Direct Simulation Methods #2

Econ 690

Purdue University

Outline

Rejection Sampling Algorithm #2

- Suppose we wish to obtain draws from some target density $f(\theta)$.
- Let Θ denote the support of $f(\theta)$.
- Suppose there exists some approximating density $s(\theta)$, called the *source density*, with support Θ^*, where $\Theta \subseteq \Theta^*$.

Rejection Sampling Algorithm #2

- In many applications requiring posterior simulation, the normalizing constant of the target density is unknown, since the joint or conditional posteriors are only given up to proportionality.
- To this end, let us work with the kernels of both the source and target densities and write:

  so that $\tilde{f}$ and $\tilde{s}$ denote the target and source kernels, respectively, and c_f and c_s denote the associated normalizing constants.
- Finally, let

Rejection Sampling Algorithm #2

Consider the following algorithm:

1.
2. Draw a candidate from the source density $s(\theta)$, i.e.,
3.

For this algorithm, we seek to answer the following questions:

- (a) Show how this algorithm includes the previous one as a special case.
- (b) What is the overall acceptance rate in this algorithm?
- (c) Sketch a proof of why this algorithm provides a draw from $f(\theta)$.

- As for our first question, consider using the algorithm above to generate a draw from $f(\theta)$ with compact support $[a, b]$.
- In addition, employ a source density $s(\theta)$ that is Uniform over the interval $[a, b]$.
- In this case we can write

$$f(\theta) = c_f g(\theta) I(a \leq \theta \leq b) = c_f \tilde{f}(\theta),$$

where $\tilde{f}(\theta) = g(\theta) I(a \leq \theta \leq b)$, $\int_a^b g(\theta) = c_f^{-1}$ and

$$s(\theta) = [b-a]^{-1} I(a \leq \theta \leq b) = [b-a]^{-1} \tilde{s}(\theta).$$

- It follows that

$$\tilde{M} = \max_{a \leq \theta \leq b} \left(\frac{\tilde{f}(\theta)}{\tilde{s}(\theta)} \right) = \max_{a \leq \theta \leq b} g(\theta) = c_f^{-1} \max_{a \leq \theta \leq b} f(\theta) = c_f^{-1} M,$$

where M is defined as the maximum of f as in our first rejection sampling algorithm.

- To implement the algorithm with the given Uniform source density, we first generate $\theta^{cand} \sim \mathcal{U}(a, b)$, which is equivalent to writing
  -
- We then generate $U_2 \sim \mathcal{U}(0, 1)$ and accept θ^{cand} provided
  -
- This decision rule and the random variables U_1 and U_2 are identical to those described in the first algorithm.

As for the second question, the overall acceptance rate is

-

As for part (c), we note that for any subset A of Θ

-

Since

-

it follows that when θ is accepted from the algorithm, it is indeed a draw from $f(\theta)$.

Let us now carry out the following exercise which illustrates the use of the algorithm, as we seek to generate draws from the triangular distribution:

-

(a) Comment on the performance of an algorithm that uses a $\mathcal{U}(-1, 1)$ source density.

(b) Comment on the performances of an alternate algorithm that uses a $N(0, \sigma^2)$ source density.

- First, consider a standard Normal source with $\sigma^2 = 1$.
- Then, investigate the performance of the acceptance/rejection method with $\sigma^2 = 2$ and $\sigma^2 = 1/6$.

- (a)For the $\mathcal{U}(-1, 1)$ source density, our previous derivation showed that our generalized algorithm reduces to the first rejection sampling algorithm discussed in the previous lecture.
- Thus, it follows that the overall acceptance rate is .5.
- We regard this as a benchmark and seek to determine if an alternate choice of source density can lead to increased efficiency.

- For the Normal source density in part (b), when $\sigma^2 = 1$, the maximum of the target/source ratio occurs at $\theta = 0$ yielding a value of $\tilde{M} = 1$.
- Since the normalizing constant of the standard Normal is $c_s = (2\pi)^{-1/2}$, it follows that the overall acceptance rate is $(2\pi)^{-1/2} \approx .40$.
- When comparing this algorithm to one with, say, $\sigma^2 = 2$, it is clear that the standard Normal source will be preferred. The maximum of the target/source ratio with $\sigma^2 = 2$ again occurs at $\theta = 0$ yielding $\tilde{M} = 1$. However, the overall acceptance probability reduces to $1/\sqrt{4\pi} \approx .28$.

- The final choice of $\sigma^2 = 1/6$ is reasonably tailored to fit this application.
- One can easily show that the mean of the target is zero and the variance is $1/6$, so that the $N(0, 1/6)$ source is chosen to match these two moments of the triangular density.
- With a little algebra, one can show that the maximum of the target/source ratio occurs at

$$\theta = \frac{1 + \sqrt{1/3}}{2} \approx .789.$$

  (Note that another maximum also occurs at -.789 since both the target and source are symmetric about zero).
- With this result in hand, the maximized value of the target/source ratio is $\tilde{M} \approx 1.37$, yielding a theoretical acceptance rate of $1/(1.37\sqrt{2\pi[1/6]}) \approx .72$.
- Thus, the $N(0, 1/6)$ source density is the most efficient of the candidates considered here.

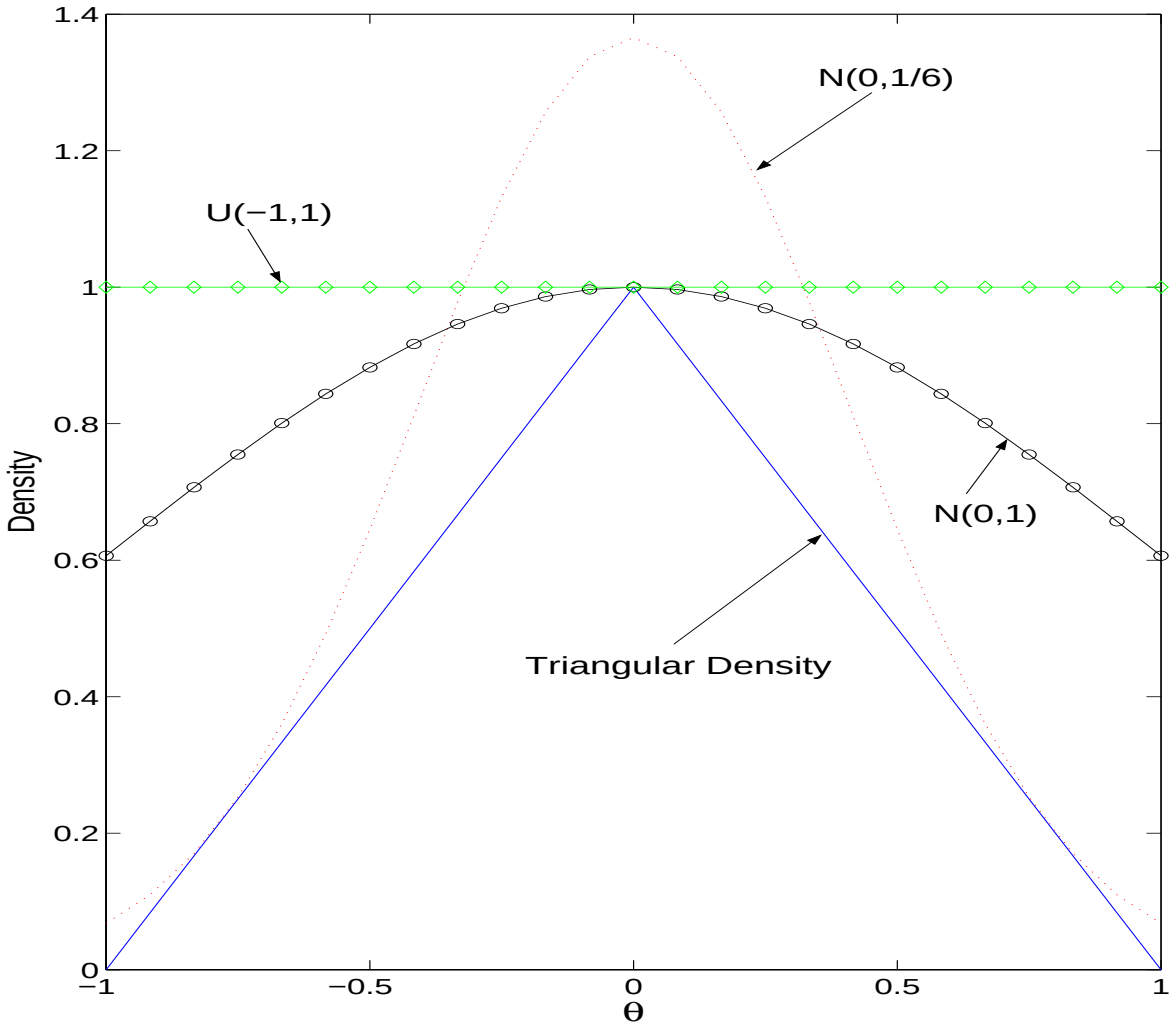

Figure: Triangular Density together with Three Different Scaled Source Densities

The Weighted Bootstrap

- A potential difficulty of the previous method is the calculation of M.

- The weighted bootstrap of Smith and Gelfand (1992, *American Statistician*) and the highly related sampling-importance resampling (SIR) algorithm of Rubin (1987 *JASA*) circumvent the need to calculate this "blanketing constant" $\tilde{M}$.
- Another alternative for *log-concave* densities [which, in the univariate case, refers to a density for which the second derivative of the log density is everywhere non-positive] is adaptive rejection sampling, which we do not discuss here. [Gilks and Wild (1992, *JRSS C*)]

- Consider the following procedure for obtaining a draw from a density of interest f:
  1. Draw $\theta_1, \theta_2, \cdots \theta_n$ from some approximating source density $s(\theta)$.
  2. Like our last rejection sampling algorithm, let us work with the kernels of the densities and write:

     Set

     and define the normalized weights

  3. Draw θ^* from the discrete set $\{\theta_1, \theta_2, \cdots, \theta_n\}$ with $\Pr(\theta^* = \theta_j) = \tilde{w}_j, \quad j = 1, 2, \cdots, n.$

We seek to show that θ^* provides an *approximate* draw from $f(\theta)$, with the accuracy of the approach improving with n, the simulated sample size from the source density. Note that

-

Continuing,

-

- The following example illustrates the performance of the weighted bootstrap, again considering the problem of sampling from the triangular density.
- For our source, we use a $\mathcal{U}(-1, 1)$ density, and generate 1,500 draws from it.
- We then calculate the density of the source at these draws (which is always 1/2 !) as well as the triangular density at these 1,500 values, and calculate the weights $\tilde{w}_i,\ 1 = 1, 2, \cdots, 1,500$.
- From this discrete distribution, we generate 50,000 draws.
- A histogram of these draws is provided on the following page:

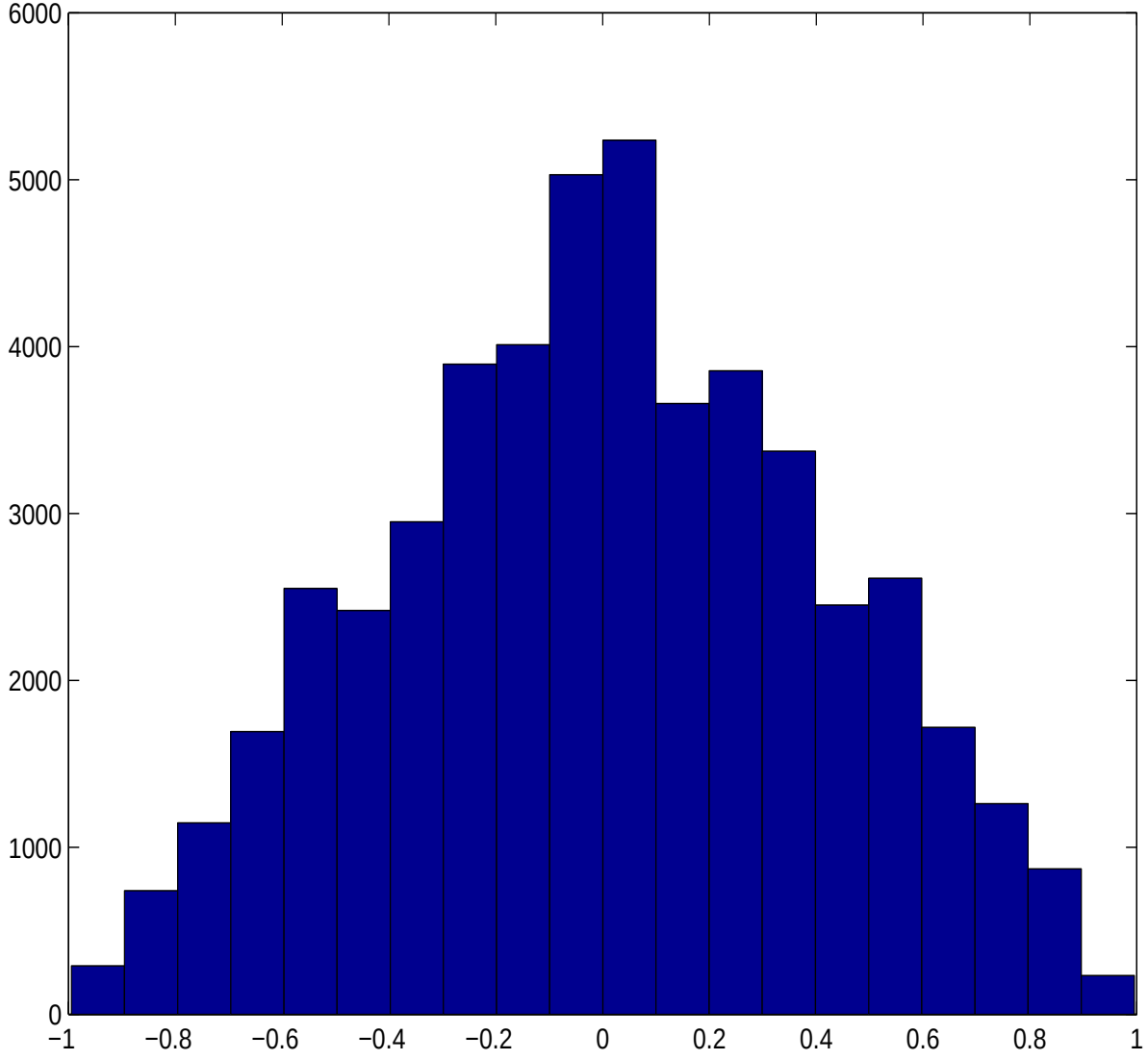

Figure: Histogram from Weighted Bootstrap Simulations

Importance Sampling

Again, consider the problem of calculating a posterior moment of the form:

-

and we assume (as is typically the case) that it is not possible to draw directly from $p(\theta|y)$.

A potentially useful method in this situation is to apply importance sampling, whose use was championed for Bayesian applications by Kloek and van Dijk (1978, *Econometrica*) and Geweke (1989 *Econometrica*).

To provide an intuitive explanation behind the importance sampling estimator, note that the posterior mean calculation can be re-written in the following way:

-

for some importance function $I(\theta)$ whose support includes Θ.

- Written this way, one can see that the original integrals have been transformed into new (though equivalent) problems of moment calculation.
- The averaging, however, is now performed with respect to $I(\theta)$ instead of $p(\theta|y)$.
- Provided one can draw from the importance function $I(\theta)$, direct Monte Carlo integration can be performed on the numerator and denominator individually to produce the importance sampling estimator (shown on the following page):

- a weighted average of the $g(\theta_i)$ with $\tilde{w}(\theta_i) = w(\theta_i)/\sum_i w(\theta_i)$ denoting the (normalized) weight and $w(\theta_i) \equiv p(\theta_i)L(\theta_i)/I(\theta_i)$.

- Importantly, note that for the case of importance sampling, $\theta_i \overset{iid}{\sim} I(\theta)$, and are not draws from $p(\theta|y)$ as was the case in direct Monte Carlo integration.
- Since the importance function $I(\theta)$ is under the control of the researcher, it can (and should) be a density from which samples can be easily obtained.
- Though the importance sampling estimator may seem like a convenient way to solve all problems of posterior moment calculation, note that if $I(\theta)$ is a poor approximation to $p(\theta|y)$, then the "weights" $\tilde{w}(\theta)$ will typically be small for most values of θ_i, resulting in a very inaccurate and unstable estimate.

- Common sense, of course, suggests that the accuracy of an importance sampling estimate will improve as $I(\theta)$ more closely approximates the target distribution $p(\theta|y)$.
- Indeed, if $I(\theta)$ and $p(\theta|y)$ coincide, then the "weights" $\tilde{w}(\theta) = 1/N$, and the importance sampling estimator reduces to the ideal case, the direct Monte Carlo estimator.
- However, this is an ideal that we cannot achieve, as direct sampling from $p(\theta|y)$ is typically not possible.

Geweke (1989) shows, under standard conditions, namely:

- $E[g(\theta)|y] \equiv \overline{g}$ exists,
- $\text{Var}[g(\theta)|y] \equiv \sigma^2$ is finite
- The support of $I(\theta)$ includes the support of $p(\theta|y)$,

then

$$E\widehat{[g(\theta)|y]} \equiv \widehat{g}_{IS} \xrightarrow{p} \overline{g}.$$

i.e., the importance sampling estimator is *consistent*. Establishing a rate of convergence is a little more involved. The important new condition is that $w(\theta)$ must be bounded above on θ which means that, in general $I(\theta)$ should be chosen to have heavier tails than $p(\theta|y)$.

Under this new condition, Geweke (1989) shows:

-

where τ^2 can be consistently estimated by:

-

Note that, if $I(\theta) = p(\theta|y)$, then $w(\theta) = 1$, and the above reduces to our estimate of $\text{Var}[g(\theta)|y] = \sigma^2$.
Also note that the numerical standard error (denoted NSE) can be approximated as follows:

-

- To evaluate the performance of a particular importance sampling estimator, Geweke (1989) suggests a number of diagnostics, including:
- Monitoring the weights $\tilde{w}_i$.If these are large for a few draws, it is suggestive of an inaccurate approximation of the posterior moment.
- Geweke (1989) also suggests keeping track of the fraction of total weight assigned to the draw receiving highest weight. If this is found to be much larger than $1/N$, it suggests that $I(\theta)$ can be improved.

- A more formal and widely-used statistic involves the calculation of a quantity termed relative numerical efficiency or RNE.
- This statistic seeks to quantify how much is lost (owing to a choice of $I(\theta)$ that is far from the target $p(\theta|y)$) by using importance sampling relative to the numerical precision that would have obtained using direct Monte Carlo integration.
- Specifically, the RNE is defined as the ratio σ^2/τ^2 which can be consistently estimated by:

$$R\hat{N}E = \frac{\sum_{i=1}^{N}[g(\theta_i) - \widehat{g}_{IS}]^2 w(\theta_i)}{\sum_{i=1}^{n} w(\theta_i)} \left(\frac{N \sum_{i=1}^{N}[g(\theta_i) - \widehat{g}_{IS}]^2 w^2(\theta_i)}{\left[\sum_{i=1}^{N} w(\theta_i)\right]^2} \right)^{-1}.$$

Since

$$NSE[\hat{g}_{IS}] \equiv \sqrt{\frac{\tau^2}{N}}$$

it follows that

-

and thus $N \times RNE$ has the interpretation of the *effective sample size*.

Since RNE tends to be smaller than 1 in practice, and may be significantly smaller when there is large variability in the weights w, this gives a sense of the "loss" relative to direct Monte Carlo integration.

- To illustrate the performance and use of importance sampling, we consider the following experiment.
- Suppose it is of interest to calculate the first and second moments of a standard normal random variable.
- As an importance function, $I(\theta)$, we employ the Laplace or double exponential distribution:
  -
- Thus, the Laplace density has two parameters: a mean μ and a parameter b which controls the variance. (Specifically, $\text{Var}(\theta) = 2b^2$.)
- Note: The Laplace distribution will have heavier tails than the normal since we have a linear rather than quadratic term in the exponential kernel.

- We will consider the use of two different Laplace distributions as importance functions: a Laplace(0,2) distribution and a Laplace(3,1) distribution.
- We will obtain 2,500 different Importance sampling estimates under each importance function, each time generating 5,000 draws from the corresponding Laplace distribution.
- This procedure will approximate the sampling distributions of the importance sampling estimators.
- In addition, we plot the sampling distribution of RNE for each case.

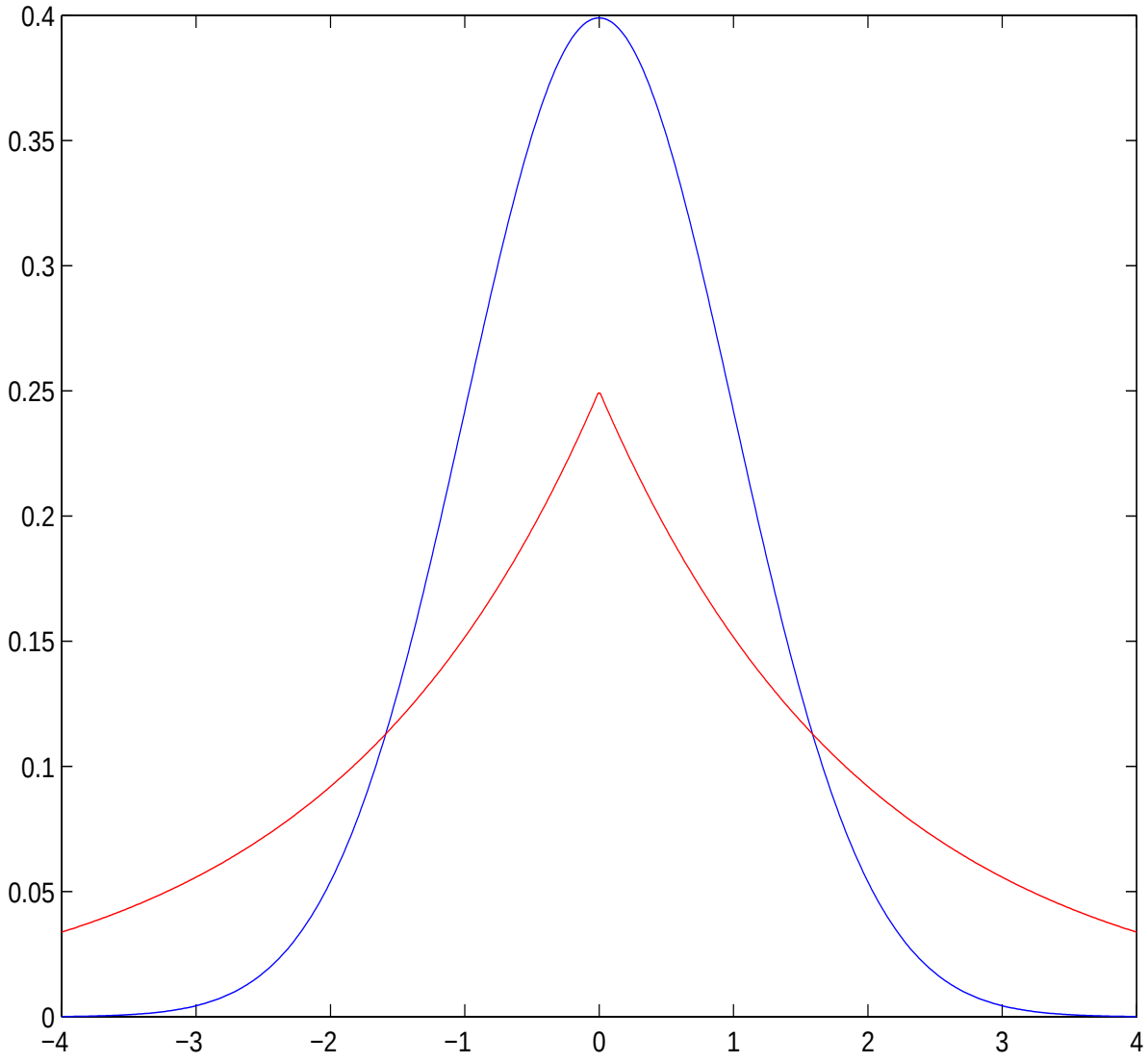

Figure: Laplace(0,2) importance function and Standard Normal Density

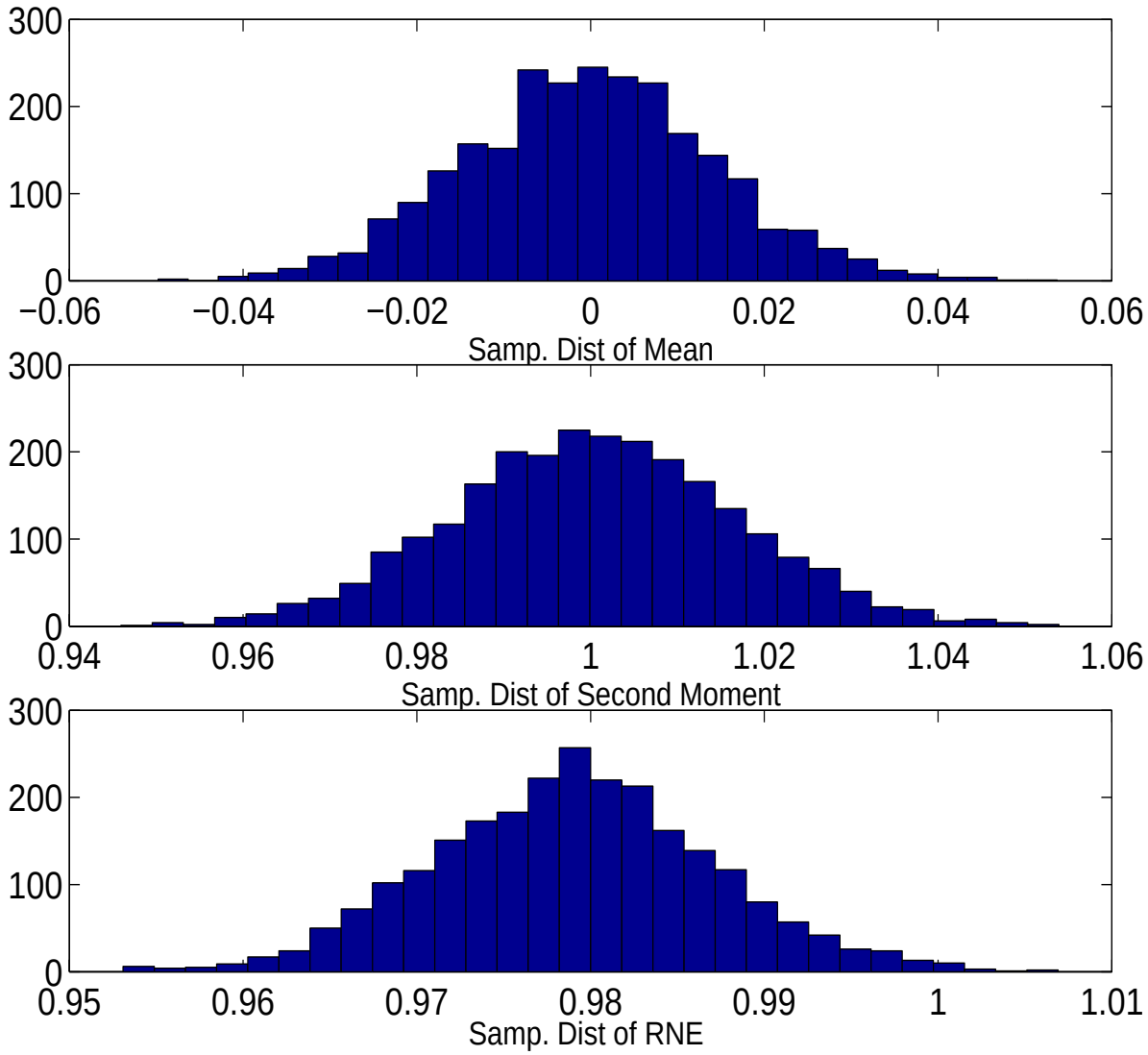

Figure: Samp. Distributions using Laplace(0,2) importance function

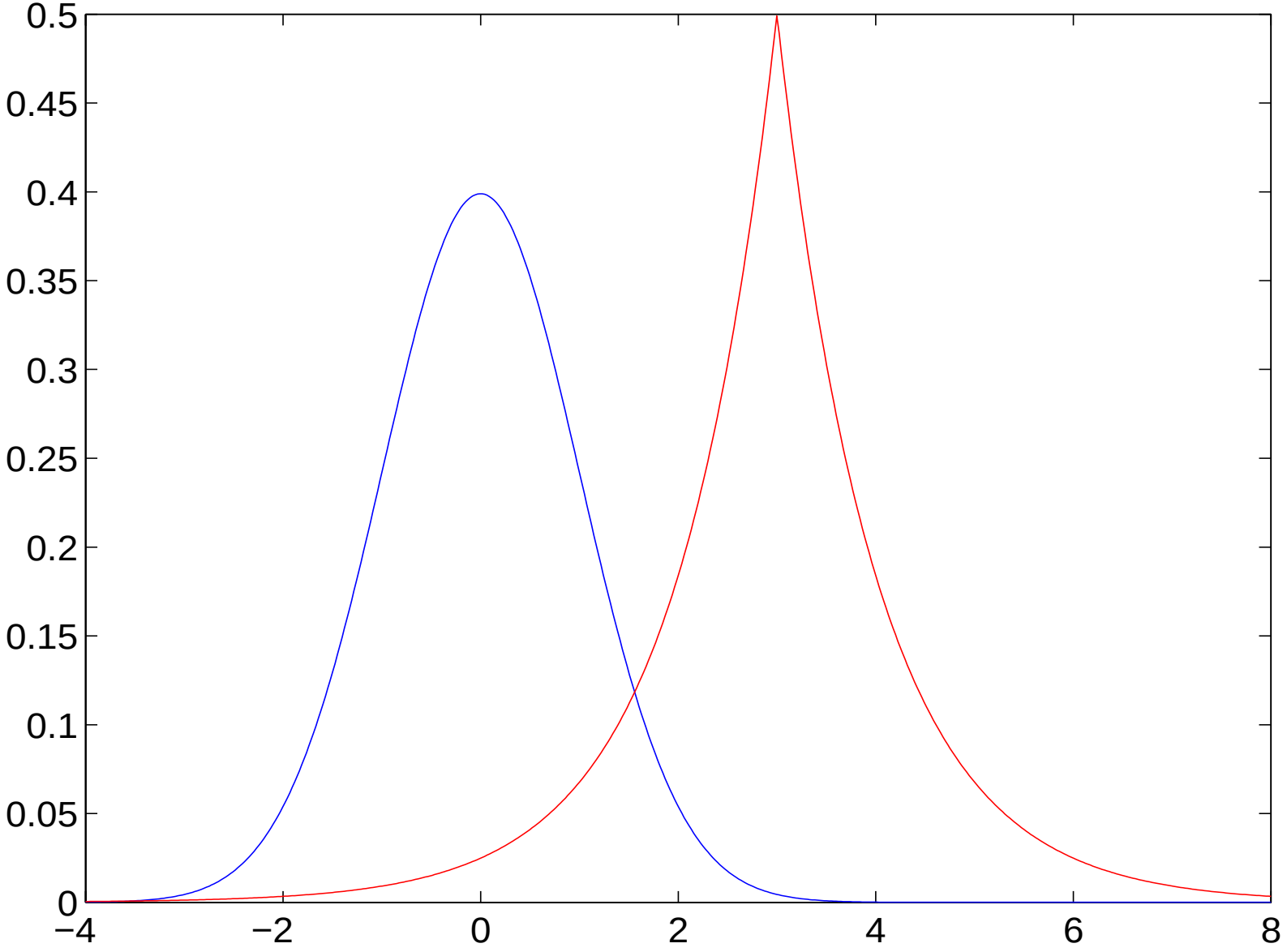

Figure: Laplace(3,1) importance function and Standard Normal Density

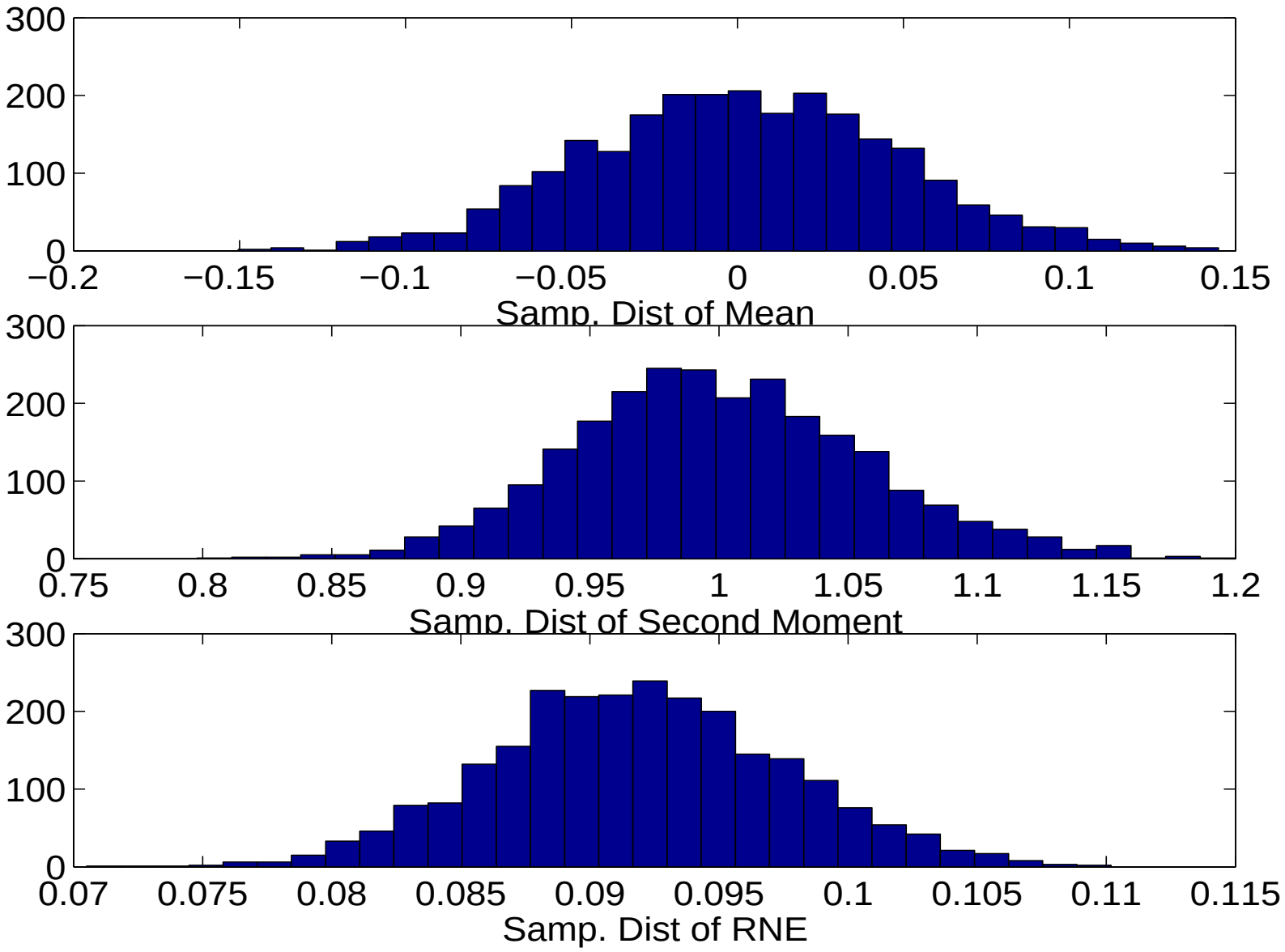

Figure: Samp. Distributions using Laplace(3,1) importance function

- Clearly, the Laplace(0,2) density performs better than the Laplace(3,1) alternative.
- Note that, in this final case, the effective sample size is approximately one-tenth of the sample size under iid sampling.
- That is, to obtain an equal level of numerical precision in the estimated mean, we will need to obtain an importance sampling sample that is ten times as large as the sample size required under direct Monte Carlo integration.