TECHNOMETRICS, MAY 1995, VOL. 37, NO. 2

Weighted Average Importance Sampling and Defensive Mixture Distributions

Tim HESTERBERG

Mathematics Department
Franklin & Marshall College
Lancaster, PA 17604-3003

Importance sampling uses observations from one distribution to estimate for another distribution by weighting the observations. Including the target distribution as one component of a mixture distribution bounds the weights and makes importance sampling more reliable. The usual importance-sampling estimate is a weighted average with weights that do not sum to 1. We discuss simple normalization and other, more efficient normalization methods. These innovations make importance sampling useful in a wider variety of problems. We demonstrate with a case study of oil-inventory reliability at a large utility.

KEY WORDS: Monte Carlo; Simulation; Variance reduction.

1. INTRODUCTION

Importance sampling is the process of estimating something about a distribution using observations from a different distribution. It was first used as a variance-reduction technique, in which the latter ("design") distribution is chosen to reduce the simulation effort required to obtain answers about the former ("target") distribution. This is particularly effective in some of the most difficult simulation problems—those in which it is necessary to obtain information associated with rare events. By increasing the design frequency of "important" events, it is possible to accurately estimate the frequency or expected magnitude of those events using fewer Monte Carlo replications. Importance sampling has been used in many fields to estimate small probabilities, beginning with nuclear shielding (Kahn 1950).

Importance sampling is also useful when it is impossible or impractical to generate samples from the target distribution—for example, in Bayesian analysis when the posterior distribution is intractable (Kloek and van Dijk 1978 and many others).

Importance sampling is especially valuable in problems with more than one target distribution. Rather than running a separate simulation for each distribution, importance sampling can be used to estimate results for many distributions in a single simulation (Beckman and McKay 1987; Tukey 1987). The number of such distributions may be infinite, as in the bootstrap tilting interval (Tibshirani 1984).

The classical importance-sampling method is designed for *Monte Carlo integration*—the estimation of an integral, or equivalently the expected value of a random variable. Let X be a random variable with known density f and let $Q(X)$ be a function of X with an unknown mean $\mu = E_f[Q(X)]$. The classical derivation of importance sampling (Hammersley and Handscomb 1964) is to express μ as

$$\mu = \int Q(x)f(x)\,dx = \int \frac{Q(x)f(x)}{g(x)}g(x)\,dx.$$

This leads to the usual importance-sampling estimate

$$\widehat{\mu}_{\text{int}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Q(X_i)f(X_i)}{g(X_i)}, \tag{1.1}$$

where $X_i \sim g$ for $i = 1, \ldots, n$. We refer to (1.1) as the *integration estimate* in the sequel.

We view importance sampling as a *sampling strategy* rather than as an integration method. Importance sampling produces an empirical distribution function consisting of observations $Q(X_i)$ and associated weights, which counteract the sampling bias. For example, if an observation X_i is twice as likely under the design distribution g as under the target distribution f, it should be given half as much weight when computing results. Means, higher moments, and percentiles can be computed from the empirical distribution function.

The integration estimate is a *weighted average with weights that do not sum to 1*. Let

$$W(x) = \frac{f(x)}{g(x)}$$

be the weight function, based on the likelihood ratio between the target and design distributions; then the empirical distribution function corresponding to the integration estimate has weight $V_{\text{int},i} = n^{-1}W_i$ on the output $Q(X_i)$ from the ith replication, where $W_i = W(X_i)$. The sum of the empirical weights is $\overline{W}$, which has expected value 1 but nonzero variance.

In one class of problems it does not matter that the weights do not sum to 1, where μ is the mean of an output variable Q that is nearly always 0. This class includes the important special case of estimating a small probability, where $Q = 1$ if a rare event occurs, otherwise $Q = 0$. The integration estimate works well for problems in this class because it is insensitive to weights assigned to replications for which Q is 0 (so it does not matter if those weights are such that the sum of all weights is 1).

The integration estimate is less successful in other problems. One example was presented by Hopmans and Kleijnen (1979), with the revealing title "Importance Sampling in Systems Simulation: A Practical Failure?" Because the weights do not add to 1, expected value estimates are not equivariant; adding a constant to Q does not increase the estimate by the same amount (Therneau 1983). Distribution-function estimates have mass not equal to 1, complementary probabilities do not add to 1, and variance estimates can be negative. An estimate can be dominated by the sum of weights, particularly if Q is nearly constant.

Normalizing the weights to sum to 1 solves those problems. The simplest normalization method gives the *ratio* weights. More efficient are the *regression*, *maximum likelihood* (ML), or *exponential* weights. We discuss the integration estimate further in Section 2, the ratio estimate in Section 3, and the other new estimates in Section 4. We compare the estimates in Section 5.

The second reason that importance sampling sometimes fails is that it may be difficult to choose a good design distribution. We propose *defensive mixture distributions* as a simple way to bound the weights, and generally to make importance sampling more reliable. We discuss the defensive mixture technique and other practical ways to choose design distributions in Section 6 and demonstrate the estimates and design techniques with a case study involving oil inventory in Section 7.

Throughout this article, X and Q may be either univariate or multivariate. If Q is multivariate, then $\mu = E(Q)$ and $\text{var}(Q)$ are vectors the same length as Q. The use of f, g, and integrals is for notational convenience only; methods and results in this article do not require continuous distributions.

2. INTEGRATION ESTIMATE

The interpretation of importance sampling offered by Hammersley and Handscomb (1964) is that they solve a new problem—instead of estimating $E_f[Q(X)]$, they estimate $E_g[Y(X)]$, where

$$Y(x) = Q(x)\frac{f(x)}{g(x)}.$$

If $g(x) > 0$ when $Q(x)f(x) \neq 0$, the estimate is unbiased and is asymptotically normally distributed with mean μ and normalized variance $\sigma^2_{\text{int}} = \text{var}_g(Y)$ as long as σ^2_{int} is finite.

The optimal design distribution for minimizing the variance of the integration estimate is

$$g^*_{\text{int}}(x) = C|Q(x)|f(x), \tag{2.1}$$

where C is a normalizing constant (Kahn and Marshall 1953).

Hammersley and Handscomb's interpretation corresponds to a transformation strategy rather than to a sampling method. The design distribution g induces a transformation $Q(x) \mapsto Y(x)$ and should be chosen so that Y is more nearly constant than is Q; indeed g^* induces the transformation that makes Y constant (if Q is nonnegative or nonpositive). When estimating the expected value of a variable that is nearly always 0, this transformation idea agrees with the heuristic of sampling the rare (nonzero) events relatively often. For other variables the opposite can be true. For example, if $Q(x) = 1$ with probability .99 and $Q = 0$ with probability .01, a good design distribution samples *less* from the region where $Q = 0$ than does the target distribution.

The integration estimate can fail when there is more than one output of a simulation because every component of Q is transformed by the same ratio f/g. For example, when X has a standard normal distribution, a design distribution $g \sim N(2.326, 1)$ reduces the variability of the estimate of the probability $\Pr(X > 2.326)$ by a factor of 37 compared to simple random sampling, but increases the variance of the estimate of $E[X]$ by a factor of 1,433.

3. RATIO ESTIMATE

The obvious solution to a set of weights that do not sum to 1 is to normalize the weights. The *ratio* estimate is obtained by simple normalization; let

$$V_{\text{ratio},i} = \frac{W_i}{\sum_{i=1}^n W_j}$$

be the ratio weights; then the ratio estimate for μ is

$$\widehat{\mu}_{\text{ratio}} = \sum_{i=1}^n V_{\text{ratio},i}\, Q(X_i) = \overline{Y}/\overline{W}. \tag{3.1}$$

This estimate is consistent iff $g(x) > 0$ whenever $f(x) > 0$ so that $E_g[W(X)] = 1$. It is asymptotically normally distributed with normalized variance

$$\sigma^2_{\text{ratio}} = \text{var}_g(Y - \mu W) \tag{3.2}$$

as long as $g(x) > 0$ whenever $f(x) > 0$ and μ and σ^2_{ratio} are finite (Hesterberg 1991). The design distribution that minimizes (3.2) is

$$g^*_{\text{ratio}}(x) = C|Q(x) - \mu|f(x), \tag{3.3}$$

where C is a normalizing constant if the estimate is consistent (i.e., if $\Pr(Q = \mu) = 0$) (Hesterberg 1988). The optimal design distribution for the ratio estimate samples heavily when Q is relatively far from its mean ($g^*_{\text{ratio}} \propto |Q - \mu|f$), which agrees with the heuristic of

sampling rare events relatively often. In contrast (2.1) samples heavily where Q is far from 0 ($g^*_{\text{int}} \propto |Q-0| f$).

Importance sampling has been used for Bayesian calculations by Kloek and van Dijk (1978) and others by writing an expected value as a ratio of two integration estimates,

$$\mu = \int Q(\theta) p(\theta \mid \text{data})\, d\theta = \frac{\int Q(\theta)\{L(\text{data} \mid \theta)\pi(\theta)/g(\theta)\}g(\theta)\, d\theta}{\int \{L(\text{data} \mid \theta)\pi(\theta)/g(\theta)\}g(\theta)\, d\theta}, \quad (3.4)$$

where θ has prior distribution π and posterior distribution $p \propto \pi L$ and where L is the density of the data given θ, then estimating both the numerator and denominator using observations from the same design distribution g. We prefer to think of (3.4) as a ratio estimate, with $f(\theta) = p(\theta \mid \text{data})$, which is known up to a normalizing constant. The normalizing constant drops out in the computation of the ratio estimate (3.1). This leads naturally to the optimal (asymptotic) design distribution given by (3.3), obtained by Hesterberg (1988) using this approach and Geweke (1989) using another approach.

4. REGRESSION, MAXIMUM LIKELIHOOD, AND EXPONENTIAL ESTIMATES

The ratio estimate differs from the integration estimate by multiplying each weight by the same factor, which is chosen to make the weights sum to 1. There is a problem with this. Suppose, for example, that $\overline{W} < 1$; the ratio estimate increases the weight assigned to all replications, relative to the integration estimate, even though small values of W are already overrepresented in the sample. The estimates in this section improve on the ratio estimate by giving smaller weight to such replications and correspondingly larger weight to replications with larger values of W.

Note that the observations (Q_i, W_i) are bivariate observations from a distribution with $E_g[W] = 1$. We assign *metaweights* to the bivariate observations so that the (meta-) weighted average of W is 1. The metaweights satisfy the *unitary constraints* $\Sigma\pi_i = 1$ and $\Sigma\pi_i W_i = 1$. Then the weights for estimating the distribution of Q under f are the product weights $V_i = \pi_i W_i$.

We consider three sets of metaweights. The *maximum likelihood* (ML) metaweights maximize the empirical likelihood function $\prod_{i=1}^n \pi_{\text{ml},i}$ and minimize the Kullback–Leibler distance $\sum_{i=1}^n u_i \ln(u_i/\pi_i)$ from the weights $\{\pi_i\}$ to the uniform weights $\{u_i = 1/n\}$. The *exponential* metaweights minimize the Kullback–Leibler distance $\sum_{i=1}^n \pi_i \ln(\pi_i/u_i)$ from the uniform weights to the metaweights, whereas the *regression* metaweights minimize both the variance of the metaweights and the Euclidean distance between the metaweights and uniform weights. The solutions can be written in the forms $\pi_{\text{ml},i} = a/(1 - b(w_i - \overline{w}))$, $\pi_{\text{exp},i} = a\exp(b(w_i - \overline{w}))$, and $\pi_{\text{reg},i} = a(1 + b(w_i - \overline{w}))$, with a and b chosen to satisfy the unitary constraints.

The constants for the regression weights can be obtained in closed-form solution. The resulting product weights are

$$V_{\text{reg},i} = W_i \frac{1 + b(W_i - \overline{W})}{n}, \quad (4.1)$$

where $b = (1 - \overline{W})/\widehat{\sigma}^2_W$ and $\widehat{\sigma}^2_W = \frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})^2$. Note that if $\overline{W} < 1$ the regression weights are greater for large W and less for small W than are the integration weights.

Parameters for the other two estimates must be found iteratively. The problem can be reduced to solving the second unitary constraint as a function of b by letting $a = a(b)$ be the value of a that satisfies the first unitary constraint. A good initial guess for b is the regression solution $b = (1 - \overline{W})/\widehat{\sigma}^2_w$.

The regression weights result in an estimate for μ of the form

$$\widehat{\mu}_{\text{reg}} = \sum_{i=1}^n V_{\text{reg},i} Q(X_i) = \overline{Y} - \widehat{\beta}(\overline{W} - 1), \quad (4.2)$$

where $\widehat{\beta}$ is the usual least squares regression slope of Y against W. This estimate can be computed using linear regression without explicitly computing the weights (4.1) and may be computed in a single pass through a simulation to avoid storing values of W and Q; in this case roundoff errors can be reduced by computing $\widehat{\beta}$ as $(\Sigma W_i^\dagger Y_i - n\overline{W}^\dagger\overline{Y})/(\Sigma W_i^{\dagger 2} - n\overline{W}^{\dagger 2})$, where $W_i^\dagger = W_i - 1$, rather than by using the algebraically equivalent formula using W.

For estimating quantiles of the distribution of Q, it is necessary to store values of W and Q, and it is more efficient to explicitly compute the weights (4.1) and use the weights in a weighted empirical distribution for Q than to use (4.2) to estimate many probabilities.

5. COMPARISON OF ESTIMATES

Despite the intuitive appeal of normalizing the weights, the ratio estimate is worse than the integration estimate for many problems, in particular for estimating small probabilities. The regression, ML, and exponential estimates are more accurate (at least for large n) than the integration estimate, by an incredible factor in many cases but by an insignificant factor for estimating small probabilities. Our basic recommendation, which we justify in this section, is to use the integration estimate for estimating small probabilities and the regression estimate for most other problems.

We begin with a simple simulation problem, estimating $\Pr(X > z_\alpha)$, $\Pr(X \le z_\alpha)$, $E[X]$, and a constant, when X has a standard normal distribution, $\alpha = .01$, and $z_\alpha = 2.236$, using $n = 40$ replications. The performance of the five estimates is shown in Table 1 (for three design distributions, which are discussed later). For estimating a small probability $\Pr\{X > z_\alpha\}$, the ratio estimate is worse than the integration estimate, counter to our intuition that weights should sum to 1, because it overreacts to weights observed for observations outside the tail of interest. The

Table 1. Accuracy by Estimate and Design, for Gaussian Example

Design	Method	$P\{X > z_u\}$	$P\{X \le z_u\}$	$E[X]$	$\mu_4 = 1$
f	SRS	2.475(*)	2.475(*)	0.025(*)	0
g_0	Integration	.068	44,800(**)	16.4(**)	4.5(**)
	Ratio	4.14	4.14	.68	0
	Regression	.23	.23	.42	0
	Exponential	.13	.13	.37	0
	ML	.16	.16	.41	0
$g_{0.1}$	Integration	.075	579.4	.096	.058
	Ratio	.240	.240	.115	0
	Regression	.073	.073	.091	0
	Exponential	.072	.072	.090	0
	ML	.073	.073	.090	0
$g_{0.5}$	Integration	.13	59.4	.031	.0057
	Ratio	.16	.16	.032	0
	Regression	.12	.12	.029	0
	Exponential	.12	.12	.029	0
	ML	.12	.12	.029	0

NOTE: Table entries give the mean squared error of 2,000 Monte Carlo experiments, with $n = 40$ observations in each experiment, for $f = N(0,1)$ and $z_u = 2.326$. The design distributions are $g_0 = N(z_u, 1)$, $g_{.1} = 90\%N(z_u, 1)+10\%N(0,1)$, $g_5 = 50\%N(z_u, 1)+50\%N(0,1)$; the latter two are stratified defensive mixtures, with $\lambda = .1$ and $\lambda = .5$. The probabilities $\Pr\{X > z_u\}$ and $\Pr\{X \le z_u\}$ are in units of %; divide entries by 100^2 to obtain mean squared errors for estimates of probabilities. (*) = analytical result; (**) = estimated standard error of this entry exceeds 10% of the entry.

other new estimates do better or worse than the integration estimate, depending on the design distribution.

For estimating quantities other than the small probability, the integration estimate is simply unacceptable because it is so dependent on the sum of the observed weights. The other estimates are all affine equivariant, so they estimate both probabilities equally well, estimate a constant correctly (this may be important for consistency of different output estimates in a simulation, e.g., $\Pr(X > a) + \Pr(X \le a) = 1$), and satisfy the usual algebraic relation for computing variances $\widehat{E}[Q^2] - \widehat{E}[Q]^2 = \widehat{E}[(Q - \widehat{E}[Q])^2]$. The regression, exponential, and ML estimates are nearly equivalent and are better than the ratio estimate.

These empirical impressions are consistent with asymptotic results (Hesterberg 1988, 1991), some of which are shown in Table 2. The regression, ML, and exponential estimates have the same asymptotic variance, which is smaller than that for the integration estimate by a factor of $(1 - \rho^2)$, where ρ is the correlation between W and Y. When estimating a small probability, ρ is usually small enough that these estimates give little improvement over the integration estimate. Furthermore, in small samples, with anticonservative design distributions like g_0 in Table 1, the variance of these estimates can be worse than for the integration estimate.

The new estimates are biased, but with large sample sizes (which are usual in simulation) the bias is negligible. In addition, for small samples the regression weights can be negative. For large-sample simulations, the regression, exponential, and ML estimates are nearly equivalent, but the regression estimate is easier to implement.

The ratio and regression estimates correspond to the estimates with the same names for estimating $E[Y]$ in the presence of a covariate W with known mean (e.g., Cochran 1977), and the regression estimate is equivalent to using W as a control variate for estimating $E[Y]$.

6. DESIGN DISTRIBUTIONS

Design distributions should be easy and fast to generate and should have good statistical properties. In this section we discuss what gives distributions good statistical properties and recommend a simple sampling technique, *defensive mixture sampling*, that helps achieve those properties. This technique overcomes the difficulty that led Bratley, Fox, and Schrage (1983, p. 66) to conclude, "Because there is no practical way to guard against gross misspecification of g, multidimensional importance sampling is risky."

We follow two general principles—the likelihood ratio g/f should be (1) relatively large in "important" regions of the sample space but (2) not too small anywhere. Consistency and robustness require the latter; efficiency requires both. The second principle means that the weight function $W(x) = f(x)/g(x)$ must always be finite and in practice should not be extremely large.

6.1 Defensive Mixture Distributions

We propose the use of *defensive mixture distributions* (Hesterberg 1987, 1988) in which the design distribution has the form

$$g_\lambda(x) = \lambda f(x) + (1 - \lambda)g_0(x), \qquad (6.1)$$

Table 2. Asymptotic Variance and Bias of Importance-Sampling Estimates

Method	Variance (order n^{-1})	Bias (order n^{-1})
Integration	$\text{var}_g(Y)$	0 (unbiased)
Ratio	$\text{var}_g(Y - \mu W)$	$-E_g[W*(Y - \mu W)]$
Regression	$\text{var}_g(Y - \beta W)$	$-E_g[(W - 1)^2(Y - \mu - \beta(W - 1))]/\text{var}_g(W)$
Exponential	$\text{var}_g(Y - \beta W)$	$-\frac{1}{2}E_g[(W - 1)^2(Y - \mu - \beta(W - 1))]/\text{var}_g(W)$
ML	$\text{var}_g(Y - \beta W)$	$0 + O(n^{-2})$

NOTE: $Y = Y(X)$, $W = W(X)$, and β is the regression slope of Y on W. The variance of the regression, exponential, and ML estimates is the smallest possible of the form $\text{var}_g(Y - cW)$ and equals $(1 - \rho^2)\text{var}_g(Y)$, where ρ is the correlation of Y and W.

where $0 < \lambda < 1$ and g_0 is another design distribution, which may be created using exponential tilting or general mixtures, described later, or other techniques of Moy (1965), Goyal, Heidelberger, and Shahabuddin (1987), and Hesterberg (1988).

Defensive mixture distributions make importance sampling robust in several ways. First, the weight function is bounded by $1/\lambda$ because

$$W(x) = \frac{f(x)}{g_\lambda(x)} \le \frac{f(x)}{\lambda f(x)} = \frac{1}{\lambda}. \tag{6.2}$$

The term "defensive" refers to the use of f as one component of the mixture to bound W.

Second, the asymptotic variance of any of the equivalent estimates is guaranteed no worse than $1/\lambda$ times the variance of the simple Monte Carlo estimate. This result follows from (3.2) and (6.2) for the ratio estimate

$$\begin{aligned}\sigma^2_{\text{ratio}} &= \int \frac{(Q(x)-\mu)^2 f(x)^2}{g(x)}\,dx \\ &\le \int \frac{(Q(x)-\mu)^2}{\lambda} f(x)\,dx = \frac{\sigma^2_{\text{SRS}}}{\lambda}.\end{aligned}$$

The same result holds for the other equivariant estimates because their asymptotic variance is less than or equal to that of the ratio estimate. This bound is useful with multiple outputs because the design distribution may be chosen to improve the estimate for one output without the danger that estimates for other outputs will fail badly, as observed in Table 1 and in Table 5, Section 7.

Third, results are much less dependent on the exact shape of g_0 than they would be if g_0 were used alone. In particular, the coverage provided by g_0 can be much less than the optimum design g^* [(2.1) and (3.3)] in some regions of the sample space without disastrous effects on efficiency because f provides some coverage everywhere. The mixture prevents accidental underrepresentation of some region of the sample space. See the discussion of the two-component mixture in the oil-inventory problem that follows for an example.

When used in a defensive mixture, g_0 should mimic g^* when g^* is greater than f (but need not mimic f when g^* is smaller than f). From (2.1) and (3.3) we see that g^*/f is proportional to $|Q(x) - c|$, where $c = 0, \mu$ for the integration and ratio estimates, respectively. So g_0 should sample heavily from "important" regions of the sample space that correspond to the tails of Q, ideally with coverage proportional to $|Q - c|$. For the regression estimate we suggest using the same heuristic, with $c = 0$ if the most important and difficult aspect of the simulation is estimating a small probability, and otherwise $c \approx \mu$.

6.2 Choice of λ

Results are not unduly sensitive to the choice of λ as long as λ is not near 0 (Hesterberg 1988). We generally use λ between .1 and .5. A large value of λ is more conservative and is appropriate in more difficult problems in which g_0 does not mimic g^* well, when there are many different output quantities to estimate or when the ratio estimate is used. A smaller value of λ is appropriate if a simulation demands accurate estimation of a small probability and the integration or regression estimate is used. In favorable circumstances the integration estimate succeeds with $\lambda = 0$. See Table 1 for empirical results in a simple problem, with $\lambda = 0, .1$, and .5.

It is also possible to choose λ adaptively or from a trial study. The asymptotic variance of each estimate (Table 2) can be written as an integral that depends on the design distribution. The asymptotic variance for any proposed design can be estimated from an available sample from some other design (a trial study, or an initial set of observations), by treating the proposed design as a target distribution. Within the class of proposed designs that differ only in the value of λ it is possible to choose the design that minimizes the estimated asymptotic variance.

Moy (1965) used a similar procedure to optimize over other classes of proposed designs. This results in some bias, however, particularly for flexible classes of designs and for the integration estimate, for which the estimated minimum-variance design is degenerate with support solely on the current observations.

6.3 General Mixture Distributions

General mixture distributions of the form

$$g(x) = \sum_{k=1}^{K} \lambda_k g_k(x), \tag{6.3}$$

where $\Sigma\lambda_k = 1$, $\lambda_k > 0$, and possibly $g_1 = f$, provide a way to create design distributions for several situations.

In simulations with multivariate output, different components of the mixture may be tailored to estimate different output variables. When both tails of a single output are important (for estimating moments or for estimating quantiles in both tails), one component of a mixture can shift all the input variables toward one tail and another component toward the other tail. Hesterberg (1988) showed that this is preferable to using an unmixed g under which the variance of each input variable is increased because of the large conditional variance of W given Q under the latter design.

If the target distribution f cannot be used in a defensive mixture (because f is unknown or it is impractical to generate observations from f), then two or more components of a mixture may combine to ensure that no region of the sample space is badly underrepresented. For example, one component may mimic f in the center of its distribution, while another has tails that are at least as heavy as those of f.

Van Dijk and Kloek (1983) used mixture distributions in Bayesian estimation; their mixtures had disjoint support and formed a piecewise approximation to the posterior distribution f. The mixtures here have overlapping support, ideally include f as one component, and mimic a distribution that has heavier tails than f.

Table 3. Standard-Error Formulas With and Without Stratification of Mixtures

Method	*Unstratified*	*Stratified*
Integration	$\frac{1}{n(n-1)}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$\frac{1}{n(n-k)}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ki}^*)^2$
Ratio	$\frac{1}{n(n-1)}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{\text{ratio}} W_i)^2$	$\frac{1}{n(n-k)}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ki}^* - \hat{\mu}_{\text{ratio}} W_{ki}^*)^2$
Regression	$\frac{1}{n(n-2)}\sum_{i=1}^{n}(Y_i - \hat{\beta} W_i - \bar{Y} + \hat{\beta}\bar{W})^2$	$\frac{1}{n(n-k-1)}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(Y_{ki}^* - \hat{\beta} W_{ki}^*)^2$

NOTE: Squared standard errors, for mixture distributions (6.3). In the stratified case exactly n_k observations are generated from distribution $g_k, k = 1, \ldots, K$. Here Y_{ki} is the ith observation from distribution k, $\bar{Y}_k = n_k^{-1}\sum_{i=1}^{n_k} Y_{ki}$, $Y_{ki}^* = Y_{ki} - \bar{Y}_k$, and W_{ki}, $\bar{W}_k$, and W_{ki}^* are similarly defined. The standard errors for the exponential and ML estimates are the same as for the regression estimate.

6.4 Stratified Mixture Proportions

For any mixture distribution, the best implementation strategy is to stratify the mixture proportions. For a defensive mixture this means to generate exactly $n\lambda$ replications from f and $n(1-\lambda)$ from g_0. The alternative to stratifying is to independently generate replications from f with probability λ, else from g_0. Stratifying is more accurate, resulting in bias and variance smaller than in Table 2. Numerical examples in this article are stratified. In either case W_i is computed without regard to which distribution was actually used to generate observation i.

Standard errors with and without stratification are given in Table 3. These standard errors may be untrustworthy, however, even for large n because the distributions of W and Y may be very skewed (this is a problem in any simulation involving rare events, not just in importance sampling). Three diagnostic procedures are available: $\overline{W}$ should be close to 1, $(\Sigma W_i)^2/(\Sigma W_i^2)$ (a measure of the effective sample size in weighted estimation) should be reasonably large, and the sum of squares in the standard-error formula should not be dominated by a small number of the n squares.

Stratification seems to be especially useful for the integration and ratio estimates. In the oil problem in Section 7, for example, stratification improves the efficiency of the integration, ratio, and regression estimates by factors of 4.0, 1.1, and 1.003, respectively, for estimating expected inventory cost and by factors of 1.3, 1.9, and 1.009 for shortage cost. It may be possible to improve the stratified regression estimate by replacing $\hat{\beta}$ in (4.2) and Table 3 with $\hat{\beta} = (\Sigma_k \Sigma_i Y_{ki}^* W_{ki}^*)/(\Sigma_k \Sigma_i W_{ki}^{*2})$, which minimizes the estimated standard deviation in Table 3.

6.5 Exponential Tilting

The most common design technique for importance sampling is exponential tilting (Clark 1966), in which g has marginal distributions of the form

$$g_j(x_j) = c_j(\alpha)\exp(\alpha t_j x_j) f_j(x_j), \tag{6.4}$$

where α determines the overall amount of tilting, t_j determines the amount of tilting for variable j relative to other variables, and c_j is a normalizing constant. If the original joint distribution f has independent marginal distributions, then (unmixed) exponential tilting produces

$$\begin{aligned} g(x) &= \prod_{j=1}^{d} g_j(x_j) = \prod_{j=1}^{d} c_j(\alpha)\exp(\alpha t_j x_j) f_j(x_j) \\ &= c(\alpha)\exp\left(\alpha \sum_{j=1}^{d} t_j x_j\right) f(x), \end{aligned} \tag{6.5}$$

where the normalizing constant is the product of the marginal normalizing constants.

Note that the weight function $W = f/g$ in (6.5) depends solely on a linear combination of the input variables. Exponential tilting is most useful for estimating tail probabilities and moments for variables that are (nearly) linear combinations of the X's because the conditional variance of W given such a variable is (nearly) 0. For nonlinear statistics the conditional variance of W makes the variance of $Y = WQ$ larger.

To estimate $\Pr(\sum_{j=1}^{d} t_j x_j < k)$, we choose α so that $E_g[\sum_{j=1}^{d} t_j X_j] = k$. This choice of α minimizes the maximum value of W over the tail region, if input distributions are normal, is nearly optimal in many applications, and is more robust to nonlinearity and nonnormality than the choice given by, for example, Johns (1988).

The t_j's can be chosen based on knowledge of the physical problem, or they may be set to the multiple regression coefficients for Q against the input variables, using $n\lambda$ replications from f as data. We use a combination of the physical method and the data-based regression method in the "two-component" design distribution in the oil problem that follows.

In practice output variables are rarely exactly linear combinations of the input variables and, because the weight function W that results from exponential tilting is unbounded, exponential tilting should be used in combination with mixture distributions.

7. OIL SIMULATION

The importance-sampling ideas presented here originated in work I performed on a Monte Carlo model designed to evaluate electric-power-plant oil-inventory stocks at a large western utility (Hesterberg 1987). The utility has a diverse electrical generation system and burns oil only when the other generation is inadequate. This hap-

pens only in winter and even then infrequently. A combination of high demand and low generation from other facilities over a sustained period, however, could exhaust the oil stocks and result in a potentially major long-term "shortage" in which there is simply not enough oil on hand to generate sufficient electricity to satisfy all customers. The size and complexity of the model and the importance of the results (inventory was valued at about $200 million, and potential shortage costs are many times greater) make an efficient estimation scheme important.

The most important question is how much oil to carry in inventory at the start of a winter: If too little oil is carried, the risk of shortage and the expected shortage costs is high; if too much is carried, the inventory costs are high. The determination of oil level involves balancing those costs and also balancing expected total cost against the insurance value of carrying higher inventory levels.

Two characteristics of this problem make importance sampling necessary:

1. Shortages are rare enough (probability about .003 in our example) that simple Monte Carlo sampling would require a prohibitively large number of replications to accurately assess shortage probability or the distribution of shortage magnitude.
2. Other variance-reduction techniques, including antithetic variates, control variables, and stratified sampling, are relatively ineffective due to the large number (about 900) of input random variables, varicty of input distributions, and nonlinearity of the problem.

We use a simplified version of the original model in this article to demonstrate the estimation and sampling methods discussed previously. We will see that a relatively simple combination of exponential tilting with defensive mixtures performs reasonably well, 25 times more efficiently than simple Monte Carlo sampling for estimating average shortage cost, but that a more complicated, general mixture model is 200 times as efficient. There are other quantities of interest that are generally easier to estimate than shortage costs and so were not considered in the importance-sampling design; even so, for these quantities the new methods do about as well as simple Monte Carlo, so there is no need to estimate those quantities from a separate simulation. In contrast, without a mixture the integration estimate would do 3,000 times worse and the regression estimate 20 times worse than simple Monte Carlo for these quantities.

We simulate a three-month winter, with random variables aggregated monthly and independence between months. The random inputs are hydro $\sim$ gamma[μ = (500, 600, 600), shape = (5, 6, 7)], nuclear $\sim$ density $\propto \exp(x/100)$ for $0 < x < 300$, temperature $\sim N[(54, 52, 55), (5^2, 5^2, 5^2)]$, degree days = max(0, 60 − temperature), electric demand $\sim N[(1{,}600, 1{,}650, 1{,}600), (100^2, 100^2, 100^2)]$ + 10 degree days, and gas demand $\sim N[(1{,}600, 1{,}700, 1{,}600), (100^2, 100^2, 100^2)]$ + 40 degree days. Variables are vector quantities unless specified, and the maximum is taken for each month $m = 1, 2, 3$ independently. There are 500 units of other electrical generation and 2,500 units of gas available each month, and the initial oil inventory is $\text{inventory}_0 = 1{,}200$ (in the original application the model was used to select an initial inventory level, by comparing simulations from different initial levels). Then gas available = max(0, 2,500 − gas-demand), balance = electric demand − 500 − hydro − nuclear − gas available, oil need = max(0, balance), and $\text{inventory}_m = \max(0, \text{inventory}_{m-1} - \text{oil need}_m)$. The following quantities are scalar, and costs are based on an inventory cost rate of 1 (dollars/barrel per month) and a shortage cost rate of 80 (dollars/barrel): inventory cost = $\sum_{m=1}^{3} \text{inventory}_m$, shortage = $1{,}200 - \sum_{m=1}^{3} \text{oil need}_m$, shortage cost = 80 shortage, and total cost = inventory cost + shortage cost.

The problem would be linear and ideal for exponential tilting if the relationship between temperature and demand were linear and if oil need would depend only on the sum of demands minus the sum of supplies (i.e., excess energy from one month could be used in another month), in which case a shortage would occur iff $\Sigma \tau_j x_j > 1{,}200 + 3(500 + 2{,}500)$, where the τ_j scale all variables to common energy units (from gigawatt-hours, barrels, cubic feet, or degrees fahrenheit; in this article electricity, oil, and gas are already in common units) and are positive for demand variables and negative for supply variables. Then τ_j can be used for t_j in (6.4) to increase the sampling frequency of shortages.

The relationship between temperature and heating demands is nonlinear, however, and there are nonlinear temporal effects because extra supplies in one month cannot be used in another month. A two-month example of this is shown in Figure 1. The shortage region corresponds to a net "balance" (Σ demands − Σ supplies) that exceeds the initial inventory level for either month individually or

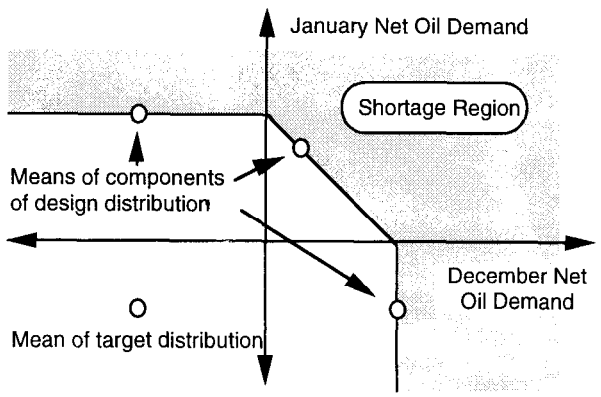

Figure 1. A Shortage Occurs if the December Balance, the January Balance, or the Sum of the Two Exceeds 1,200, Where Balance = Net Oil Demand = Σ Demands − Σ Supplies. The design distribution covers the shortage region by using one component for each set of months that can combine to create a shortage.

for the sum of both months. A linear problem would have a linear boundary for the shortage region.

Simple (unmixed) exponential tilting moves the mean of the design distribution up and to the right in Figure 1 and the weight function increases down and to the left; this weight function is unbounded, even within the shortage region. This unboundedness may be impossible to diagnose in a simulation using observed values of W because there might not be any replications observed from the underrepresented part of the shortage region. But it can be easily prevented using a defensive mixture distribution.

We consider two designs here. The first is a relatively quick-and-dirty design, consisting of a mixture of the target distribution with $\lambda = .5$ and a single exponentially tilted alternate distribution g_0 of the form (6.5), with $t_j = r_m \tau_j$, where r_m is a factor that depends on the likelihood that month m will contribute to a shortage; for example, January values are tilted more than February values, and the τ values normalize to common energy units (assuming cold weather for the temperature variables). We choose the r_m to maximize the empirical correlation of oil requirements with the linear combination $\Sigma r_m \tau_j x_j$ of input factors, using data from the initial $\lambda n = 250$ observations, and choose the overall tilting parameter α so that the mean of the linear combination under g_0 is approximately at the 99th percentile of distribution of the linear combination under f.

A more efficient, though more complicated, design uses an alternate distribution, which is itself a mixture of seven component distributions, which correspond to the $2^3 - 1$ nonempty subsets of months. For each such component, only the variables in the corresponding subset are tilted. This gives g of the form $g(x) = \lambda f(x) + (1 - \lambda)\Sigma P_M g_M(x)$, where M indexes the subsets (e.g., "Dec&Jan"), P_M are mixing proportions, and $g_M(x) = C_M \exp\{\alpha_M \Sigma t_j x_j\} f(x)$ is an exponentially tilted distribution in which only variables in the corresponding subset are tilted; that is, $t_j = \tau_j$ if variable j occurs in one of the months in M, otherwise $t_j = 0$.

The tilting parameters α_M in this design tilt the input distributions so that the mean oil requirement for the corresponding subset of months is on the boundary of the shortage region, $\mu_M \equiv E_{g_M}[X] \in \partial\{\text{shortage region}\}$. The three points on the border of the shortage region in Figure 1 correspond to the values of μ_M for the $2^2 - 1 = 3$ combinations of months in that two-month figure. The mean of the target distribution is in the lower left quadrant because in both January and December the expected balance is negative. The topmost dot is the mean of a (component) design distribution in which January variables are tilted to increase demands and reduce supplies so that the (tilted) mean January balance is 1,200, on the border of the shortage region. The other component distributions have just December, or both January and December, variables tilted so that the mean balance is on the border of the shortage region.

The mixture proportions for the larger design are proportional to the likelihood ratio between the component density and the target density evaluated at the mean, $P_M \propto f(\mu_M)/g_M(\mu_M)$. This approximately minimizes the maximum W value over the shortage region, though not exactly because the components do not have disjoint support. The mixing proportions are shown in Table 4. Note that the mixing proportions reflect the relative likelihoods of shortages; for example, shortages are much more likely to occur from unfavorable circumstances in December and January than from unfavorable circumstances in February alone. The smaller components could be eliminated to simplify the simulation.

Numerical results for both designs and for all estimates are shown in Table 5. All estimates perform well for estimating shortage cost and shortage probability—these quantities depend solely on rare events, and fit into the traditional importance-sampling scheme. For example, the regression estimate is about 238 times as efficient as simple Monte Carlo sampling for estimating the expected shortage cost, and the integration estimate is about 212 times as efficient. The new estimates are also reasonably

Table 4. Mixing Proportions for the Eight-Component Mixture Design

Monthly tilted	*Mean of balance*	*St. dev. of balance*	*# st. dev.'s from shortage*	*Proportion of g_0*	*Proportion of design*	*Observations when $n = 500$*
Dec	−216	371	3.82	.007	.007/2	2
Jan	−66	384	3.30	.056	.056/2	15
Feb	−416	397	4.20	.001	.001/2	1
Dec & Jan	−282	534	2.72	.472	.472/2	117
Dec & Feb	−632	543	3.46	.036	.036/2	9
Jan & Feb	−482	552	3.13	.127	.127/2	32
Dec–Feb	−698	665	2.93	.302	.302/2	74
Original(f)					.5	250

NOTE: "Balance" is the sum of demands minus the sum of supplies. The initial oil inventory is 1,200, so a December mean balance of −216 implies that the balance must be 1,416 units from its mean (under the target distribution) to cause a shortage in that month. The first component of the mixture has (only) December variables tilted to modify the mean balance by 1,416 units, which represents 3.82 times the standard deviation of the December balance. The design is a defensive mixture with $\lambda = .5$. The number of observations from each component are stratified by multiplying each proportion by $n = 500$ and rounding smaller numbers up and larger numbers down; values of λ are adjusted accordingly prior to computing W.

Table 5. Efficiency by Design and Estimate for the Oil Simulation

Simulation output	Integration	Ratio	Regression, ML, and exponential
Two-component mixture			
Shortage cost	26.3	25.1	26.8
Inventory cost	.09	1.14	1.32
Total cost	.27	2.95	3.19
December inventory	.05	0.90	.93
January inventory	.14	1.21	1.44
February inventory	.16	1.15	1.35
Shortage probability	12.3	11.8	12.6
Eight-component mixture			
Shortage cost	212	175	238
Inventory cost	.14	1.01	1.04
Total cost	.42	2.85	2.89
December inventory	.07	.83	.84
January inventory	.21	.99	1.02
February inventory	.24	1.05	1.09
Shortage probability	55	50	58

NOTE: Entries in this table represent the estimated efficiency relative to simple random sampling, $\text{var}_f(\bar{Q})/\text{var}_g(\hat{\mu})$, based on 1,600 experiments with $n = 500$ replications in each experiment. The design distributions are described in the text. The standard errors of the estimates in this table are less than 1/26 times the corresponding estimate, except for the two-component mixture estimates for shortage cost and shortage probability, which are around 1/12. The regression, ML, and exponential results are the same, to the number of digits shown in the table.

accurate for estimating other outputs, but the integration estimate is not; for example, for estimating one key output, expected total cost, the regression and integration methods are about 2.89 and .42 times as efficient as simple Monte Carlo, respectively.

The two-component and eight-component designs are roughly comparable for estimating most quantities, but the eight-component design is more accurate for shortage cost and probability. In the original application it is important to estimate the derivative of total cost with respect to the initial inventory level accurately because the minimum cost point is where the derivative is equal to 0. The derivative is closely related to the shortage probability so that the extra effort required for the eight-component design is justified by its accuracy for estimating the shortage probability.

8. SUMMARY

Importance sampling has traditionally been used for estimating probabilities of rare events. In rare-event problems with nice structure, there may be no need to use either the new estimates or design methods described in this article. The methods in this article are intended for more general simulation problems in which the structure of a problem is not nice, in which more than one output quantity is to be estimated, or in which importance sampling is used by necessity rather than choice.

Of the estimation techniques discussed in this article, the classical *integration* estimate is adequate for estimating small probabilities, and the new *regression* estimate is preferred in most other problems. The importance sampling found in the Bayesian literature is equivalent to use of the *ratio* method, which proved to be less efficient; I conjecture that adaptations of the regression method may be valuable in Bayesian analysis.

The relatively simple *defensive mixture* design technique makes importance sampling robust. General mixture distributions can also play a role in simulation designs, particularly in simulations in which more than one tail of a variable or more than one variable are of interest.

COMPUTATIONAL DETAILS

All simulations were run in S-*PLUS* (Becker and Chambers 1984; Statistical Sciences Inc. 1991). Programs are available via electronic mail from the author, T_Hesterberg@FandM.edu.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grant DMS89-05874 and Foundation for the Improvement of Post-Secondary Education Grant G008730432 of project 116BH71444. I thank Brad Efron, Art Owen, Vernon Johns, Robert Tibshirani, and Paul Gribik for helpful discussions, Kiril Selverov for programming assistance, and the referees for comments that considerably improved the presentation of this work.

[*Received August 1992. Revised September 1994.*]

REFERENCES

Becker, R. A., and Chambers, J. M. (1984), *S: An Interactive Environment for Data Analysis and Graphics*, Belmont, CA: Wadsworth.

Beckman, R. J., and McKay, M. D. (1987), "Monte Carlo Estimation Under Different Distributions Using the Same Simulation," *Technometrics*, 29, 153–160.

Bratley, P., Fox, B. L., and Schrage, L. E. (1983), *A Guide to Simulation*, New York: Springer-Verlag.

Clark, F. H. (1966), "The Exponential Transform as an Importance-Sampling Device: A Review," Technical Report ORNL-RSIC-14, Oak Ridge National Laboratory.

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 74, 1317–1339.

Goyal, A., Heidelberger, P., and Shahabuddin, P. (1987), "Measure Specific Dynamic Importance Sampling for Availability Simulations," in *Proceedings of the 1987 Winter Simulation Conference* (Atlanta, December 14–16), eds. A. Thesen, H. Grant, and W. D. Kelton, New York: Institute of Electrical and Electronic Engineers, pp. 351–357.

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

Hesterberg, T. C. (1987), "Importance Sampling in Multivariate Problems," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 412–417.

——— (1988), "Advances in Importance Sampling," unpublished Ph.D. thesis, Stanford University, Statistics Dept.

——— (1991), "Importance Sampling for Bayesian Estimation," in *Computing And Graphics In Statistics*, (Vol. 36 of the Institute for Mathematics and Its Applications Volumes in Mathematics and Its

Applications), eds. A. Buja and P. Tukey, New York: Springer-Verlag, pp. 63–75.

Hopmans, A., and Kleijnen, J. P. C. (1979), "Importance Sampling in Systems Simulation: A Practical Failure?" *Mathematics and Computers in Simulation*, 21, 209–220.

Johns, M. V. (1988), "Importance Sampling for Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 83, 709–714.

Kahn, H. (1950), "Random Sampling (Monte Carlo) Techniques in Neutron Attenuation Problems, I and II," *Nucleonics*, 6(5), 27–37, and 6(6), 60–65.

Kahn, H., and Marshall, A. W. (1953), "Methods of Reducing Sample Size in Monte Carlo Computations," *Journal of the Operations Research Society of America*, 1, 263–278.

Kloek, T., and Van Dijk, H. K. (1978), "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica*, 46, 1–19.

Moy, W. A. (1965), "Sampling Techniques for Increasing the Efficiency of Simulations of Queueing Systems," unpublished Ph.D. dissertation, Northwestern University, Industrial Engineering and Management Science.

Statistical Sciences, Inc. (1991), *S-PLUS Reference Manual, Version 3.0*, Seattle: Author.

Therneau, T. M. (1983), "Variance Reduction Techniques for the Bootstrap," Technical Report 200 (Ph.D. thesis), Stanford University, Dept. of Statistics.

Tibshirani, R. J. (1984), "Bootstrap Confidence Intervals," Technical Report LCS-3, Stanford University, Dept. of Statistics.

Tukey, J. W. (1987), "Configural Polysampling," *SIAM Review*, 29, 1–20.

Van Dijk, H. K., and Kloek, T. (1983), "Experiments With Some Alternatives for Simple Importance Sampling in Monte Carlo Integration," in *Bayesian Statistics 2*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Amsterdam, Elsevier, pp. 511–530.