Topic 9: The Law of Large Numbers

October, 2009

If take a sequence of random variables each independent from the other and having a common distribution $X_1, X_2, \ldots$, and plot the running average

$$\frac{1}{n}S_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

If the common mean for the X_i's is μ, then

$$E[\frac{1}{n}S_n] = \frac{1}{n}(EX_1 + EX_2 + \cdots + EX_n) = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{1}{n}n\mu = \mu.$$

If the common variance of the X_i's is σ^2, then

$$\mathrm{Var}(\frac{1}{n}S_n) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)) = \frac{1}{n^2}(\sigma + \sigma + \cdots + \sigma) = \frac{1}{n^2}n\sigma = \frac{1}{n}\sigma^2.$$

So the mean of these running averages remains at μ but the variance is inversely proportional to the number of terms in the sum.

The result is the **law of large numbers**:

For a sequence of random variables each independent from the other and having a common distribution $X_1, X_2, \ldots$,

$$\lim_{n\to\infty} \frac{1}{n}S_n$$

has a limit if and only if the X_i's have a common mean μ. In this case the limit is μ.

Example 1. *We can look an example by simulating 100 independent normal random variables, mean 68 and standard deviation 3.5. This is meant to simulate the running average of the heights of independently chosen European males.*

```
> x<-rnorm(100,68,3.5)
> s<-cumsum(x)
> plot(s/n,xlab="n",ylim=c(60,70),type="l")
```

Example 2. *We now simulate 1000 independent Cauchy random variables. These random variables have no mean as you can see that their running averages do not seem to be converging.*

```
> x<-rcauchy(1000)
> s<-cumsum(x)
> plot(s/n,xlab="n",ylim=c(-6,6),type="l")
```

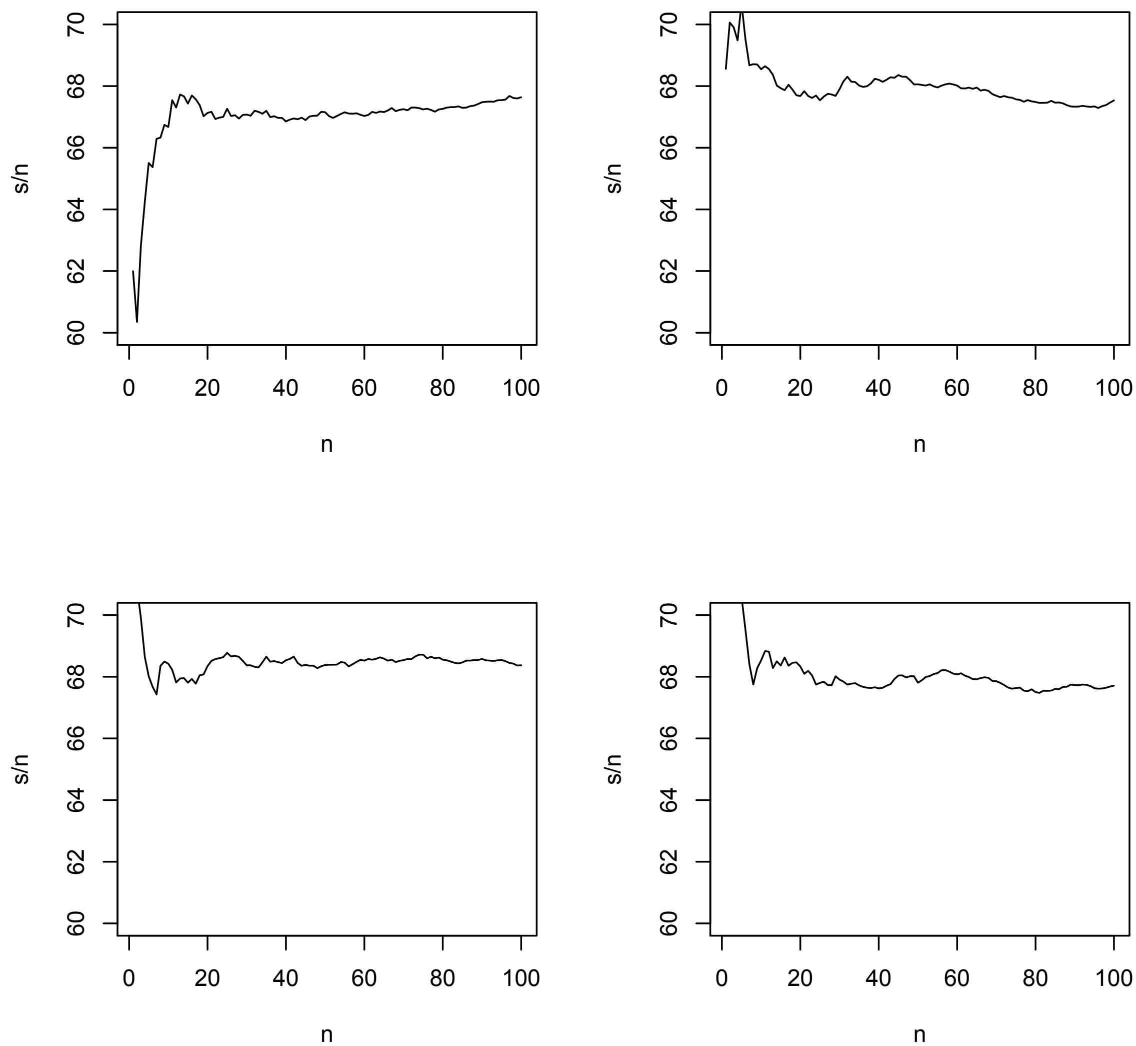

Figure 1: Four simulations of the running average S_n/n, $n = 1, 2, \ldots, 100$ for independent normal random variables, mean 68 and standard deviation 3.5.

1 Monte Carlo Integration

Monte Carlo methods use stochastic simulations to approximate solutions to questions too difficult to solve analytically.

For example, if $X_1, X_2, \ldots$ be independent random variables uniformally distributed on the interval $[0, 1]$. Then

$$\overline{g(X)}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \int_0^1 g(x)\, dx = I(g)$$

with probability 1 as $n \to \infty$.

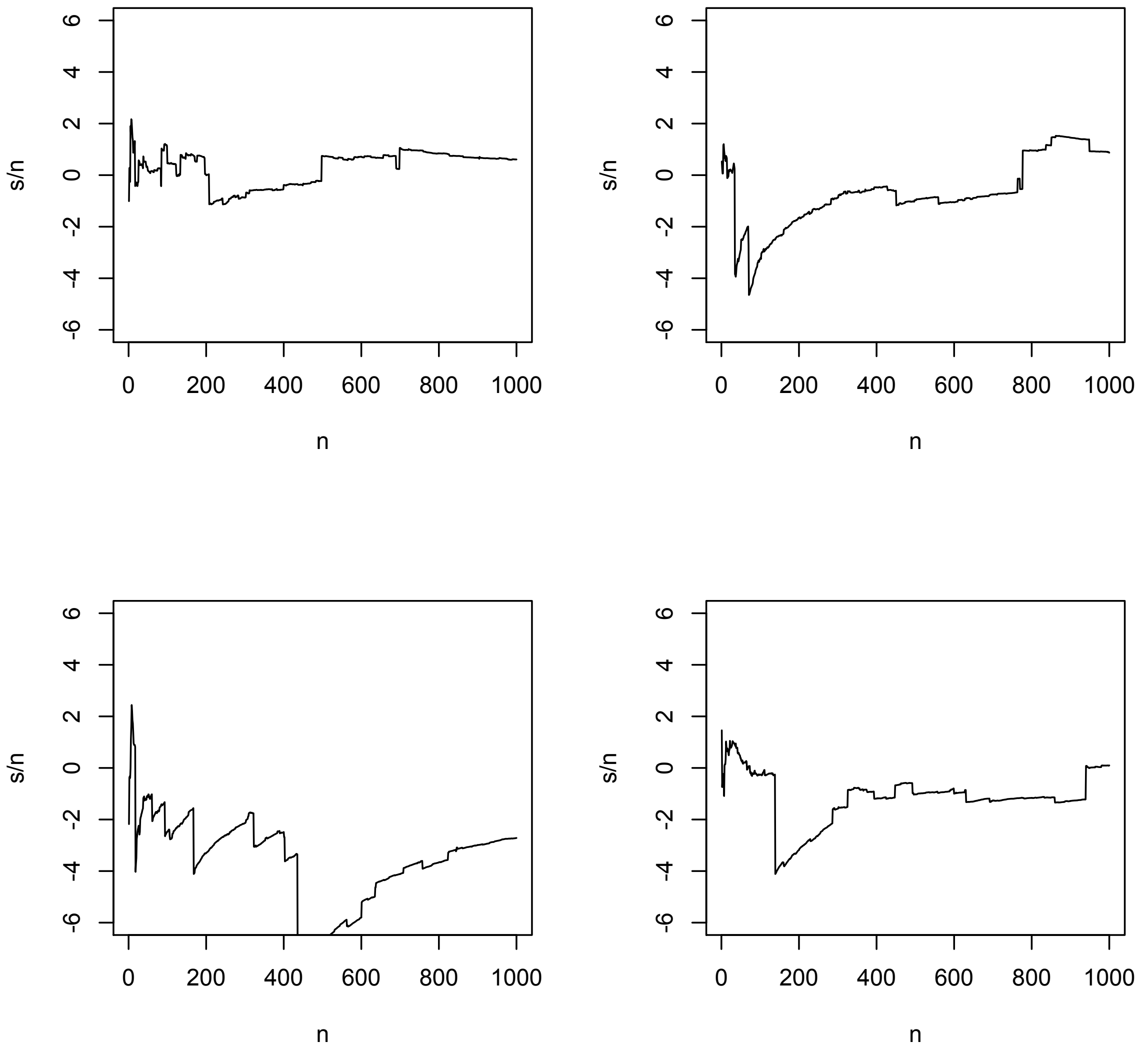

Figure 2: Four simulations of the running average S_n/n, $n = 1, 2, \ldots, 1000$ for independent Cauchy random variables. Note that the running averate does not seem to be settling down and is subject to "shocks".

Exercise 3. *Extend this idea to integrals on the interval* $[a, b]$,

The error in the estimate of the integral can be estimated by the variance

$$\text{Var}(\overline{g(X)}_n) = \frac{1}{n}\text{Var}(g(X_1)).$$

where $\sigma^2 = \text{Var}(g(X_1)) = \int_0^1 (g(x) - I(g))^2 \, dx$.

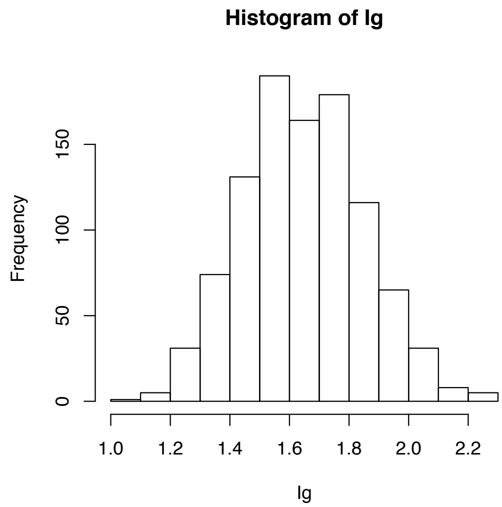

Figure 3: Histogram of 1000 Monte Carlo estimates for the integral $\int_0^1\int_0^1\int_0^1 32x^3/(y+z^4+1)\,dx\,dy\,dz$. The sample standard deviation $\sigma = 0.187$.

We can also use this to evaluate multivariate integrals. For example,

$$I(g) = \int_0^1\int_0^1\int_0^1 g(x,y,z)\,dx\,dy\,dz$$

can be estimated using Monte Carlo integration by generating three sequences of uniform random variables, $X_1, X_2, \ldots, X_n$, $Y_1, Y_2, \ldots, Y_n$, and $Z_1, Z_2, \ldots Z_n$.

Then,

$$I(g) \approx \frac{1}{n}\sum_{i=1}^n g(X_i, Y_i, Z_i).$$

```
> Ig<-rep(0,1000)
> for(i in 1:1000){x<-runif(100);y<-runif(100);z<-runif(100);g<-32*x^3/(3*(y+z^4+1));
  Ig[i]<-mean(g)}
> hist(Ig)
> summary(Ig)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.045   1.507   1.644   1.650   1.788   2.284
```

To modify this technique for a region $[a_1,b_1]\times[a_2,b_2]\times[a_3,b_3]$ use indepenent uniform random variables X_i, Y_i, and Z_i on the respective intervals, then

$$\frac{1}{n}\sum_{i=1}^n g(X_i,Y_i,Z_i) \to Eg(X_1,Y_1,Z_1) = \frac{1}{b_1-a_1}\frac{1}{b_2-a_2}\frac{1}{b_3-a_3}\int_{a_1}^{b_1}\int_{a_2}^{b_2}\int_{a_3}^{b_3} g(x,y,z)\,dz\,dy\,dx.$$

Thus, the estimate for the integral is $(b_1-a_1)(b_2-a_2)(b_3-a_3)\sum_{i=1}^n g(X_i,Y_i,Z_i)/n$.

2 Importance Sampling

Importance sampling methods begin with the observation that we could perform the Monte Carlo integration beginning with $Y_1, Y_2, \ldots$ independent random variables with common density f_Y, then define the **importance sampling**

weights

$$w(y) = \frac{g(y)}{f_Y(y)}.$$

Then

$$\overline{w(Y)}_n = \frac{1}{n}\sum_{i=1}^n w(Y_i) \to \int_{-\infty}^{\infty} w(y)f_Y(y)\,dy = \int_{-\infty}^{\infty} \frac{g(y)}{f_Y(y)} f_Y(y)\,dy = I(g).$$

This is an improvement if the variance in the estimator decreases, i.e.,

$$\int_{-\infty}^{\infty} (w(y) - I(g))^2 f_Y(y)\,dy = \sigma_f^2 << \sigma^2.$$

The density f_Y is called the **importance sampling function** or the **proposal density**.

Example 4. *For the integral*

$$\int_0^1 \frac{e^{-x/2}}{\sqrt{x(1-x)}}dx,$$

we can use Monte Carlo simulation based on uniform random variables.

```
> Ig<-rep(0,1000)
> for(i in 1:1000){x<-runif(100);g<-exp(-x/2)*1/sqrt(x*(1-x));Ig[i]<-mean(g)}
> summary(Ig)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.970   2.277   2.425   2.484   2.583   8.586
> sqrt(var(Ig))
[1] 0.3938047
```

Based on a 1000 simulations, we find a sample mean value of 2.425 and a sample standard deviation of 0.394. If we use as the proposal density a $Beta(1/2, 1/2)$*, then*

$$f_Y(y) = \frac{1}{\pi} y^{1/2-1}(1-y)^{1/2-1}$$

on the interval $[0, 1]$*. Thus the weight*

$$w(y) = \pi e^{-y/2}.$$

```
> IS<-rep(0,1000)
[1] 0.0002105915
> for(i in 1:1000){y<-rbeta(100,1/2,1/2);w<-pi*exp(-y/2);IS[i]<-mean(w)}
> summary(IS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.321   2.455   2.483   2.484   2.515   2.609
> sqrt(var(IS))
[1] 0.04377021
```

Based on a 1000 simulations, we find a sample mean value of 2.484 and a sample standard deviation of 0.044, about 1/9th the size of the Monte Carlo weright. Part of the gain is illusory. Beta random variables take longer to simulate. If they require a factor more than 81 to simulate, then the extra work needed to create a good importance sample is not helpful in producing a more accurate estimate for the integral.