Biostatistics 615/815 Lecture 19: Importance sampling

Hyun Min Kang

November 17th, 2011

Monte-Carlo Methods

Informal definition

- Approximation by random sampling
- Randomized algorithms to solve deterministic problems approximately.

An example problem

Calculating

$$\theta = \int_0^1 f(x)\,dx$$

where $f(x)$ is a complex function with $0 \leq f(x) \leq 1$
The problem is equivalent to computing $E[f(u)]$ where $u \sim U(0,1)$.

The crude Monte-Carlo method

Algorithm

- Generate $u_1, u_2, \cdots, u_B$ uniformly from $U(0,1)$.
- Take their average to estimate θ

$$\hat{\theta} = \frac{1}{B}\sum_{i=1}^{B} f(u_i)$$

The crude Monte-Carlo method

Algorithm

- Generate $u_1, u_2, \cdots, u_B$ uniformly from $U(0,1)$.
- Take their average to estimate θ

$$\hat{\theta} = \frac{1}{B}\sum_{i=1}^{B} f(u_i)$$

Desirable properties of Monte-Carlo methods

- Consistency : Estimates converges to true answer as B increases
- Unbiasedness : $E[\hat{\theta}] = \theta$
- Minimal Variance

Analysis of crude Monte-Carlo method

Bias

$$E[\hat{\theta}] = \frac{1}{B}\sum_{i=1}^{B} E[f(u_i)] = \frac{1}{B}\sum_{i=1}^{B} \theta = \theta$$

Analysis of crude Monte-Carlo method

Bias

$$E[\hat{\theta}] = \frac{1}{B}\sum_{i=1}^{B} E[f(u_i)] = \frac{1}{B}\sum_{i=1}^{B} \theta = \theta$$

Variance

$$\begin{aligned} \mathrm{Var}[\hat{\theta}] &= \frac{1}{B}\int_0^1 (f(u)-\theta)^2 du \\ &= \frac{1}{B}E[f(u)^2] - \frac{\theta^2}{B} \end{aligned}$$

Analysis of crude Monte-Carlo method

Bias

$$E[\hat{\theta}] = \frac{1}{B}\sum_{i=1}^{B} E[f(u_i)] = \frac{1}{B}\sum_{i=1}^{B} \theta = \theta$$

Variance

$$\begin{aligned} \mathrm{Var}[\hat{\theta}] &= \frac{1}{B}\int_0^1 (f(u)-\theta)^2 du \\ &= \frac{1}{B}E[f(u)^2] - \frac{\theta^2}{B} \end{aligned}$$

Consistency

$$\lim_{B\to\infty} \hat{\theta} = \theta$$

Accept-reject (or hit-and-miss) Monte Carlo method

Algorithm

1. Define a rectangle R between $(0,0)$ and $(1,1)$
   - Or more generally, between (x_m, x_M) and (y_m, y_M).
2. Set $h = 0$ (hit), $m = 0$ (miss).
3. Sample a random point $(x, y) \in R$.
4. If $y < f(x)$, then increase h. Otherwise, increase m
5. Repeat step 3 and 4 for B times
6. $\hat{\theta} = \frac{h}{h+m}$.

Analysis of accept-reject Monte Carlo method

Bias

Let u_i, v_i follow $U(0,1)$, then $\Pr(v_i < f(u_i)) = \theta$

$$\begin{aligned} E[\hat{\theta}] &= E\left[\frac{h}{h+m}\right] \\ &= \frac{\sum_{i=1}^{B} I(v_i < f(u_i))}{B} \\ &= \theta \end{aligned}$$

Analysis of accept-reject Monte Carlo method

Bias

Let u_i, v_i follow $U(0,1)$, then $\Pr(v_i < f(u_i)) = \theta$

$$\begin{aligned} E[\hat{\theta}] &= E\left[\frac{h}{h+m}\right] \\ &= \frac{\sum_{i=1}^{B} I(v_i < f(u_i))}{B} \\ &= \theta \end{aligned}$$

Variance

$h \sim \text{Binom}(B, \theta)$.

$$\text{Var}[\hat{\theta}] = \frac{\theta(1-\theta)}{B}$$

Which method is better?

$$\begin{aligned}
\sigma^2_{AR} - \sigma^2_{crude} &= \frac{\theta(1-\theta)}{B} - \frac{1}{B}E[f(u)^2] + \frac{\theta^2}{B} \\
&= \frac{\theta - E[f(u)]^2}{B} \\
&= \frac{1}{B}\int_0^1 f(u)(1-f(u))\,du \geq 0
\end{aligned}$$

The crude Monte-Carlo method has less variance then accept-rejection method

Revisiting The Crude Monte Carlo

$$\theta = E[f(u)] = \int_0^1 f(u)\,du$$

$$\hat{\theta} = \frac{1}{B}\sum_{i=1}^{B} f(u_i)$$

More generally, when x has pdf $p(x)$, if x_i is random variable following $p(x)$,

$$\theta_p = E_p[f(x)] = \int f(x)p(x)\,dx$$

$$\hat{\theta}_p = \frac{1}{B}\sum_{i=1}^{B} f(x_i)$$

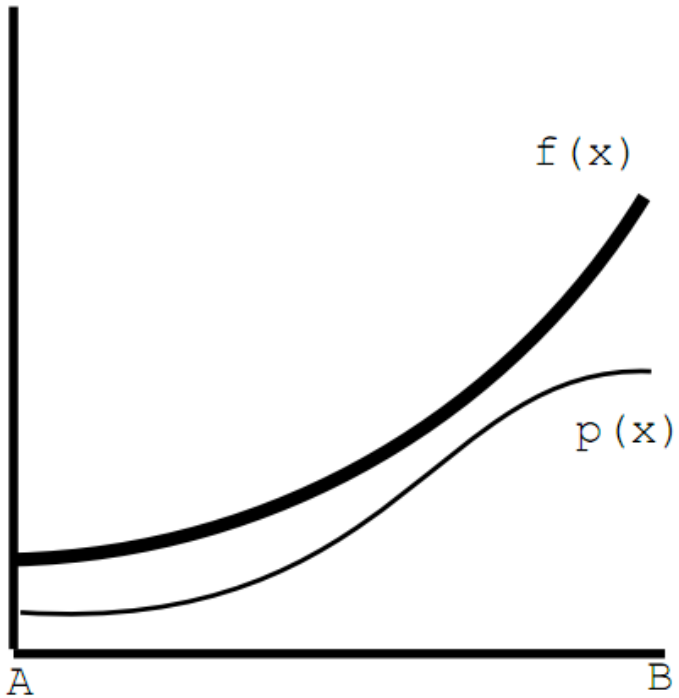
Importance sampling

Let x_i be random variable, and let $p(x)$ be an arbitrary probability density function.

$$\begin{aligned}\theta &= E_u[f(x)] = \int f(x)\,dx = \int \frac{f(x)}{p(x)} p(x)\,dx = E_p\left[\frac{f(x)}{p(x)}\right] \\ \hat{\theta} &= \frac{1}{B}\sum_{i=1}^{B} \frac{f(x_i)}{p(x_i)}\end{aligned}$$

where x_i is sampled from distribution representd by pdf $p(x)$

Key Idea



- When $f(x)$ is not uniform, variance of $\hat{\theta}$ may be large.
- The idea is to pretend sampling from (almost) uniform distribution.

Analysis of Importance Sampling

Bias

$$E[\hat{\theta}] = \frac{1}{B}\sum_{i=1}^{B} E_p\left[\frac{f(x_i)}{p(x_i)}\right] = \frac{1}{B}\sum_{i=1}^{B}\theta = \theta$$

Analysis of Importance Sampling

Bias

$$E[\hat{\theta}] = \frac{1}{B}\sum_{i=1}^{B} E_p\left[\frac{f(x_i)}{p(x_i)}\right] = \frac{1}{B}\sum_{i=1}^{B}\theta = \theta$$

Variance

$$\begin{aligned}\mathrm{Var}[\hat{\theta}] &= \frac{1}{B}\int\left(\frac{f(x)}{p(x)}-\theta\right)^2 p(x)\,dx \\ &= \frac{1}{B}E_p\left[\left(\frac{f(x)}{p(x)}\right)^2\right] - \frac{\theta^2}{B}\end{aligned}$$

The variance may or may not increase. Roughly speaking, if $p(x)$ is similar to $f(x)$, $f(x)/p(x)$ becomes flattened and will have smaller variance.

Simulation of rare events

Problem

- Consider a random variable $X \sim N(0,1)$
- What is $\Pr[X \geq 10]$?

Simulation of rare events

Problem

- Consider a random variable $X \sim N(0,1)$
- What is $\Pr[X \geq 10]$?

Possible Solutions

- Let $f(x)$ and $F(x)$ be pdf and cdf of standard normal distribution.
- Then $\Pr[X \geq 10] = 1 - F(10) = 7.62 \times 10^{-24}$, and we're all set.

Simulation of rare events

Problem

- Consider a random variable $X \sim N(0,1)$
- What is $\Pr[X \geq 10]$?

Possible Solutions

- Let $f(x)$ and $F(x)$ be pdf and cdf of standard normal distribution.
- Then $\Pr[X \geq 10] = 1 - F(10) = 7.62 \times 10^{-24}$, and we're all set.
- But what if we don't have $F(x)$ but only $f(x)$?
  - In many cases, cdf is not easy to obtain compared to pdf or random draws.

If we don't have CDF: ways to calculate $\Pr[X \geq 10]$

Accept-reject sampling

Sample random variables from $N(0,1)$, and count how many of them are greater than 10

If we don't have CDF: ways to calculate $\Pr[X \geq 10]$

Accept-reject sampling

Sample random variables from $N(0,1)$, and count how many of them are greater than 10

- How many random variables should be sampled to observe at least one $X \geq 10$?
- $1/\Pr[X \geq 10] = 1.3 \times 10^{23}$

If we don't have CDF: ways to calculate $\Pr[X \geq 10]$

Accept-reject sampling

Sample random variables from $N(0,1)$, and count how many of them are greater than 10

- How many random variables should be sampled to observe at least one $X \geq 10$?
- $1/\Pr[X \geq 10] = 1.3 \times 10^{23}$

Monte-Carlo Integration

- If we have pdf $f(x)$, $\Pr[X \geq 10] = \int_{10}^{\infty} f(x)\,dx$
- Use Monte-Carlo integration to compute this quantity
  1. Sample B values uniformly from $[10, 10+W]$ for a large value of W (e.g. 50).
  2. Estimate $\hat{\theta} = \frac{1}{B}\sum_{i=1}^{B} f(u_i)$.

An Importance Sampling Solution

1. Transform the problem into an unbounded integration problem (to make it simple)

$$\Pr[X \geq 10] = \int_{10}^{\infty} f(x)\,dx = \int I(x \geq 10) f(x)\,dx$$

2. Sample B random values from $N(\mu, 1)$ where μ is a large value nearby 10, and let $f_\mu(x)$ be the pdf.
3. Estimate the probability as an weighted average

$$\hat{\theta} = \frac{1}{B}\left[I(x_i \geq 10) \frac{f(x)}{f_\mu(x)} \right]$$

An Example R code

```
## pnormUpper() function to calculate Pr[x>t] using n random samples
pnormUpper <- function(n, t) {
  lo <- t
  hi <- t + 50   ## hi is a reasonably large number

  ## accept-reject sampling
  r <- rnorm(n)        ## random sampling from N(0,1)
  v1 <- sum(r > t)/n   ## count how many meets the condition

  ## monte-carlo integration
  u <- runif(n,lo,hi)            ## uniform sampling [t,t+50]
  v2 <- mean(dnorm(u))*(hi-lo)   ## monte-carlo integration

  ## importance sampling using N(t,1)
  g <- rnorm(n,t,1)     ## sample from N(t,1)
  v3 <- sum( (g > t) * dnorm(g)/dnorm(g,t,1)) / n;  ## take a weighted average

  return (c(v1,v2,v3))  ## return three values
}
```

Evaluating different methods

```
## test pnormUpperTest(n,t) function using r times of repetition
pnormUpperTest <- function(r, n, t) {
  gold <- pnorm(t,lower.tail=FALSE)  ## gold standard answer
  emp <- matrix(nrow=r,ncol=3)  ## matrix containing empirical answers
  for(i in 1:r) { emp[i,] <- pnormUpper(n,t) }  ## repeat r times
  m <- colMeans(emp)        ## obtain mean of the estimates
  s <- apply(emp,2,sd)      ## obtain stdev of the estimates
  print("GOLD :")
  print(gold);              ## print gold standard answer
  print("BIAS (ABSOLUTE) :")
  print(m-gold)             ## print bias
  print("STDEV (ABSOLUTE) :")
  print(s)                  ## print stdev
  print("BIAS (RELATIVE) :")
  print((m-gold)/gold)      ## print relative bias
  print("STDEV (RELATIVE) :")
  print(s/gold)             ## print relative stdev
}
```

An example output

```
> pnormUpperTest(100,1000,10)
[1] "GOLD :"
[1] 7.619853e-24
[1] "BIAS (ABSOLUTE) :"
[1] -7.619853e-24 -5.596279e-26  4.806933e-26
[1] "STDEV (ABSOLUTE) :"
[1] 0.000000e+00 3.917905e-24 7.559024e-25
[1] "BIAS (RELATIVE) :"
[1] -1.000000000 -0.007344339  0.006308433
[1] "STDEV (RELATIVE) :"
[1] 0.0000000 0.5141707 0.0992017
```

Another example output

```
> pnormUpperTest(100,10000,10)
[1] "GOLD :"
[1] 7.619853e-24
[1] "BIAS (ABSOLUTE) :"
[1] -7.619853e-24  2.202168e-26  1.972362e-26
[1] "STDEV (ABSOLUTE) :"
[1] 0.000000e+00 1.186711e-24 2.935474e-25
[1] "BIAS (RELATIVE) :"
[1] -1.000000000  0.002890040  0.002588451
[1] "STDEV (RELATIVE) :"
[1] 0.00000000 0.15573932 0.03852402
```

1,000 importance sampling gives smaller variance than monte-carlo integration with 10,000 random samples.

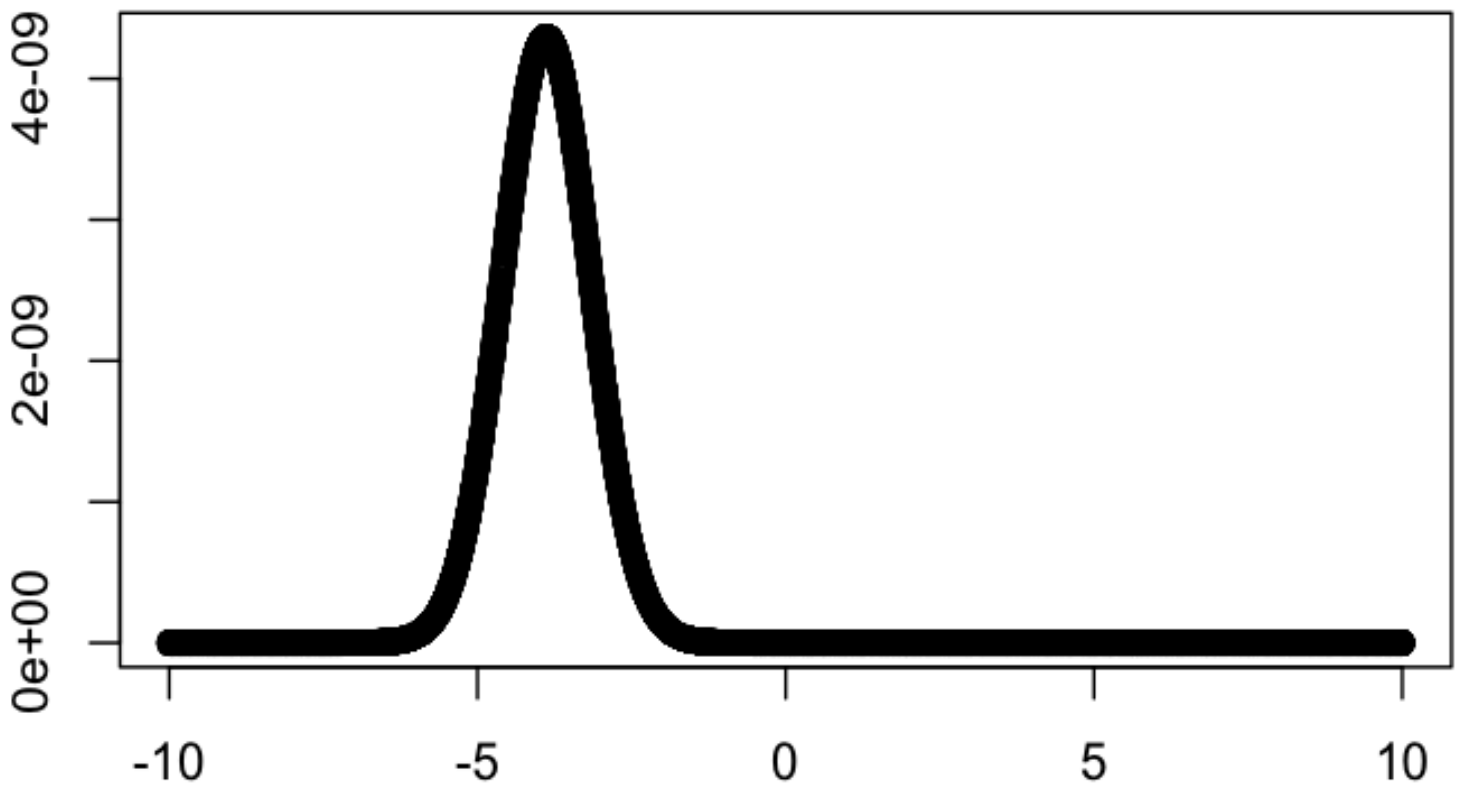
Integral of probit normal distribution

- Disease risk of an individual follows $X \sim N(\mu, \sigma^2)$.
- Probability of disease $\Pr(Y = 1) = \Phi(X)$, where $\Phi(x)$ is CDF of standard normal distribution.
- Want to compute the disease prevalence across the population.

$$\theta = \int_{-\infty}^{\infty} \Phi(x) \mathcal{N}(x; \mu, \sigma^2) \, dx$$

where $\mathcal{N}(\cdot; \mu, \sigma^2)$ is pdf of normal distribution.

Plot of $\Phi(x)\mathcal{N}(x; -8, 1^2)$



Monte-Carlo integration using uniform samples

1. Sample x uniformly from a sufficiently large interval (e.g. $[-50, 50]$).
2. Evaluate integrals using

$$\hat{\theta} = \frac{1}{B} \sum_{i=1}^{B} \Phi(x_i) \mathcal{N}(x_i; \mu, \sigma^2)$$

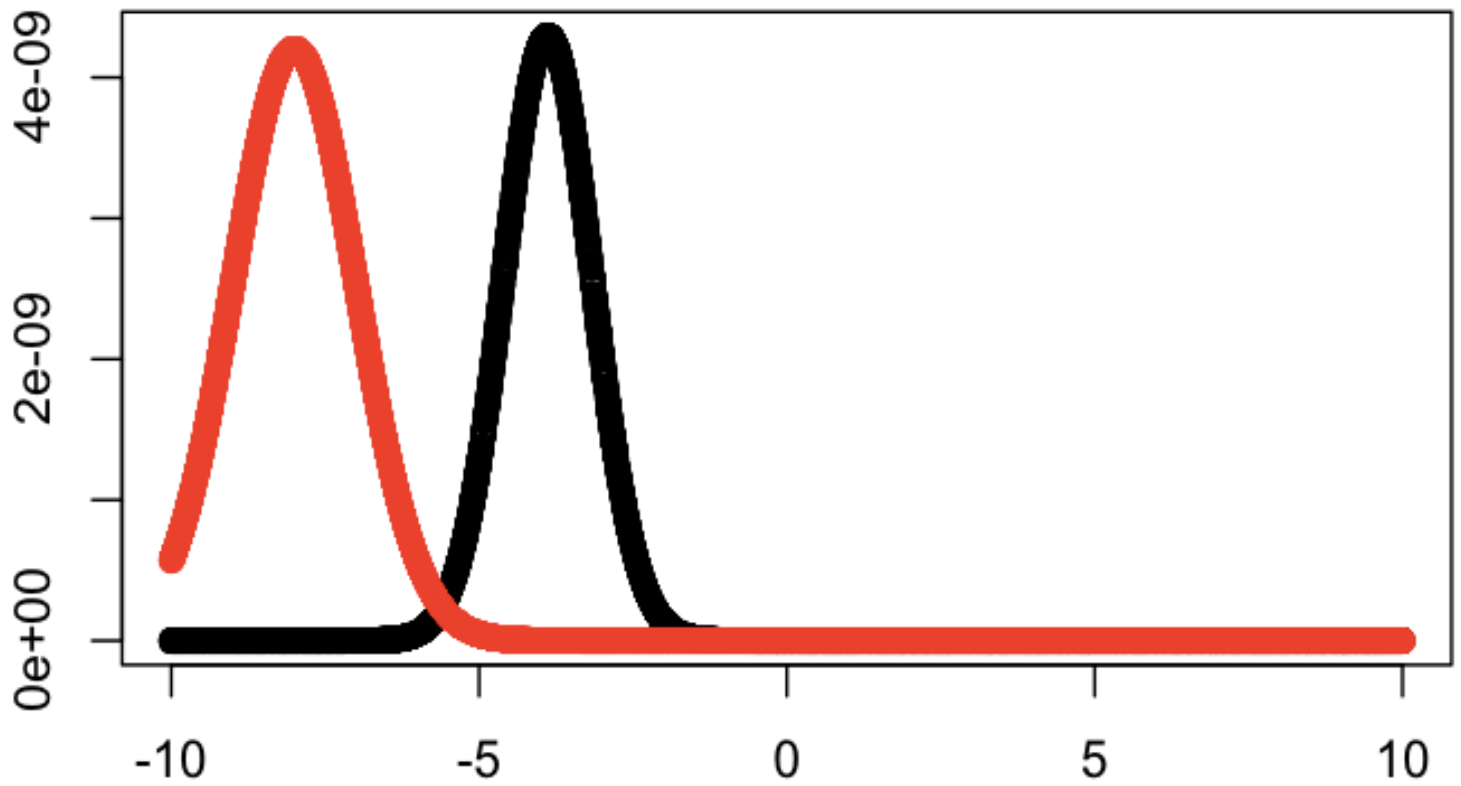
Note that, for some μ and σ^2, $[-50, 50]$ may not be a sufficiently large interval.

Monte-Carlo integration using normal distribution

1. Sample x from $N(\mu, \sigma^2)$
2. Evaluate integrals by

$$\hat{\theta} = \frac{1}{B}\sum_{i=1}^{B}\Phi(x_i)$$

$\mathcal{N}(x; -8, 1^2)$ (red) and $\Phi(x)\mathcal{N}(x; -8, 1^2)$ (black)



Two distributions are quite different – $\mathcal{N}(x; -8, 1^2)$ may not be an ideal distribution for monte-carlo integration

Monte-Carlo integration by importance sampling

1. Sample x from a new distribution
   - For example, $N(\mu', \sigma'^2)$
   - $\mu' = \frac{\mu}{\sigma^2+1}$
   - $\sigma' = \sigma$.
2. Evaluate integrals by weighting importance samples

$$\hat{\theta} = \frac{1}{B}\sum_{i=1}^{B}\left[\Phi(x_i)\frac{\mathcal{N}(x;\mu,\sigma^2)}{\mathcal{N}(x;\mu',\sigma'^2)}\right]$$

An Example R code

```
probitNormIntegral <- function(n,mu,sigma) {
  ## integration across uniform distribution
  lo <- -50
  hi <- 50
  u <- runif(n,lo,hi)
  v1 <- mean(dnorm(u,mu,sigma)*pnorm(u))*(hi-lo)

  ## integration using random samples from N(mu,sigma^2)
  g <- rnorm(n,mu,sigma)
  v2 <- mean(pnorm(g))

  ## importance sampling using N(mu',sigma^2)
  adjm <- mu/(sigma^2+1)
  r <- rnorm(n,adjm,sigma)
  v3 <- mean(pnorm(r)*dnorm(r,mu,sigma)/dnorm(r,adjm,sigma))
  return (c(v1,v2,v3))
}
```

Testing different methods

```
probitNormTest <- function(r, n, mu,sigma) {
  emp <- matrix(nrow=r,ncol=3)
  for(i in 1:r) {
    emp[i,] <- probitNormIntegral(n,mu,sigma)
  }
  m <- colMeans(emp)
  s <- apply(emp,2,sd)
  print("MEAN :")
  print(m)
  print("STDEV :")
  print(s)
  print("STDEV (RELATIVE) :")
  print(s/m)
}
```

Example Output

```
> probitNormTest(100,1000,-8,1)
[1] "MEAN :"
[1] 7.643951e-09 6.205931e-09 7.701978e-09
[1] "STDEV :"
[1] 1.579951e-09 1.239459e-08 1.019870e-10
[1] "STDEV (RELATIVE) :"
[1] 0.20669298 1.99721608 0.01324166
```

Importance sampling show smallest variance.

Summary

- Crude Monte Carlo method
  - Use uniform distribution (or other original generative model) to calculate the integration
  - Every random sample is equally weighted.
  - Straightforward to understand
- Rejection sampling
  - Estimation from discrete count of random variables
  - Larger variance than crude monte-carlo method
  - Typically easy to implement
- Importance sampling
  - Reweight the probability distribution
  - Possible to reduce the variance in the estimation
  - Effective for inference involving rare events
  - Challenge is how to find the good sampling distribution.