Statistics 580

Monte Carlo Methods

Introduction

Statisticians have used simulation to evaluate the behaviour of complex random variables whose precise distribution cannot be exactly evaluated theoretically. For example, suppose that a statistic T based on a sample $x_1, x_2, \ldots, x_n$ has been formulated for testing a certain hypothesis; the test procedure is to reject H_0 if $T(x_1, x_2, \ldots, x_n) \geq c$. Since the exact distribution of T is unknown, the value of c has been determined such that the asymptotic type I error rate is $\alpha = .05$, (say), i.e., $Pr(T \geq c | H_0 \text{ is true}) = .05$ as $n \to \infty$. We can study the actual small sample behavior of T by the following procedure:

a. generate $x_1, x_2, \ldots, x_n$ from the appropriate distribution, say, the Normal distribution and compute $T_j = T(x_1, x_2, \ldots, x_n)$,

b. repeat N times, yielding a sample $T_1, T_2, \ldots, T_N$, and

c. compute proportion of times $T_j \geq c$ as an estimate of the error rate i.e., $\hat{\alpha}_{(N)} = \frac{1}{N}\sum_{j=1}^{N} I(T_j \geq c) = \frac{\#(T_j \geq c)}{N}$, where $I(\cdot)$ is the indicator function.

d. Note that $\hat{\alpha}_{(N)} \overset{a.s.}{\to} \alpha$, the type I error rate.

Results of such a study may provide evidence to support the asymptotic theory even for small samples. Similarily, we may study the power of the test under different alternatives using Monte Carlo sampling. The study of Monte Carlo methods involves learning about

(i) methods available to perform step (a) above correctly and efficiently, and

(ii) how to perform step (c) above efficiently, by incorporating "variance reduction" techniques to obtain more accurate estimates.

In 1908, Monte Carlo sampling was used by W. S. Gossett, who published under the pseudonym "Student", to approximate the distribution of $z = \sqrt{n}(\bar{x} - \mu)/s$ where $(x_1, x_2, \ldots, x_n)$ is a random sample from a Normal distribution with mean μ and variance σ^2 and s^2 is the sample variance. He writes:

> "Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically [i.e., by simulation]. The material used was a ... table containing the height and left middle finger measurements of 3000 criminals.... The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random ... each consecutive set of 4 was taken as a sample ... [i.e.,n above is = 4] ... and the mean [and] standard deviation of each sample determined.... This provides us with two sets of ... 750 z's on which to test the theoretical results arrived at. The height and left middle finger ... table was chosen because the distribution of both was approximately normal..."

The object of many simulation experiments in statistics is the estimation of an expectation of the form $E[g(\mathbf{X})]$ where $\mathbf{X}$ is a random vector. Suppose that $f(\mathbf{x})$ is the joint density of $\mathbf{X}$. Then

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

Thus, virtually any Monte Carlo procedure may be characterized as a problem of evaluating an integral (a multidimensional one, as in this case). For example, if $T(\mathbf{x})$ is a statistic based on a random sample $\mathbf{x}$, computing the following quantities involving functions of T may be of interest:

$$\begin{aligned} \text{the mean, } \theta_1 &= \int T(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}, \\ \text{the tail probability, } \theta_2 &= \int I(T(\mathbf{x}) \le c) f(\mathbf{x})\, d\mathbf{x}, \\ \text{or the variance, } \theta_3 &= \int (T(\mathbf{x}) - \theta_1)^2 f(\mathbf{x})\, d\mathbf{x} \\ &= \int T(\mathbf{x})^2 f(\mathbf{x})\, d\mathbf{x} - \theta_1^2, \end{aligned}$$

where $I()$ is the indicator function.

Crude Monte Carlo

As outlined in the introduction, applications of simulation in Statistics involve the computation of an integral of the form $\int g(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}$, where $g(\mathbf{x})$ is a function of a statistic computed on a random sample $\mathbf{x}$. A simple estimate of the above integral can be obtained by:

- sample N time from the distribution of $\mathbf{X}$
- compute $T(\mathbf{x})$ from each sample: $T_1, T_2, \ldots, T_N$
- if $g(T)$ is the function of interest, form $g(T_1), g(T_2), \ldots, g(T_N)$ and $\hat{\theta}_{(N)} = \frac{1}{N} \sum_{j=1}^{N} g(T_j)$
- Note that $\hat{\theta}_{(N)} \stackrel{a.s.}{\to} \theta$
- if $\text{var}(g(T)) = \sigma^2$, the usual estimate of σ^2 is

$$s^2 = \frac{\sum_{j=1}^{N} (g(T_j) - (\hat{\theta}))^2}{N-1};$$

thus S.E.$(\hat{\theta}) = s/\sqrt{N}$ and for large N, the Central Limit Theorem could be used to construct approximate confidence bounds for θ. Clearly, the precision of $\hat{\theta}$ is proportional to $1/\sqrt{N}$ (in other words, the error in $\hat{\theta}$ is $O(N^{-\frac{1}{2}})$), and it depends on the value of s^2. Monte Carlo methods attempt to reduce S.E.$(\hat{\theta})$, that is obtain more accurate estimates without increasing N. These techniques are called *variance reduction techniques*. When considering variance reduction techniques, one must take into account the cost (in terms of difficulty level) of implementing such a procedure and weigh it against the gain in precision of estimates from incorporating such a procedure.

Before proceeding to a detailed description of available methods, a brief example is used to illustrate what is meant by variance reduction. Suppose the parameter to be estimated is an upper percentage point of the standard Cauchy distribution:

$$\theta = Pr(X > 2)$$

where

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty$$

is the Cauchy density. To estimate θ, let $g(x) = I(x > 2)$, and express $\theta = Pr(x > 2)$ as the integral $\int_{-\infty}^{\infty} I(x > 2) f(x) dx$. Note that x is a scalar here. Using crude Monte Carlo, θ is estimated by generating N Cauchy variates $x_1, x_2, \ldots, x_N$ and computing the proportion that exceeds 2, i.e.,

$$\hat{\theta} = \frac{\sum_{j=1}^{N} I(x_i > 2)}{N} = \frac{\#(x_i > 2)}{N}.$$

Note that since $N\hat{\theta} \sim$ Binomial (N, θ), the variance of $\hat{\theta}$ is given by $\theta(1-\theta)/N$. Since θ can be directly computed by analytical integration i.e.,

$$\theta = 1 - F(2) = \frac{1}{2} - \pi^{-1} \tan 2 \approx .1476$$

the variance of the crude Monte Carlo estimate $\hat{\theta}$ is given approximately as $0.126/N$. In order to construct a more efficient Monte Carlo method, note that for $y = 2/x$,

$$\theta = \int_2^{\infty} \frac{1}{\pi(1+x^2)} dx = \int_0^1 \frac{2}{\pi(4+y^2)} dy = \int_0^1 g(y) dy \,.$$

This particular integral is in a form where it can be evaluated using crude Monte Carlo by sampling from the uniform distribution $U(0,1)$. If $u_1, u_2, \ldots, u_N$ from $U(0,1)$, an estimate of θ is

$$\tilde{\theta} = \frac{1}{N} \sum_{j=1}^{N} g(u_j).$$

For comparison with the estimate of θ obtained from crude Monte Carlo the variance of $\tilde{\theta}$ can again be computed by evaluating the integral

$$\text{var}(\tilde{\theta}) = \frac{1}{N} \int \{g(x) - \theta\}^2 f(x) dx$$

analytically, where $f(x)$ is the density of $U(0,1)$. This gives $\text{var}(\tilde{\theta}) \approx 9.4 \times 10^{-5}/N$, a variance reduction of 1340.

In the following sections, several methods available for variance reduction will be discussed. Among these are

- stratified sampling
- control variates
- importance sampling
- antithetic variates
- conditioning swindles

Stratified Sampling

Consider again the estimation of

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

For stratified sampling, we first partition the domain of integration S into m disjoint subsets S_i, $i = 1, 2, \ldots, m$. Define θ_i as

$$\begin{aligned} \theta_i &= E\Big(g(\mathbf{X})\Big|\mathbf{X}\epsilon\, S_i\Big) \\ &= \int_{S_i} g(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} \end{aligned}$$

for $i = 1, 2, \ldots, m$. Then

$$\theta = \int_S g(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} = \sum_{i=1}^{m} \int_{S_i} g(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} = \sum_{i=1}^{m} \theta_i$$

The motivation behind this method is to be able to sample more from regions of S that contribute more variability to the estimator of θ, rather than spread the samples evenly across the whole region S. This idea is the reason why stratified sampling is used often as a general sampling techniques; for e.g., instead of random sampling from the entire state, one might use stratified sampling based on subregions such as counties, or districts, sampling more from those regions that contribute more information about the quantity being estimated.

Define

$$p_i = \int_{S_i} f(\mathbf{x})\, d\mathbf{x}$$

and

$$f_i(\mathbf{x}) = \frac{1}{p_i}\, f(\mathbf{x})\,.$$

Then $\sum_{i=1}^{m} p_i = 1$ and

$$\int_{S_i} f_i(\mathbf{x})\, d\mathbf{x} = 1\,.$$

By expressing

$$g_i(\mathbf{x}) = \begin{cases} g(\mathbf{x}), & \text{if } \mathbf{x}\,\epsilon\, S_i \\ 0, & \text{otherwise} \end{cases}$$

we can show that

$$\begin{aligned} \theta_i &= \int_{S_i} p_i\, g(\mathbf{x})\, \frac{f(\mathbf{x})}{p_i}\, d\mathbf{x} \\ &= p_i \int_S g_i(\mathbf{x})\, f_i(\mathbf{x})\, d\mathbf{x} \end{aligned}$$

and the crude Monte Carlo estimator of θ_i is then

$$\hat{\theta}_i = \frac{p_i \sum_{j=1}^{N_i} g_i(\mathbf{x}_j)}{N_i}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_{N_i}$ is a random sample from the distribution with density $f_i(\mathbf{x})$ on S_i, with variance given by

$$\mathrm{Var}(\hat{\theta}_i) = \frac{p_i^2\,\sigma_i^2}{N_i},$$

where $\sigma_i^2 = \mathrm{Var}\Big(g(\mathbf{x})\Big|\mathbf{X}\epsilon\, S_i\Big)$. Thus the combined estimator of θ is

$$\tilde{\theta} = \sum_{i=1}^{m} \hat{\theta}_i = \sum_{i=1}^{m} \frac{p_i}{N_i} \sum_{j=1}^{N_i} g_i(\mathbf{x}_j),$$

with variance given by

$$\mathrm{Var}(\tilde{\theta}) = \sum_{i=1}^{m} \frac{p_i^2\,\sigma_i^2}{N_i}.$$

For the best stratification, we need to choose N_i optimally i.e., by minimizing $\mathrm{Var}(\tilde{\theta})$ with respect to $N_i,\ i = 1, 2, \ldots, m$ Minimize

$$\sum_{i=1}^{m} \frac{p_i^2\,\sigma_i^2}{N_i} + \lambda(N - \sum_{i=1}^{m} N_i)$$

Equating partial derivatives with respect to N_i to zero it has been shown that the optimum occurs when

$$N_i = \frac{p_i\,\sigma_i}{\sum_{i=1}^{m} p_i\,\sigma_i} N$$

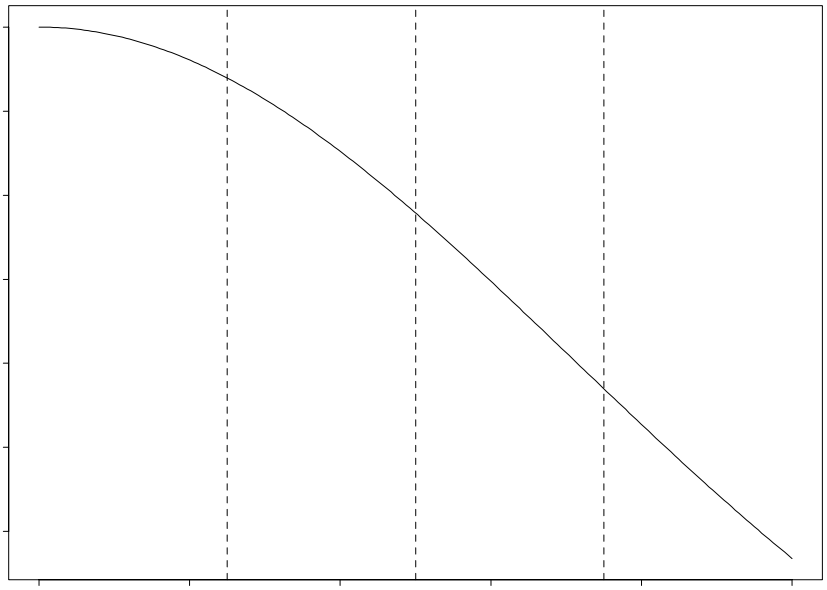
for $N = \sum_{i=1}^{m} N_i$. Thus after the strata are chosen, the samples sizes N_i should be selected proportional to $p_i\,\sigma_i$. Of course, since σ_i are unknown, they need to be estimated by some method, such as from past data or in absence of any information, by conducting a pilot study. Small samples of points are taken from each S_i. The variances are estimated from these samples. This allows the sample sizes for the stratified sample to be allocated in such a way that more points are sampled from those strata where $g(\mathbf{x})$ shows the most variation.

Example: Evaluate $I = \int_0^1 e^{-x^2} dx$ using stratified sampling.

Since in this example $f(x)$ is the $U(0,1)$ density we will use 4 equispaced partitions of the interval $(0,1)$ giving $p_1 = p_2 = p_3 = p_4 = .25$. To conduct a pilot study to determine strata sample sizes, we first estimated the strata variances of $g(x)$:

$$\sigma_i^2 = \mathrm{var}\,(g(x)|x \in S_i)\,\text{for } i = 1,\ 2,\ 3,\ 4$$

by a simulation experiment. We generated 10000 $u(0,1)$ variables u and compute the standard deviation of e^{-u^2} for u's falling in each of the intervals (0, .25), (.25, .5), (5, .75), and (.75, 1)



These came out to be $\sigma_i = .018,\ .046,\ .061$, and $.058$. Since $p_1 = p_2 = p_3 = p_4$, $N_i \propto \sigma_i / \sum \sigma_i$, we get $\sigma_i / \sum \sigma_i = .0984,\ .2514,\ .3333$, and $.3169$ implying that for $N = 10000$,

$$N_1 = 984,\ N_2 = 2514,\ N_3 = 3333,\ \text{and}\ N_4 = 3169.$$

We will round these and use

$$N_1 = 1000,\ N_2 = 2500,\ N_3 = 3500,\ \text{and}\ N_4 = 3000.$$

as sample sizes of our four strata.

A MC experiment using stratified sampling for obtaining an estimate of I and its standard error can now be carried out as follows. For each stratum $(m_i,\ m_{i+1})$, $m_1 = 0$, $m_2 = .25$, $m_3 = .50$, $m_4 = .75$, and, $m_5 = 1.0$, generate N_i numbers $u_j \sim U(m_i,\ m_{i+1}))$ for $j = 1, \ldots, N_i$ and compute $x_j = e^{-u_j^2}$ for the N_i u_j's. Then compute estimates

$$\begin{aligned} \hat{\theta}_i &= \sum x_j / N_i \\ \hat{\sigma}_i^2 &= \sum (x_j - \overline{x})^2 / (N_i - 1). \end{aligned}$$

for $i = 1,\ 2,\ 3,\ 4$. Pool these estimates to obtain $\tilde{\theta} = \sum \hat{\theta}_i / 4$ and $\text{var}(\tilde{\theta}) = \frac{\sum \hat{\sigma}_i^2 / N_i}{4^2}$. We obtained

$$\tilde{\theta} = 0.7470707 \text{ and var } (\tilde{\theta}) = 2.132326e - 007$$

A crude MC estimation of I is obtained by ignoring strata and using the 10000 $u(0,1)$ variables to compute means and variance of $e^{-u_j^2}$. We obtained

$$\hat{\theta} = 0.7462886 \text{ and var } (\hat{\theta}) = 4.040211e - 006$$

Thus the variance reduction is $4.040211e - 006 / 2.132326e - 007 \approx 19$.

Control Variates

Suppose it is required to estimate $\theta = E(g(\mathbf{x})) = \int g(\mathbf{x})f(\mathbf{x})d\mathbf{x}$ where $f(\mathbf{x})$ is a joint density as before. Let $h(\mathbf{x})$ be a function "similar" to $g(\mathbf{x})$ such that $E(h(\mathbf{x})) = \int h(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \tau$ is known. Then θ can be written as

$$\theta = \int [g(\mathbf{x}) - h(\mathbf{x})]f(\mathbf{x})d\mathbf{x} + \tau = E[g(\mathbf{x}) - h(\mathbf{x})] + \tau$$

and the crude Monte Carlo estimator of θ now becomes

$$\tilde{\theta} = \frac{1}{N}\sum_{j=1}^{N}[g(\mathbf{x}_j) - h(\mathbf{x}_j)] + \tau \ ,$$

and

$$\text{var}(\tilde{\theta}) = \frac{\text{var}(g(\mathbf{x})) + \text{var}(h(\mathbf{x})) - 2\text{cov}(g(\mathbf{x}), h(\mathbf{x}))}{N}.$$

If $\text{var}(h(\mathbf{x})) \approx \text{var}(g(\mathbf{x}))$ which is possibly so because h mimics g, then there is a reduction in variance if

$$corr(g(\mathbf{x}), h(\mathbf{x})) > \frac{1}{2}.$$

In general, θ can be written as

$$\theta = E[g(\mathbf{x}) - \beta\{h(\mathbf{x}) - \tau\}]$$

in which case the crude Monte Carlo estimator of θ is

$$\tilde{\theta} = \frac{1}{N}[\sum_{j=1}^{N} g(\mathbf{x}_j) - \beta\{\sum_{j=1}^{N} h(\{\mathbf{x_j}) - \tau\}] \ ,$$

As a trivial example consider a Monte Carlo study to estimate $\theta = E(\text{median}(x_1, x_2, \ldots, x_n))$ where $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ are i.i.d. random variables from some distribution. If $\mu = E(\bar{x})$ is known, then it might be a good idea to use $\bar{x}$ as a control variate. Crude Monte Carlo is performed on the difference $\{\text{median}(\mathbf{x}) - \bar{x}\}$ rather than on $\text{median}(\mathbf{x})$ giving the estimator

$$\tilde{\theta} = \frac{1}{N}\sum\{\text{median}(\mathbf{x}) - \bar{x}\} + \mu$$

with

$$\text{var}(\tilde{\theta}) = \frac{\text{var}\{\text{median}(\mathbf{x}) - \bar{x}\}}{N}.$$

The basic problem in using the method is identifying a suitable variate to use as a control variable. Generally, if there are more than one candidate for a control variate, a possible way to combine them is to find an appropriate linear combination of them. Let $Z = g(\mathbf{x})$ and let $W_i = h_i(\mathbf{x})$ for $i = 1, ..., p$ be the possible control variates, where as above it is required to estimate $\theta = E(Z)$ and $E(W_i) = \tau_i$ are known. Then θ can be expressed as

$$\begin{aligned} \theta &= E[g(\mathbf{x}) - \{\beta_1(h_1(\mathbf{x}) - \tau_1) + \ldots + \beta_p(h_p(\mathbf{x}) - \tau_p)\}] \\ &= E[Z - \beta_1(W_1 - \tau_1) - \ldots - \beta_p(W_p - \tau_p)]. \end{aligned}$$

This is a generalization of the application with a single control variate given above, i.e., if $p = 1$ then

$$\theta = E[Z - \beta(W - \tau)].$$

An unbiased estimator of θ in this case is obtained by averaging observations of $Z - \beta(W_1 - \tau)$ as before and β is chosen to minimize $\text{var}(\hat{\theta})$. This amounts to regressing observations of Z on the observations of W. In the general case, this leads to multiple regression of Z on $W_1, \ldots, W_p$. However note that W_i's are random and therefore the estimators of θ and $\text{var}(\hat{\theta})$ obtained by standard regression methods may not be unbiased. However, if the estimates $\hat{\beta}_i$ are found in a preliminary experiment and then held fixed, unbiased estimators can be obtained.

To illustrate this technique, again consider the example of estimating $\theta = P(x > 2) = \frac{1}{2} - \int_0^2 f(x)dx$ where $x \sim$ Cauchy. Expanding $f(x) = 1/\pi(1 + x^2)$ suggests control variates x^2 and x^4, i.e., $h_1(x) = x^2$, $h_2(x) = x^4$. A small regression experiment based on samples drawn from standard Cauchy gave

$$\hat{\theta} = \frac{1}{2} - [f(x) + 0.15(x^2 - 8/3) - 0.025(x^4 - 32/5)].$$

with $\text{var}(\hat{\theta}) \approx 6.3 \times 10^{-4}/N$.

Koehler and Larntz (1980) consider the likelihood ratio (LR) statistic for testing of goodness-of-fit in multinomials:

$$H_0 : p_i = p_{i0}, \quad i = 1, \ldots, k$$

where $(X_1, ..., X_k) \sim \text{mult}(n, \mathbf{p})$, where $\mathbf{p} = (p_1, ..., p_k)'$. The chi-squared statistic is given by

$$\chi^2 = \sum_{j=1}^{k} \frac{(x_j - np_{j0})^2}{np_{j0}},$$

and the LR statistic by

$$G^2 = 2 \sum_{j=1}^{k} x_j \log \frac{x_j}{np_{j0}}.$$

The problem is to estimate $\theta = E(G^2)$ by Monte Carlo sampling and they consider using the χ^2 statistic as a control variate for G^2 as it is known that $E(\chi^2) = k - 1$ and that G^2 and χ^2 are often highly correlated. Then $\theta = E(G^2 - \chi^2) + (k - 1)$ and the Monte Carlo estimator is given by

$$\tilde{\theta} = \frac{1}{N} \sum \{G^2 - \chi^2\} + k - 1$$

and an estimate of $\text{var}(\tilde{\theta})$ is given by

$$\widehat{\text{var}(\tilde{\theta})} = \frac{\text{var}\{G^2 - \chi^2\}}{N}.$$

Now, since in general, $\text{var}(G^2)$ and $\text{var}(\chi^2)$ are not necessarily similar in magnitude, an improved estimator may be obtained by considering

$$\theta = E[G^2 - \beta\,(\chi^2 - (k-1))]$$

and the estimator is

$$\tilde{\theta}_\beta = \frac{1}{N}\sum\{G^2 - \hat{\beta}\,\chi^2\} + \hat{\beta}\,(k-1)$$

where $\hat{\beta}$ is estimate of β obtained through an independent experiment. To do this, compute simulated values of G^2 and χ^2 statistics from a number of independent samples (say, 100) obtained under the null hypothesis. Then regress the G^2 values on χ^2 values to obtain $\hat{\beta}$. An estimate of $\mathrm{var}(\tilde{\theta}_\beta)$ is given by

$$\hat{\mathrm{var}}(\tilde{\theta}_\beta) = \frac{\mathrm{var}\left\{G^2 - \hat{\beta}\,\chi^2\right\}}{N}\,.$$

Importance Sampling

Again, the interest is in estimating $\theta = E(g(\mathbf{x})) = \int g(\mathbf{x})f(\mathbf{x})d\mathbf{x}$ where $f(\mathbf{x})$ is a joint density. Importance sampling attempts to reduce the variance of the crude Monte Carlo estimator $\hat{\theta}$ by changing the distribution from which the actual sampling is carried out. Recall that in crude Monte Carlo the sample is generated from the distribution of $\mathbf{x}$ (with density $f(\mathbf{x})$). Suppose that it is possible to find a distribution $H(\mathbf{x})$ with density $h(\mathbf{x})$ with the property that $g(\mathbf{x})f(\mathbf{x})/h(\mathbf{x})$ has small variability. Then θ can be expressed as

$$\theta = \int g(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \int g(\mathbf{x})\frac{f(\mathbf{x})}{h(\mathbf{x})}h(\mathbf{x})d\mathbf{x} = E\left[\phi(\mathbf{x})\right]$$

where $\phi(\mathbf{x}) = g(\mathbf{x})f(\mathbf{x})/h(\mathbf{x})$ and the expectation is with respect to H. Thus, to estimate θ the crude Monte Carlo estimator of $\phi(\mathbf{x})$ with respect to the distribution H could be used, i.e.,

$$\tilde{\theta} = \frac{1}{N}\sum_{j=1}^{N}\phi(\mathbf{x}_j)$$

where $\mathbf{x}_j = (x_1, \ldots, x_N)'$ is sampled from H. The variance of this estimator is

$$\mathrm{var}(\tilde{\theta}) = \frac{1}{N}\int\{\phi(\mathbf{x}_j) - \theta\}^2 h(\mathbf{x})d\mathbf{x} = \frac{1}{N}\int[\frac{f(\mathbf{x})g(\mathbf{x})}{h(\mathbf{x})}]^2 h(\mathbf{x})d\mathbf{x} - \frac{\theta^2}{N}.$$

Thus $\mathrm{var}(\tilde{\theta})$ will be smaller than $\mathrm{var}(\hat{\theta})$ if

$$\int[\frac{f(\mathbf{x})g(\mathbf{x})}{h(\mathbf{x})}]^2 h(\mathbf{x})d\mathbf{x} \leq \int[g(\mathbf{x})]^2 f(\mathbf{x})d\mathbf{x}$$

since

$$\mathrm{var}(\hat{\theta}) = \int[g(\mathbf{x})]^2 f(\mathbf{x})d\mathbf{x} - \frac{\theta^2}{N}\,.$$

It can be shown using Cauchy-Schwarz's inequality for integrals that the above inequality holds if

$$h(\mathbf{x}) = \frac{|g(\mathbf{x})|f(\mathbf{x})}{\int |f(\mathbf{x})|g(\mathbf{x})d\mathbf{x}},$$

i.e., if $h(\mathbf{x}) \propto |g(\mathbf{x})|f(\mathbf{x})$.

The choice of an appropriate density h is usually difficult and depends on the problem at hand. A slightly better idea is obtained by looking at the expression for the left hand side of the above inequality a little differently:

$$\int [\frac{f(\mathbf{x})g(\mathbf{x})}{h(\mathbf{x})}]^2 h(\mathbf{x})d\mathbf{x} = \int [f(\mathbf{x})^2 g(\mathbf{x})][\frac{f(\mathbf{x})}{h(\mathbf{x})}]h(\mathbf{x})d\mathbf{x}$$

To ensure a reduction in var$(\tilde{\theta})$), this suggests that $h(\mathbf{x})$ should be chosen such that $h(\mathbf{x}) > f(\mathbf{x})$ for large values of $f(\mathbf{x})^2 g(\mathbf{x})$ and $h(\mathbf{x}) < f(\mathbf{x})$ for small values of $f(\mathbf{x})^2 g(\mathbf{x})$.

Since sampling is now done from $h(\mathbf{x})$, $\tilde{\theta}$ will be more efficient if $h(\mathbf{x})$ is selected so that regions of $\mathbf{x}$ for which $|g(\mathbf{x})|f(\mathbf{x})$ is large are more frequently sampled than when using $f(\mathbf{x})$. By choosing the sampling distribution $h(\mathbf{x})$ this way, the probability mass is redistributed according to the relative importance of $\mathbf{x}$ as measured by $|g(\mathbf{x})|f(\mathbf{x})$. This gives the name *importance sampling* to this method since *important* $\mathbf{x}$'s are sampled more often and $h(\mathbf{x})$ is called the *importance* function.

Example 1:
Evaluate $\theta = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \int_{-\infty}^{\infty} I(y < x)\phi(y)dy$, where $\phi(y)$ is the Standard Normal density. A density curve similar in shape to $\phi(y)$ is the logistic density:

$$f(y) = \frac{\pi \exp(-\pi\, y/\sqrt{3})}{\sqrt{3}(1 + \exp(-\pi y/\sqrt{3}))^2}$$

with mean 0 and variance 1. The cdf of this distribution is:

$$F(y) = \frac{1}{1 + \exp(-\pi\, y/\sqrt{3})}.$$

A crude Monte Carlo estimate of θ is

$$\hat{\theta} = \frac{\sum_{i=1}^{N} I(z_i < x)}{N} = \frac{\#(z_i < x)}{N}$$

where $z_1, \ldots, z_N$ is a random sample from the standard normal. An importance sampling estimator of θ can be obtained by considering

$$\theta = \int_{-\infty}^{\infty} \frac{I(y < x)\phi(y)}{f(y)} f(y)dy = \int_{-\infty}^{\infty} \frac{k\phi(y)}{f(y)} \, \frac{I(y < x)f(y)}{k} \, dy$$

where k is determined such that $f(y)I(y < x)/k$ is a density i.e., $f(y)/k$ is a density over $(-\infty, x)$:

$$k = \frac{1}{1 + \exp(-\pi x/\sqrt{3})} \, .$$

The importance sampling estimator of θ is:

$$\tilde{\theta} = \frac{k}{N} \sum_{i=1}^{N} \frac{\phi(y_i)}{f(y_i)}$$

where $y_1, \ldots, y_N$ is a random sample from the density $f(y)I(y < x)/k$. The sampling from this density is easily done through the method of inversion: Generate a variate u_i from uniform $(0, 1)$ and transform by

$$y_i = -\frac{\sqrt{3}}{\pi} \log\left\{(1 + \exp(-\pi\, x/\sqrt{3}))/u_i - 1\right\}. \qquad \square$$

A recommended procedure for selecting an appropriate $h(\mathbf{x})$ given $g_L \leq g(\mathbf{x}) \leq g_U$ for all $\mathbf{x}$, is to choose $h(\mathbf{x})$ so that, for all $\mathbf{x}$,

$$h(\mathbf{x}) \geq \frac{g(\mathbf{x}) - g_L}{g_U - g_L}$$

While this does not guarantee variance reduction, it can be used to advantage when crude Monte Carlo will require large sample sizes to obtain a required accuracy. In many cases, this occurs usually when $g(\mathbf{x})$ is unbounded, i.e., $g_U \equiv \infty$ in the region of integration. In this case, we may chose $h(\mathbf{x})$ such that $\phi(\mathbf{x}) = \frac{f(\mathbf{x})g(\mathbf{x})}{h(\mathbf{x})}$ is bounded.

Example 2:
Consider the evaluation of

$$\theta = \int_0^1 x^{\alpha-1} e^{-x} dx, \qquad \frac{1}{2} < \alpha \leq 1$$

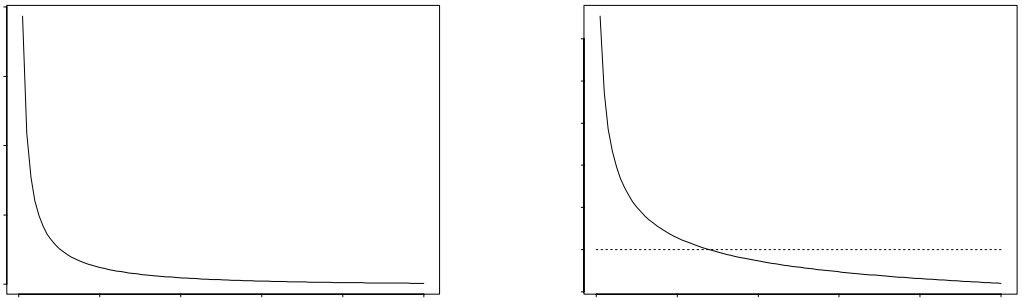
Here if $g(x) = x^{\alpha-1}e^{-x}$ and $f(x) = I_{(0,1)}(x)$, we find that $g(x)$ is unbounded at $x = 0$. Suppose we cosider the case when $\alpha = .6$ Then the crude MC estimate of θ is

$$\hat{\theta} = \frac{\sum_{j=1}^{N} g(u_j)}{N} = \frac{\sum_{j=1}^{N} (u_j^{-.4} e^{-u_j})}{N}$$

where $u_1, u_2, \ldots, u_N$ are from $U(0, 1)$. On the left below is a plot of $g(x)^2 f(x)$ for values of $x \in (0, 1)$. This plot shows that we need to select $h(x) > 1$ (say) for $x < .1$ (say) and $h(x) < 1$ for $x > .1$. Suppose we chose $h(x) = \alpha x^{\alpha-1}$. The plot on the right is a plot of $h(x)$ for $\alpha = .6$. The dashed line is a plot of $f(x) = I_{(0,1)}(x)$. This plot shows clearly that $h(x)$ would give more weight to those values of $x < .1$, the values for which $g(x)$ gets larger, than would $f(x)$, which assigned equal weights to all values of $g(x)$. We also see that since

$$\theta = \int_0^1 x^{\alpha-1} e^{-x} dx = \int_0^1 \phi(x) \alpha x^{\alpha-1} dx$$

where $\phi(x) = \frac{e^{-x}}{\alpha}$, $\phi(x)$ is bounded at $x = 0$.



Example 3: A more complex application
In *Sequential Testing*, the likelihood ratio test statistics λ is used to test

$$H_0 : \text{Population density is } f_0(x) \quad \text{vs.} \quad H_A : \text{Population density is } f_1(x)$$

where

$$\begin{aligned} \lambda &= \frac{f_1(x_1, \ldots, x_m)}{f_0(x_1, \ldots, x_m)} = \prod_{j=1}^{m} \frac{f_1(x_j)}{f_0(x_j)} \\ &= \exp\left\{\sum_{j=1}^{M} \log \frac{f_1(x_j)}{f_0(x_j)}\right\} = \exp\left\{\sum_{j=1}^{m} z_j\right\} \end{aligned}$$

and, $z_j = \log\{f_1(x_j)/f_0(x_j)\}$. Using the decision rule:

Reject H_0	if	$\sum z_j \geq b$
Accept H_0	if	$\sum z_j \leq a$
continue sampling	if	$a < \sum z_j < b$

It is usually assumed that $a < 0 < b$. Here the stopping time, m, is a random variable and also, note that $E_{H_0}(z_j) < 0$. The theory of sequential tests needs to evaluate probabilities such as $\theta = Pr\{\sum z_j \geq b | H_0 \text{ is true}\}$.

As an example, consider $X \sim N(\mu, 1)$ and suppose that $H_0 : \mu = \mu_0$ vs. $H_a : \mu = \mu_1$. For simplicity, let $\mu \leq 0$. Consider estimating the type I error rate

$$\theta = Pr_{H_0}\left\{\sum z_j \geq b\right\}.$$

where $z_j = x_j$, i.e. $\sum x_j$ is the test statistic used. Since θ in integral form is

$$\theta = \int_{-\infty}^{b} g_0(x)dx$$

where g_0 is the density of $\sum z_j$ under the null hypothesis H_0, which can also be expressed in the form

$$\theta = \int I(\sum z_j \geq b) f_0(x)dx,$$

where f_0 is the density of X_j under H_0, the usual Monte Carlo estimator of θ is

$$\hat{\theta} = \frac{\sum_{i=1}^{N} I(y_i \geq b)}{N} = \frac{\#(y_i \geq b)}{N}$$

where $y_1, \ldots, y_N$ are N independently generated values of $\sum z_j$. An estimator of θ under importance sampling is obtained by considering

$$\theta = \int I(\sum z_j \geq b) \frac{f_0(x)}{f_1(x)} f_1(x)\, dx$$

where f_1 is the density of X_j under the alternative hypothesis H_A.

Generating N independent samples of $y_1, y_2, \ldots, y_N$ under H_A the estimator of θ under importance sampling is

$$\tilde{\theta} = \frac{\sum_{i=1}^{N} I(y_i \geq b) f_0(y_i)/f_1(y_i)}{N}\,.$$

For this example,

$$\frac{f_0(y_i)}{f_1(y_i)} = \frac{e^{-1/2(y_i-\mu_0)^2}}{e^{-1/2(y_i-\mu_1)^2}} = e^{-1/2(\mu_0-\mu_1)(2y_i-(\mu_0+\mu_1))}\,.$$

If μ_1 is chosen to be $-\mu_0$ (remember μ_0 is ≤ 0 and thus $\mu_1 > 0$), the algebra simplifies to

$$\frac{f_0(y_i)}{f_1(y_i)} = e^{2\mu_0 y_i}.$$

Thus the importance sampling estimator reduces to $\sum I(y_i \geq b) \exp(2\mu_0 y_i)$ where y_i are sampled under H_A. The choice of $-\mu_0$ for μ_1 is optimal in the sense that it minimizes $\mathrm{Var}(\tilde{\theta})$ asymptotially. The following results are from Siegmund (1976):

$-\mu_0$	a	b	$\tilde{\theta}$	$\{N\mathrm{Var}(\hat{\theta})\}^{1/2}$	$\{N\mathrm{Var}(\tilde{\theta})\}^{1/2}$	R.E.
.0	any	any	.5	.5	.5	1
.125	-5	5	.199	.399	.102	15
.25	-5	5	.0578	.233	.0200	136
.50	-5	5	.00376	.0612	.00171	1280
.125	-9	9	.0835	.277	.0273	102
.25	-9	9	.00824	.0904	.00207	1910
.50	-9	9	.0000692	.00832	.0000311	7160

Antithetic Variates

Again, the problem of estimating $\theta = \int g(\mathbf{x}) f(\mathbf{x}) d\,\mathbf{x}$ is considered. In the control variate and importance sampling variance reduction methods the averaging of the function $g(\mathbf{x})$ done in crude Monte Carlo was replaced by the averaging of $g(\mathbf{x})$ combined with another function. The knowledge about this second function brought about the variance reduction of the resulting Monte Carlo method. In the antithetic variate method, two estimators are found by averaging two different functions, say, g_1 and g_2. These two estimators are such that they are negatively correlated, so that the combined estimator has smaller variance.

Suppose that $\hat{\theta}_1$ is the crude Monte Carlo estimator based on g_1 and a sample of size $N/2$ and $\hat{\theta}_2$, that based on g_2 and a sample of size $N/2$. Then let the combined estimator be

$$\tilde{\theta} = 1/2 \left(\hat{\theta}_1 + \hat{\theta}_2 \right) .$$

where $\hat{\theta}_1 = 2 \sum g_1(\mathbf{x}_j)/N$ and $\hat{\theta}_2 = 2 \sum g_2(\mathbf{x}_j)/N$. Then

$$\begin{aligned} \mathrm{Var}(\tilde{\theta}) &= 1/4 \left[\mathrm{Var}(\hat{\theta}_1) + \mathrm{Var}(\hat{\theta}_2) + 2\mathrm{cov}(\hat{\theta}_1, \hat{\theta}_2) \right] \\ &= 1/2 \left[2\sigma_1^2/N + 2\sigma_2^2/N + 4\rho\sigma_1\sigma_2/N \right] \end{aligned}$$

where $\rho = corr(\hat{\theta}_1, \hat{\theta}_2)$. A substantial reduction in $\mathrm{Var}(\tilde{\theta})$ over $\mathrm{Var}(\hat{\theta}_1)$ will be achieved if $\rho \ll 0$. If this is so, then g_2 is said to be an antithetic variate for g_1. The usual way of generating values for g_2 so that the estimate of $\hat{\theta}_2$ will have high negative corelation with $\hat{\theta}_1$ is by using uniform variates $1-u$ to generate samples for $\hat{\theta}_2$ of the same u's used to generate samples for $\hat{\theta}_1$. In practice this is achieved by using streams u_j and $1 - u_j$ to generate 2 samples from the required distribution by inversion.

The following example will help clarify the idea. Consider estimation of $\theta = E[g(u)]$ where u is uniformly distributed between 0 and 1. Then for any function $g(\cdot)$, let

$$\theta = E(g(u)) = \int_0^1 g(u) du \, ,$$

so $\frac{1}{N} \sum g(u_j)$ is an unbiased estimator of θ for a random sample $u_1, \ldots, u_N$. But since $1-u$ is also uniformly distributed on $(0,1)$, $\frac{1}{N} \sum g(1-u_j)$ is also an unbiased estimator of θ. If $g(x)$ is a monotonic function in x, $g(u)$ and $g(1-u)$ are negatively correlated and thus are "antithetic variates."

Actually, the process of using an antithetic variate is that of combining 2 estimates of θ in the best way possible. If the two estimates have the same variance then the best combination is the average and compared to crude Monte Carlo with N trials, antithetic Monte Carlo with $N/2$ trials decreases the variance of the estimator if $\rho < 0$. In general, if the estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ are available with a covariance structure $\sum$, they can be combined using generalized least squares estimator. First, model $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)'$ as

$$\hat{\theta} = \mathbf{1}\theta + \epsilon \quad , \quad \mathrm{Cov}(\epsilon) = \Sigma \, .$$

The best combined estimate is

$$\tilde{\theta} = (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} \mathbf{1}' \Sigma^{-1} \hat{\theta}$$

and $\mathrm{Var}(\tilde{\theta}) = \left(\mathbf{1}'\Sigma^{-1}\mathbf{1}\right)^{-1}$. Σ is usually estimated from the data; provided $\hat{\Sigma}$ is independent of $\hat{\theta}$, the resulting estimate $\tilde{\theta}$ remains unbiased for θ and may have lower variance. Just averaging $\hat{\theta}_1, \ldots, \hat{\theta}_k$ will give the optimal estimate if $\Sigma\,\mathbf{1} = c\,\mathbf{1}$ for some c, i.e., if the row sums of Σ are the same. Schruben and Margolin(1978) , *Journal of the American Statistical Association*, 73, 504-525, illustrates the use of antithetic streams for a rather complex simulation.

Conditioning Swindles

For simplicity of discussion, the problem of estimating $\theta = E(X)$ is considered. The extension to the problem of estimating $\theta = \int g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = E(g(\mathbf{X}))$ is straightforward. For a random variable W defined on the same probability space as X

$$E(X) = E\{E(X|W)\}$$

and

$$\mathrm{var}(X) = \mathrm{var}\{E(X|W)\} + E\{\mathrm{var}(X|W)\}$$

giving $\mathrm{var}(X) \geq \mathrm{var}\{E(X|W)\}$. The crude Monte Carlo estimate of θ is

$$\hat{\theta} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

and its variance

$$\mathrm{var}(\hat{\theta}) = \frac{\mathrm{var}(X)}{N}$$

where $x_1, x_2, \ldots, x_N$ is a random sample from the distribution of X. The conditional Monte Carlo estimator is

$$\tilde{\theta} = \frac{1}{N}\sum_{j=1}^{N} q(w_j)$$

and its variance

$$\mathrm{var}(\tilde{\theta}) = \frac{\mathrm{var}\{E(X|W)\}}{N}$$

where $w_1, w_2, \ldots, w_N$ is a random sample from the distribution of W and $q(w_j) = E(X|W_j)$. The usual situation in which the *conditioning swindle* an be applied is when $q(w) = E(X|W)$ can be easily evaluated analytically. In general, we can write

$$\mathrm{var}(g(\mathbf{X})) = \mathrm{var}\{E(g(\mathbf{X})|W)\} + E\{\mathrm{var}(g(\mathbf{X})|W)\}$$

and techniques such as importance sampling and stratified sampling were actually based on this conditioning principle.

Example 1:
Suppose $W \sim$ Poisson(10) and that $X|W \sim$ Beta(w, w^2+1). The problem is to estimate $\theta = E(X)$. Attempting to use crude Monte Carlo, i.e., to average samples from X will be inefficient. To use the conditioning swindle, first compute $E(X|W)$ which is shown to be $w/(w^2+w+1)$. Thus the conditional Monte Carlo estimator is

$$\tilde{\theta} = \frac{1}{N}\sum_{j=1}^{N} \frac{w_j}{w_j^2 + w_j + 1}$$

where $w_1, w_2, \ldots, w_N$ is a random sample from Poisson (10).

In this example, the conditioning variable was rather obvious; sometimes real ingenuity is required to discover a good random variable to condition on.

Example 2:
Suppose $x_1, x_2, \ldots, x_n$ is a random sample from $N(0,1)$ and let M = median $(X_1, \ldots, X_n$. The object is to estimate $\theta = \text{var}(M) = E[M^2]$. The crude Monte Carlo estimate is obtained by generating N samples of size n from $N(0,1)$ and simply averaging the squares of medians from each sample:

$$\hat{\theta} = \frac{1}{N}\sum_{j=1}^{N} M_j^2 \,.$$

An alternative is to use the *independence decomposition swindle*. If $\bar{X}$ is the sample mean, note that

$$\begin{aligned} \text{var}(M) &= \text{var}(\bar{X} + (M - \bar{X})) \text{ where } \bar{X} = \text{sample mean} \\ &= \text{var}(\bar{X}) + \text{var}(M - \bar{X}) + 2\text{cov}(\bar{X}, M - \bar{X}) \end{aligned}$$

However, $\bar{X}$ is independent of $M - \bar{X}$ giving $\text{cov}(\bar{X}, M - \bar{X}) = 0$. Thus

$$\text{var}(M) = \frac{1}{n} + E(M - \bar{X})^2$$

giving the swindle estimate of θ to be

$$\tilde{\theta} = \frac{1}{n} + \frac{1}{N}\sum_{j=1}^{N} (m_j - \bar{X}_j)^2.$$

One would expect the variation in $M - \bar{X}$ to be much less than the variation in M. In general, the idea is to decompose the stimator X into $X = U + V$ where U and V are independent; thus $\text{var}(X) = \text{var}(U) + \text{var}(V)$. If $\text{var}(U)$ is known, then $\text{var}(V) \leq \text{var}(X)$, and therefore V can be estimated more precisely than X.

Princeton Robustness Study (Andrews et al. (1972))

The simulation study of robust estimators combines aspects of the conditioning and independence swindles. The problem is evaluating the performance of a number of possible robust estimators of location. Let $X_1, X_2, \ldots, X_n$ be a random sample from an unknown distribution which is assumed to be symmetric about its median m. Consider $T(X_1, \ldots, X_n)$ to be an unbiased estimator of m, i.e., $E(T) = m$. Additionally, the estimators will be required to have the property that

$$T(aX_1 + b, aX_2 + b, \cdots, aX_n + b) = a\,T(X_1, X_2, \ldots, X_n) + b$$

i.e., T is a location and scale equivariant statistic. A good estimator would be one with small variance over most of the possible distributions for the sample $X_1, X_2, \ldots, X_n$. Thus, in general, it will be required to estimate θ where

$$\theta = E\{g(T)\}$$

If the interest is in estimating the variance of T, then $g(T) = T^2$. So the problem can be essentialy reduced to estimating $\theta = E\{T^2\}$.

Next, a method to construct a reasonable class of distributions to generate sample data is considered. It was decided to represent the random variable X as the ratio Z/V where $Z \sim N(0,1)$ and V is positive random variable independent of Z. This choice is motivated by the fact that conditional on V, X will have a simple distribution; a fact that might be useful for variance-reduction by employing a conditioning swindle. Some examples of possible distribution that may be represented are:

(i) If $V = 1/\sigma$, then $X \sim N(0, \sigma^2)$

(ii) If $V = \sqrt{\chi^2(\nu)/\nu}$, then $X \sim t(\nu)$

(iii) If $V = Abs$(Std. Normal Random Variable) then $X \sim$ Cauchy

(iv) If V has the two point distribution:

$$\begin{aligned} V &= c \quad \text{with probability } \alpha \\ &= 1 \quad \text{otherwise,} \end{aligned}$$

then $X \sim$ contaminated Normal, i.e., $N(0,1)$ with probability $1-\alpha$ and $N(0, 1/c^2)$ with probability α. To take advantage of this relatively nice conditional distribution, X_j's must be generated using ratio's of sequences of independent Z_j's and V_j's. Consider $X_j = Z_j/V_j$. Conditional on V_j, $X_j \sim N(0, 1/V_j^2)$. To incorporate conditioning and independence swindles to the problem of Monte Carlo estimation of θ, it is found convenient to express X_j in the form

$$X_j = \hat{\beta} + s\,C_j \quad \text{or} \quad \mathbf{X} = \hat{\beta}\,\mathbf{1} + s\ \mathbf{C}$$

where $\hat{\beta}, s$, and C_j are random variables to be defined later, such that $\hat{\beta}, s$ and $\mathbf{C}$ are independent given V. Note that because of the way T was defined,

$$T(\mathbf{X}) = \hat{\beta} + s\,T(\mathbf{C})\,.$$

Under these conditions it is possible to reduce the problem of estimating $\theta = E(T^2)$ to

$$\begin{aligned}
\theta &= E\{T^2(\mathbf{X})\} \\
&= E_{\mathbf{V}}\left[E_{\mathbf{X}|\mathbf{V}}\{T^2(\mathbf{X})\}\right] \\
&= E_{\mathbf{V}}\left[E_{\hat{\beta},s,\mathbf{C}}\{T^2(\hat{\beta},s,\mathbf{C})\}\right] \\
&= E_{\mathbf{V}}\left[E_{\mathbf{C}|\mathbf{V}}\left[E_{\hat{\beta}|\mathbf{V}}\left[E_{s|\mathbf{V}}\left\{\hat{\beta}^2 + 2\,\hat{\beta}\,s\,T(\mathbf{C}) + s^2\,T^2(\mathbf{C})\right\}\right]\right]\right] \\
&= E_{\mathbf{V}}\left[E_{\hat{\beta}|\mathbf{V}}\left(\hat{\beta}^2\right) + 2\,E_{\hat{\beta}|\mathbf{V}}(\hat{\beta})E_{s|\mathbf{V}}(s)E_{\mathbf{C}|\mathbf{V}}(T(\mathbf{C})) + E_{s|\mathbf{V}}(s^2)E_{\mathbf{C}|\mathbf{V}}(T^2(\mathbf{C}))\right]
\end{aligned}$$

The random variables $\hat{\beta}$ and s^2 will be defined so that conditional on $\mathbf{V}$, $\hat{\beta} \sim N(0, 1/\Sigma v_j^2)$, $(n\ -\ 1\ )s^2 \sim \chi^2(n-1)$ and $\hat{\beta}$ is independent of s^2. Thus

$$E_{\hat{\beta}|\mathbf{V}}(\hat{\beta}) = 0\ ,\ E_{\hat{\beta}|\mathbf{V}}(\hat{\beta}^2) = 1/\Sigma V_j^2 \text{ and } E_{s|\mathbf{V}}(s^2) = 1$$

Therefore

$$\theta = E\left[1/\Sigma V_j^2\right] + E\left[T^2(\mathbf{C})\right]\ .$$

Thus, the problem of estimation of θ has been split into two parts. In many cases, it is possible to evaluate $E(1/\mathbf{V}'\mathbf{V})$ analytically. For example, in generating samples from t-distribution with ν d.f., $V_j \sim \sqrt{\chi^2/\nu}$ and thus $E_{\mathbf{V}}(1/\Sigma V_j^2) = E(\nu/\chi^2(n\nu)) = \nu/(n\nu - 2)$. In other cases, the Delta method may be used to obtain an approximation (see the Appendix). The problem of estimating θ thus reduces to estimating $E(T^2(\mathbf{C}))$, which can be easily accomplished by averaging over N samples: $\{T^2(\mathbf{C}_1) + \cdots + T^2(\mathbf{C}_N)\}/N$. The variance of this estimate should be lower than an estimate found by averaging $T^2(\mathbf{X})$ since $T^2(\mathbf{C})$ is much less variable than $T^2(\mathbf{X})$.

Now, $\hat{\beta}$, s and $\mathbf{C}$ will be defined such that

$$X_j = \hat{\beta} + sC_j$$

and given $\mathbf{V}$, $\hat{\beta}$, s, and $\mathbf{C}$ are independent. Consider regressing Z_j on V_j through the origin. The slope is given by

$$\hat{\beta} = \frac{\Sigma V_j Z_j}{\Sigma V_j^2},$$

and $\hat{\beta} \sim N(0, 1/\mathbf{V}'\mathbf{V})$ given $\mathbf{V}$. The MSE from regression is

$$s^2 = \frac{\sum_{j=1}^{n}(Z_j - \hat{\beta}V_j)^2}{n-1}$$

and $(n-1)s^2 \sim \chi^2(n-1)$ given $\mathbf{V}$. Since Z_j are independent and normally distributed, s^2 and $\hat{\beta}$ are independent. Now form the standardized residuals from the regression.

$$\left\{ \frac{Z_j - \hat{\beta} V_j}{s} \right\}.$$

It can be shown that these standardized residuals are independent of $(\hat{\beta}, s^2)$ (Simon (1976)). Now let

$$C_j = \frac{Z_j - \hat{\beta} V_j}{V_j s} = \frac{Z_j / V_j - \hat{\beta}}{s} = \frac{X_j - \hat{\beta}}{s}.$$

The C_j are independent of $(\hat{\beta}, s^2)$ conditional on V_j since they are functions of the standardized residuals above. Then the needed form for X_j is obtained:

$$X_j = \hat{\beta} + sC_j.$$

The essence of this transformation is to average analytically over as much of the variation as possible. The construction of the distribution of X as a Normal-over-independent random variable is somewhat restrictive; however it does include t, Cauchy, Laplace and contaminated normal as possible distribution for X.

A procedure for a Monte Carlo experiment for comparing variances of several estimators of location under a specified sampling distribution using the *location and scale swindle* described above follows. The variances are used to measure the standard of performance of the estimators and are parameters of interest in the study. A good estimators are those which would have comparably low variances over a range of possible parent distributions. In some cases theoretical results can be used to compare the performance of estimators; for example, if the sample in Normal then $\bar{X}$ the sample mean is certainly the best since it has minimum variance. However, if the sample is from Cauchy $\bar{X}$ has infinite variance thus not a contender.

The estimators considered are the *mean*, *median*, and *10% trimmed mean* computed from independent samples from the t-distribution with k.d.f. generated as ratios of Normal-over-independent random variables. Note that these estimators satisfy the location and scale equivariant property. The problem is thus one for which the swindle could be applied. This experiment is to be repeated for various sample sizes (done here for 2 sample sizes n=10, 20). In order to facilitate comparisons across different sample size the actual quantity tabulated is $n \times \text{var}(T)$ instead of just $\text{var}(T)$. Since the data were generated from a distribution with median zero the parameter to be estimated is $\theta = E(T^2)$. Both a crude Monte Carlo and the Swindle are employed. In the first case, for each replication,

1. Sample $z_1, \ldots, z_n \sim N(0,1)$ and $v_1, \ldots, v_n \sim \chi^2(k)$,independently; Set $x_i = z_i / \sqrt{v_i/k}$, $i = 1, \ldots, n$.

2. Form $W = T(\mathbf{x})^2$.

then average W over 1000 replications, to obtain a crude MC estimate.

For the Swindle, for each replication,

1. Sample $z_1, \ldots, z_n \sim N(0,1)$ and $y_1, \ldots, y_n \sim \chi^2(k)$, independently; Set $x_i = z_i/v_i$ where $v_i = \sqrt{y_i/k}, i = 1, \ldots, n$.

2. Compute $\hat{\beta}$ and s^2 from (v_i, z_i), $i = 1, \ldots, n$.

3. Set $c_i = (x_i - \hat{\beta})/s$ for $i = 1, \ldots, n$, and form $W^* = k/(nk-2) + T(\mathbf{c})^2$

then average W^* over the 1000 replications. Recall that W and W^* are the Monte Carlo estimates of var(T) from crude Monte Carlo and from the swindle. In order to be able to find out whether there is an actual variance reduction, the sampling variances of both these estimates are needed. Although formulas can be developed for computing these more

Table 1: Crude Monte Carlo and swindle estimates of $n \times \text{var}(T)$ based on 1000 simulations for estimators T = Mean, Median, and 10% Trimmed Mean of samples of size $n = 10$ from the t-distribution with k d.f.

		k				
T		3	4	10	20	100
Mean						
	Crude MC estimate	2.53	1.94	1.16	1.17	1.04
	Standard error	.16	.10	.049	.054	.049
	Swindle estimate	2.84	1.93	1.24	1.11	1.02
	Standard error	.16	.065	.0098	.0047	.00082
	Variance reduction	1	2	25	132	3571
Median						
	Crude MC estimate	1.65	1.72	1.31	1.57	1.46
	Standard error	.08	.094	.059	.073	0.066
	Swindle estimate	1.76	1.68	1.45	1.45	1.42
	Standard error	.034	.033	.020	.019	.017
	Variance reduction	6	8	9	15	15
10% Trimmed Mean						
	Crude MC estimate	1.59	1.54	1.11	1.19	1.11
	Standard error	.073	.075	.047	.055	.052
	Swindle estimate	1.75	1.52	1.19	1.13	1.07
	Standard error	.038	.024	.0077	.0049	.0029
	Variance reduction	4	10	37	126	322

efficiently, estimates can be obtained form the actual sampling procedure itself. Writing

$$\theta = \text{Var}(T) = E(T^2)$$

Table 2: Crude Monte Carlo and swindle estimates of $n \times \mathrm{var}(T)$ based on 1000 simulations for estimators T = Mean, Median, and 10% Trimmed Mean of samples of sizes $n = 20$ from the t-distribution with k d.f.

		k				
T		3	4	10	20	100
Mean						
	Crude MC estimate	2.75	1.76	1.24	1.12	0.99
	Standard error	.19	.087	.056	.056	.045
	Swindle estimate	2.96	1.87	1.25	1.11	1.02
	Standard error	.20	.062	.010	.0046	.00085
	Variance reduction	1	2	31	148	2803
Median						
	Crude MC estimate	1.71	1.75	1.67	1.59	1.46
	Standard error	.080	.075	.074	.077	.061
	Swindle estimate	1.79	1.72	1.58	1.54	1.46
	Standard error	.036	.035	.025	.022	.019
	Variance reduction	5	5	9	12	10
10% Trimmed Mean						
	Crude MC estimate	1.55	1.40	1.23	1.15	1.04
	Standard error	.071	.060	.055	.057	.047
	Swindle estimate	1.64	1.46	1.20	1.13	1.07
	Standard error	.027	.019	.0082	.0055	.0029
	Variance reduction	7	10	45	107	263

we have

$$\hat{\theta} = \frac{1}{N} \sum_{j=1}^{N} T_j^2$$

where, say, $T_j^2 \equiv T(\mathbf{x}_j)^2$. Then

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}\left(\frac{1}{N} \sum_{j=1}^{N} T_j^2 \right) = \frac{1}{N^2} \sum_{j=1}^{N} \mathrm{Var}(T_j^2).$$

So an estimate of $\mathrm{Var}(\hat{\theta})$ is

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{N^2} \times N \times \widehat{\mathrm{Var}}(T^2)$$

Since an estimate of $\mathrm{Var}(T^2)$ can be obtained from the sample as

$$\widehat{\mathrm{Var}}(T^2) = \frac{\sum T_j^4}{N} - \left(\frac{\sum T_j^2}{N} \right)^2$$

The standard errors reported in Table 1 and 2 are thus computed using:

$$\text{s.e.}(n \times \widehat{\text{var}(T)}) = n \times \sqrt{\left\{ \frac{\sum T_j^4}{N^2} - \frac{(\sum T_j^2)^2}{N^3} \right\}},$$

.

The Location-Scale Swindle for a Scale Estimator

Here the problem is less straightforward than the location parameter case, where the interest was in the unbiased estimator of the median of a symmetric distribution. However, it is less clear what the scale parameter of interest is, although the samples would still be drawn from a symmetric distribution. Suppose it is imposed that the scale estimator considered is *location-invariant* and *scale-transparent*, i.e.,

$$S(cX_1 + d, cX_2 + d, \cdots, cX_n + d) = |\, c \,|\, T(X_1, X_2, \ldots, X_n) + d$$

or, in vector form

$$S(c\mathbf{X} + d = |\, c \,|\, S(\mathbf{X})$$

then

$$\log S(c\mathbf{X} + \mathbf{d} = \log |\, c \,| + \log S(\mathbf{X})\,.$$

Thus if the interest is in comparing scale estimators, $Q(\mathbf{X}) = \log S(\mathbf{X})$ is a good function of $S(\mathbf{X})$ to use and $Var\{Q(\mathbf{X})\}$ provides a good measure of the performance of the estimator $S(\mathbf{X})$ since it is scale-invariant (i.e., will not depend on c). Thus estimators $S(\mathbf{X})$ such that $Var\{\log S(\mathbf{X})\}$ is well-behaved over a large class of distributions would be better.

By our presentation on the location estimator we have $\mathbf{X} = \hat{\beta}\,\mathbf{1} + s\,\mathbf{C}$ and it can also be shown easily that $S(\mathbf{X}) = s\,S(\mathbf{C})$. Thus

$$\begin{aligned}
\text{Var}\{Q(\mathbf{X})\} &= \text{Var}\{\log S(\mathbf{X})\} \\
&= \text{Var}\{\log s + \log S(\mathbf{C})\} \\
&= \text{Var}[E\{(\log s + \log S(\mathbf{C}))|\mathbf{V}\}] + E[\text{Var}\{(\log s + \log S(\mathbf{C}))|\mathbf{V}\}] \\
&= \text{Var}\{\log \sqrt{\frac{W}{k-1}} + \text{Var}\{\log S(\mathbf{C})\}
\end{aligned}$$

where $W \sim \chi^2(k-1)$ and note that $\text{Var}\{E(\log s|\mathbf{V})\} = 0$.

Appendix

A Note on approximating $E(1 \mid \sum_{i=1}^{n} V_C^2)$ using the Delta method

First order approximation using the Taylor series approximation does not provide sufficient accuracy for the computation of the expectation of a function of random variable; at least the second order term is needed. Consider the computation of $E(g(X))$. Assume $X \dot{\sim} (\mu,\ \sigma^2)$ and expand $g(x)$ around μ:

$$g(x) \approx g(\mu) + (x-\mu)g'(\mu) + \frac{1}{2}(x-\mu)^2 g''(\mu).$$

Thus

$$E[g(X)] \approx g(\mu) + \frac{1}{2}g''(\mu)\sigma^2$$

This is a standard Delta method formula for the mean of a function of X. For $g(x) = \frac{1}{x}$ $g''(x) = \frac{2}{x^3}$ implying that $E\left(\frac{1}{X}\right) \approx \frac{1}{\mu} + \frac{\sigma^2}{\mu^3}$. Thus the first order approximation for $E(\frac{1}{X})$ is $\frac{1}{\mu}$, and the second order approximation is $\frac{1}{\mu} + \frac{\sigma^2}{\mu^3}$. In the swindle problem we have $X = \sum_{i=1}^{n} V_i^2$

$\underline{N(0,\ 1)/U(0,\ 1) \text{ case:}}$
Straight forward application of the Delta method, taking $V \sim U(0,\ 1)$ gives the first order approximation $E(\frac{1}{\Sigma V_i^2}) = \frac{3}{n}$, and the second order approximation

$$E\left(\frac{1}{\Sigma V_i^2}\right) = \frac{3}{n} + \frac{12}{5n^2}$$

$\underline{.9N(0,\ 1) + .1N(0,\ 9) \text{ case:}}$
Not as straightforward as above. First compute

$$\begin{aligned}
\mu = E(\Sigma V_i^2) &= n\left(1(.9) + \frac{1}{9}(.1)\right) = \frac{8.2}{9}n \\
\sigma^2 = Var\left(\sum_{i=1}^{n} V_i^2\right) &= nVar(V_i^2) \\
&= nE(V_i^2 - \mu)^2 \\
&= n\left\{\left(1 - \frac{8.2}{9}\right)^2 (.9) + \left(\frac{1}{9} - \frac{8.2}{9}\right)^2 (.1)\right\} \\
&= \frac{5.76n}{81}
\end{aligned}$$

Thus the first order approximation for $E(\frac{1}{\Sigma V_i^2})$ is $\frac{9}{8.2n}$ and the second order approximation

$$E\left(\frac{1}{\Sigma V_i^2}\right) = \frac{9}{8.2n} + \frac{5.76n}{81} \cdot \left(\frac{9}{8.2}\right)^3 \frac{1}{n^2} \approx \frac{9}{8.2n} + \frac{6480}{68921n^2}$$

References

Andrew, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey (1972) *Robust Estimation of Location.* Princeton University Press: Princeton, N.J.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods.* Springer-Verlag: NY.

Hammersley, J. M. and Handscomb, D. C. (1964) *Monte Carlo Methods.* Methuen: London.

Johnstone, Iain and Velleman, Paul (1984) " Monte Carlo Swindles Based on Variance Decompositions,"*Proc. Stat. Comp. Section, Amer. Statist. Assoc.(1983)*,52-57.

Koehler, Kenneth J. (1981) "An Improvement of a Monte Carlo Technique using Asymptotic Moments with an Application to the Likelihood Ratio Statistics," *Communications in Statistics – Simulation and Computation*, 343-357. 73, 504-519.

Lange, K. (1998). *Numerical Analysis for Statisticians.* Springer-Verlag: NY.

Ripley, B. D. (1987). *Stochastic Simulation.* Wiley: NY.

Rothery, P. (1982) "The Use of Control Variates in Monte Carlo Estimation of Power," *Applied Statistics*, 31, 125-129.

Rubinstein, R. Y. (1981) *Simulation and the Monte Carlo Method.* Wiley: New York.

Schruben, Lee W. and Margolin, Barry H. (1978) "Pseudo-random number assignment in statistically designed simulation and distribution sampling experiments," *J. Amer. Statist. Assoc.*,

Siegmund, D. (1976) "Importance Sampling in the Monte Carlo study of Sequential Tests," *Annals of Statistics*, 4, 673-684.

Simon, Gary (1976) "Computer Simulation Swindles, with Applications to Estimates of Location and Dispersion," *Applied Statistics*, 25, 266-274.