Systems Simulation
ECE 597/697 S

Variance Reduction Techniques

Prof. Michael Zink
Overview

- Introduction
- Potential Pitfalls
- Antithetic Variables
- Control Variates
- Common Random Numbers
Introduction

- “Statistical” efficiency is related to the variance of the output random variables from a simulation.
- If we can somehow reduce the variance of an output random variable of interest without disturbing its expectation, we can obtain greater precision.
- That is, smaller confidence intervals, for the same amount of simulating, or, alternatively, achieve a desired precision with less simulating.
Introduction

- The method of applying Variance Reduction Techniques (VRTs) usually depends on the particular model(s) of interest.
- Hence, a complete understanding of the model is required for proper use of VRTs.
- Typically, it is impossible to know beforehand how great a variance reduction might be realized.
- Or worse, whether the variance will be reduced at all in comparison with straight-forward simulation.
Introduction

- In typical simulation study one is interested in determining $\theta$ (parameter connected with some statistical model).
- To obtain $\theta$ a simulation is performed to obtain output datum $\theta = E[X]$. Repeated simulation runs are performed.
- Simulation study is terminated after $n$ runs have been performed and the estimate of $\theta$ is given by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$
Introduction

- Because this result is an unbiased estimate of $\theta$, it follows its mean square error is equal to its variance:
  $$MSE = E \left[ (\overline{X} - \theta)^2 \right] = Var(\overline{X}) = \frac{Var(X)}{n}$$

- Hence, obtaining a different unbiased estimate of $\theta$ having a smaller variance than $\overline{X}$, would result in an improved estimator.

- In this lecture, we look at different methods that can be used to reduce the variance of the (so-called raw) simulation estimate $\overline{X}$. 
Quality Control

- First, let us look at some pitfalls, even in quite simple models, of using the raw simulation estimator.
- Consider a process that produces items sequentially.
- Measurable value attached to them and when process is “in control” these (normalized) values come from a standard normal distribution.
- When process goes “out of control” process changed to some other distribution.
Quality Control

- To help detect when process goes out of control, an *exponentially weighted moving-average* control rule is often used.
- Let $X_1, X_2, \ldots$ denotes the sequence of data values.
- For a fixed value $\alpha$, $0 \leq \alpha \leq 1$, sequence $S_n$ is defined by
  \[
  S_0 = 0 \\
  S_n = \alpha S_{n-1} + (1 - \alpha) X_n, \quad n \geq 1
  \]
Quality Control

- When process is in control all $X_n$ have mean 0 and exponentially weighted moving-average $S_n$ also.
- Random process is declared out of control at random time $N$, where
  
  $$N = \text{Min}\left\{ n : |S_n| > B \right\}$$

- $|S_n|$ will eventually exceed $B$ and process will be declared out of order even if it is still working properly (even when data values are being generated by standard normal distribution).
Quality Control

- Choose $a$ and $B$ so that when $X_n, n \geq 1$, are coming from standard normal distribution, $E[N]$ is large.
- E.g., $E[N]=800$ is acceptable.
- Claim: $a=0.9$ and $B=0.8$ achieve that value.
- How can this be checked?
- One way is to use simulation.
Quality Control

- Generate standard normals $X_n$, $n \geq 1$, until $|S_n|$ exceeds 0.8.
- Output for first run $N_1$ denotes number of normals needed until this occurs.
- We generate other runs and $E[N]$ is the average value of output data over all runs.
- Let us assume we want to be 99% confident that estimate of $E[N]$ is accurate to within ±0.1.
- 99% of the time a normal random variable is within ±2.58 standard deviations of its mean, the number of runs needed is:

$$\frac{2.58\sigma_n}{\sqrt{n}} \approx 0.1$$
Quality Control

- $\sigma_n$ is the sample standard deviation based on first $n$ data values.
- Now $\sigma_n$ will approx. be equal to $\sigma(N)$, the standard deviation on $N$, which we argue is equal to $E[N]$.
- It appears that the distribution of time until the moving average exceeds the control limit is approx. memoryless $\Rightarrow$ approx. exponential random variable.
- But for an exponential random variable $Y$, $\text{Var}(Y) = (E(y))^2 \Rightarrow \sigma(N) \approx E[N]$. 
Quality Control

- Hence, if the original claim that $E[N] \approx 800$ is correct, the number of runs needed is such that $\sqrt{n} \approx 25.8 \times 800$ or $n \approx (25.8 \times 800)^2 \approx 4.26 \times 10^8$

- In addition, each run requires approx. 800 random variables.

- Thus, this simulation would require $800 \times 4.26 \times 10^8 \approx 3.41 \times 10^{11}$ normal random variables!
The Use of Antithetic Variables

- Antithetic variables (AV) is a variance reduction technique that is more applicable in simulating a single system.
- Try to induce correlation between separate runs, but now we seek negative correlation.
- Make pairs of runs of the model such that a “small” observation on one of the runs in the pair tends to be offset by a “large” observation on the other.
- So the two observations are negatively correlated.
- Then, if we use the average of the two observations in the pair as a basic data point for analysis, it will tend to be closer to the common expectation $\mu$ of an observation than it would be if the two observations in the pair were independent.
Antithetic Variables

- Antithetic variable induces negative correlation by using complementary random numbers to drive the two runs of the pair.

- If $U$ is a particular random number used for a particular purpose in the first run, we use $1 - U$ for the same purpose in the second run.

- This number $1 - U$ is valid because if $U \sim U(0,1)$ then $(1 - U) \sim U(0,1)$.

- Note that synchronization is required in AV too – use of complementary random numbers for the same purpose in a pair.
Antithetic Variables

- Since, $X_j$’s are i.i.d,

$$\text{Var}(\bar{X}_n) = \frac{\text{Var}(X_j)}{n} = \frac{\text{Var}(X_j^{(1)}) + \text{Var}(X_j^{(2)}) + 2\text{Cov}(X_j^{(1)}, X_j^{(2)})}{4n}.$$

- If the two runs within a pair were made independently, then $\text{Cov}(X_j^{(1)}, X_j^{(2)}) = 0$.

- On the other hand, if we could induce negative correlation between $X_j^{(1)}$ are $X_j^{(2)}$ then $\text{Cov}(X_j^{(1)}, X_j^{(2)}) < 0$, which reduces $\text{Var}[\bar{X}_n]$.

- This is the goal of Antithetic Variables.
Example I: Simulating the Reliability Function

- Consider a system of $n$ components each of which is either functioning or failed. Letting

$$S_i = \begin{cases} 
1 & \text{if component } i \text{ works} \\
0 & \text{otherwise}
\end{cases}$$

$s = (s_1, \ldots, s_n)$ is called the state vector.

- Suppose there is a nondecreasing function $\Phi(s_1, \ldots, s_n)$ such that

$$\Phi(s_1, \ldots, s_n) = \begin{cases} 
1 & \text{if the system works under state vector } s_1, \ldots, s_n \\
0 & \text{otherwise}
\end{cases}$$

This is called the structure function.
Example I

- Some common structure functions are the following:
  a) The *series structure*, which works only if all components function
     \[ \phi(s_1, \ldots, s_n) = \text{Min } s_i \]
  b) The *parallel structure*, which works if at least one component works
     \[ \phi(s_1, \ldots, s_n) = \text{Max } s_i \]
  c) The *k-of-n system*, which works if at least \( k \) of \( n \) systems are working
     \[ \phi(s_1, \ldots, s_n) = \begin{cases} 
     1 & \text{if } \sum_{i=1}^{n} s_i \geq k \\
     0 & \text{otherwise} 
     \end{cases} \]
Example I

d) The *bridge structure*, a five-component system for which

\[ \phi(s_1, s_2, s_3, s_4, s_5) = \text{Max}(s_1s_3s_5, s_2s_3s_4, s_1s_4, s_2s_5) \]

idea is to have is that the system functions if a signal can go from left to right through it.
Example I

- Suppose the states of the components are independent random variables
  \[ P\{S_i = 1\} = p_i = 1 - P\{S_i = 0\} \quad i = 1, \ldots, n \]

- The function \( r(p_1, \ldots, p_n) \) is called the reliability function. It represents the probability that the system will work when the components are independent with component \( i \) working with probability \( p_i \).
  \[ r(p_1, \ldots, p_n) = P\{\phi(S_1, \ldots, S_n) = 1\} \]
  \[ = E[\phi(S_1, \ldots, S_n)] \]
Example I

- For a series system
  \[ r(p_1, \ldots, p_n) = P\{S_i = 1 \text{ for all } i = 1, \ldots, n\} \]
  \[ = \prod_{i=1}^{n} P\{S_i = 1\} = \prod_{i=1}^{n} p_i \]

- And for a parallel system
  \[ r(p_1, \ldots, p_n) = P\{S_i = 1 \text{ for at least one } i, i = 1, \ldots, n\} \]
  \[ = 1 - P\{S_i = 0 \text{ for all } i = 1, \ldots, n\} \]
  \[ = 1 - \prod_{i=1}^{n} P\{S_i = 0\} = 1 - \prod_{i=1}^{n} (1 - p_i) \]
Example I

- For most other systems computing the reliability function is complex.
- Thus, for given nondecreasing structure function $\Phi$ and given probabilities $p_1, \ldots, p_n$, we are interested in using simulation to estimate

$$r(p_1, \ldots, p_n) = E\left[ \phi(S_1, \ldots, S_n) \right]$$

- Now $S_i$ can be simulated by generating random numbers $U_1, \ldots, U_n$ and then setting

$$S_i = \begin{cases} 
1 & \text{if } U_i < p_i \\
0 & \text{otherwise}
\end{cases}$$
Example I

- Hence we see that

\[ \phi(s_1, \ldots, s_m) = h(U_1, \ldots, U_n) \]

where \( h \) is a decreasing function of \( U_1, \ldots, U_n \).

- Therefore, \( \text{Cov}(h(U), h(1-U)) \leq 0 \) and so the antithetic variable approach of using \( U_1, \ldots, U_n \) to generate both \( h(U_1, \ldots, U_n) \) and \( h(1-U_1, \ldots, 1-U_n) \) results in a smaller variance than if an independent set of random numbers were used to generate the second value of \( h \).
Example II: Simulating a Queuing System

- Consider a given queuing system, let $D_i$ denote the delay in queue of $i$th customer.
- We are interested in simulating the system to estimate $\theta = E[X]$, where $X = D_1 + \ldots + D_n$ is the sum of the delays in the queue of the first $n$ arrivals.
- Let $I_1, \ldots, I_n$ denote the first $n$ interarrival times and $S_1, \ldots, S_n$ denote the first $n$ service times.
- These random variables are independent.
- Now $X$ is a function of the $2n$ random variables.

$$X = h(I_1, \ldots, I_n, S_1, \ldots, S_n)$$
Example II: Simulating a Queuing System

- Delay of a certain customer usually increases as the service times of other customers increase and vice versa.
- It follows, that $h$ is a monotone function of its coordinates.
- Hence, if inverse transform method is used to generate the random variables $I_1, ..., I_n$ and $S_1, ..., S_n$, the antithetic variable method results in a smaller variance.
Example II: Simulating a Queuing System

- Initially use the $2n$ random numbers $U_i, i=1,...,2n$, to generate the interarrival and service times.

\[ I_i = F_i^{-1}(U_i), \quad S_i = G_i^{-1}(U_{n+i}) \]

- Then second simulation run should be done in the same fashion but using random numbers $1-U_i, i=1,...,2n$.

- This results in a smaller variance than if a new set of $2n$ random numbers were generated for the second run.
Example III

- Use simulation to estimate $\theta = E\left[e^U\right] = \int_0^1 e^x \, dx$

- We know the solution is $\Theta = e - 1$, but the point of this example is to see what kind of improvement is possible by use of antithetic variables.

- To begin note that

$$Cov\left(e^U, e^{1-U}\right) = E\left[e^U e^{1-U}\right] - E\left[e^U\right]E\left[e^{1-U}\right]$$

$$= e - (e - 1)^2 = -0.2342$$
Example III

- Also because

\[ \text{Var}(e^U) = E[e^{2U}] - \left( E[e^U] \right)^2 \]

\[ = \int_0^1 e^{2x} \, dx - (e - 1)^2 \]

\[ = \frac{e^2 - 1}{2} - (e - 1)^2 = 0.2420 \]
Example III

- We see that the use of independent random numbers results in a variance of
  \[
  \text{Var}\left(\frac{\exp\{U_1\} + \exp\{U_2\}}{2}\right) = \frac{\text{Var}(e^U)}{2} = 0.1210
  \]

- Whereas the use of the antithetic variables $U$ and $1-U$ gives a variance of
  \[
  \text{Var}\left(\frac{e^U + e^{1-U}}{2}\right) = \frac{\text{Var}(e^U)}{2} + \frac{\text{Cov}(e^U, e^{1-U})}{2} = 0.0039
  \]

- A variance reduction of 96.7%
The Use of Control Variates

- The method of Control Variates (CV) attempts to take advantage of correlation between certain random variables to obtain a variance reduction.
- Suppose we are interested in the output parameter $X$. Particularly, we want $\mu = E[X]$.
- Suppose $Y$ be another random variable involved in the same simulation that is thought to be correlated with $X$ – either positively or negatively.
- Also suppose that we know the value $\nu = E[Y]$.
Control Variates

- If $X$ and $Y$ are positively correlated, then it is highly likely that in a particular simulation run, $Y > \nu$ would lead to $X > \mu$.
- Thus, if in a run, we notice that $Y > \nu$, we might suspect that $X$ is above its expectation $\mu$ as well, and accordingly we adjust $X$ downward by some amount.
- Alternatively, if we find $Y < \nu$, then we would suspect $X < \mu$ as well and adjust $X$ upward accordingly.
- This way, we use our knowledge of $Y$’s expectation to pull $X$ (up or down) towards its expected value $\mu$, thus reducing variability about $\mu$ from one run to next.
- We call $Y$ a control variate of $X$. 

Control Variates

- Unlike AV, success of CV does not depend on the correlation being of a particular sign.
- If $Y$ and $X$ are negatively correlated, CV would still work.
- Now, we would simply adjust $X$ upward if $Y > \nu$ and downward if $Y < \nu$.
- To implement this idea, we need to quantify the amount of the upward or downward adjustment of $X$.
- We will express this quantification in terms of the deviation $Y - \nu$, of $Y$ from its expectation.
Control Variates

- Let $c$ be a constant that has the same sign as correlation between $Y$ and $X$.
- We use $c$ to scale the deviation $Y - \nu$ to arrive at an adjustment to $X$ and thus define the “controlled” estimator: $X_c = X - c(Y - \nu)$.
- Since $\mu = E[X]$ and $\nu = E[Y]$, then for any real number $c$, $E(X_c) = \mu$.
- So above is an unbiased estimator of $\mu$ and may have lower variance.
  \[ Var(X_c) = Var(X) + c^2 Var(Y) + 2c Cov(X, Y). \]
- So has less variance if and only if: $c^2 Var(Y) < 2c Cov(X, Y)$. 
Control Variates

- We need to choose the value of $c$ carefully so that the condition is always satisfied.
- The optimal value is:
  $$c^* = -\frac{Cov(X,Y)}{Var(Y)}.$$
- In practice, though, it is more difficult than it appears!
- Depending on the source and nature of the control variate $Y$, we may not know $Var(Y)$ and certainly not know $Cov(X, Y)$.
- Hence obtaining the optimal value of $c$ might be difficult.
Control Variates

- For the value \( c^* = -\frac{\text{Cov}(X,Y)}{\text{Var}(Y)} \)

the variance of the estimator is

\[
\text{Var}(X + c^* (Y - \mu)) = \text{Var}(X) - \frac{[\text{Cov}(X,Y)]^2}{\text{Var}(Y)}
\]

- Upon dividing the equation above by \( \text{Var}(X) \), we obtain that

\[
\frac{\text{Var}(X + c^* (Y - \mu))}{\text{Var}(X)} = 1 - \text{Corr}^2(X,Y)
\]

- Hence, the variance reduction obtained in using the control variate \( Y \) is \( 100\text{Corr}^2(X,Y) \) percent.
Control Variates

- The quantities Cov(X,Y) and Var(Y) are usually not known in advance and must be estimated.

\[
Cov_e(X,Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}
\]

\[
Var_e(Y) = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{(n-1)}
\]

- We can approximate \(c^*\) by \(\hat{c}^*\), where

\[
\hat{c}^* = -\frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
\]
Control Variates

- The variance of the controlled estimator

\[
\text{Var}\left( \bar{X} + c^*(\bar{Y} - \mu) \right) = \frac{1}{n} \left( \text{Var}(X) - \frac{\text{Cov}^2(X,Y)}{\text{Var}(Y)} \right)
\]

can then be estimated by using the estimator of \( \text{Cov}(X,Y) \) along with the sample variance estimators of \( \text{Var}(X) \) and \( \text{Var}(Y) \).
Example IV

- Suppose as in Example II that we wanted to use simulation to estimate the reliability function

\[ r(p_1, \ldots, p_n) = E\left[ \phi(S_1, \ldots S_n) \right] \]

where

\[ S_i = \begin{cases} 
1 & \text{if } U_i < p_i \\
0 & \text{otherwise} 
\end{cases} \]

- Since \( E[S_i] = p_i \), it follows that

\[ E\left[ \sum_{i=1}^{n} S_i \right] = \sum_{i=1}^{n} p_i \]
Example IV

- Hence, we can use the number of working components
  \[ Y = \sum S_i \]
  as a control variate of the estimator
  \[ X = \phi(S_1, \ldots, S_n) \]

- Since \[ \sum_{i=1}^{n} S_i \] and \[ \phi(S_1, \ldots, S_n) \] are both increasing functions of the \( S_i \), they are positively correlated, and thus the sign of \( c^* \) is negative.
Example V

- Queuing system in which customers arrive in accordance to with nonhomogeneous Poisson process with intensity function $\lambda(s), s>0$.
- Service times are independent random variables with distribution $G$ and are also independent of arrival times.
- We are interested in estimating total time spent in the system by all customers arriving before time $t$.
- We are interested in $\theta=E[X]$, where $W_i$ denote the amount of time the $i$th customer spends in the system and $X = \sum_{i=1}^{N(t)} W_i$. 


Example V

- The total service time of all these customers is a natural quantity to use as a control in this case.

- With $S_i$ being the service time of the $i$th customer and set
  
  \[ Y = \sum_{i=1}^{N(t)} S_i \]
  
- Since these service times are independent of $N[t]$ it follows $E[Y] = E[S]E[N(t)]$

- Where $E[S]$ is the mean service time and $E[N(t)]$ is the mean number of arrivals by $t$, are both known quantities.
Example VI

- As in Example III, suppose we are interested in simulation to compute $\theta = E\left[ e^U \right]$

- A natural variate in this case is the random number $U$.

\[
\text{Cov}(e^U, U) = E[Ue^U] - E[U]E[e^U] = \int_0^1 xe^x dx - \frac{(e-1)}{2} = 1 - \frac{(e-1)}{2} = 0.14086
\]
Example VI

- Because $\text{Var}(U) = \frac{1}{12}$ it follows

$$\text{Var}\left(e^U + c^* \left(U - \frac{1}{2}\right)\right) = \text{Var}(e^U) - 12(0.14086)^2$$

$$= 0.2420 - 0.2380 = 0.0039$$

- In this case the use of control variate $U$ can lead to a variance reduction of up to 98.4%. 

Remarks

- Variance of the controlled estimator is not known in advance.
- Simulations are often performed in two stages.
- In the first stage a small number of runs are performed so as to give a rough estimate of $\text{Var}(X+c^*(Y-\mu_y))$.
- Number of trials needed in the second run can be fixed so variance for final estimator is in acceptable bound.
Remarks

- A valuable way of interpreting control variable approach is that it combines estimators of $\theta$.
- Assume both values $X$ and $W$ and are both determined by the simulation and $E[W]=E[X]=\theta$.
- Then an unbiased estimator of the following form may be considered
  \[ \alpha X + (1 - \alpha)W \]
- The best estimator to minimize the variance is given by
  \[ \alpha^* = \frac{\text{Var}(W) - \text{Cov}(X,W)}{\text{Var}(X) + \text{Var}(W) - 2\text{Cov}(X,W)} \]
Variance Reduction by Conditioning

- Conditional Variance:
  \[ Var(X) = E[Var(X \mid Y)] + Var[E(X \mid Y)] \]

- Both terms on the right are non-negative since the variance is always non-negative, one can see that
  \[ Var(X) \geq Var[E(X \mid Y)] \]

- Now assume we are interested in a simulation study to ascertain the value of \( \theta = E[X] \), with \( X \) being an output variable of the simulation run.
Variance Reduction by Conditioning

- Suppose there is a second variable, such that $E[X|Y]$ is known and has a value that can be determined from simulation runs.
- Since $E[E[X|Y]] = E[X] = \theta$, it follows that $E[X|Y]$ is also an unbiased estimator of $\theta$.
- Thus from $\text{Var}(X) \geq \text{Var}[E(X|Y)]$

it follows that $E[X|Y]$ is a superior estimator of $\theta$ over the raw estimator $X$. 
Variance Reduction by Conditioning

- Why is the conditional estimator superior to the raw estimator?
- We are performing a simulation to estimate the unknown value of $E[X]$.
- Simulation run proceeds in two stages:
  1. Observe simulated value of variable $Y$.
  2. Then simulate value of $X$.
- However, after observing $Y$, we are able to compute the (conditional) expected value of $X$.
- Using this value as an estimate of $E[X]$, eliminates the additional variance involved in simulating the actual value $X$. 
Variance Reduction by Conditioning

- One might consider further improvements by using an estimator of type $aX + (1-a)E[X|Y]$.
- The best estimator of this type has $a = a^*$, with

$$
\alpha^* = \frac{\text{Var}(E[X|Y]) - \text{Cov}(X, E[X|Y])}{\text{Var}(X) + \text{Var}(E[X|Y]) - 2\text{Cov}(X, E[X|Y])}
$$

- Now it will be shown that using $a^* = 0$, showing that combining the estimators $X$ and $E[X]$ does not improve just using $E[X|Y]$.
- Note that

$$
\text{Var}(E[X|Y]) = E\left[\left(E[X|Y]\right)^2\right] - \left(E[E[X|Y]]\right)^2
$$

$$
= E\left[\left(E[X|Y]\right)^2\right] - \left(E[X]\right)^2
$$
Variance Reduction by Conditioning

- On the other hand

\[
\text{Cov}(X, E[X|Y]) = E[X E[X|Y]] - E[X]E[E[X|Y]]
\]

\[
= E[X E[X|Y]] - (E[X])^2
\]

\[
= E[E[X E[X|Y]|Y]] - (E[X])^2
\]

\[
= E[E[X|Y]E[X|Y]] - (E[X])^2
\]

\[
= \text{Var}(E[X|Y])
\]

since given \( Y \), \( E[X|Y] \) is a constant.
Variance Reduction by Conditioning

- Thus, we see that no conditional variance reduction is possible by combining the estimators $X$ and $E[X|Y]$. 
Example VII

- Reconsider use of simulation to estimate \( \pi \).
- In the lecture on Random Number Generation, we have seen that \( \pi \) can be estimated by determining how often a randomly chosen point in the square of area 4 is falls within the inscribed circle of radius 1.
- Specifically, if we let \( V_i = 2U_i - 1 \), where \( U_i, i = 1, 2 \) are random numbers and set

\[
I = \begin{cases} 
1 & \text{if } V_1^2 + V_2^2 < 1 \\
0 & \text{otherwise}
\end{cases}
\]

then \( E[I] = \pi / 4 \).
Example VII

- The use of the average of successive values of $I$ to estimate $\pi/4$ can be improved upon using $E[I|V_1]$ rather than $I$.
- Now

$E[I | V_1 = v] = P\left\{ V_1^2 + V_2^2 \leq 1 | V_1 = v \right\}$

$= P\left\{ v^2 + V_2^2 \leq 1 | V_1 = v \right\}$

$= P\left\{ V_2^2 \leq 1 - v^2 \right\}$ by the independence of $V_1$ and $V_2$

$= P\left\{ -\left(1 - v^2\right)^{1/2} \leq V_2 \leq \left(1 - v^2\right)^{1/2} \right\}$
Example VII

\[
(1-v^2)^{1/2} = \int_{-(1-v^2)^{1/2}}^{(1-v^2)^{1/2}} \left(\frac{1}{2}\right) dx \quad \text{since } V_2 \text{ is uniform over } (-1,1)
\]

\[
= \left(1 - v^2\right)^{1/2}
\]

- Hence, \( E[I|V_1]=(1-V_1^2)^{1/2} \) and so the estimator \( (1-V_1^2)^{1/2} \) also has mean \( \pi/4 \) and has smaller variance than \( I \).
Example VII

- Since

\[
E\left[\left(1 - V_1^2\right)^{1/2}\right] = \int_{-1}^{1} \left(1 - x^2\right)^{1/2} \left(\frac{1}{2}\right) dx
\]

\[
= \int_{0}^{1} \left(1 - x^2\right)^{1/2} dx
\]

\[
= \int_{0}^{1} \left(1 - U^2\right)^{1/2} dx
\]

- We can simplify somewhat by using the estimator \((1 - U^2)^{1/2}\), where \(U\) is a random number.
Example VII

- The improvement in variance by using estimator $(1-U^2)^{1/2}$ over the estimator $I$ is easily determined by

$$Var\left[(1 - U^2)^{1/2}\right] = E[1 - U^2] - \left(\frac{\pi}{4}\right)^2$$

$$= \frac{2}{3} - \left(\frac{\pi}{4}\right)^2 \approx 0.0498$$

- The first equality used the identity $Var(W) = E[W^2] - (E[W])^2$. 
Example VII

- Because $I$ is a Bernoulli random variable having mean $\pi/4$, we have
  \[ Var(I) = \left( \frac{\pi}{4} \right) \left( 1 - \frac{\pi}{4} \right) \approx 0.1686 \]

- Conditioning results in a 70.44% reduction in variance.

- Since the function $(1-u^2)^{1/2}$ is a monotone decreasing function of $u$ in the region $0<u<1$, the estimator $(1-U^2)^{1/2}$ can be improved by using antithetic variables.
Example VII

- That is, the estimator
  \[ \frac{1}{2} \left[ (1 - U^2)^{1/2} + \left( 1 - (1 - U)^2 \right)^{1/2} \right] \]
  has smaller variance than
  \[ \frac{1}{2} \left[ (1 - U_1^2)^{1/2} + \left( 1 - U_2^2 \right)^{1/2} \right] \]

- Another way of improving the estimator \((1 - U^2)^{1/2}\)
  is by using a control variable.
Example VII

- A natural control variable in this case is $U^2$ and, because $E[U^2]=1/3$ an estimator of type

$$\left(1 - U^2\right)^{1/2} + c\left(U^2 - \frac{1}{3}\right)$$

can be used.

- The best $c$ — namely, $c^*=-\text{Cov}\left[(1-U^2)^{1/2},U^2\right]/\text{Var}(U^2)$ — can be estimated by using the simulation to estimate the covariance term.
Stratified Sampling

- Suppose we want to estimate \( \theta = E[X] \), and suppose there is some random variable \( Y \), with possible values \( y_1, \ldots, y_k \) such that
  - a) the probabilities \( p_i = P\{Y = y_i\} \), \( i = 1, \ldots, k \), are known; and
  - b) for each \( i = 1, \ldots, k \), we can simulate the value of \( X \) conditional on \( Y = y_i \).

- Now, if we are planning to estimate \( E[X] \) by \( n \) simulation runs, the usual approach would be to generate \( n \) independent replications of the random variable \( X \).

- Then use \( \bar{X} \) their average, as the estimate of \( E[X] \).
Stratified Sampling

- The variance of this estimator is
  \[ \text{Var}(\bar{X}) = \frac{1}{n} \text{Var}(X) \]

- However, writing
  \[ E[ X ] = \sum_{i=1}^{k} E[ X \mid Y = y_i ] p_i \]
  one can see that another way of estimating \( E[X] \)
  is by estimating the \( k \) quantities above.

- E.g., rather than generating \( n \) independent
  replications of \( X \), \( np_i \) simulations conditional on
  the event that \( Y = y_i \) can be performed.
Stratified Sampling

- If \( \bar{X}_i \) is the average of the \( np_i \) observed values of \( X \) generated conditional on \( Y=y_i \) then we would have the unbiased estimator

\[
\epsilon = \sum_{i=1}^{k} \bar{X}_i p_i
\]

- The estimator \( \epsilon \) is called a *stratified sampling* estimator of \( E[X] \).

- Because \( \bar{X}_i \) is the average of \( np_i \) independent random variables whose distribution is equal to the condition distribution of \( X \) given that \( Y=y_i \), it follows

\[
Var(\bar{X}_i) = \frac{Var(X | Y = y_i)}{np_i}
\]
Consequently, using the preceding and that $\bar{X}_i$, $i=1,...,k$, are independent, we see that

$$Var(\varepsilon) = \sum_{i=1}^{k} p_i^2 Var(\bar{X}_i)$$

$$= \frac{1}{n} \sum_{i=1}^{k} p_i^2 Var(X | Y = y_i)$$

$$= \frac{1}{n} E[Var(X | Y)]$$
Stratified Sampling

- The variance savings in using stratified sampling estimator $\varepsilon$ over the usual raw simulation estimator is

  $$Var(\bar{X}) - Var(\varepsilon) = \frac{1}{n} Var(E[X|Y])$$

- The variance savings per run is $Var(E[X|Y])$ which can be substantial when the value of $Y$ strongly affects the conditional expectation of $X$.

- Variance of stratified sampling estimator can be estimated by letting $S_i^2$ be the sample variance of the $np_i$ runs done conditionally on $Y=y_i$. 
Example VIII

- Customers arrive at infinite server queue:
  - Good day: Poisson process with rate 12/hour
  - Other days: Poisson process with rate 4/hour
- Service times, on all days, are exponentially distributed with rate 1/hour.
- Every day, system is shut down at hour 10.
- All those presently in service are forced to leave without completing service.
- Each day is independently a good day with probability 0.5.
- We want to use simulation to estimate $\theta$, the mean number of customers that do not have service completed.
Example VIII

- Let $X$ denote the # of customers whose service is not completed on a randomly selected day.
- Let $Y$ be 0 for an ordinary day and 1 for a good day.
- Then, it can be shown that the conditional distributions of $X$ given that $Y=0$ and that $Y=1$ are both Poisson with means
  
  $$ E[X | Y = 0] = 4(1 - e^{-10}), \quad E[X | Y = 1] = 12(1 - e^{-10}) $$

- Because the variance of a Poisson random variable is equal to its mean, the preceding show that
  
  $$ \text{Var}(X | Y = 0) = E[X | Y = 0] \approx 4 $$
  $$ \text{Var}(X | Y = 1) = E[X | Y = 1] \approx 12 $$
Example VIII

- Thus, $E\left[\text{Var}(X \mid Y)\right] \approx \frac{1}{2} (4 + 12) = 8$

  and

  \[ \text{Var}(E[X \mid Y]) = E\left[\left( E[X \mid Y]\right)^2\right] - (E[X])^2 \approx \frac{4^2 + 12^2}{2} - 8^2 = 16 \]

- Consequently, $\text{Var}(X) \approx 8 + 16 = 24$ which is 3 times as large as $E[\text{Var}(X \mid Y)]$, the variance of the stratified sampling estimator that simulates exactly half the days as good days and the other half as ordinary days.
Importance Sampling

- Let \( \mathbf{X} = (X_1, \ldots, X_n) \) denote a vector of random variables having a joint density function \( f(\mathbf{x}) = f(x_1, \ldots, x_2) \), and we are interested in estimating

\[
\theta = E[h(\mathbf{X})] = \int h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}
\]

where the preceding is an \( n \)-dimensional integral over all possible values of \( \mathbf{x} \).

- Direct simulation of random vector \( \mathbf{X} \), in order to compute values of \( h(\mathbf{X}) \) is inefficient because:
  a) difficult to simulate random vector having density function \( f(\mathbf{x}) \)
  b) variance of \( h(\mathbf{X}) \) is large
  c) combination of a) and b)
Importance Sampling

- Another way to use simulation to estimate $\theta$ is to note if $g(x)$ is another probability density such that $f(x) = 0$ wherever $g(x) = 0$ than $\theta$ can be expressed as

$$
\theta = \int \frac{h(x)f(x)}{g(x)} g(x) dx
$$

$$
= E_g \left[ \frac{h(x)f(x)}{g(x)} \right]
$$

where $E_g$ emphasizes that the random vector $X$ has joint density $g(x)$. 
Importance Sampling

- It follows that $\theta$ can be estimated by successively generating values of a random vector $\mathbf{X}$ having density function $g(\mathbf{x})$.
- Then using as the estimator the average of the values of $h(\mathbf{x})f(\mathbf{x})/g(\mathbf{x})$.
- If a density function $g(\mathbf{x})$ can be chosen so that random variable $h(\mathbf{x})f(\mathbf{x})/g(\mathbf{x})$ has a small variance, then the importance sampling approach can result in an efficient estimator of $\theta$. 
Importance Sampling

- Why can importance sampling be useful?
- $f(\mathbf{x})$ and $g(\mathbf{x})$ represent likelihood of obtaining vector $\mathbf{X}$ ($\mathbf{X}$ is a random vector with respective densities $f$ and $g$).
- If $\mathbf{X}$ is distributed according to $g$, then usually $f(\mathbf{x})$ will be small in relation to $g(\mathbf{x})$.
- When $\mathbf{X}$ is simulated according to $g$ likelihood ratio $f(\mathbf{x})/g(\mathbf{x})$ will be $<<1$.
- Its mean is 1:

$$E_g \left[ \frac{f(\mathbf{x})}{g(\mathbf{x})} \right] = \int \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = \int f(\mathbf{x}) d\mathbf{x} = 1$$
Importance Sampling

- Despite $f(x)/g(x)$ is usually smaller than 1, its mean is 1.
- Results in large variance.
- How can $h(x)f(x)/g(x)$ have a small variance?
- Arrange to choose density $g$ such that those values of $x$ for which $f(x)/g(x)$ is large are precisely the values for which $h(x)$ is exceedingly small.
- Thus the ratio $h(x)f(x)/g(x)$ is always small.
Importance Sampling

- How to select an appropriate density \( g \).
- *Tilted densities* are useful.
- Let \( M(t) = E_f[e^{tx}] = \int e^{tx} f(x) dx \) be the moment generating function corresponding to a one-dimensional density \( f \).

**Definition:** A density function

\[
  f_t(x) = \frac{e^{tx} f(x)}{M(t)}
\]

is called a tilted density of \( f \), \(-\infty < t < \infty\).

- Density \( f_t \) tends to be larger than one with density \( f \) when \( t > 0 \) and tends to be smaller when \( t < 0 \).
Example IV

- If \( f \) is the exponential density with rate \( \lambda \), then
  \[
  f_t(x) = Ce^{tx} \lambda e^{-\lambda x} = Ce^{-(\lambda-t)x}
  \]
  where \( C=1/M(t) \) does not depend on \( x \).
- Therefore, for \( t<\lambda \), \( f_t \) is an exponential density with rate \( \lambda-t \).
- If \( f \) is a Bernoulli probability mass function with parameter \( p \), then
  \[
  f(x) = p^x (1-p)^{1-x}, \quad x = 0,1
  \]
Example IV

- Hence, \( M(t) = E_f \left[ e^{tX} \right] = pe^t + 1 - p \)
  and so
  \[
  f_t(x) = \frac{1}{M(t)} \left( pe^t \right)^x (1 - p)^{1-x}
  \]
  \[
  = \left( \frac{pe^t}{pe^t + 1 - p} \right)^x \left( \frac{1 - p}{pe^t + 1 - p} \right)^{1-x}
  \]

- That is \( f_t \) is the probability mass function of a Bernoulli random variable with parameter
  \[
  p_t = \left( pe^t \right) / \left( pe^t + 1 - p \right)
  \]
Example IV

- If $f$ is a normal density with parameter $\mu$ and $\sigma^2$ then $f_t$ is a normal density having mean $\mu + \sigma^2$ and variance $\sigma^2$. 
Common Random Numbers

- More commonly used for comparing multiple systems rather than for analyzing a single system.
- Basic principle: “we should compare alternate configurations under similar experimental conditions.”
- Hence we will be more confident that the observed differences are due to differences in the system configuration rather than the fluctuations of the “experimental conditions.”
- In our simulation experiment, these experimental conditions are the generated random variates that are used to drive the model through the simulated time.
Common Random Numbers

- The name for this technique is because of the possibility in many situations of using the same basic $U(0,1)$ random numbers to drive each of the alternate configurations through time.
- To see the rationale for the use of CRN, consider two systems to be compared with output performance parameters $X_{1j}$ and $X_{2j}$, respectively for replication $j$.
- Let $\mu_i = E[X_i]$ be the expected output measure for system $i$.
- We are interested in $\zeta = \mu_1 - \mu_2$.
- Let $Z_j = X_{1j} - X_{2j}$ for $j = 1, 2 \ldots n$.
- Then, $E[Z_j] = \zeta$. That is, $\bar{Z}_n = \frac{\sum_{j=1}^{n} Z_j}{n}$ is an unbiased estimator of $\zeta$.
Common Random Numbers

- Since $Z_j$’s are IID variables:
  \[
  Var[\bar{Z}_n] = \frac{Var(Z_j)}{n} = \frac{Var(X_{1j}) + Var(X_{2j}) - 2Cov(X_{1j}, X_{2j})}{n}.
  \]

- If the simulations of two different configurations are done independently, with different random numbers, $X_{1j}$ and $X_{2j}$ will be independent. So the covariance will be zero.

- If we could somehow simulate the two configurations so that $X_{1j}$ and $X_{2j}$ are positively co-related, then $Cov(X_{1j}, X_{2j}) > 0$ so that the variance of the sample mean is reduced.

- So its value is closer to the population parameter $\zeta$. 
Common Random Numbers

- CRN is a technique where we try to introduce this positive correlation by using the same random numbers to simulate all configurations.
- However, success of using CRN is not guaranteed.
- We can see that as long as the output performance measures for two configurations $X_{1j}$ and $X_{2j}$ react monotonically to the common random numbers, CRN works.
- However, if $X_{1j}$ and $X_{2j}$ react in opposite directions to the random variables, CRN backfires.
- Another drawback of CRN is that formal statistical analyses can be complicated by the induced correlation.
Common Random Numbers

Synchronization

- To implement CRN, we must match up, or synchronize, the random numbers across different configurations on a particular replications.
- Ideally, a specific random variable should be used for a specific purpose on one configuration is used for exactly same purpose on all configurations.
- For example, say we are comparing different configurations of a queuing system.
- If a random number is used to generate service time for one system, the same random number should be used to generate service times for the other systems.