Independent Monte Carlo

Interested in

$$E\left[f(X)\right] = \int f(x)\,d\nu(x) = \mu_f$$

Approximate μ_f with

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

where $x_1, \ldots, x_n$ is sampled from the probability measure $\nu(X)$.

Under certain regularity conditions,

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

If $x_1, \ldots, x_n$ are an iid sample from $\nu(X)$, then

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

converges by the law of large numbers, assuming that $E\left[\left|f(X)\right|\right] < \infty$.

In addition, if

$$E\left[f(X)^2\right] = \int f(x)^2 \, d\nu(x) < \infty$$

then by the CLT

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

is approximately normally distributed, with mean $E\left[f(X)\right]$ and variance $\mathrm{Var}\left(f(X)\right)/n$.

The usual estimate of $\sigma_f^2 = \mathrm{Var}\left(f(X)\right)$ is

$$s_f^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(f(x_i) - \bar{f}\right)^2,$$

the usual unbiased estimate of $\mathrm{Var}\left(f(X)\right)$.

Example: Confidence interval properties

Want to look at properties of the normal theory 95% CI for μ

$$\bar{x} \pm t_{0.025}\frac{s}{\sqrt{m}}$$

1) Coverage probability (assuming $\mu = 0$)

$$C = E\left[I\left(\bar{x} - t_{0.025} \frac{s}{\sqrt{m}} \le 0 \le \bar{x} - t_{0.025} \frac{s}{\sqrt{m}} \right) \right]$$

$$= E\left[I\left(\frac{|\bar{x}|\sqrt{m}}{s} < t_{0.025} \right) \right]$$

2) Mean interval width

$$E[w] = E\left[2t_{0.025} \frac{s}{\sqrt{m}} \right] = \frac{2t_{0.025}}{\sqrt{m}} E[s]$$

when $m = 10$ for the following distributions

1) $N(0,1)$

2) $\text{Cauchy}(0,1)$

3) t_3

4) $U(-1,1)$

For the $N(0,1)$, it is known that the true coverage rate is 95% and the mean interval width is

$$2t_{0.025}\frac{\Gamma\left(m/2\right)}{\Gamma\left(m-1/2\right)}\frac{\sqrt{2}}{\sqrt{m(m-1)}}$$

For m = 10, the mean width is 1.391597.

Also for the Cauchy$(0,1)$, the mean interval width is ∞.

In the other cases, determining the exact values is difficult since the distributions of $\bar{x}$ and s are not tractable.

In each case, m = 1000 imputations will be generated

Estimates:

1) Coverage probability

$$\hat{C}=\frac{1}{n}\sum_{i=1}^{n}I\left(\frac{|\bar{x}_i|\sqrt{m}}{s_i}<t_{0.025}\right)$$

2) Mean interval width

$$\begin{aligned}\bar{w} &= \frac{1}{n}\sum_{i=1}^{n} w_i \\ &= \frac{2t_{0.025}}{\sqrt{m}}\frac{1}{n}\sum_{i=1}^{n} s_i \\ &= \frac{2t_{0.025}}{\sqrt{m}}\bar{s}\end{aligned}$$

Coverage Rate:

Distribution	$\hat{C}$	$SE_{\hat{C}}$
$N(0,1)$	0.954	0.0066245
Cauchy$(0,1)$	0.981	0.0043173
t_3	0.958	0.0063432
$U(-1,1)$	0.967	0.0056490

where

$$SE_{\hat{C}} = \sqrt{\frac{\hat{C}\left(1-\hat{C}\right)}{n}}$$

Mean Interval Width:

Distribution	$\bar{w}$	$SE_{\bar{w}}$
$N(0,1)$	1.372	0.010122
Cauchy$(0,1)$	34.339	9.631
t_3	2.149	0.032455
$U(-1,1)$	0.824	0.004204

The estimation errors for the $N(0,1)$ cases are

Coverage: 0.004

Mean width: 0.0193

both of which are within 2 SE's.

Cauchy example:

The assumption that $E\left[\left|f(X)\right|\right] < \infty$ is important. Since the Cauchy has no finite moments, $E[s] = \infty$, and thus the mean interval width is ∞. Thus the reported sample average and standard error is not meaningful.

However even though the mean width is not defined, the coverage rate is well defined. For every dataset, the indicator function is well defined and the integral of any indicator function of probable interest is finite.

Sample size:

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

Usual approach is by bounding the SE.

Want

$$SE \leq \frac{\sigma_f}{\sqrt{n}}$$

which gives

$$n \geq \frac{\sigma_f^2}{SE^2}$$

where SE is the desired standard error and $\sigma_f^2 = \mathrm{Var}\left(f\left(X\right)\right)$.

There is the same problem here as with determining the sample size necessary to bound the size of a confidence interval: What is σ_f^2?

Sometimes you can guess on what σ_f^2 might be.

For example, in the coverage rate case

$$\sigma_f^2 = C(1-C)$$

Since for the examples, C will be around 0.95, use that to pick n.

It can be tougher for other problems. For the width example, the question comes down to determining $\mathrm{Var}(s_i)$. While this could be done for the normal (and the Cauchy), it is tougher for the other distributions.

One approach is to do a small test sample to get a guess of σ_f^2 and use this to figure out how many more samples need to be added.

Single sample – Multiple questions

In the example, a single sample was used to answer both questions (i.e. $\bar{x}_i$ and s_i are the same in averages). I could have used these same samples to answer many more questions (e.g. $\mathrm{Var}(w)$, $E\left[\left|\bar{x}\right|\right]$, $E\left[s^2 + 4.2\sqrt{\left|m\bar{x}\right|}\right]$, etc)

When dealing with multiple quantities to be studied, you need to pick a sample size that meets the requirements for all quantities (at least the important ones).

Implementation in S-Plus/R & Matlab

When possible use vectorized calculations, not loops, particularly with S-Plus.

Vectorized	Loop

`rnorm.vec <- function(n, mu=0, sigma=1) {`
 	
`ndat<-matrix(rnorm(10*n, mu, sigma), ncol=10)`
`xbar <- apply(ndat, 1, mean)`
`s <- sqrt(apply(ndat, 1, var))`
 	
`cover <- abs(xbar) * sqrt(10) / s <= qt(0.975, 9)`
`width <- 2 * qt(0.975, 9) * s / sqrt(10)`
 	
`C <- mean(cover)`
`wbar <- mean(width)`
 	
`list(cover=cover, C=C, width=width, wbar=wbar) }`
	`rnorm.loop <- function(n, mu=0, sigma=1) {`
	
	`xbar<-rep(0,n)`
	`s <- rep(0,n)`
	`cover <- rep(0,n)`
	`width <- rep(0,n)`
	
	`for(i in 1:n) {`
	
	`  x <- rnorm(10, mu, sigma)`
	`  xbar[i] <- mean(x)`
	`  s[i] <- sqrt(var(x))`
	
	`  cover[i] <- abs(xbar[i]) * sqrt(10) / s[i] <= qt(0.975, 9)`
	`  width <- 2 * qt(0.975, 9) * s[i] / sqrt(10)`
	`}`
	
	`C <- mean(cover)`
	`wbar <- mean(width)`
	
	`list(cover=cover, C=C, width=width, wbar=wbar) }`

Run times when n = 10,000

	R	S-Plus
Vectorized	1.5 sec	35 sec
Loop	2.5 sec	48 sec
Loop/Vector	1.67	1.37

Tests done on 1.6GHz Pentium 4 running Windows XP

R version: 1.8.1

S-Plus version: 6.0 Release 2

General consensus about S-Plus would have suggested that the loop/vector ratio should have been higher with S-Plus than with R.

While I'm not sure how to quantify it with this setup, the memory use for loops is usually worse than for vectorized setups, particularly with S-Plus.

In Matlab, looping isn't as bad, though if a procedure can be done with vectorized calculations its, usually preferable.

Getting more precise estimates

1) increase n
2) different sampling scheme

Stratified Sampling

Break the sample space S into disjoint regions $S_1, \ldots, S_K$

Sample points $x_{k1}, \ldots, x_{kn_k}$ in region k

Within each region get sample average

$$\bar{f}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} f(x_{ki})$$

Then estimate μ_f by

$$\hat{\mu}_f = \sum_{k=1}^{K} P[S_k] \bar{f}_k$$

This estimator is based on the idea

$$E[f(X)] = E[E[f(X)|S_k]]$$

The variance of this estimator is

$$\mathrm{Var}(\hat{\mu}_f) = \sum_{k=1}^{K} (P[S_k])^2 \frac{\mathrm{Var}(f(X)|X \in S_k)}{n_k}$$

If the regions are picked reasonably, this will have a smaller variance than

$$\frac{\mathrm{Var}\left(f\left(X\right)\right)}{n}$$

If $n_k = nP\left[S_k\right]$ (proportional sampling), the variance of the stratified estimator reduces to

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_f\right) &= \frac{1}{n}\sum_{k=1}^{K} P\left[\mathrm{S}_k\right]\mathrm{Var}\left(f\left(X\right)\middle|X \in \mathrm{S}_k\right) \\ &= \frac{1}{n}E\left[\mathrm{Var}\left(f\left(X\right)\middle|Z\right)\right]\end{aligned}$$

where Z is a random variable satisfying $Z = k$ when the single random point drawn falls in S_k.

Since

$$\begin{aligned}\mathrm{Var}\left(f\left(X\right)\right) = E\left[\mathrm{Var}\left(f\left(X\right)\middle|Z\right)\right] \\ + \mathrm{Var}\left(E\left(f\left(X\right)\middle|Z\right)\right)\end{aligned}$$

The stratified estimator has a smaller variance that the sample average estimator.

Nonproportional sampling can give even more efficiency

The optimal sample size choices, subject to $\sum n_k = n$ is

$$n_k = n \frac{P[S_k]\sqrt{\mathrm{Var}\left(f(X)\middle|X \in S_k\right)}}{\sum_{j=1}^{K} P\left[S_j\right]\sqrt{\mathrm{Var}\left(f(X)\middle|X \in S_j\right)}}$$

This implies that regions with high variability should get more samples than regions of small variability.

Antithetic Variates

Combines 2 correlated estimators to achieve a more precise estimator.

Based on the idea

$$\mathrm{Var}\left(\frac{V+W}{2}\right) = \frac{1}{4}\mathrm{Var}(V) + \frac{1}{4}\mathrm{Var}(W) + \frac{1}{2}\mathrm{Cov}(V,W)$$

If V and W are negatively correlated, then you get a more precise estimate than if they were uncorrelated (or positively correlated).

So we need to generate coupled, negatively correlated random variables.

The following proposition gives us an approach for doing this

Proposition 21.4.1. Suppose X is a random variable and the functions $f(x)$ and $g(x)$ are both increasing or both decreasing. If the random variables $f(X)$ and $g(X)$ have finite second moments, then

$$\text{Cov}\left(f(X), g(X)\right) \geq 0$$

If $f(x)$ is increasing and $g(x)$ is decreasing (or vice-versa), then the covariance ≤ 0.

Proof (See Lange, page 291)

Suppose we wish to calculate

$$\int f(x)g(x)dx$$

where $f(x)$ is an increasing function and the density $g(x)$ has CDF $G(x)$. Then the function $f\left(G^{-1}(u)\right)$ is increasing and the function $f\left(G^{-1}(1-u)\right)$ is decreasing when $u \in [0,1]$.

If $U_1, \ldots, U_n$ is and iid sample from $U(0,1)$, then

$$E\left[f\left(G^{-1}(U)\right)\right] = E\left[f\left(G^{-1}(1-U)\right)\right]$$
$$= \int f(x)g(x)dx$$

and $f\left(G^{-1}(U)\right)$ and $f\left(G^{-1}(1-U)\right)$ are negatively correlated.

Thus the estimator

$$\frac{1}{2n}\sum_{i=1}^{n}\left\{f\left(G^{-1}(U_i)\right) + f\left(G^{-1}(1-U_i)\right)\right\}$$

has a smaller variance than

$$\frac{1}{2n}\sum_{i=1}^{2n} f\left(G^{-1}(U_i)\right)$$

The idea behind this estimator is that if $U_1,\ldots,U_n$ are uniform, so are $1-U_1,\ldots,1-U_n$. Then this implies that $G^{-1}(U_1),\ldots,G^{-1}(U_n)$ and $G^{-1}(1-U_1),\ldots,G^{-1}(1-U_n)$ are both sets of draws from $X \sim G$.

What this estimator is doing is making sure that if the p^{th} quantile is in the sample of X, so is the $1 - p^{th}$ quantile.

In a sense, its giving a more balance sampled.

Note that this also works if $f(x)$ is a decreasing function.

Example:

Let $X \sim Exp(2)$ and we want to find

$$E\left[\sqrt{X}\right] = \int 0.5\sqrt{x}e^{-x/2}dx$$
$$= \sqrt{2}\Gamma(1.5) = 1.253314$$

Generate n = 1000 values from $Exp(2)$ and use Antithetic variates

Sampler	Estimate	SE
U sample	1.239673	0.0199
1 – U sample	1.258362	0.0205
Antithetic	1.249017	0.0032

The error with antithetic estimate is -0.0043.

If a single sample of n = 2000 was taken, the standard error would be approximately 0.0143.

The gain in efficiency due to antithetic variates is approximately 20.25 (the square of the ratio of the standard errors).

To get the same efficiency out of a single sample, almost 40,000 samples would be needed.