Adaptive Allocation for Importance Sampling*

Ronald F. Peierls
Brookhaven National Laboratory

and

Joseph A. Yahav
Hebrew University

The problem of estimating optimally a linear combination of means from populations with different and unknown variances is considered. An asymptotically pointwise optimal solution of sequential allocation of the observations is provided.

1. Introduction.

The problem of interest is that of estimating a linear combination of means from several populations. The estimator is the linear combination of the sample means and the question addressed is how to allocate a fixed number of allowed observations between the different populations.

The particular context which motivated this work was Monte Carlo quadrature, where efficiency can be increased by partitioning the integration region and then sampling with different densities in different regions. The cost of each evaluation of the integrand leads to a constraint on the total

---

*Research supported under contract DE-AC02-76CH00016 with the U.S. Department of Energy.

AMS 1980 subject classifications. Primary 62F12; Secondary 62F35.

Key words: sequential, allocation, estimation, linear combination of means.

number of samples. However, the discussion is of more general relevance.

To determine the allocation we will adopt minimum variance as a criterion. When the variances of the populations are known this problem has a well-known solution. When the variances are not known the problem becomes somewhat more complicated. In our framework we allow sequential allocation. The solution to the problem is an extension of work by Robbins, Simons and Starr (1967). In Section 2 we give an allocation rule that yields the optimal rule when the variances are known and is asymptotically pointwise optimal if the variances are not known and we use consistent estimators of the variances for each population.

In Section 3 we construct the linear empirical Bayes estimators for estimating the variances simultaneously given by Robbins (1982). These estimators are consistent and yield somewhat better results for moderate number of observations per population, especially for populations where the fourth moment is very large or does not exist. In Section 4 we describe three Monte Carlo experiments which demonstrate the operating characteristic of our procedure, we compare our procedure with the same procedure using the classical estimators for the variances, and with one for known variances.

The problem of sequential allocation was addressed in a technical report by Halton and Zeidman (1971) from the point of view of achieving a given required accuracy at a certain confidence level. In their procedure the number of observations required is a random variable.

In our paper the number of observations are predetermined and the problem is to minimize the mean square error of the estimator. Here the accuracy achieved is a random variable.

2. The allocation rule.

We start by defining our problem in mathematical terms.

Let X_{ij}, $i=1,2,\cdots, m$ $j=1,2,\cdots, n_i$ be independent random variables, where X_{ij}, $j=1,\cdots, n_i$, are identically distributed with unknown distribution

F_i. We assume that F_i is non-degenerate and strictly increasing on its support, that $E(X_{ij}^2)$ are finite and that

$$n_i^{-2} \sum_{i=1}^{n_i} X_{ij}^4 \to 0 \quad \text{as } n_i \to \infty \text{ for } i=1,2,\cdots,m \quad \text{almost surely.}$$

Let $E[X_{ij}] = \mu_i$ and $V(X_{ij}) = \sigma_i^2$.

We are interested in estimating $\mu = \sum_{i=1}^{m} c_i \mu_i$ where $c_1, c_2, \cdots, c_m$ are known constants. Without loss of generality we take the c_i's to be non negative. (If c_i is negative then in all formulas relating to variances and standard derivations we use $\sqrt{c_i^2} = |c_i|$.) Let N_0 denote the number of observations we are permitted to use. That is,

$$\sum_{i=1}^{m} n_i \leq N_0 . \tag{2.1}$$

If we use as our estimator:

$$\hat{\mu} = \sum_{i=1}^{m} c_i X_i, \text{ where } X_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \tag{2.2}$$

then we have

$$E[\hat{\mu}] = \mu, \quad \text{and} \quad \mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{m} \frac{\sigma_i^2 c_i^2}{n_i} . \tag{2.3}$$

Minimising Var $(\hat{\mu})$ with respect to the n_i subject to (2.1), we obtain

$$\mathrm{Var}(\hat{\mu}) \geq \frac{1}{N_0} \left(\sum_{i=1}^{m} \sigma_i c_i\right)^2. \tag{2.4}$$

Defining $\theta_i = \sigma_i c_i / (\sum_{j=1}^{m} \sigma_j c_j)$, the condition for the equality in (2.4) to be satisfied is $n_i = \theta_i N_0$. This cannot be satisfied for integer n_i in general, but we can approximate this most efficient allocation when N >> m as in Lemma 2.1.

LEMMA 2.1. Let $N = N_0 - m$ and define $n_i^* = \min\{n_i : n_i \geq \theta_i N\}$.

Then

(2.5) $$N \leq \sum_{i=1}^{m} n_i^* \leq N_0$$

and for this allocation

(2.6) $$\frac{(\sum_{i=1}^{m} \sigma_i c_i)^2}{N_0} \leq \mathrm{var}(\hat{\mu}) \leq \frac{(\sum_{i=1}^{m} \sigma_i c_i)^2}{N}$$

Proof: By definition

$$n_i^* \geq \theta_i N \text{ and } n_i^* - 1 < \theta_i N$$

Hence

$$\sum_{i=1}^{m} n_i^* \geq N \sum_{i=1}^{m} \theta_i = N \text{ and } \sum_{i=1}^{m} n_i^* - m < N \sum_{i=1}^{m} \theta_i = N ,$$

which yields (2.5).

To prove (2.6) we note that

$$\mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{m} \frac{\sigma_i^2 c_i^2}{n_i^*} \leq \sum_{i=1}^{m} \frac{\sigma_i^2 c_i^2}{\theta_i N} = \frac{(\sum_{i=1}^{m} \sigma_i c_i)^2}{N} .$$

The other inequality in (2.6) is simply the bound (2.4).

For $m = N_0 - N << N_0$, (2.6) shows that the allocation n_i^* approaches the optimal allocation. It is not easily achieved, however, since it requires prior knowledge of all the variances σ_i^2 , whereas in most problems the variances are not known initially. To overcome this difficulty we propose a sequential allocation rule which requires only consistent estimators of the variances at each stage and which approaches the same result. Let k be a positive integer such that km << N. The allocation rule is then:

"At the first stage take k observations from each population, estimate σ_i by $\hat{\sigma}_i$; at each subsequent stage w take one observation from each population for which

$$n_i^{(w)} \leq \hat{\theta}_i \min \{ \sum_{i=1}^{m} n_i(W), N\}$$

where
$$\hat{\theta}_i = \hat{\sigma}_i c_i / (\sum_{j=1}^{m} c_j \hat{\sigma}_j):$$

continue as long as $\Sigma_{i=1}^{m} n_i^{(w)} < N + m$. If the constraint on the number of observations does not permit taking all the observations required by the above rule, then take as many as possible according to the order i=1,2,$\cdots$,m."

THEOREM 2.1: Assume $\hat{\sigma}_{in_i}$ to be consistent and positive estimators of σ_i. Let $V_N^* = \frac{1}{N} (\Sigma_{i=1}^{m} c_i \sigma_i)^2$ and let $V_N = E_N[(\hat{\mu} - \mu)^2]$ where $\hat{\mu}$ is obtained by the above rule. Then,

$$\lim_N \frac{V_N}{V_N^*} = 1 \tag{2.7}$$

Convergence and limits are taken to be almost sure unless otherwise stated.

Proof: We will proceed by proving 4 lemmas.

LEMMA 2.2. Let T_i^* denote the resulting number of observations allocated to population i by the allocation rule. Then

$$T_i^* \geq N\hat{\theta}_i, \; i=1,2,...m \tag{2.8}$$

and

$$(2.9) \qquad \sum_{i=1}^{m} \frac{\hat{\sigma}_i^2 c_i^2}{T_i^*} \leq \frac{1}{N} (\Sigma \hat{\sigma}_i c_i)^2$$

Proof: It is sufficient to show (2.8) since (2.9) follows directly. Suppose (2.8) is not correct for some i: without loss of generality we assume

$$(2.10) \qquad T_i^* \geq N\hat{\theta}_i \quad i = 1,2,\ldots s$$

$$(2.11) \qquad T_i^* < N\hat{\theta}_i \quad i = s+1,\ldots m$$

Let w_i be the stage at which the final allocation was made to population i: the allocation rule implying

$$(2.12) \qquad T_i^* - 1 \leq \hat{\theta}_i \,.\, \min \{ \sum_{j=1}^{m} n_j(w_i), N\} \leq \hat{\theta}_i \,.\, N$$

Using (2.11) for i=1,2,...s, and (2.10) for i=s+1,...m and summing, we obtain

$$(2.13) \qquad \sum_{i=1}^{m} T_i^* - s < N$$

and hence

$$\sum_{i=1}^{m} T_i^* < N+s < N_0$$

Thus, since the allowed number of observations is not exhausted, the stopping rule implies that

$$(2.14) \qquad T_i^* > \hat{\theta}_i \min \{ \sum_{j=1}^{m} T_j^* , N\} \quad i=1,2,\ldots m.$$

If $N < \sum_{j=1}^{m} T_i^*$, then this contradicts (2.11), otherwise,

$$(2.15) \qquad T_i^* > \hat{\theta}_i \sum_{j=1}^{n} T_j^*$$

and summing over i we again get a contradiction. Thus the assumption that (2.8) is false always leads to a contradiction, which proves the lemma.

LEMMA 2.3. Let X_j, j=1,2,... be an infinite sequence of i.i.d random variables with $E[X_i^2] < \infty$, $E[X_i] = \mu$ and $V(X_i) = \sigma^2$. For each stopping time T we define

$$S_T = \sum_{i=1}^{T} X_i \; , \; \overline{X}_T = S_T / T$$

It follows that

$$E[S_T] = \mu E[T], \; E[(S_T - T\mu)^2] = \sigma^2 E[T].$$

Let T_1, T_2,..., be a sequence of stopping times such that

$\frac{T_n}{n} \leqslant d$, $\frac{T_n}{n} \longrightarrow c$ as $n\to\infty$. Then,

(2.16) $$\lim_n \; T_n E[(\overline{X}_{T_n} - \mu)^2] = \sigma^2.$$

Proof of lemma 2.3: We have

$$\lim_n \frac{T_n E[(\overline{X}_{T_n} - \mu)^2]}{T_n E[(\frac{T_n}{nc})^2 \cdot (\overline{X}_{T_n} - \mu)^2]} = 1 \; ;$$

we note that

$$T_n E[(\frac{T_n}{nc})^2 \; (\overline{X}_{T_n} - \mu)^2 \;] = T_n (\frac{1}{nc})^2 \cdot E[(S_{T_n} - T_n \mu)^2]$$

$$= T_n \; (\frac{1}{nc})^2 \cdot \sigma^2 \; E(T_n).$$

but,

$$\frac{T_n E[T_n]}{(nc)^2} \longrightarrow 1$$

hence (2.16) holds.

LEMMA 2.4. If $\hat{\sigma}_{in} > 0$ and $\lim_n \hat{\sigma}_{in} = \sigma_i$, for i=1,2,···, n, then for T_{iN} defined by our rule $T_{iN} \to \infty$ as $N \to \infty$ and $\hat{\sigma}_{iT_i} \to \sigma_i$.

Proof of Lemma 2.4: As $N \to \infty$, $\frac{\sum_{i=1}^{m} c_i \hat{\sigma}_i}{N} \to 0$. So for any fixed n_i there is N large enough so that $\frac{\hat{\sigma}_i}{n_i} > \frac{\sum_{j=1}^{m} c_j \hat{\sigma}_j}{N}$. This will force us to take at least one more observation from population i. Hence $T_{iN} \to \infty$ which yields $\hat{\sigma}_{iT_i} \to \sigma_i$.

LEMMA 2.5. For the T_i as defined by the above rule and for $\hat{\sigma}_{in_i}$ consistent sequences of positive estimators of σ_i, we have

$$\lim_N \frac{T_{iN}}{N} = \frac{c_i \sigma_i}{\sum_{j=1}^{m} \sigma_j c_j} \quad \text{for } i=1,2,\cdots, m.$$

Proof of lemma 2.5. By lemma 2.2 we have at the last stage of sampling, for large enough N,

$$\frac{\hat{\sigma}_{iT_i} c_i}{T_i} \leq \frac{\sum_{j=1}^{m} \hat{\sigma}_j c_j}{N}$$

and

$$\frac{\sum_{j=1}^{m} \hat{\sigma}_{jT_j} c_j}{N} \leq \frac{\hat{\sigma}_{iT_i-1} c_i}{T_i - 1}$$

hence

$$\frac{\hat{\sigma}_{iT_i} c_i}{\sum_{j=1}^{m} \hat{\sigma}_j c_i} \leq \frac{T_i}{N} \leq \frac{\hat{\sigma}_{iT_i} c_i}{\sum_{j=1}^{m} \hat{\sigma}_j c_j} + \frac{1}{N}$$

as $N \to \infty$, we get the desired result.

Now we proceed with the proof of the theorem. We have:

$$\frac{E[(\hat{\mu}-\mu)^2]}{V_N^*} = \frac{E[(\sum_{i=1}^{m} c_i X_{iT_i} - \sum_{i=1}^{m} c_i\mu_i)^2]}{(\sum_{i=1}^{m} c_i\sigma_i)^2/\ N} = \frac{\sum_{i=1}^{m} c_i^2 E[(X_{iT_i} - \mu_i)^2]}{(\sum_{i=1}^{m} c_i\sigma_i)^2} \cdot N$$

$$= \frac{\sum_{i=1}^{m} c_i^2\ T_i E[(X_{iT_i} - \mu_i)^2]\ \cdot\frac{N}{T_i}}{(\sum_{i=1}^{m} c_i\sigma_i)^2}$$

and taking limits using lemmas 2.3, 2.4, and 2.5 we get (2.16).

3. Estimating many variances simultaneously. We use the notation and set up of Section 2. The classical estimators of the variances are

$$(3.1)\qquad S_{in_i}^2 = \frac{\sum_{i=1}^{n_i} (X_{ij} - X_i)^2}{n_i - 1}$$

These estimators are unbiased and consistent and under the assumption that F_i are strictly increasing on their support the S_i^2's are strictly positive for all n_i with probability one.

One can therefore use these estimators in the rule of Section 2. However, when we deal with a problem where the fourth moment is large or infinite the S_i^2 tend to be somewhat erratic for moderate sample sizes. From M-C experiments we learn that the linear empirical Bayes (l.e.B.) estimators behave somewhat better and we suggest their usage.

We follow Robbins (1982).

Let

{F} be a collection of distribution functions,

$\{F_i$: l=1,2,...m} be i.i.d. from {F},

$\{X_{ij}: j=1,2,\ldots n_i\}$ be i.i.d. from F_i, and assume $E[X_{ij}^4] < \infty$.

Define

$$X_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \ , \quad S_i^2 = \frac{1}{n_{i-1}} \Sigma_{j=1}^{n_i} (X_{ij} - X_i)^2 \ ,$$

(3.3) $$\mu_i = E[X_{ij} | F_i], \ s_i^2 = V(X_{ij} | F_i) \ ,$$

$$D_i = V((X_{ij} - \mu_i)^2 | F_i) \ .$$

It follows that

$$E[X_i | F_i] = \mu_i \ , \ E[S_i^2 | F_i] = \sigma_i^2 \ ,$$

(3.4) $$V(S_i^2 | F_i) = \frac{1}{n_i} (D_i + \frac{2}{n_i - 1} \sigma_i^4) \ ,$$

and hence that

(3.5) $$\begin{aligned} V(S_i^2) &= E[V(S_i^2 | F_i)] + V(E[S_i^2 | F_i]) \\ &= \frac{1}{n_i} E[D_i] + \frac{2}{n_i(n_i - 1)} E[\sigma_i^4] + V(\sigma_i^2) \ . \end{aligned}$$

Consider the linear regression (best in terms of squared error) of σ_i^2 on S_i^2:

(3.6) $$t(S_i^2) = \frac{\mathrm{Cov}(S_i^2, \sigma_i^2)}{V(S_i^2)} (S_i^2 - E[S_i^2]) + E[\sigma_i^2] \ .$$

We note that $\{\sigma_i^2 : l = 1,2,\ldots m\}$ and $\{D_i : i=1,2,\ldots m\}$ are sequences of i.i.d. random variables, whereas $\{S_i^2 : i=1,2,\ldots m\}$ are not i.i.d., since they depend on n_i. Using (3.4), (3.5) and the fact that

$$\mathrm{cov}(S_i^2, \sigma_i^2) = V(\sigma_i^2) \ ,$$

we can remove any explicit averaging involving S_i^2 and then we can drop the subscript i to write

$$V(\sigma_i^2) = V(\sigma^2),\ E[D_i] = E[D],\ E[\sigma_i^4] = E[\sigma_i^4],\ E[\sigma_i^2] = E[\sigma^2],$$

$$(3.7) \qquad t(S_i^2) = \frac{V(\sigma^2)}{V(\sigma^2) + \frac{1}{n_i} E[D] + \frac{2}{n(n-1)} E[\sigma^4]} (S_i^2 - E(\sigma^2)] + E[\sigma^2].$$

The parameters in equation (3.7) are not known but we may estimate them in a consistent manner as $m \to \infty$. This will enable us to use equations (3.7) for linear empirical Bayes (l.e.B.) estimation of the σ_i^2.

For this estimation we deal with the moments

$$M_{ip} = E[X_{ij}^P | F_i]$$

In particular,

$$\sigma_i^2 = \frac{n_i(M_{i2} - M_{i1}^2)}{n_i - 1}$$

$$\sigma_i^4 =$$

$$(3.8) \quad \frac{n_i[n_i^2(M_{i1}^4 - 2M_{i2}M_{i1}^2) + (n_i^2 - 3n_i + 3)\ M_{i2}^2 + (n_i - 1)(4M_{i3}M_{i1} - M_{i4})]}{(n_i - 1)\ (n_i - 2)\ (n_i - 3)}$$

$$D_i =$$

$$\frac{n_i[4n_i^2(2M_{i2}M_{i1}^2 - M_{i1}^4) + (n_i^2 + 3n_i - 6)M_{i2}^2 + (n_i^2 - n_i + 2)(M_{i4} - 4M_{i3}M_{i1}^2)]}{(n_i - 1)\ (n_i - 2)\ (n_i - 3)}$$

Estimates $\hat{\sigma}_i^2$, $\hat{\sigma}_i^4$, $\hat{D}_i$ are obtained by estimating the M_{ip} in (3.8) as the sample moments

$$\hat{M}_{ip} = \frac{1}{n_i} \Sigma_{j=1}^{n_i} \hat{X}_{ij}^P .$$

Then $E[\sigma^2]$, $E[\sigma^4]$, $E[D]$ are estimated as

$$\hat{E}[\sigma^2] = \frac{1}{m} \sum_{i=1}^{m} \hat{\sigma}_i^2 (= \frac{1}{m} \sum_{i=1}^{m} \hat{S}_i^2 = \overline{S}^2)$$

$$\hat{E}[\sigma^4] = \frac{1}{m} \sum_{i=1}^{m} \hat{\sigma}_i^4$$

(3.9) $$\hat{E}[D] = \frac{1}{m} \sum_{i=1}^{m} \hat{D}_i$$

$$\hat{V}(\sigma^2) = \hat{E}[\sigma^4] - \hat{E}^2[\sigma^2]$$

A series of Monte Carlo calculations was carried out to investigate the effectiveness of the proposed allocation scheme. For each calculation a set of m populations was chosen and an estimation procedure specified. Then r independent estimation runs were made each involving N_0 samples in all drawn from the combined populations according to the particular estimation scheme. The mean μ and variance σ^2 were computed theoretically and the r estimates $\hat{\mu}$, $\hat{\sigma}^2$ were recorded. In order to compare the different cases we work with the normalized parameters

$$x = (\hat{\mu} - \mu)/\sigma; \; y = (\hat{\mu} - \mu)/\hat{\sigma}; \; z^2 = \hat{\sigma}^2/\sigma^2$$

For each case the following quantities were computed:

$$\bar{x} = (\Sigma x)/r$$
$$\overline{z^2} = (\Sigma z^2)/r$$
$$\overline{\Delta^2} = \Sigma(x-\bar{x})^2/(r-1)$$
$$x_+ = \mathrm{Max}|x|$$
$$y_+ = \mathrm{Max}|y|$$

where the sums and maxima are taken over the r independent runs.

Three population sets were used:

(a) Normal: $m = 10$, $F_i = N(i,4i^2)$; $\mu = 55$, $\sigma^2 = 24.2$

(b) Polynomial: $m = 10$, $F_i = G(i)$; $\mu = 82.5$, $\sigma^2 = 4.54$

(c) Mixed: $m = 20$, $F_i = \begin{cases} N(i,4i^2), & 1 \leq i \leq 5 \\ G(i), & 6 \leq i \leq 10 \\ \tfrac{1}{2} N(i,\tfrac{1}{i}) + \tfrac{1}{2} G(\tfrac{1}{i}), & 11 \leq i \leq 20 \end{cases}$ $\mu = 150.5$, $\sigma^2 = 32.09$

where G(a) is the distribution with density

$$\rho(x) = \begin{cases} 3a^3/x^4 & x > a \\ 0 & x < a \end{cases}$$

For each population set five different estimation procedures were tried

(1) Uniform: (N/m points drawn from each population)

(2a) Sequential: $n_o = 5$, $m_1 = 1$; classical variance estimates

(2b) Sequential: $n_o = 5$, $m_1 = 10$; classical variance estimates

(3a) Sequential: $n_o = 5$, $m_1 = 1$ l.e.B. variance estimates

(3b) Sequential: $n_o = 5$, $m_1 = 10$ l.e.B. variance estimates

For uniform allocation the results did not depend on the method of variance estimations.

For all cases we took $r = 250$, $N_0 = 500$. The value of $\bar{x}$ should be 0 with a standard deviation of $r^{-1/2}$, in this case .07 The robustness of the estimate is indicated by x_+ and y_+, the latter being appropriate in circumstances where good error estimates are as important as a good estimate of the mean.

TABLE I

Results of Monte Carlo Studies

		Uniform Allocation	Sequential Allocation: n=1 (n=10)			
			Classical variances		l.e.B. variances	
Normal	$\bar{x}$	.041	.085	(. 081)	.073	(.01)
	Δ	1.07	.93	*	.97	(1.08)
	$\bar{z}$	1.12	.99	*	1.03	(1.02)
	x_+	2.8	2.50	(2.52)	3.5	(2.99)
	y_+	1.47	1.49	(1.48)	2.27	(1.49)
Polynomial	$\bar{x}$	.07	-.96	(-.78)	-.38	(-.22)
	Δ	1.32	.88	(.91)	.89	(.95)
	$\bar{z}$	1.40	.87	(.78)	1.01	(.99)
	x_+	12.4	5.07	(3.6)	2.9	(2.7)
	y_+	2.95	4.95	(5.)	3.5	(3.7)
Mixed	$\bar{x}$	.49	-.11	(.002)	.13	(.25)
	Δ	1.14	1.47	(1.54)	1.11	(1.10)
	$\bar{z}$	1.15	.96	(1.01)	1.07	(1.08)
	x_+	4.	4.2	*	3.5	(3.35)
	y_+	1.79	3.4	(3.9)	1.92	(2.2)

*n_1 = 10 identical to n_1 = 1

The sequential allocation method using l.e.B. variance estimates is seen to be the most successful strategy. For the normal case there is no strong distinction between the different methods. For the polynomial case, in which no fourth moment exists (making conventional Monte Carlo error estimates suspect), the uniform allocation gives an unbiased result for $\bar{x}$, but with undesirably large variances and large fluctuations. The sequential scheme with classical variance estimates introduces a severe bias in $\bar{x}$, unreasonably low variances, and extremely poor error estimates as indicated by the large value of y_+. The l.e.B. procedures show significantly less bias, much better variances and estimates and quite acceptable values of x_+ and y_+.

For the mixed case the uniform allocation gives a biased result and the classical variance estimates again lead to misleading small error estimates, and large fluctuations.

In executing the algorithm, using $n_1 = 1$, (i.e. recomputing variances at each step), seemed to have no particular advantage over using $n_1 = 10$.

REFERENCES

Halton, J. H. and Zeidman, E. A. (1971). "The evaluation of multidimensional integrals by the Monte Carlo sequential stratification technique". Technical Report #137. The University of Wisconsin Computer Sciences Department.

Robbins, H., Simons G. and Starr N. (1967). "A sequential analogue of the Behrens-Fisher problem". Annals of Mathematical Statistics **38**, 1384-1391.

Robbins, H. (1982). "Estimating many variances". Statistical Decision Theory and Related Topics III. Academic Press **2**, 218-226.