

Importance Sampling

John Mount*

January 1, 2012

Abstract

We describe an application of using a change of sampling density to get easier access to rare events during numeric simulations (this is called importance sampling). Our emphasis is on the derivation of the change of density instead of the algorithmic details. We work a small example to make the technique concrete.

Contents

1	Introduction	1
2	The Problem	2
3	Importance Sampling	2
3.1	Some Notation and Terminology	2
3.2	Change of Density	3
3.3	The Method	4
3.4	A Wrong Estimate	5
3.5	Variance	6
3.6	Picking $q()$	6
4	Worked Example	7
4.1	The Steps	8
4.2	The Steps Applied to Our Example	9
4.3	Results	10
5	Conclusion	14

1 Introduction

We describe briefly the powerful simulation technique known as “importance sampling.” Importance sampling is a technique that lets you use numerical simulation to explore events that, at first look, appear too rare to be reliably approximated numerically. The correctness of importance sampling follows almost immediately from the definition of a change of density. Like most mathematical techniques, importance sampling brings in its own concerns and controls that were not obvious in the original problem. To deal with these concerns (like picking the re-weighting to use) we will largely appeal to the ideas from “A Tutorial on the Cross-Entropy Method” [dBKMR05]. The writeup is hosted on the Win-Vector blog: <http://www.win-vector.com/blog/2012/01/importance-sampling/> ([link](#)).

*email: <mailto:jmount@win-vector.com>

2 The Problem

Importance sampling is concerned with the simulation and characterization of rare events. To make this discussion specific we will settle for one simplified rare event: simulated mortgage default. For our simulation we will assume we have a portfolio of four commercial mortgages. Each of the four mortgage payers has a monthly income that is log-normally distributed with mean \$2000 (and $\sigma = 1$, for a maximal variance for the given mean; see http://en.wikipedia.org/wiki/Log-normal_distribution ([link](#))) plus a common income component (also log normal with mean \$2000 and $\sigma = 1$). The common income component is the same for all four mortgage payers (like a payment from a large shared income source).

An example monthly income statement for our four mortgage payers could be as follows. The unique incomes are: \$1237, \$4313, \$1215, \$687 and the common component as \$846. Thus mortgage payer one has income of \$1237 + \$846, mortgage payer two has an income of \$4313 + \$846 and so on. We will further assume (as a very unrealistic simplification) that these payers never save, have no other assets and must each make a \$500 payment each month to avoid defaulting on their mortgage. This would seem easy as the expected incomes are far greater than the expenses, but we have picked income distributions with very high variances. So it is not unheard of for one of these payers to have a very small income in a given month. In this model default is a rare but possible event. Also the shared income component introduces the statistical problem of a correlation- the mortgages are more likely to default together (a fact that has been often ignored to ruinous effect).

Using a straightforward simulation (repeated 1,000,000 times to get a reliable view of our rare event) we find that in about 2.7% of the simulations at least one mortgage payer defaults (due to the large variance in income and complete lack of sharing or savings). We also estimate that around 0.76% of individual mortgages default (which is more than $2.7\%/4 = 0.68\%$). In fact the expected number of defaults from our four payers given that we have at least one payer defaulting is around 1.13 (which is much higher than we would expect if defaults were independent). We would like to measure the correlation between mortgage defaults. The problem is that for very rare events an achievable number of samples may not be enough for reliable estimates (especially for estimating things like expected number of defaults given at least one mortgage default has occurred, as this is calculated from only a 2.7% portion of the data in this case). For illustration purposes we have deliberately picked a not too rare “rare event” so it is accessible through standard sampling- but for truly rare events standard sampling will eventually break down.

3 Importance Sampling

The solution we discuss is called importance sampling. The main idea of importance sampling is what I call a change of variables theorem for sampling: change of density.

3.1 Some Notation and Terminology

To work correctly with probabilities we need to be fairly precise about how the random variables that we are working with behave. We are going to avoid some technicalities by being a little non-standard in our definitions and treatment of probability and measure.

We define Ω as our range of all possible values (typically Ω will be \mathbb{R}^n , $[0-1]^n$, \mathbb{Z}^n , $\{0, 1\}^n$, $\{1, \dots, k\}^n$ or some combination of these). As is common in probability theory we will write $\int_{x \in S} f(x) dx$ as the integration or summation of $f()$ over $S \subseteq \Omega$.¹

¹In probability theory it is acceptable to use the integral notation even for sums over finite probability spaces (or probability spaces with atomic masses).

A function $f()$ and a set $S \subseteq \Omega$ is called a *summable pair* if $\int_{x \in S} f(x)dx$ exists and we say this as: “ $f()$ is summable over S .” In this write-up we will only be working over pairs of functions and sets assumed to be summable.

A function $p()$ is that is non-negative and summable over Ω with $\int_{\substack{x \in \Omega \\ p(x) \neq 0}} p(x)dx > 0$ will be called an “un-normalized density” (and for conciseness we will omit mentioning Ω). An un-normalized density $p()$ is called a normalized density (or just a density or a distribution) if we also have $\int_{\substack{x \in \Omega \\ p(x) \neq 0}} p(x)dx = 1$.

For a density (normalized or un-normalized) $p()$: a random variable “drawn according to $p()$ ” is a selection of an $x \in \Omega$ such that the probability that $x = v$ is proportional to $p(v)$ and we *never* select x such that $p(x) = 0$.²

The expectation of an arbitrary function $f()$ with respect to a density (normalized or un-normalized) $p()$ is written as $E_{x \sim p()}[f(x)]$ and in this paper is defined as:

$$E_{x \sim p()}[f(x)] = \int_{\substack{x \in \Omega \\ p(x) \neq 0}} f(x)p(x)dx / \int_{\substack{x \in \Omega \\ p(x) \neq 0}} p(x)dx. \quad (1)$$

The intent is to have:

$$E_{x \sim p()}[f(x)] \approx \frac{1}{m} \sum_{\substack{i=1 \\ x_i \sim p()}}^m f(x)$$

but to have this as a consequence of a versatile integration framework (not to have to use this as a definition).

For a set $S \subseteq \Omega$ we define the function $\Phi_S()$ as the function that $\Phi_S(x)$ is 1 if $x \in S$ and 0 otherwise. This sort of function is called an indicator and for a density $p()$ (normalized or un-normalized) we write $p(S) = p(\Phi_S()) = E_{x \sim p()}[\Phi_S(x)]$ as the probability or measure of S under $p()$.

3.2 Change of Density

The reason we defined our terms so laboriously is so that we can now clearly write the change of density theorem. The problem is that change of density is both subtle (so it is a bit hard to believe) and an immediate consequence of definitions (so it doesn’t have a satisfying derivation if you are not sufficiently careful with your definitions).

Theorem 1 (Change of Density). *Let $p()$ be an un-normalized density and $f()$ a function on Ω . Let $q()$ be a second un-normalized density such that $q(x) = 0$ implies $f(x)p(x) = 0$. Then:*

$$E_{x \sim p()}[f(x)] = E_{x \sim q()}[f(x)p(x)/q(x)] \left(\int_{\substack{x \in \Omega \\ q(x) \neq 0}} q(x)dx / \int_{\substack{x \in \Omega \\ p(x) \neq 0}} p(x)dx \right)$$

(when both expectations exist).

Proof. By definition we have:

$$E_{x \sim q()}[f(x)p(x)/q(x)] = \int_{\substack{x \in \Omega \\ q(x) \neq 0}} (f(x)p(x)/q(x))q(x)dx / \int_{\substack{x \in \Omega \\ q(x) \neq 0}} q(x)dx.$$

²This is not standard. In standard measure and probability theory you can have events occur where $p(x) = 0$, but we are only going to depend on $p(x) = 0$ not occurring on its own for random variables that we are drawing a finite number of times so we can get away with this additional condition (despite the suggestive appearance of the $E[\]$ notation and associated integral.) Also the statement “ x is drawn with odds $p(x)$ ” is meant as a naive stand-in for the the more formal definition that we observe $x \in S$ with probability $\int_{\substack{x \in S \\ p(x) \neq 0}} p(x)dx / \int_{\substack{x \in \Omega \\ p(x) \neq 0}} p(x)dx$ for all measurable $S \subseteq \Omega$.

This simplifies to:

$$\int_{\substack{x \in \Omega \\ q(x) \neq 0}} f(x)p(x)dx / \int_{\substack{x \in \Omega \\ q(x) \neq 0}} q(x)dx$$

(and note we are not canceling out any 0/0 cases due to the $q(x) \neq 0$ condition). Also by definition:

$$\mathbb{E}_{x \sim p()}[f(x)] = \int_{\substack{x \in \Omega \\ p(x) \neq 0}} f(x)p(x)dx / \int_{\substack{x \in \Omega \\ p(x) \neq 0}} p(x)dx.$$

So we only need to confirm that:

$$\int_{\substack{x \in \Omega \\ q(x) \neq 0}} f(x)p(x)dx = \int_{\substack{x \in \Omega \\ p(x) \neq 0}} f(x)p(x)dx.$$

And this is true because $q(x) = 0$ implies $f(x)p(x) = 0$ (so the two integrals differ only in regions of Ω where $f(x)p(x)$ is zero).³

□

The exciting case is when $q()$ and $p()$ are in fact normalized densities and we have

$$\mathbb{E}_{x \sim p()}[f(x)] = \mathbb{E}_{x \sim q()}[f(x)p(x)/q(x)]. \quad (2)$$

Suppose we are trying to estimate $\mathbb{E}_{x \sim p()}[f(x)]$. By the change of density theorem we know:

$$\mathbb{E}_{x \sim p()}[f(x)] = \mathbb{E}_{x \sim q()}[f(x)p(x)/q(x)].$$

And the straightforward way to estimate $\mathbb{E}_{x \sim q()}[f(x)p(x)/q(x)]$ is to say:

$$\mathbb{E}_{x \sim q()}[f(x)p(x)/q(x)] \approx \frac{1}{m} \sum_{\substack{i=1 \\ x_i \sim q()}}^m f(x_i)p(x_i)/q(x_i).$$

Yielding the overall estimate:

$$\mathbb{E}_{x \sim p()}[f(x)] \approx \frac{1}{m} \sum_{\substack{i=1 \\ x_i \sim q()}}^m f(x_i)p(x_i)/q(x_i) \quad (3)$$

which we will call the change of density estimate. We are using our non-standard convention that we do not experience any $x \sim q()$ where $q(x) = 0$ to ensure there are never any terms of the form 0/0 in Equation 3.

3.3 The Method

We define an indicator function $\Phi()$ such that $\Phi(x) = 1$ exactly for $x \in \Omega$ that have the rare event we are interested in (and 0 for other x). For example suppose $\Phi(x) = 1$ if at least one of the mortgages in our portfolio of four mortgages defaults (and 0 otherwise). In our original example we were interested in estimating $p(\Phi())$: the probability that our portfolio of four mortgages encounters at least one default (where $p()$ was the income distribution we defined in Section 2). Our problem is

³Note that requiring $q(x) = 0$ implies $f(x)p(x) = 0$ is a necessary condition. As an example of what can go wrong: set $\Omega = \{0, 1\}$, $p(x) = \frac{1}{2}$, $f(0) = 0$, $f(1) = 1$, $q(0) = 1$ and $q(1) = 0$. In this case $\mathbb{E}_{x \sim p()}[f(x)] = \frac{1}{2}$ but $\mathbb{E}_{x \sim q()}[f(x)p(x)/q(x)] = 0$. Whereas even a $q()$ with $q(0) = 1 - \epsilon$ and $q(1) = \epsilon$ works as $\mathbb{E}_{x \sim q()}[f(x)p(x)/q(x)] = (f(0)p(0)/q(0))q(0) + (f(1)p(1)/q(1))q(1) = (0 \times \frac{1}{2})/(1 - \epsilon)(1 - \epsilon) + (1 \times \frac{1}{2}/\epsilon)\epsilon = \frac{1}{2}$ (though would be a poor choice for speeding things up).

that $\Phi()$ might be a very rare event under the $p()$ density. What we do is pick another density $q()$ where $\Phi()$ is a more common event. We then use Equation 3 to estimate $p(\Phi())$ as:

$$p(\Phi()) \approx \frac{1}{m} \sum_{\substack{i=1 \\ x_i \sim q()}}^m \Phi(x_i)p(x_i)/q(x_i) \quad (4)$$

This estimate is always correct in expectation, but we need additional conditions on the variance of the estimate to ensure it is in fact useful. For example it is traditional to insist that $q()$ has been picked such that $q(x) > \epsilon$ for all x such that $\Phi(x) = 1$ (otherwise we may have terms where unseen terms where $\Phi(x_i)p(x_i)/q(x_i)$ is arbitrarily high, leading to huge variance).

3.4 A Wrong Estimate

Notice that we are not using an estimate of the more intuitive conditional form:

$$\begin{aligned} &\text{Generate } x_1, \dots, x_m \sim q() \\ p(\Phi()) &\approx \frac{\sum_{i=1}^m \Phi(x_i)p(x_i)/q(x_i)}{\sum_{i=1}^m p(x_i)/q(x_i)} \end{aligned}$$

which *looks* very much like a conditional expectation or probability. This estimate is often incorrect (are therefor can't be justified in general). We will call this the "pseudo-conditional estimate."

The pseudo-conditional estimate seems more intuitive than Equation 4 (as it looks more familiar) but it is incorrect. Notice that we have just said "it looks familiar" we have not derived the pseudo-conditional estimate or given any reason to expect it is correct.

We give an extreme example. Suppose Ω is the range of integers $\{1, \dots, 10\}$, $\Phi(x) = 1$ only when $x \leq 5$ (otherwise $\Phi(x) = 0$), $p(x) = \frac{1}{10}$ and $q(x) = \frac{\Phi(x)}{5}$. Both $p()$ and $q()$ are normalized densities and we have $q(x) = 0$ implies $\Phi(x) = 0$. As usual the task is to estimate $p(\Phi())$ (the probability that $x \sim p()$ is has $\Phi(x) = 1$). We apply Equation 4 to try and speed up estimation. For this $q()$ Equation 4 is in fact perfect- a single sample drawn according to $q(x)$ plugged into Equation 4 always exactly (and correctly) estimates $p(\Phi()) = \frac{1}{2}$ because $p(x)/q(x) = \frac{1/10}{1/5} = \frac{1}{2}$ when $q(x) \neq 0$.

Now consider these same Ω , $\Phi()$, $p()$, $q()$ plugged into the pseudo-conditional estimate. For this example we also see $q(x) \neq 0$ implies $\Phi(x) = 1$. Therefore $\Phi(x) = 1$ for all x_i we see in all $q()$ -drawn samples. So we always have the numerator identical to the denominator in our pseudo-conditional estimate. Thus the estimate is always 1 (which is wrong). The problem is: while the expected value of the numerator of the pseudo-conditional estimate is $m \times p(\Phi())$ (by Equation 4) the expected value of the denominator is not always m .⁴

Why do we even discuss this erroneous estimate? Because it can easily arise in practice (either due to looking like a conditional expectation or mis-applying Equation 4 to the denominator).

Under some conditions the pseudo-conditional estimate will work. For example if $q()$ is the indicator function for a random subset of Ω (independent of $p()$ and $\Phi()$) the pseudo conditional estimate is a correct estimator. So you can't always detect this math error by testing or debugging.

Overall we do not recommend the pseudo-conditional estimate as a heuristic. If you need to estimate something that does not have an obvious change of density estimator we recommend finding a way to write it as a function of other values that do have such estimates. For example: to estimate the expected number of mortgage defaults given we have at least one mortgage default we recommend writing this as: $E[\text{number defaults}]/\text{Prob}[\text{at least one default}]$ and plugging in estimates for these two quantities. Of course an equation that is true for actual values may not hold when estimates are plugged, in and in some cases such an equation will algebraically reduce to the pseudo-conditional estimate; but at least you know where your sins are.

⁴ If Equation 4 could be applied to $\sum_{i=1}^m 1p(x_i)/q(x_i)$ we would have the expectation is indeed m . But in this case we do not meet the conditions required to apply the change as $q(x) = 0$ does not imply $1p(x) = 0$.

3.5 Variance

Up until now we have given importance sampling a free ride. We have praised it and criticized competing estimates. What we need to correctly state the conditions required to make the importance sample estimate usable: low variance. We have proven that the importance sample (despite its funny looking form) is correct in expectation. But what can go wrong is that our empirical sample may not be a good estimate of the *unknown* true expected value. The quantity we need to bound in order to ensure this is: the variance $\text{Var}_{x \sim q()}[\Phi(x)p(x)/q(x)]$ defined as

$$\mathbb{E}_{x \sim q()}[(\Phi(x)p(x)/q(x) - \mathbb{E}_{y \sim q()}[\Phi(y)p(y)/q(y)])^2].$$

We can draw a larger number of $x_i \sim q()$ and get an empirical estimate of the variance. But we have the usual circularity we find in statistics- our empirical estimate of variance is not reliable unless we have a prior bound on the *unknown* true variance.

Un-experienced variance is a big danger. Suppose there is an $x \in \Omega$ with $0 < q(x) = \epsilon \ll 1$, $\Phi(x) = 1$ and $1 \geq p(x) \gg \epsilon^2/2$. This x would be generated with probability ϵ and contributes a term of $\epsilon(p(x)/\epsilon - \mathbb{E}_{y \sim q()}[\Phi(y)p(y)/q(y)])^2$ to the true variance. Since we have assumed enough to say $\mathbb{E}_{y \sim q()}[\Phi(y)p(y)/q(y)] = p(\Phi()) \leq 1$ and $p(x)/\epsilon \gg 2$ we have: $\text{Var}_{x \sim q()}[\Phi(x)p(x)/q(x)] \geq p(x)^2/(4\epsilon)$. This is doubly bad. If ϵ is sufficiently small we will be unlikely to see these ϵ -rare events, yet they can be driving the true variance arbitrarily high (making the empirical estimate un-detectably unreliable).

In practice we sometimes do see terms with $p(x)/q(x) \gg m$ when estimating $p(\Phi())$ (yielding silly estimates of $p(\Phi()) > 1$). We can't just throw these terms out as that would then down-bias our estimate.

Overall we need to control variance. This is one of the reasons that importance sampling is under-used in theoretical computer science. Without prior bounds on variance it is hard to prove things about the estimator. But the method has a long history. Some notable uses of importance sampling in theoretical computer science are:

- Knuth's estimation of the size of search trees.[Knu74]
- The polynomial time integration and sampling techniques of Applegate, Dyer, Frieze, Kannan, and Lovasz.[KLS97, AK91, DFK91]
- Lower bound techniques in circuit theory.[AS92]

3.6 Picking $q()$

Of course the new concern is how to pick a new $q()$. In some sense the perfect $q()$ is of the form $q(x) = p(x)\Phi(x) / \int_{x \in \Omega} p(x)\Phi(x)dx$. With this $q()$ a sample of size 1 always perfectly estimates $p(\Phi())$. Of course the denominator of this perfect $q()$ is exactly the quantity we are trying to estimate (so if we knew that value we would not even have to work the problem, this is common in probability theory; if you had access to the so called "partition function" you would have everything). But let us (as in [dBKMR05]) read off some of the desirable properties the perfect $q()$ has so we can see if we can practically design for some of these properties.

1. $q(x)$ is zero where $\Phi(x)$ is zero.

We can't always hope to replicate this, but the idea is $q()$ can afford to ignore any x with $\Phi(x) = 0$. The more concentrated we are on the rare events where $\Phi(x) = 1$ we are interested in, the better.

2. $q(x)$ is non-vanishing where $\Phi(x)$ is non-zero.

This is doubly important. This was used in our proof of the change of density theorem. And, without an ϵ such that $q(x) \geq \epsilon$ when $\Phi(x) = 1$ we have difficulty bounding the variance of our estimator.

3. $p(x)/q(x)$ is a constant where $q(x)$ non-zero.

This is interesting in two ways. First we are again minimizing variance (constant functions having the least variance). Second this is the basis for the Cross Entropy Method which says picking $q()$ (from some implementable parameterized family) to maximize

$\int_{x \in \Omega, \Phi(x)=1} p(x) \ln(q(x)) dx$ is a good idea (and such a $q()$ minimizes the cross entropy [CT91] between $q(x)$ and $p(x)\Phi(x)$, or makes $q(x)$ as similar to $p(x)\Phi(x)$ as we can given our degrees of freedom). This integral is not usually available so the cross entropy method picks a $q()$ maximizing the observed empirical estimate:

$$\frac{1}{m} \sum_{\substack{i=1 \\ x_i \sim r(), \Phi(x_i)=1}}^m (p(x_i)/r(x_i)) \ln(q(x_i)) \quad (5)$$

where $r()$ is our previous best estimate for $q()$ and we have yet again applied Equation 4 to get a correct change of density.

We leave many of the details of picking $q()$ to [dBKMR05]. We will however, continue on our specific example problem using these techniques in a later section. For our example we will not explicitly worry about minimizing the empirical entropy (Equation 5) but just pick $q()$ with most of the properties of that solution.⁵

So a good $q()$ to sample from has:

1. Margins matching the summary of rare events found in the last sample (last sample is m -points generated $\sim r()$):

$$E_{x \sim q()} [x] \approx \frac{1}{m} \sum_{\substack{i=1 \\ x_i \sim r(), \Phi(x_i)=1}}^m x_i p(x_i)/r(x_i)$$

2. Maximal entropy: in this case $\sigma = 1$ and the simulated coordinate distributions independent of each other.

It depends on the nature of the problem how rich you want to make the modeling possibilities for $q()$ and how long you want to work on cross entropy concerns (versus matching margins while maintaining high entropy). What we have had good luck with is matching margins (getting moments/expectations right) and maximizing entropy. That matching margins is essentially the bulk of both logistic regression and maximum entropy modeling is touched on in both:

<http://www.win-vector.com/blog/2011/09/the-simpler-derivation-of-logistic-regression/> ([link](#)) and

<http://www.win-vector.com/blog/2011/09/the-equivalence-of-logistic-regression-and-maximum-entropy-models/> ([link](#)).

We find that if we alter $\Phi(x)$ to return values other than than 0 and 1 (for example to count the number of defaults instead of indicating just the presence of one mortgage default, or even probabilities to indicate being nearer to default) we can more quickly find a good $q()$. `score()` in our worked example (Section 4) is such a $\Phi()$.

4 Worked Example

We return to our simplistic example of a portfolio of four mortgages. The family of distributions we chose to use to model were log-normal distributions with σ fixed at 1 (maximal entropy for a given variance). The technique is essentially that of [dBKMR05] except we alter the bias schedule. For each portfolio simulated we compute the `score()` of the portfolio as the number of defaults (ranging from 0 to 4). Our sample space Ω is \mathbb{R}^{+5} (the first four coordinates being the unique incomes and the last one being the common income).

⁵These conditions are often enough to in fact impose the correct minimal cross-entropy solution.

4.1 The Steps

We start with our simulation distribution $r()$ equal to the true distribution $p()$ (where defaults are, as we mentioned, rare). We then simulate 1000 formations of the portfolio of four mortgages, draw detailed incomes and see who defaults to compute $\text{score}()$.

We generate samples $x_1, \dots, x_{1000} \sim r()$ (all in \mathbb{R}^5). Using these samples we estimate new margins that are biased towards default:

$$\bar{x} = \frac{\sum_{\substack{i=1 \\ \text{score}(x_i) \geq \gamma}}^{1000} x_i p(x_i)/r(x_i)}{\sum_{\substack{i=1 \\ \text{score}(x_i) \geq \gamma}}^{1000} 1}$$

where γ is the minimum $\text{score}()$ we consider interesting (in this case we start with $\gamma = 1$ but we could raise it as we progress).⁶ This can be recognized as the change and density estimate using $q()$ to estimate the expectations under $p()$ of the five income distributions conditioned on $\text{score}(x_i) \geq \gamma$ (i.e. what types of incomes are seen if there is at least one default).

We then define a slowed-down estimate \tilde{x} :

$$\hat{x} = 0.1 \times \bar{x} + 0.9 \times \tilde{x}$$

where \tilde{x} is our last set of margins (so we are moving our estimates a bit slower than what we see empirically).

We then use these estimated margins to build a new distribution biased towards default. We pick our next distribution $q()$ to be such that

$$E_{x \sim q()}[x] = \hat{x}.$$

So we pick $q()$ to have expectations more like as we have seen in default situations⁷ In our case we update our log-normal model keeping coordinates independent, $\sigma = 1$ and just matching the means to \hat{x}). This new $q()$ is more biased to low-income situations that lead to defaults. Defaults tend to happen when one of the first four coordinates (the individual incomes) are low (so these means are driven down in the re-estimate) and defaults happen even more often when the common income is low (so this mean is driven down faster in the re-estimate).

We iterate by setting $r() = q()$, drawing a new sample and then re-fitting $q()$. After a few iterations we stop and then use our final $q()$ and Equation 3 to generate any estimates we are interested in. In our case we stop iteration once 25% of the samples we are working with show a default. The reason we slow down the update is we are worried about empirical accidents early in

⁶Notice this is not the maligned pseudo-conditional estimate:

$$\bar{x} = \frac{\sum_{\substack{i=1 \\ \text{score}(x_i) \geq \gamma}}^{1000} x_i p(x_i)/r(x_i)}{\sum_{\substack{i=1 \\ \text{score}(x_i) \geq \gamma}}^{1000} p(x_i)/r(x_i)}.$$

We have actually been experimenting with an altered change of density estimate with the additional bias that we assume each observation x_i with $\text{score}(x_i) \geq \gamma$ is *replicated* $\text{score}(x_i)$ times in our sample:

$$\bar{x} = \frac{\sum_{\substack{i=1 \\ \text{score}(x_i) \geq \gamma}}^{1000} x_i \text{score}(x_i) p(x_i)/r(x_i)}{\sum_{\substack{i=1 \\ \text{score}(x_i) \geq \gamma}}^{1000} \text{score}(x_i)}.$$

Either way it is good to have a solid idea what probability model your estimate is simulating.

⁷We generate the samples according to $r()$ as in later rounds of simulation $r()$ will give us easier access to defaults. And we use the change of density theorem to weight these defaults according to $p()$ so that they are more representative of our original problem.

our simulation introducing unwanted bias into later samples. The reason we stop updating before we go too far is that we are worried about variance in the $p(x_i)/r(x_i)$ terms. This variance starts at 0 when $p(x_i)/r(x_i) = 1$ (the initial round where $r() = p()$) and we want to not move too fast or too far way from this ratio being a constant.⁸

Problems the analyst needs to solve to use the cross entropy version of importance sampling include:

1. Picking the parameterized probability model to work with (we strongly suggest independent variables with a high level of detail (like our example of generating five incomes that are independent instead of the formally equivalent model of generating four correlated incomes).
2. Implementing the score function $\text{score}()$ (and its possible extension beyond returning only 0 or 1).
3. Picking the minimum interesting score γ (and its possible updates between simulation rounds).
4. Picking procedure to update the $q()$ distribution (be it solving cross entropy, or just matching means).
5. Picking the slow-down rate to mix new estimated means into old means.
6. Picking the stopping condition (how prevalent we will allow the rare event to become in our final $q()$ distribution).

But trying variations of these settings is a *very* small price to pay to get a systematic method to pick new $q()$ densities for us.

4.2 The Steps Applied to Our Example

Explained again, with numbers: we start our probability model with $r() = p()$ with all five log-normal distributions set with expectation \$2000.

$$r() \sim [\ln(u = 7.1, \sigma = 1), \ln(u = 7.1, \sigma = 1), \ln(u = 7.1, \sigma = 1), \ln(u = 7.1, \sigma = 1), \ln(u = 7.1, \sigma = 1)]$$

(as 7.1 is the the log-normal parameter that have mean \$2000 when $\sigma = 1$). We then generate 1000 samples from this distribution and compute the pseudo-conditional expectation of the incomes weighted by the number of defaults in the portfolio. In our case the observed pseudo-conditioned incomes are far lower than the unconditioned expected values of \$2000. We get

$$\bar{x} = [1464.08, 1426.2, 768.94, 1294.96, 211.02].$$

We weight-average these with our last expectations to get new “slowed down” target margins:

$$\tilde{x} = [1946.41, 1942.62, 1876.89, 1929.5, 1821.1]$$

We then build the distribution $q()$ with these margins, set $r() = q()$ and repeat a few times. In our example we stopped when at least 25% of the samples we are working with showed at least one mortgage default (which was after 10 rounds of improvement).

Our final distribution $q()$ was:

$$q() \sim [\ln(u = 6.54, \sigma = 1), \ln(u = 6.55, \sigma = 1), \ln(u = 6.53, \sigma = 1), \ln(u = 6.57, \sigma = 1), \ln(u = 6.21, \sigma = 1)].$$

This is quite different than our original distribution $p()$ and has means:

$$[1141.89, 1157.65, 1129.18, 1174.47, 823.03]$$

⁸Note: it is the variation (seen or unseen) of this ratio that is a problem, not the size of the ratio.

which have fallen quite a way from \$2000. Notice that the common income (the last coordinate) has fallen fastest (as this drives the most defaults) and the other means stayed near each other. Our advice on how to ensure such informative behavior is to re-start a few times (and average outcomes) and to copy any symmetries of the problem into the model (for example use only one distribution for each of the individual incomes).

All of the steps outlined above are implemented in small example Java program found at <https://github.com/WinVector/Importance-Sampling> ([link](#)) looking at in particular: <https://github.com/WinVector/Importance-Sampling/blob/master/src/com/winvector/sample/example/DefaultExample.java> ([link](#)).

4.3 Results

Even though our final $q()$ is very different from the initial true distribution $p()$ we can use it to rapidly get good estimates of the structures of mortgage defaults. The facts of interest we will illustrate are:

1. The probability of at least one default in the portfolio of four mortgages ($p(\Phi())$).
2. The expected default rate in the portfolio of four mortgages (number of defaults in the portfolio divided by four and then averaged of the number of portfolios simulated).
3. The expected default rate in the portfolio of four mortgages (number of defaults in the portfolio divided by four and then averaged of the number of portfolios simulated) conditioned on at least one mortgage in the portfolio has defaulted.

Our empirical experiments are as follows:

- “Large”

We generate a collection of 10,000,000 portfolios of four mortgages from the original $p()$ distribution (each time by generating the five incomes and inspecting for defaults). We consider this a large sample and we will use the estimates from this sample as the “truth” we are trying to compare to. We do not expect this estimate to vary considerably with repetition (and we have confirmed this by re-running ten times).

- “Standard”

We generate a collection of 100 portfolios of four mortgages from the original $p()$ distribution. This is our standard or small sample (what we might do once if we did not have much time). We expect this estimate to vary a lot. We repeat this experiment 1,000 times to show how the estimate is varying per experiment.

- “ImportanceSample”

We generate a collection of 100 portfolios of four mortgages from the final biased $q()$ distribution. We expect this estimate to vary less than the standard estimate with the same sample size. We repeat this experiment 1,000 times to show how the estimate is varying per experiment.

In practice you would not run 1,000 repeats of summarizing data from collections of 100 portfolios. You would instead just summarize over all 100,000 simulated portfolios.⁹ What we are trying to show is the variation experienced in trying to use small sized collections to estimate rare events and how importance sampling improves the estimate even for small experiment sizes (like 100). So the repeating 1,000 times is not part of the described methods, it is part of the experiment to see how the methods work and how stable the methods are in practice.

⁹Though some bootstrap estimates do use repeated groups to estimate variance among other summaries.

In Figure 1 we see the estimated probabilities of having at least one of the four portfolios default. The vertical line (green, small dashes and matches legend key “Large”) illustrates the estimated correct answer of 2.7% and the two density plots (legend keys “ImportanceSample” and “Standard”) illustrate the distribution estimated probability of at least one mortgage in the portfolio defaulting in samples of size 100 repeated 1000 times. We would consider the importance sample a small improvement over the standard estimate. The importance sample variance is a bit lower and and more importantly the importance sample never estimated the probability of default as being zero. But in fact improvement in both estimates would be also attained by increasing sample size.

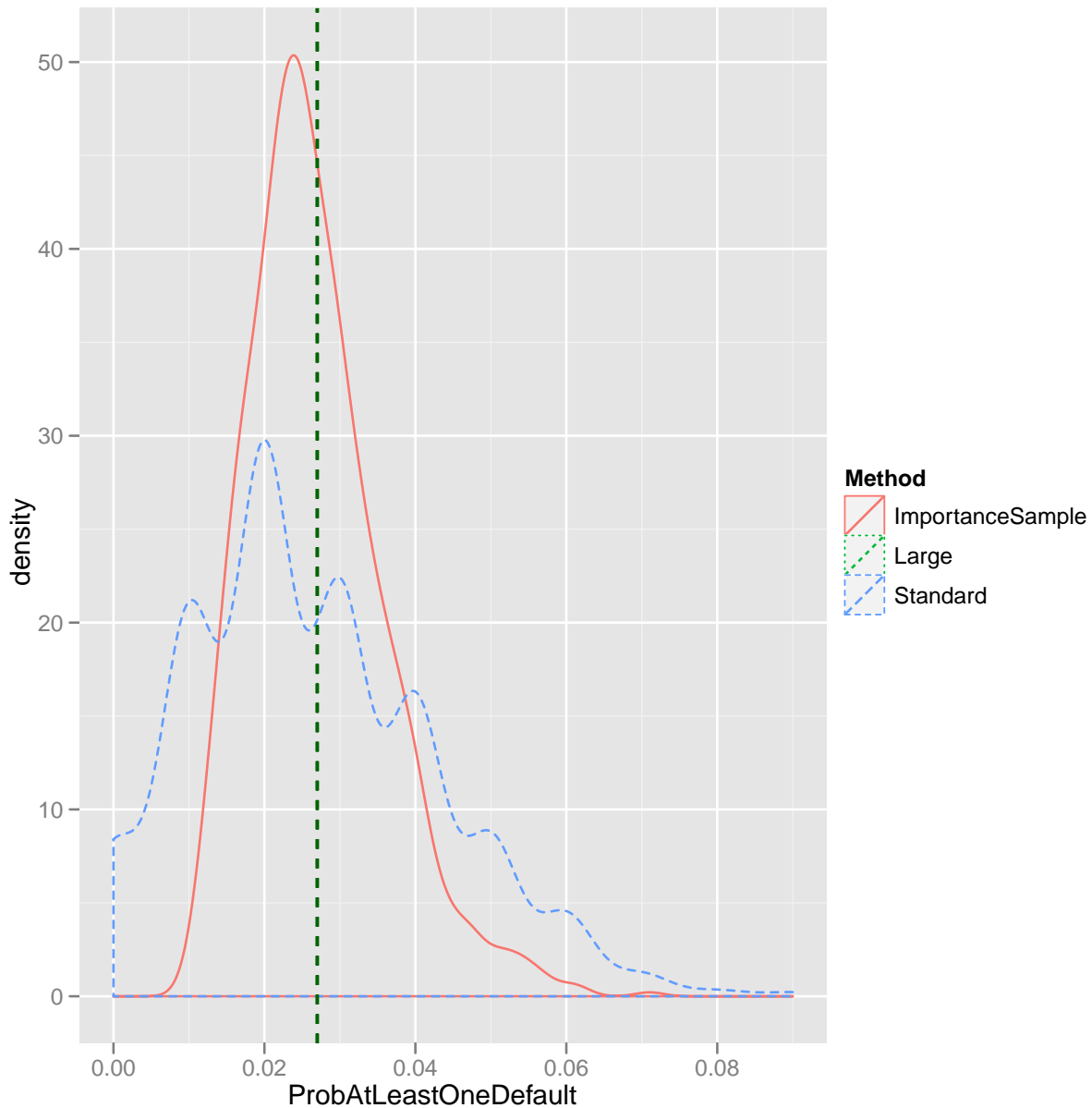


Figure 1: Estimated Probability of at Least One Mortgage Default

Figure 2 shows the estimated default rate (number of defaults/4 averaged over 100 samples in each of 1000 runs). The result is (as you would expect) similar to the last outcome.

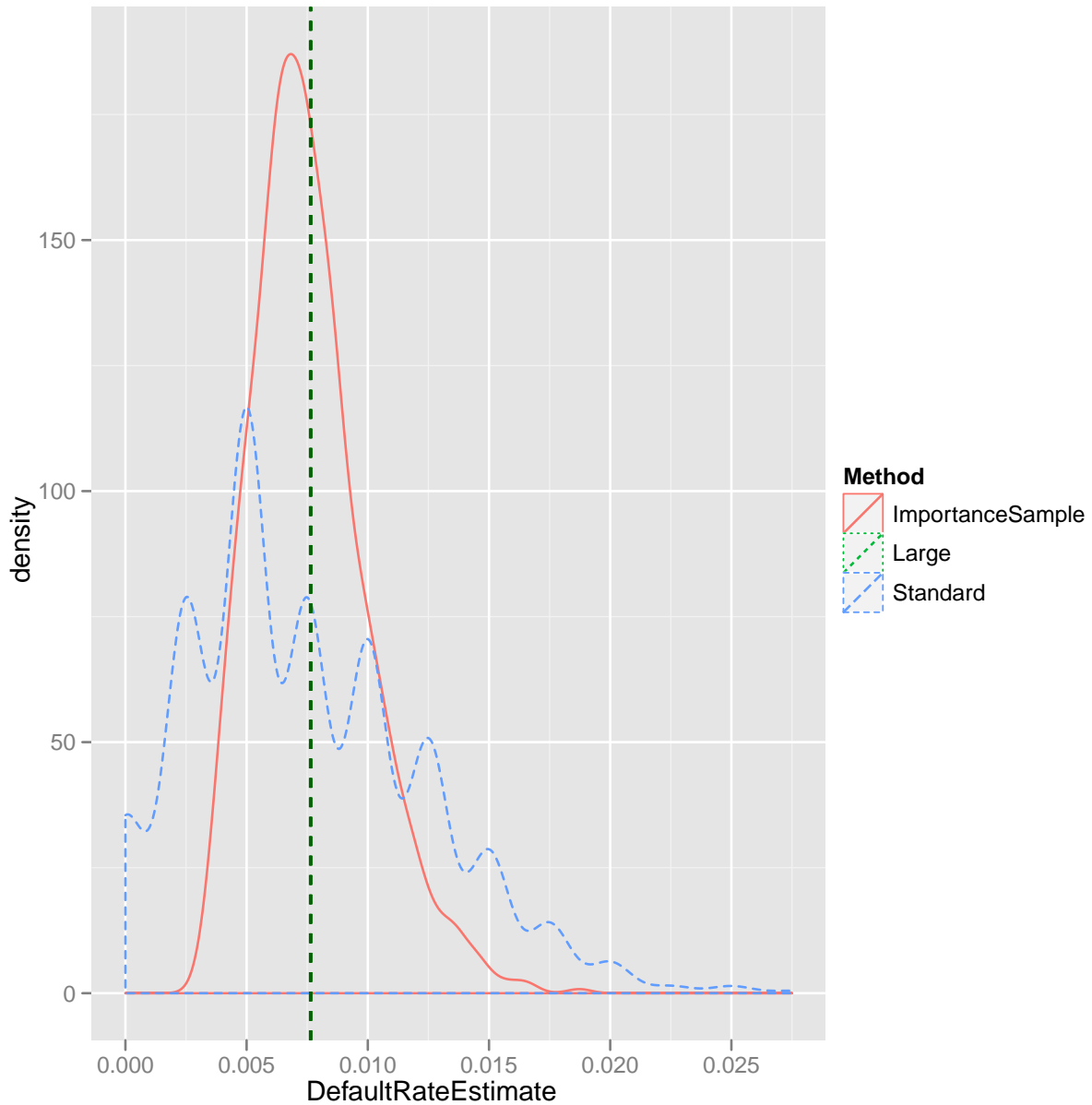


Figure 2: Estimated Default Rate

Figure 3 shows the expected number of defaults given there was at least one mortgage default. This was computed as the ratio of two other estimates: $E[\text{number defaults}] / \text{Prob}[\text{at least one mortgage default}]$. Here is where the importance sample is starting to shine- it is much better at characterizing the features present during the rare event (like the contagion or correlation between mortgage defaults). This is an improvement that would be hard to get by merely increasing sample size.

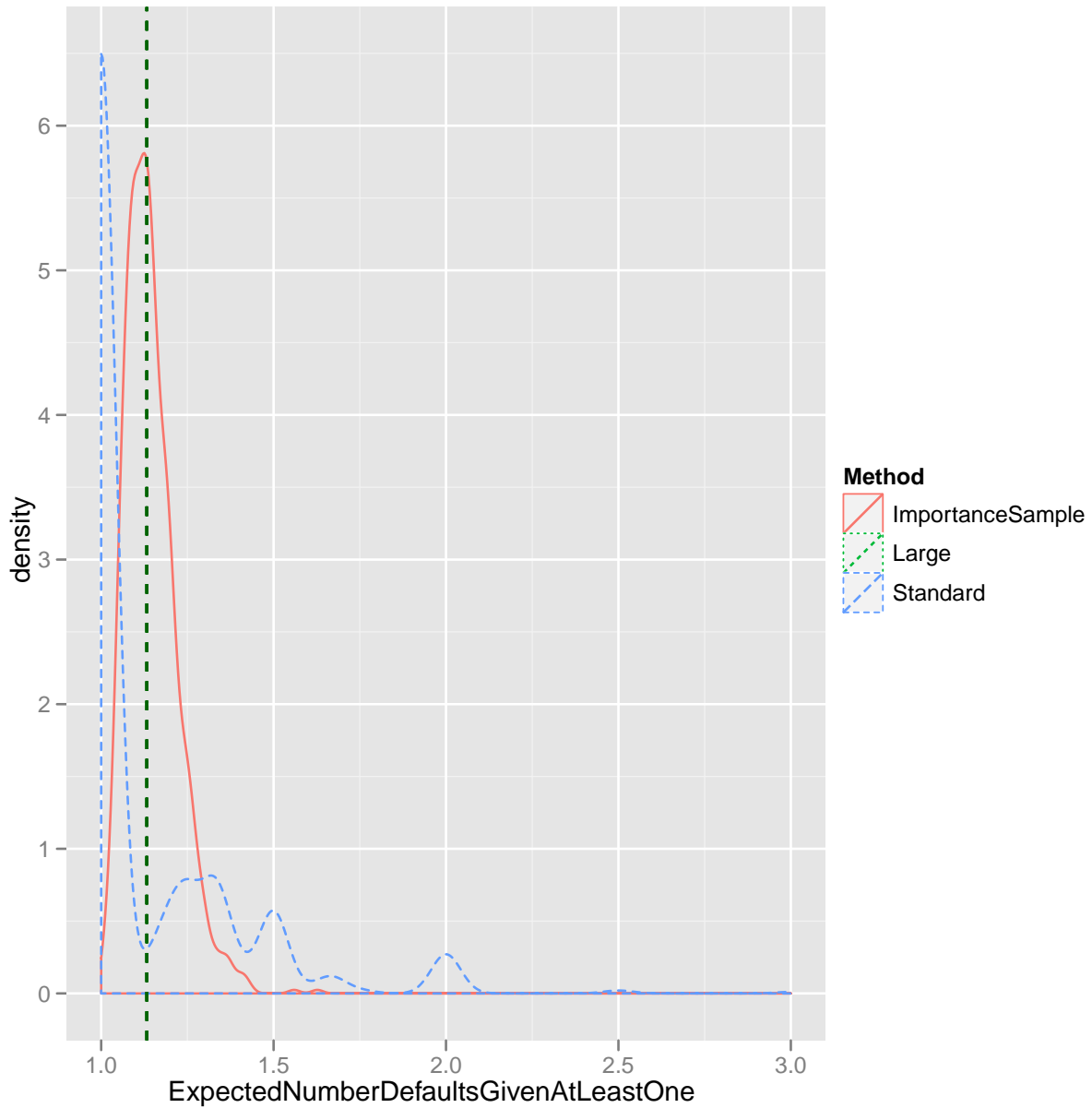


Figure 3: Estimated Number of Defaults Given at Least One Mortgage Default

Figure 4 shows how different the true (original $p()$) distribution and the final $q()$ are behaving. Under the true distribution only around 2.7% of the draws have even one default, and under the final $q()$ around 26% of the draws have at least one mortgage default (making characterizing the nature of defaults a bit easier). By “draws” we mean portfolios in our sample without regard to the weights assigned to them.

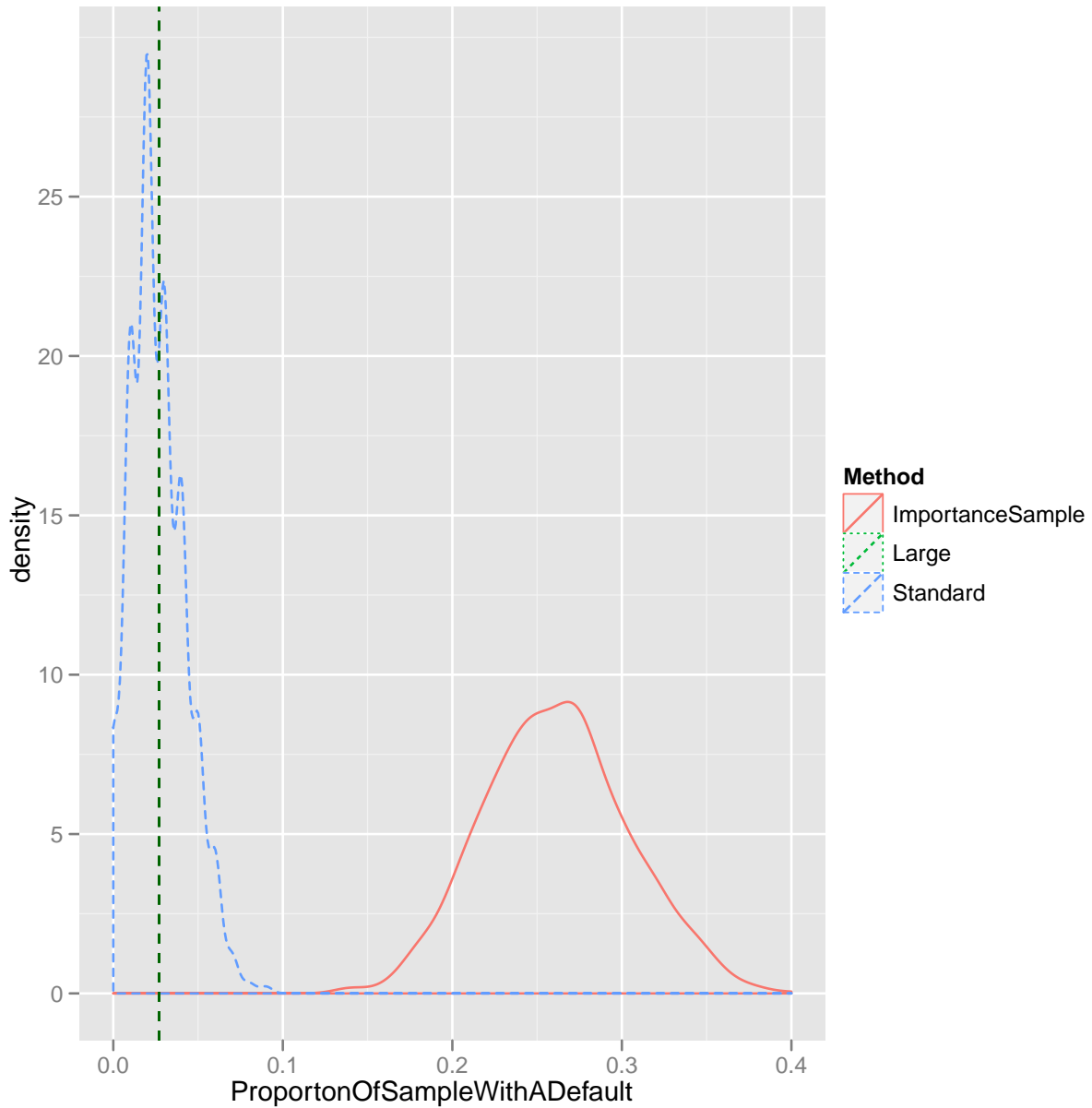


Figure 4: Proportion of Defaults in Simulation Sample

5 Conclusion

Importance sampling is nearly a turn-key improvement you can add to simulations. Some care must be taken in designing the bias for the $q()$'s, but there is good guidance on how to achieve this. Also we suggest re-simulating with several trajectories and enforcing any symmetries known to be in the problem (such as identical expected income, which we did not enforce here) to prevent the system from running to a weird particular solution subset.

The cross entropy method (an important version of importance sampling) can be thought of as a generative dual to maximum entropy modeling¹⁰. That is: it is not considered unusual to believe a

¹⁰See: <http://www.win-vector.com/blog/2011/09/the-equivalence-of-logistic-regression-and-maximum-entropy-models/> ([link](#)).

maximum entropy (i.e. flat) distribution that reproduces the marginals of a category is good at predicting the category (as this is how maximum entropy or logistic regression classifiers are built). Therefore a generator that is also maximal entropy should be considered a good simulation (even if the distribution it is simulating is in fact very different). This efficiency of simulation (that most of the information is in the marginals and maximum entropy is a good principle of least commitment) is what yields the surprising power and applicability of importance sampling.

References

- [AK91] David Applegate and Ravi Kannan, *Sampling and integration of near log-concave functions*, Proceedings of the twenty-third annual ACM symposium on Theory of computing (New York, NY, USA), STOC '91, ACM, 1991, pp. 156–163.
- [AS92] Noga Alon and Joel H. Spencer, *The probabilistic method*, Wiley, New York, 1992.
- [CT91] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, John Wiley & sons, 1991.
- [dBKMR05] Pieter-TJerk de Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein, *A Tutorial on the Cross-Entropy Method*, Annals of Operations Research **134** (2005), no. 1, 19–67.
- [DFK91] Martin Dyer, Alan Frieze, and Ravi Kannan, *A random polynomial-time algorithm for approximating the volume of convex bodies*, J. ACM **38** (1991), 1–17.
- [KLS97] Ravi Kannan, László Lovász, and Miklós Simonovits, *Random walks and an $o^*(n^5)$ volume algorithm for convex bodies*, Random Struct. Algorithms **11** (1997), 1–50.
- [Knu74] Donald E. Knuth, *Estimating the efficiency of backtrack programs.*, Tech. report, Stanford University, Stanford, CA, USA, 1974.