MATH 598: TOPICS IN STATISTICS

IMPORTANCE SAMPLING AND VARIANCE REDUCTION

Importance Sampling: To compute the integral

$$I(g) = \int g(y)f(y)dy = \mathbb{E}_f[g(Y)]$$

where $f(y)$ is a probability density, we use the Monte Carlo (MC) estimator

$$\widehat{I}_N(g) = \frac{1}{N}\sum_{i=1}^{N} g(Y_i)$$

where $Y_1, \ldots, Y_N$ are sampled (independently) from $f(.)$. However, we may also write

$$I(g) = \int \frac{g(y)f(y)}{f_0(y)} f_0(y)dy = \mathbb{E}_{f_0}\left[\frac{g(Y)f(Y)}{f_0(Y)}\right]$$

provided that the support of $f_0(y)$ includes the support of $f(y)$. This suggests the *Importance Sampling* (IS) strategy where $Y_1, \ldots, Y_N$ are sampled independently from f_0, with the estimator

$$\widehat{I}_N^{(f_0)}(g) = \frac{1}{N}\sum_{i=1}^{N} \frac{g(Y_i)f(Y_i)}{f_0(Y_i)}.$$

The choice $f_0(y) \equiv f(y)$ for all y returns the original MC estimator. By construction, the IS estimator is unbiased for $I(g)$. We might use IS if f is hard to sample from whereas f_0 is straightforward, or if the variance of $\widehat{I}_N^{(f_0)}(g)$ is smaller than that of $\widehat{I}_N(g)$. We have that

$$\text{Var}_{f_0}\left[\widehat{I}_N^{(f_0)}(g)\right] = \frac{1}{N}\text{Var}_{f_0}\left[\frac{g(Y)f(Y)}{f_0(Y)}\right]$$

and

$$\text{Var}_{f_0}\left[\frac{g(Y)f(Y)}{f_0(Y)}\right] = \int \left(\frac{g(y)f(y)}{f_0(y)} - I(g)\right)^2 f_0(y)\, dy = \int \left\{\frac{g(y)f(y)}{f_0(y)}\right\}^2 f_0(y)\, dy - \{I(g)\}^2$$

which shows that the variance is finite if and only if

$$\int \left\{\frac{g(y)f(y)}{f_0(y)}\right\}^2 f_0(y)\, dy < \infty.$$

We have that by Jensen's inequality

$$\int \left\{\frac{g(y)f(y)}{f_0(y)}\right\}^2 f_0(y)\, dy \geq \left\{\int \frac{|g(y)|f(y)}{f_0(y)} f_0(y)\, dy\right\}^2 = \left\{\int |g(y)|f(y)\, dy\right\}^2$$

and so we have that the right-hand side is a lower bound on the left-hand side. We can make the two sides equal, and therefore achieve the lower bound by setting

$$f_0(y) = \frac{|g(y)|f(y)}{\int |g(t)|f(t)\, dt}$$

However, this is infeasible as we cannot explicitly compute the pdf in most cases. It does inform us that we should choose $f_0(y)$ to be close (in 'shape') to $|g(y)|f(y)$.

Note also that the statistic

$$\frac{1}{N}\sum_{i=1}^{N} \frac{f(Y_i)}{f_0(Y_i)} \xrightarrow{a.s.} \int \frac{f(y)}{f_0(y)} f_0(y)\, dy = 1$$

so we may instead use the estimator

$$\frac{\sum_{i=1}^{N} \frac{g(Y_i)f(Y_i)}{f_0(Y_i)}}{\sum_{i=1}^{N} \frac{f(Y_i)}{f_0(Y_i)}} = \sum_{i=1}^{N} w(Y_i)g(Y_i)$$

where

$$w(Y_i) = \frac{\frac{f(Y_i)}{f_0(Y_i)}}{\sum_{j=1}^{N} \frac{f(Y_j)}{f_0(Y_j)}} \qquad i = 1, \ldots, N.$$

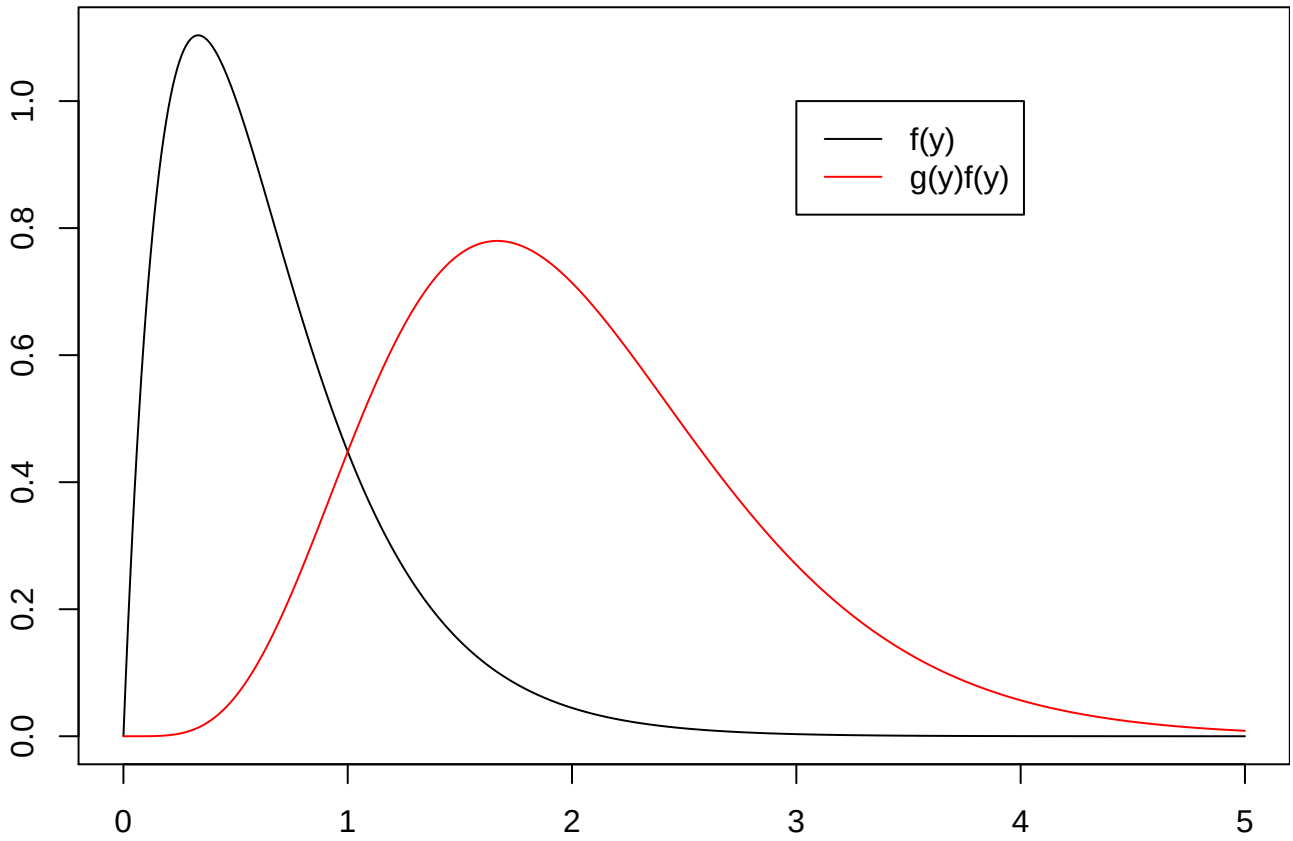
This approach can be useful if the densities $f(y)$ and $f_0(y)$ are only known up to proportionality.

EXAMPLE: Suppose $Y \sim Gamma(2, 3)$, and aim to compute $\mathbb{E}[Y^4]$. For $r > 0$, with $Y \sim Gamma(\alpha, \beta)$

$$\mathbb{E}[Y^r] = \frac{1}{\beta^r} \frac{\Gamma(\alpha + r)}{\Gamma(\alpha)}$$

so therefore $\mathbb{E}[Y^4] = (5 \times 4 \times 3 \times 2)/3^4 = 1.481481$.

```
al<-2;be<-3;r<-4
true.val<-(gamma(al+r)/gamma(al))/be^r
xv<-seq(0,5,by=0.01)
yv<-dgamma(xv,al,be)
yv4<-xv^4
par(mar=c(3,3,2,0))
plot(xv,yv,type='l',xlab='y',ylab=' ')
lines(xv,yv*yv4,col='red')
legend(3,1,c('f(y)','g(y)f(y)'),lty=1,col=c('black','red'))
```

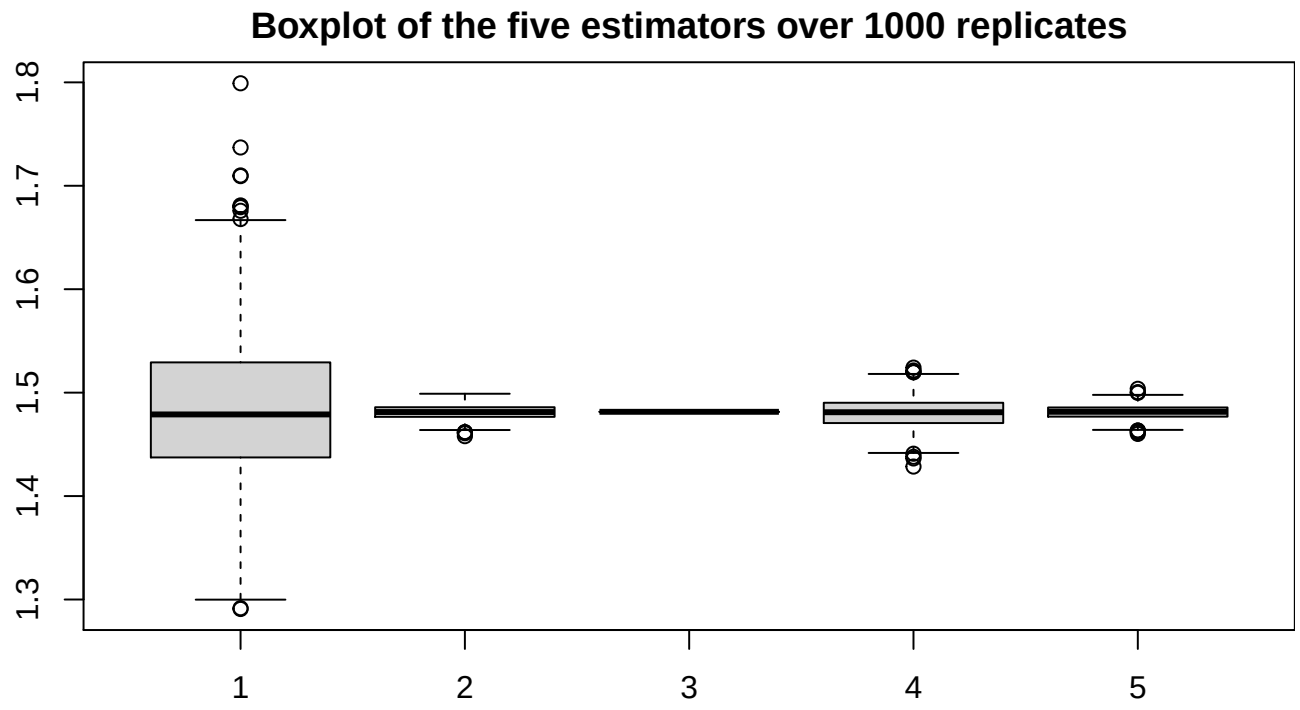


We test the importance sampling estimator with the following choices of f_0

- $f_0(y) \equiv Gamma(2,3)$ (that is, ordinary MC estimation);
- $f_0(y) \equiv Gamma(5,2)$;
- $f_0(y) \equiv Gamma(6,3)$;
- $f_0(y) \equiv Normal(2,2)$;
- $f_0(y) \equiv Student(5)$ centered at $y = 2$;

Replicate runs show the variability in each case

```
set.seed(64)
N<-10000
nreps<-1000
IS.ests<-matrix(0,nrow=nreps,ncol=5)
for(irep in 1:nreps){
    Y1<-rgamma(N,2,3)
    IS.ests[irep,1]<-mean(Y1^4)
    Y2<-rgamma(N,5,2)
    IS.ests[irep,2]<-mean(Y2^4*dgamma(Y2,2,3)/dgamma(Y2,5,2))
    Y3<-rgamma(N,6,3)
    IS.ests[irep,3]<-mean(Y3^4*dgamma(Y3,2,3)/dgamma(Y3,6,3))
    Y4<-rnorm(N,2,2)
    IS.ests[irep,4]<-mean(Y4^4*dgamma(Y4,2,3)/dnorm(Y4,2,2))
    Y5<-rt(N,df=5)+2
    IS.ests[irep,5]<-mean(Y5^4*dgamma(Y5,2,3)/dt(Y5-2,df=5))
}
par(mar=c(3,3,2,0))
boxplot(IS.ests);title('Boxplot of the five estimators over 1000 replicates')
```




```
#True.value
true.val

+ [1] 1.481481

apply(IS.ests,2,mean)    #Mean of estimator

+ [1] 1.484467 1.481253 1.481481 1.480800 1.481221

N*apply(IS.ests,2,var)   #Variance of estimator

+ [1] 50.9544051  0.4634694  0.0000000  2.1962180  0.4276574
```

Here it is evident that the IS estimators perform much better (in terms of estimator variance) than the MC estimator. Note that for the case $f_0(y) \equiv Gamma(6,3)$ we have exactly matched the integrand

$$f_0(y) \propto y^{6-1}\exp\{-3y\} = y^4 y^{2-1}\exp\{-3y\} \propto g(y)f(y).$$

Rejection Sampling: Rejection sampling is a form of direct sampling from $f(y)$ that uses an alternate distribution f_0 for sampling which can be used under the condition

$$\frac{f(y)}{f_0(y)} < M < \infty$$

for all y. To apply rejection sampling

(i) generate $Y \sim f_0(y)$ and $U \sim Uniform(0,1)$ independently

(ii) if

$$U \leq \frac{f(Y)}{Mf_0(Y)}$$

accept Y as a sample from $f(y)$; if

$$U > \frac{f(Y)}{Mf_0(Y)}$$

discard Y and return to (i).

We have that

$$\begin{aligned}
\Pr[Y \text{ is accepted}] &= \Pr\left[U \leq \frac{f(Y)}{Mf_0(Y)}\right] \\
&= \int_{-\infty}^{\infty} \left\{ \int_0^{f(y)/(Mf_0(y))} du \right\} f_0(y)\, dy \\
&= \int_{-\infty}^{\infty} \frac{f(y)}{Mf_0(y)} f_0(y)\, dy \\
&= \frac{1}{M} \int_{-\infty}^{\infty} f(y)\, dy = \frac{1}{M}
\end{aligned}$$

and then that for $y \in \mathbb{R}$,

$$\begin{aligned}
\Pr[Y \leq y | Y \text{ is accepted}] &= \frac{\Pr[Y \leq y, Y \text{ is accepted}]}{\Pr[Y \text{ is accepted}]} \\
&= M \int_{-\infty}^{y} \left\{ \int_0^{f(t)/(Mf_0(t))} du \right\} f_0(t)\, dt \\
&= \int_{-\infty}^{y} \frac{f(t)}{f_0(t)} f_0(t)\, dt \\
&= \int_{-\infty}^{y} f(t)\, dt
\end{aligned}$$

and therefore the accepted points have distribution $f(y)$.

To illustrate the use of rejection sampling, consider the following model:

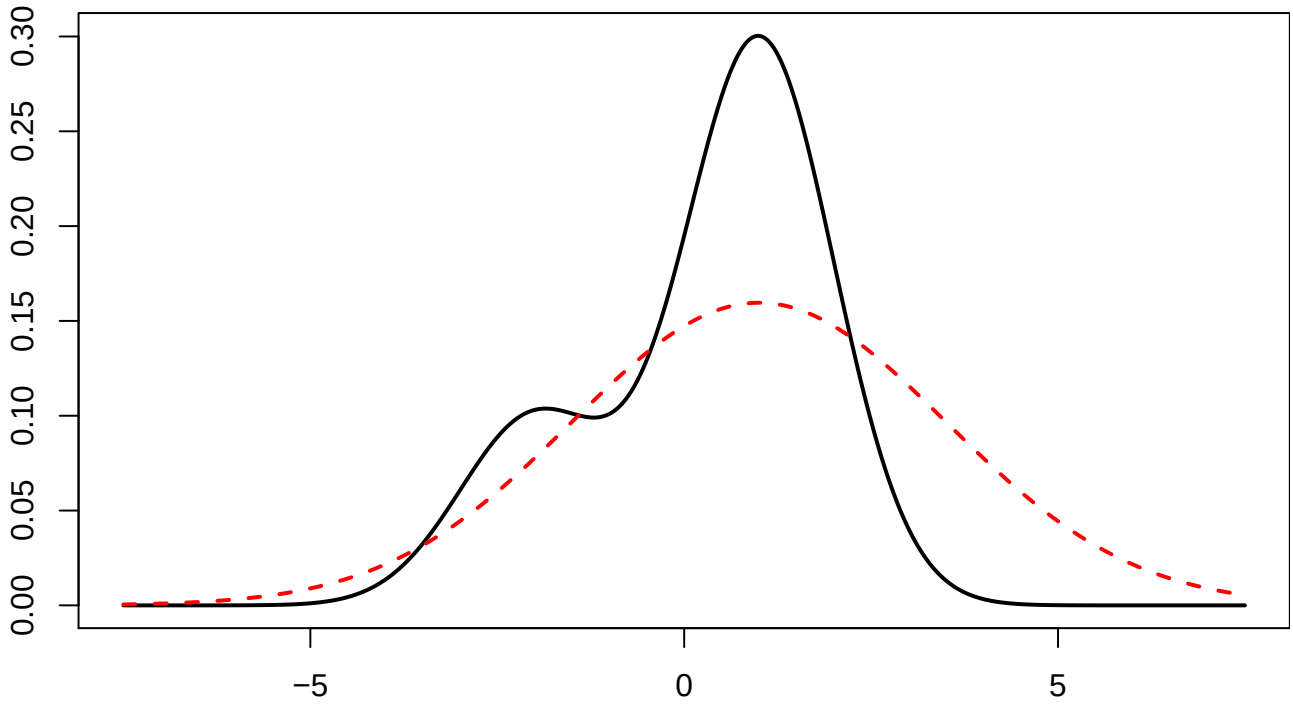
$$f(y) = \frac{1}{4}\frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right) + \frac{3}{4}\frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right)$$

that is, a mixture of two Normal densities, where $\phi(.)$ is the standard Normal pdf. In the example below

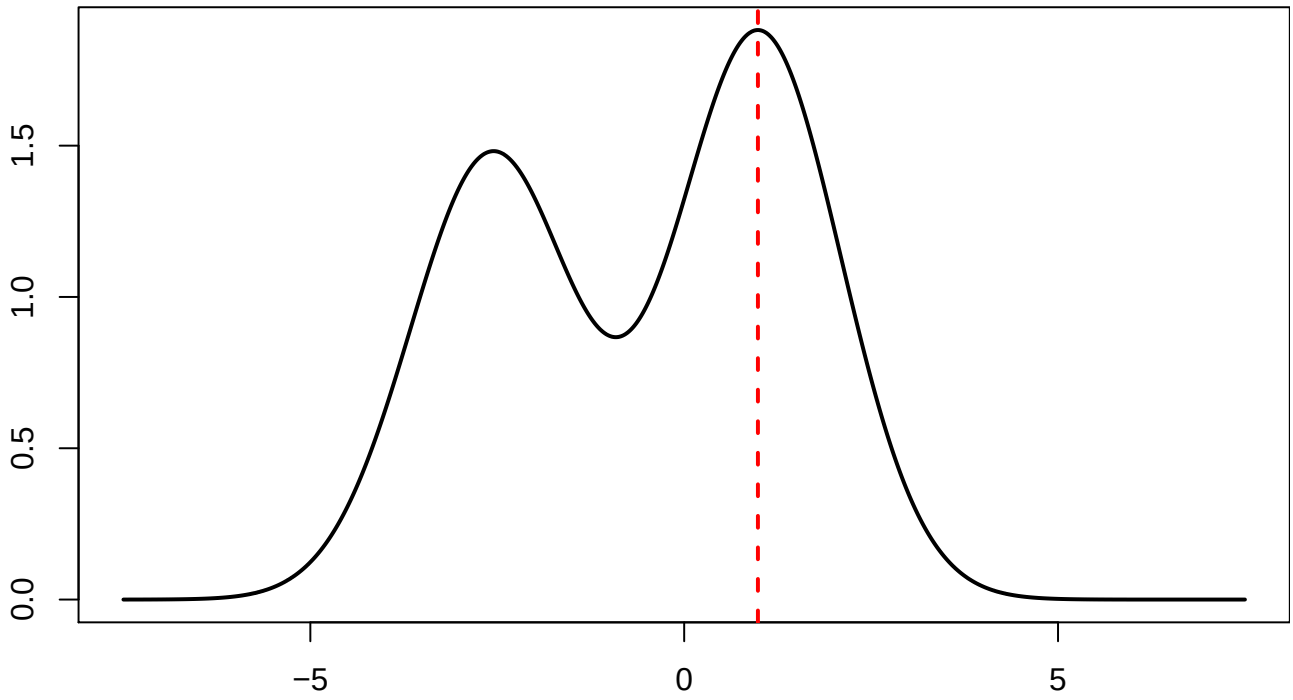
$$\mu_1 = -2 \qquad \sigma_1 = 1 \qquad \mu_2 = 1 \qquad \sigma_2 = 1.$$

For $f_0(y)$ we propose to use the $Normal(1, 2.5)$ distribution.

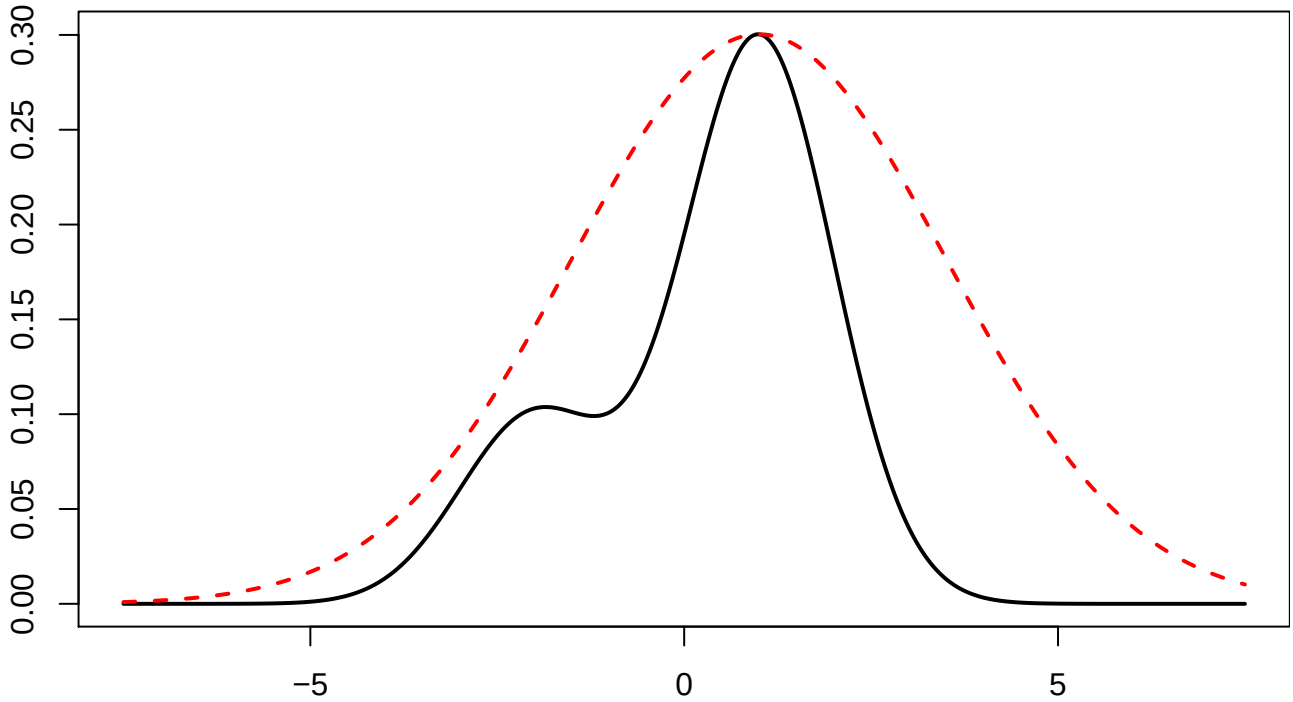
```r
xv<-seq(-7.5,7.5,by=0.001)
length(xv)

+ [1] 15001

fx<-0.25*dnorm(xv,-2,1)+0.75*dnorm(xv,1,1)
par(mar=c(3,3,2,0))
plot(xv,fx,type="l",lwd=2)
f0x<-dnorm(xv,1,2.5)
lines(xv,f0x,col="red",lwd=2,lty=2)
```



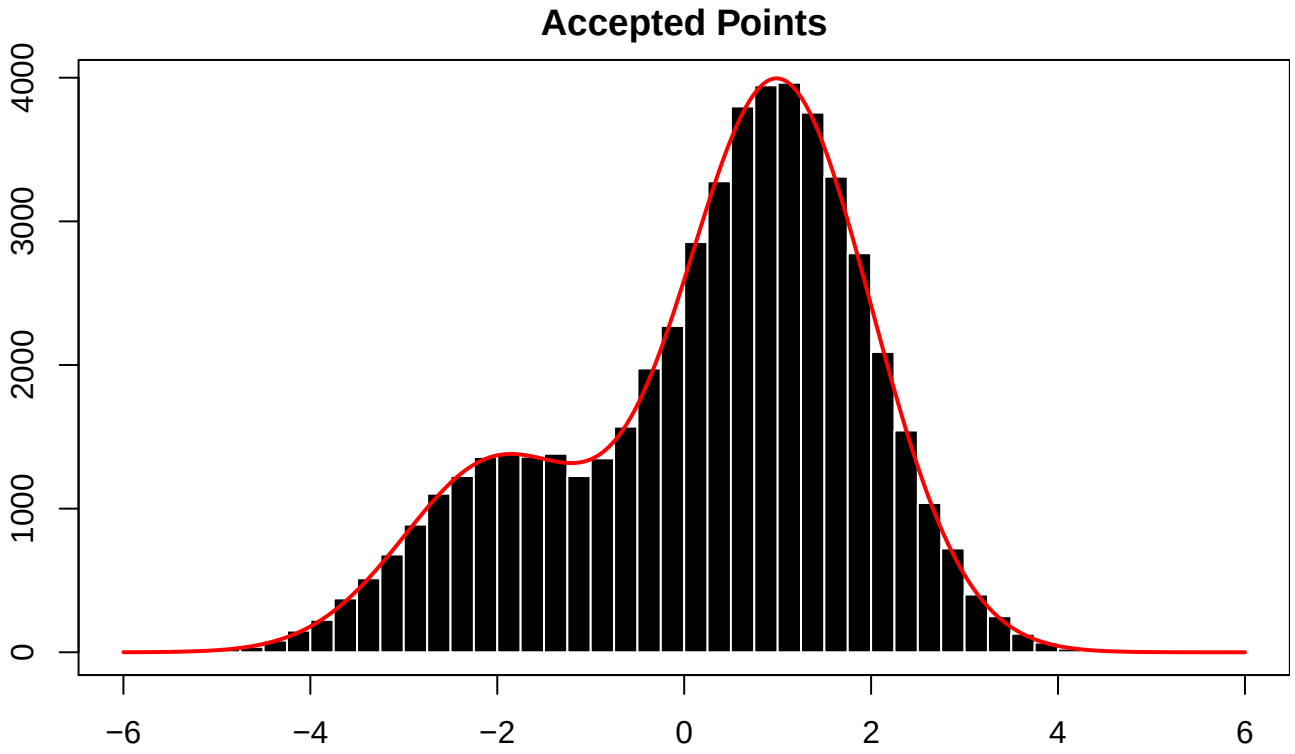

```r
f.ratio<-function(xs){
        fxv<-0.25*dnorm(xs,-2,1)+0.75*dnorm(xs,1,1)
        f0xv<-dnorm(xs,1,2.5)
        return(fxv/f0xv)
}

mval<-optim(c(0),fn=f.ratio,control=list(fnscale=-1),method="BFGS")

M<-max(fx/f0x)


plot(xv,fx/f0x,type="l",xlab="x",ylab=expression(f(x)/f[0](x)),lwd=2)
abline(v=mval$par,col="red",lty=2,lwd=2)
```




```
plot(xv,fx,type="l",lwd=2)
lines(xv,mval$val*f0x,col="red",lwd=2,lty=2)
```

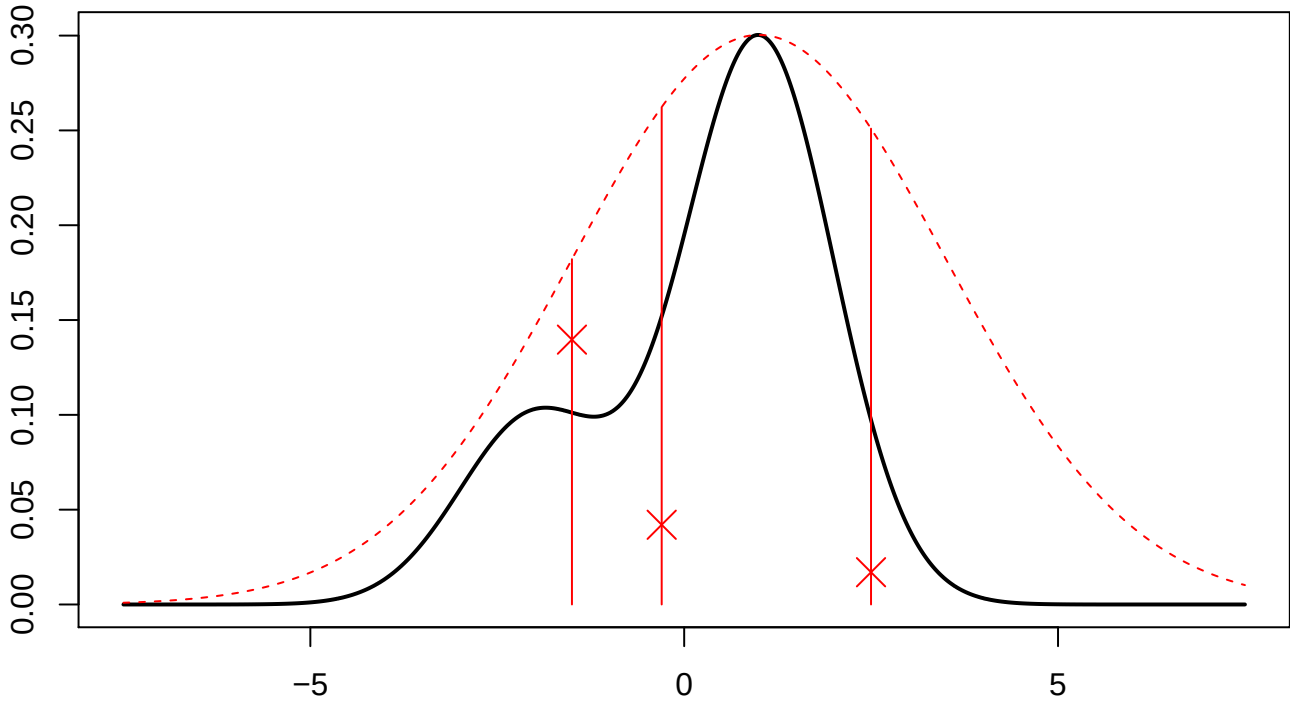


```
M<-mval$val

1/M

+ [1] 0.5313253

#####################################################
N<-100000
U<-runif(N)
Y<-rnorm(N,1,2.5)
pY<-(0.25*dnorm(Y,-2,1)+0.75*dnorm(Y,1,1))
p0Y<-dnorm(Y,1,2.5)

index<-c(1:N)
ival<-U<(pY/(M*p0Y))
Y.acc<-Y[ival]
acc.rate<-sum(ival)/N

hist(Y.acc,br=seq(-6,6,by=0.25),main="Accepted Points",col="black",border="white",xlab="x")
Xv<-seq(-6,6,by=0.01)
Yv<-(0.25*dnorm(Xv,-2,1)+0.75*dnorm(Xv,1,1))
lines(Xv,Yv*length(Y.acc)*0.25,col="red",lwd=2)
box()
```




```
plot(xv,fx,type="l",lwd=2,ylab="Density",xlab="x")
lines(xv,mval$val*f0x,col="red",lty=2)
id<-7200
xh<-xv[id]
yh<-mval$val*f0x[id]
lines(c(xh,xh),c(0,yh),col="red")
u<-runif(1)
points(xh,u*yh,col="red",pch=4,cex=2)
```

```
id<-6000
xh<-xv[id]
yh<-mval$val*f0x[id]
lines(c(xh,xh),c(0,yh),col="red")
u<-runif(1)
points(xh,u*yh,col="red",pch=4,cex=2)

id<-10000
xh<-xv[id]
yh<-mval$val*f0x[id]
lines(c(xh,xh),c(0,yh),col="red")
u<-runif(1)
points(xh,u*yh,col="red",pch=4,cex=2)
```



Antithetic Variables: The approach of antithetic variables exploits the fact that the average of two negatively correlated random variables can have a variance that is lower than the average variance of the variables. For example, if Y_1 and Y_2 have the same distribution, but are negatively correlated, then

$$\text{Var}\left[\frac{Y_1+Y_2}{2}\right] = \frac{1}{4}\text{Var}[Y_1] + \frac{1}{4}\text{Var}[Y_2] + \frac{1}{4}\text{Cov}[Y_1,Y_2] = \frac{1}{2}\text{Var}[Y_1] + \frac{1}{4}\text{Cov}[Y_1,Y_2] < \frac{1}{2}\text{Var}[Y_1].$$

Applying this principle to Monte Carlo calculations, we try to construct samples $Y_{1i},\ldots,Y_{1N}$ and $Y_{2i},\ldots,Y_{2N}$ such that $(g(Y_{1i}), g(Y_{2i}))$ are negatively correlated so that

$$\frac{1}{2N}\sum_{i=1}^{N}(g(Y_{1i}) + g(Y_{2i}))$$

has a lower variance than the ordinary Monte Carlo estimator. Consider the following example: suppose we want to compute

$$\int_0^1 (y+y^2)\,dy = \frac{5}{6}$$

that is with $g(y) = y + y^2$, using Monte Carlo by sampling independently from a $Uniform(0,1)$ distribution. Note that

$$Y_1 = Y \qquad Y_2 = 1 - Y$$

have the same distribution, and are negatively correlated, and therefore the estimator

$$\frac{1}{2N}\sum_{i=1}^{N}(g(Y_{1i}) + g(Y_{2i}))$$

should have a lower variance than the estimator

$$\frac{1}{2N}\sum_{i=1}^{2N} g(Y_i)$$

```
set.seed(64)
N<-5000
nreps<-1000
AV.ests<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
    Y<-runif(2*N)
    AV.ests[irep,1]<-mean(Y+Y^2)
    Y1<-runif(N)
    Y2<-1-Y1
    AV.ests[irep,2]<-mean(Y1+Y1^2 + Y2+Y2^2)/2
}
cor(Y1+Y1^2,Y2+Y2^2)

+ [1] -0.9680439

N*apply(AV.ests,2,var)

+ [1] 0.16309497 0.00539412

var(AV.ests[,1])/var(AV.ests[,2])

+ [1] 30.23569
```

In this example there is a 30-fold decrease in variance of the estimator.