefficient. So lets try to get some $z_i$s around $a$.

Sample $x_1, x_2, \ldots, x_m$ from $N(a, 1)$ and calculate

$$\hat{p}_{aIS} = \frac{1}{m} \sum_{i=1}^{m} I(z_i \geq a) \frac{\phi(x_i)}{\phi(x_i - a)}$$

The variance of this estimator is

$$\mathrm{Var}(\hat{p}_{aIS}) = \frac{e^{a^2} p_{2a} - p_a^2}{m}$$

It can be shown that

$$\frac{\mathrm{Var}(\hat{p}_a)}{\mathrm{Var}(\hat{p}_{aIS})} > 1$$

for any $a > 0$ (importance sampling always does better). If $a = 5$

$$\frac{\mathrm{Var}(\hat{p}_a)}{\mathrm{Var}(\hat{p}_{aIS})} = 614475.8$$

so 1 importance sample draw is worth over 600,000 draws from the $N(0, 1)$ distribution.

Note that this procedure can be improved slightly by sampling from a distribution with a mean slightly larger than $a$. The optimal choice of the mean $b$ satisfies

$$2b = \frac{\phi(a + b)}{\Phi(-a - b)}$$

For the genetics problem mentioned, it implies that we need to simulate pedigree data with the gene located a bit closer to the marker than the data suggests it is.

# Markov Chain Monte Carlo (MCMC)

Instead of generating independent samples, generate dependent samples via a Markov Chain

$$\theta^0 \rightarrow \theta^1 \rightarrow \theta^2 \rightarrow \theta^3 \rightarrow \ldots$$

where the stationary distribution of the chain is the desired distribution $p(\theta)$.

The Markov Chain is defined by a transition distribution $T_t(\theta^t|\theta^{t-1})$, which describes the possible moves when you are in state $\theta^{t-1}$

Useful for a wide range of problems.

Popular for Bayesian analyses, but it is a general sampling procedure. For example, it has been used for calculating likelihoods in genetic linkage analysis.

# Gibbs Sampling

One example of MCMC

Idea: Break the random variable $\theta$ in $k$ pieces ($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$) and sample the pieces sequentially. (The pieces $\theta_i$ could be univariate or multivariate.)

1. Initialize chain: $\theta^0 = \{\theta_1^0, \theta_2^0, \ldots, \theta_k^0\}$ by some mechanism.

2. At time $t$, sample $\theta^t = \{\theta_1^t, \theta_2^t, \ldots, \theta_k^t\}$ by

   - Step 1: sample $\theta_1^t \sim p(\theta_1 | \theta_2^{t-1}, \ldots, \theta_k^{t-1})$
   - Step 2: sample $\theta_2^t \sim p(\theta_2 | \theta_1^t, \theta_3^{t-1} \ldots, \theta_k^{t-1})$
   - Step $j$: sample $\theta_j^t \sim p(\theta_j | \theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1} \ldots, \theta_k^{t-1})$
   - Step $k$: sample $\theta_k^t \sim p(\theta_k | \theta_1^t, \ldots, \theta_{k-1}^t)$

Under certain regularity conditions, the realizations $\theta^1, \theta^2, \theta^3, \ldots$ form a Markov chain with stationary distribution $p(\theta)$. Thus the realizations can be treated as dependent samples from the desired distribution.

Example: Nuclear Pump Failure

Gaver & O'Muircheartaigh (Technometrics, 1987)

Gelfand & Smith (JASA, 1990)

Observed 10 nuclear reactor pumps and counted the number of failures for each pump.

| Pump | Failures ($s_i$) | Observation Time ($t_i$) | Observed Rate ($l_i$) |
|---|---|---|---|
| 1 | 5 | 94.320 | 0.053 |
| 2 | 1 | 15.720 | 0.064 |
| 3 | 5 | 62.880 | 0.080 |
| 4 | 14 | 125.760 | 0.111 |
| 5 | 3 | 5.240 | 0.573 |
| 6 | 19 | 31.440 | 0.604 |
| 7 | 1 | 1.048 | 0.954 |
| 8 | 1 | 1.048 | 0.954 |
| 9 | 4 | 2.096 | 1.910 |
| 10 | 22 | 10.480 | 2.099 |

Observation time in 1000's of hours

Observed Rate = # Failure / 1000 hours

Want to determine the true failure rate for each pump with the following hierarchical model

$$s_i|\lambda_i \overset{ind}{\sim} Poisson(\lambda_i t_i)$$

$$\lambda_i|\beta \overset{iid}{\sim} Gamma(\alpha, \beta)$$

$$\beta \sim Gamma(\gamma, \delta)$$

Note that this is a slightly different parameterization but the same model that Gelfand and Smith used.

In this example, $\alpha$ will be assumed to be a fixed parameter. We could put a prior on it, or as Gelfand and Smith do, estimate it from the data and take and empirical Bayes solution.

Want to determine the following posterior distributions

1. $p(\lambda_i|s)$ for each pump

2. $p(\beta|s)$ and $p\left(\frac{1}{\beta}|s\right)$

Note the both sets of these distributions are difficult to get analytically. It is possible to show that

$$p(\lambda|s) \propto \frac{1}{(\delta + \sum \lambda_i)^{10\alpha+\gamma}} \prod \frac{t_i^{\alpha+s_i} \lambda^{\alpha+s_i-1} e^{-\lambda_i t_i}}{\Gamma(\alpha + s_i)}$$

Note that the $\lambda$'s are correlated and trying to get the marginal for each looks to be intractable analytically.

Instead lets run a Gibbs sampler to determine $p(\lambda, \beta|s)$ from which we can get the desired posteriors.

One possible Gibbs scheme is

- Step 1: sample $\lambda_1 \sim p(\lambda_1|\lambda_{(-1)}, \beta, s)$
- Step 2: sample $\lambda_2 \sim p(\lambda_2|\lambda_{(-2)}, \beta, s)$

  ...

- Step 10: sample $\lambda_{10} \sim p(\lambda_{10}|\lambda_{(-10)}, \beta, s)$
- Step 11: sample $\beta \sim p(\beta|\lambda, s)$

where $\lambda_{(-j)} = \{\lambda_1, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_{10}\}$

Need the following conditional distributions

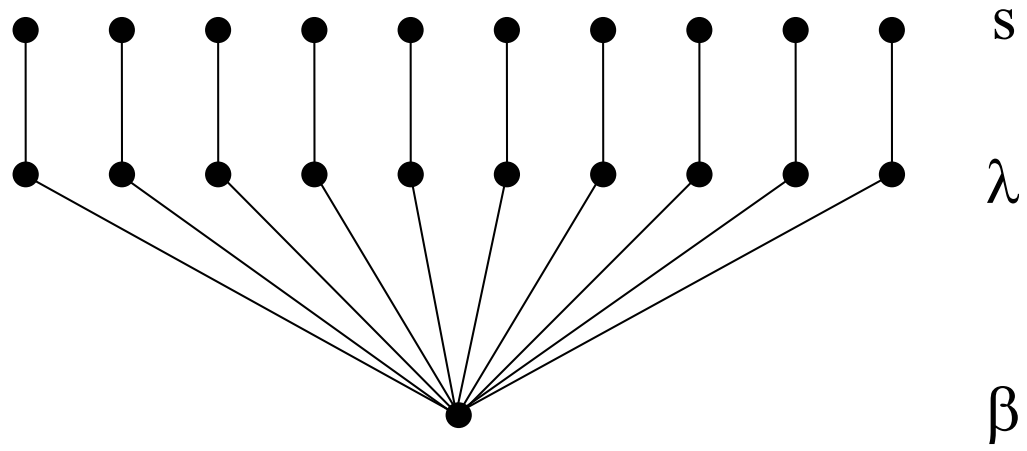
$$\begin{aligned}\lambda_j \sim p(\lambda_j|\lambda_{(-j)}, \beta, s) &= p(\lambda_j|\beta, s_j) \\ &= Gamma(\alpha + s_j, \beta + t_j)\end{aligned}$$

$$\begin{aligned}\beta \sim p(\beta|\lambda, s) &= p(\beta|\lambda) \\ &= Gamma(\gamma + 10\alpha, \delta + \sum \lambda)\end{aligned}$$

These can be gotten from the joint distribution by including only the terms in the product that contain the random variable of interest

$$p(s, \lambda, \beta) = \left(\prod_{i=1}^{10} \frac{(\lambda_i t_i)^{s_i} e^{-\lambda_i t_i}}{s_i!}\right) \left(\prod_{i=1}^{10} \lambda_i^{\alpha-1} \beta^{\alpha} \frac{e^{-\lambda_i \beta}}{\Gamma(\alpha)}\right) \beta^{\delta-1} \gamma^{\delta} \frac{e^{-\beta\delta}}{\Gamma(\gamma)}$$

Equivalently, you can do this by looking at the graph structure of the model by only including terms that correspond to edges joining to the node of interest. (e.g. for $\beta$, which edges connect with the node for $\beta$.)




In these graphs, for every factor of the joint distribution, the nodes for the variables in the factor are joined.

$$p(s, \lambda, \beta) = \left( \prod_{i=1}^{10} \frac{(\lambda_i t_i)^{s_i} e^{-\lambda_i t_i}}{s_i!} \right) \left( \prod_{i=1}^{10} \lambda_i^{\alpha-1} \beta^{\alpha} \frac{e^{-\lambda_i \beta}}{\Gamma(\alpha)} \right) \beta^{\delta-1} \gamma^{\delta} \frac{e^{-\beta\delta}}{\Gamma(\gamma)}$$

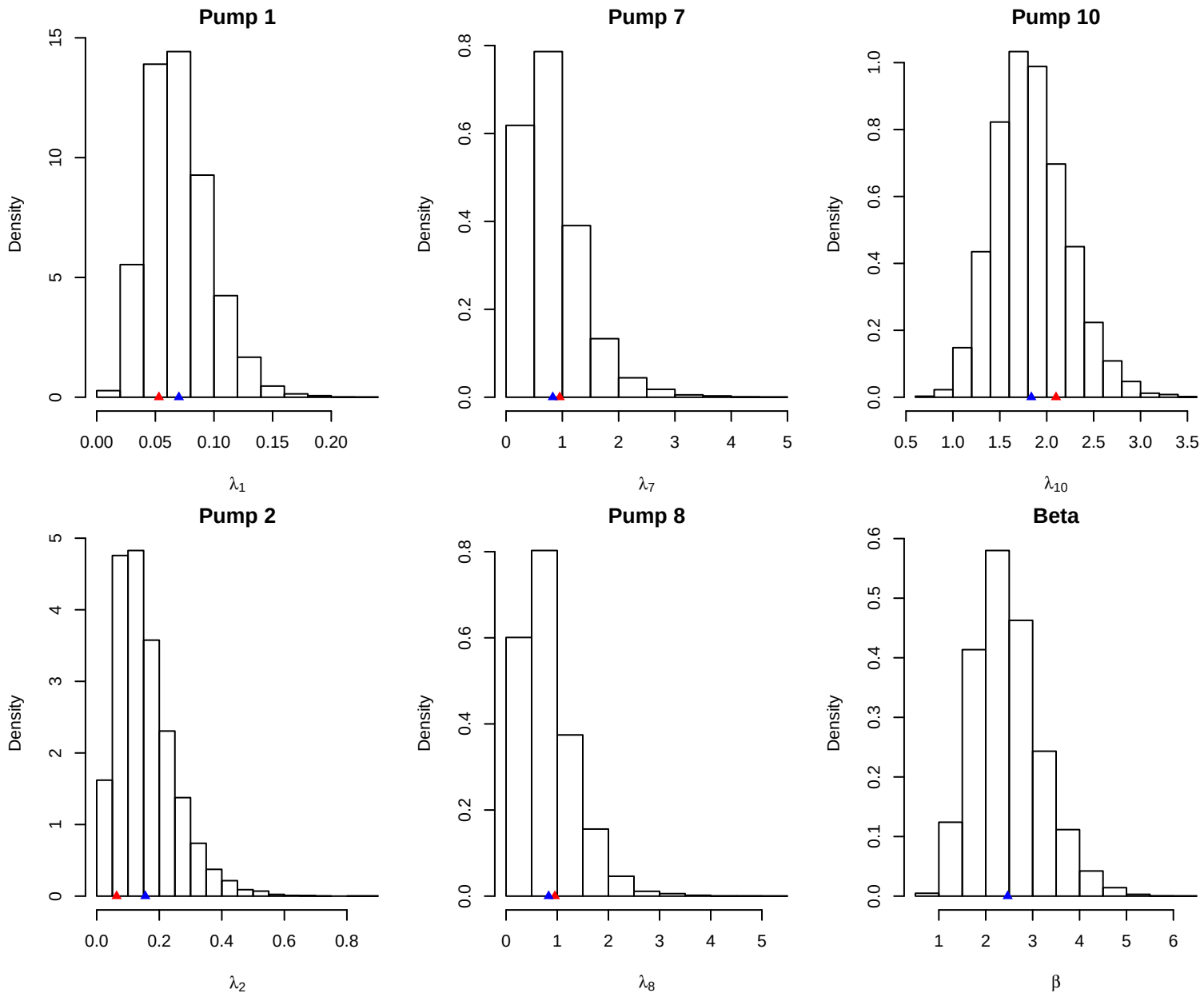
Example Run:

The following values were used for the prior parameters

$$\alpha = 1.8 \qquad \delta = 1 \qquad \gamma = 0.1$$

$n = 10000$ imputations were generated after a burn in of $1000$ imputations.

The starting values for the chain were $\beta^0 = \bar{l} = 1.33, \lambda_i = l_i$.


Pump 1
Density
0
5
10
15
0.00
0.05
0.10
0.15
0.20
λ1
Pump 7
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
λ7
Pump 10
Density
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
λ10
Pump 2
Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
λ2
Pump 8
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
λ8
Beta
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
β

| Pump | $l_i$ | $t_i$ | $E[\lambda_i \mid s]$ | $\mathrm{Med}(\lambda_i \mid s)$ | $\mathrm{SD}(\lambda_i \mid s)$ |
|---|---|---|---|---|---|
| 1 | 0.0530 | 94.320 | 0.0700 | 0.0667 | 0.0267 |
| 2 | 0.0636 | 15.720 | 0.1553 | 0.1368 | 0.0935 |
| 3 | 0.0795 | 62.880 | 0.1044 | 0.0991 | 0.0401 |
| 4 | 0.1113 | 125.760 | 0.1231 | 0.1204 | 0.0307 |
| 5 | 0.5725 | 5.240 | 0.6283 | 0.5861 | 0.2914 |
| 6 | 0.6043 | 31.440 | 0.6167 | 0.6071 | 0.1343 |
| 7 | 0.9541 | 1.048 | 0.8298 | 0.7188 | 0.5379 |
| 8 | 0.9541 | 1.048 | 0.8316 | 0.7178 | 0.5302 |
| 9 | 1.9083 | 2.096 | 1.3020 | 1.2150 | 0.5744 |
| 10 | 2.0992 | 10.480 | 1.8358 | 1.8087 | 0.3873 |

$$E\left[\frac{\alpha}{\beta}|s\right] = 0.7929$$

| $E[\beta\|s]$ | $\mathrm{Med}(\beta\|s)$ | $\mathrm{SD}(\beta\|s)$ |
|---|---|---|
| 2.4678 | 2.3970 | 0.7074 |