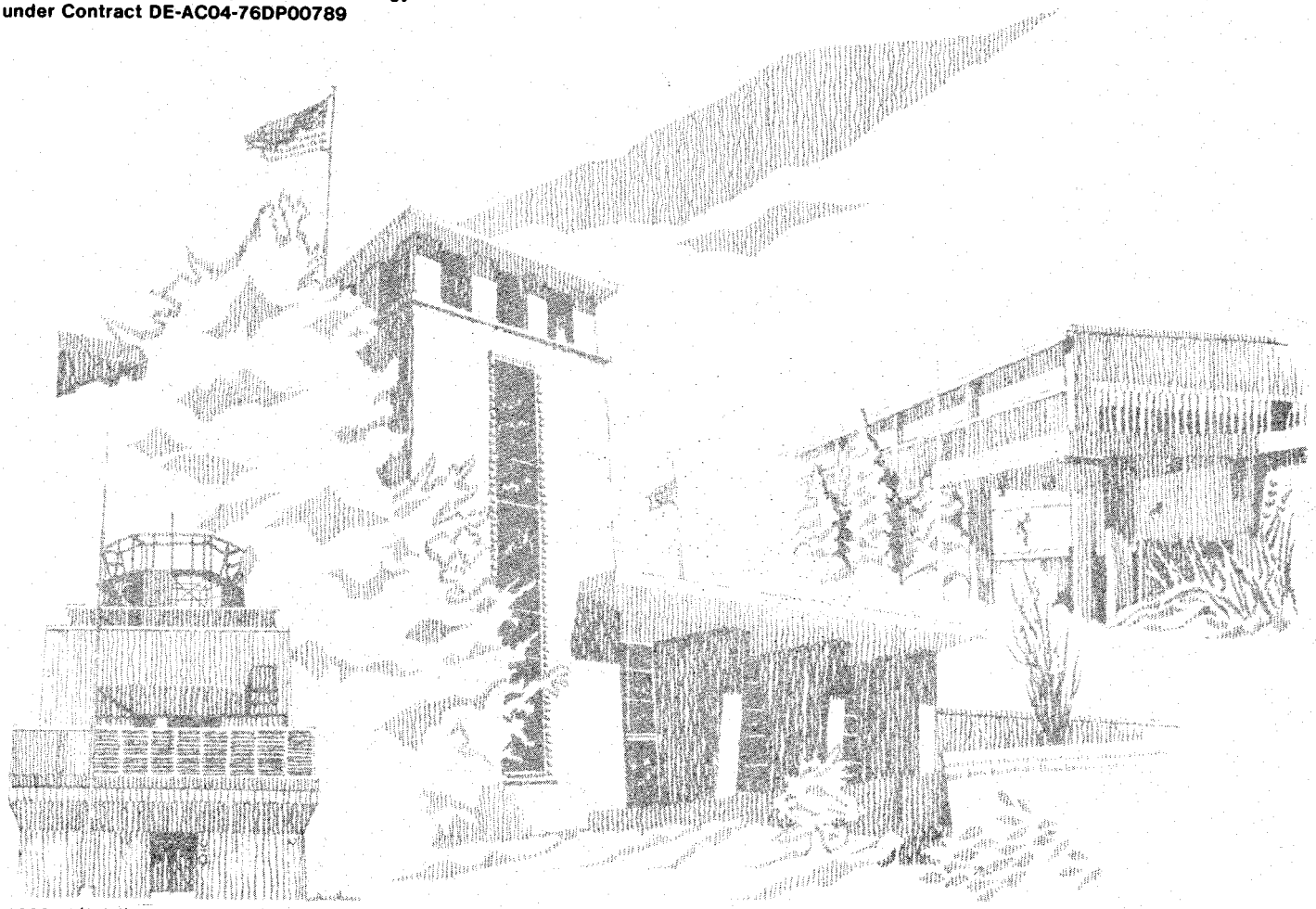
SANDIA REPORT SAND81-1200 • Unlimited Release • UC-13
Printed November 1981

A Sample Preselection Process Designed to Enhance Early Planning Information

Frank W. Müller, Floyd W. Spencer

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550
for the United States Department of Energy
under Contract DE-AC04-76DP00789



SF2900Q(8-81)

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof or any of their contractors or subcontractors.

Printed in the United States of America
Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, VA 22161

NTIS price codes
Printed copy: $5.00
Microfiche copy: A01

SAND81-1200
Unlimited Release
Printed November 1981

Distribution
Category UC-13

A Sample Preselection Process Designed to Enhance Early Planning Information

Frank W. Müller, Floyd W. Spencer
Quality Assurance Advanced Planning Division 1417
Sandia National Laboratories
Albuquerque, NM 87185

Abstract

The DOE provides for the continuing evaluation of the nuclear weapon stockpiles through a stockpile sampling program in which randomly selected weapons are withdrawn for testing from the stockpiles each year. For some time, DOE has used a preselection scheme to obtain early identification of certain characteristics of the sample weapons for planning purposes, but which does so without jeopardizing the necessary randomization of sample selection. A DOD desire for additional and more detailed planning information to minimize weapon movements has led to an improvement of the original preselection scheme that enhances the planning information and its accuracy, while still preserving randomization.

Contents

Introduction 7
Original Preselection Scheme 7
Preselection 8
Actual Selection 8
Improved Preselection 8
Identifying Actual Sample Units 11
Uses of the Planning Information 12
Use of Two Classification Characteristics 12
APPENDIX — Development of a Statistical Algorithm 15

Illustrations

Figure

1 Preselection of Five Samples From Four Subclassifications 9
2 Actual Selection of Five Samples From Four Subclassifications 9
3 Moving a Sample From Subclassification 4 to Subclassification 2 10
4 Comparison of New and Old Schemes 11

Tables

1 Original Preselection Random Numbers Compared With New Scheme Random Numbers 9
2 Weapon Subclassifications When Mod and Location Are of Interest 13

A Sample Preselection Process Designed to Enhance Early Planning Information

Introduction

The US Department of Energy (DOE) maintains responsibility for the condition of the nation's nuclear stockpiles, even after the weapons are in the custody of the Military Services. To discharge that responsibility, DOE conducts stockpile sampling programs in which stockpiled weapons are randomly selected each year for testing.

However, the constitution of a stockpile for testing purposes is not always the same as that used for nuclear weapon accountability. In particular, DOE accounts for nuclear weapons by Mod and yield, but stockpile samples are usually drawn without regard for Mod or yield. In these cases, information regarding Mods and yields normally would not be available until the sample was selected. Consequently, before the sample was selected, planners had to use the "expected numbers" of various yields and Mods in a weapon sample. These rarely provided an accurate breakdown of the actual sample. This caused a planning nuisance, requiring continual changes in inventory documentation and obtaining the necessary approvals for such changes.

Several years ago, Sandia National Laboratories (SNL) Quality Assurance Advanced Planning Division developed a scheme to preselect samples. The method identified the number of weapons in each Mod and their yield, without impairing the random, statistical properties of the sampling process. When we could forecast, with even modest accuracy, the number of weapons in each Mod and in each yield that would exist at the time the sample was to be selected, it greatly reduced the nuisance work. This preselection scheme has been used ever since it was developed. That's what you think!

The US Department of Defense (DOD) is also expressing interest in early planning information, but in subclassifications of weapons different from those that interest DOE. The DOD is interested in combining weapon movements, hence wants to know which sites will have to surrender weapons to the sampling process in any given year.* Theoretically, the same preselection scheme could be applied, except for two factors:

- The number of subclassifications of interest is much greater now than before, meaning that the number of weapons per subclassification is much lower than before
- The subclassifications of weapons can change much faster and easier than before

These two factors can significantly amplify both the number and the magnitude of forecasting errors, possibly rendering useless the finely detailed planning information that we could provide DOD.

There are two ways to improve the accuracy — hence the value — of early planning information that a preselection process might provide to DOE and DOD:

- Improve the accuracy of forecasts — simply a matter of obtaining information from DOD and DOE regarding planned weapon movements, production, retirement, etc; since this information can be requested at any time by those involved, we will not discuss it further in this report.
- Improve the preselection scheme — if changes are made in the stockpile between preselection and the time the sample is actually selected, ensure that the number of changes in the planning information is minimized. In this report, we discuss such an improvement.

Original Preselection Scheme

Let us assume that a certain number, k, of different weapon subclassifications can be identified for a

*DOD would like to know particular serial numbers, but we do not here address that option, since it could lead to preferential treatment of weapons.

given stockpile. To be of practical significance, each subclassification must exist either in the population forecast for preselection or in the population existing at the time the sample is selected, or both. For the moment, let us consider the subclassifications without worrying where they come from; later we will consider classifications based on more than one characteristic.

First, let us designate A_i and B_i to represent the number of units in the ith subclassification in the forecast and actual populations, respectively:

$$\sum A_i \quad \text{and} \quad \sum B_i$$

to represent the total sizes of the two populations; and

$$A_i / \sum A_i = \alpha_i$$

and

$$B_i / \sum B_i = \beta_i,$$

to represent the fractions of the respective populations found in the ith subclassification. Thus,

$$\sum \alpha_i = \sum \beta_i = 1.0,$$

where the index i reflects an arbitrary ordering of the k subclassifications which, in the original scheme, remains fixed during both processes of preselecting and actual sampling.

Preselection

Let us first discuss the preselection process. Suppose a certain number, n, of samples is to be selected. We first generate n uniform [0,1) random numbers. We then predict the number of sample units to be selected from each subclassification by comparing the random numbers against the k intervals* that may be identified on a line of unit length by plotting the partial sums,

$$\sum_{i=1}^{s} \alpha_i, \quad s = 1, 2, \ldots, k.$$

*k intervals correspond directly with the k subclassifications.

Figure 1 illustrates a case where k = 4 and n = 5; (a) represents the four subclassifications and (b) the random numbers. By observing into which interval each random number falls, we determine that the sample size from each of the four subclassifications is 1, 2, 0, 2, respectively. This represents the early planning information, which can be provided to DOE or DOD.

Actual Selection

Now consider actual selection of the sample at the prescribed date. The same random numbers (saved from the preselection process) are now compared against the new set of k intervals determined by plotting the partial sums

$$\sum_{i=1}^{s} \beta_i, \quad s = 1, 2, \ldots, k.$$

In Figure 2, (a) shows how the four subclassifications actually wound up on the sample select date, in contrast to how we thought they would wind up (Figure 1); (b) represents the same random numbers as in Figure 1.

The sample sizes actually selected from each subclassification are 1, 2, 1, 1, respectively, rather than those predicted at preselection. This difference is due to the change in intervals from the time of preselection to that of actual selection. In this case, the number of units forecast to be in subclassifications 1, 2, and 4 were wrong. Curiously, correcting the error had the effect of moving the interval of subclassification 3 over a random number, even though the proportion of the total population in that subclassification had not changed.

Half the early planning numbers turned out to be wrong, which, for our purposes, is much too inaccurate. On the other hand, half the planning information was correct, meaning that we did eliminate a noticeable part of the nuisance of adjusting inventories; the earlier practice of using expected numbers was almost always wholly incorrect.

Improved Preselection

For DOE, the original scheme was adequate because the stability (or predictability) of stockpile constitution, in terms of subclassifications that interested DOE, was reasonably high, whereas the number of these subclassifications was low. New DOD interests, however, now force us to attend to many more subclassifications, greatly reducing their stability and predictability, and we find it expedient to modify the

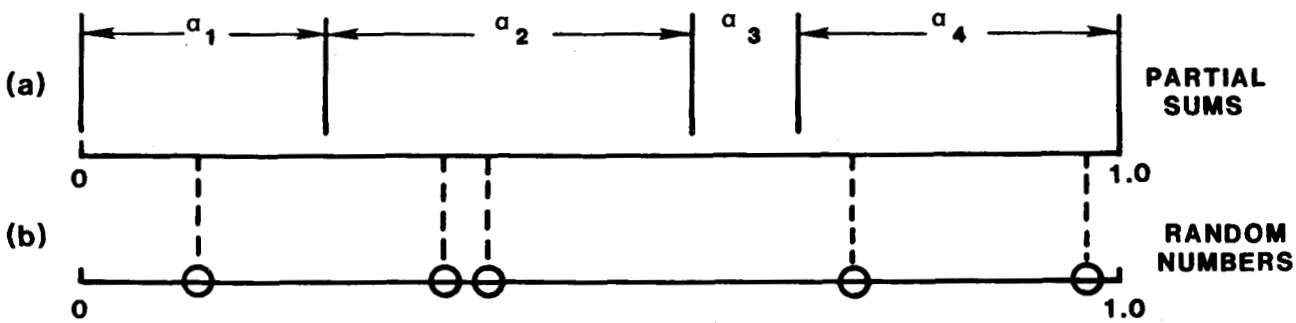


Figure 1. Preselection of Five Samples From Four Subclassifications

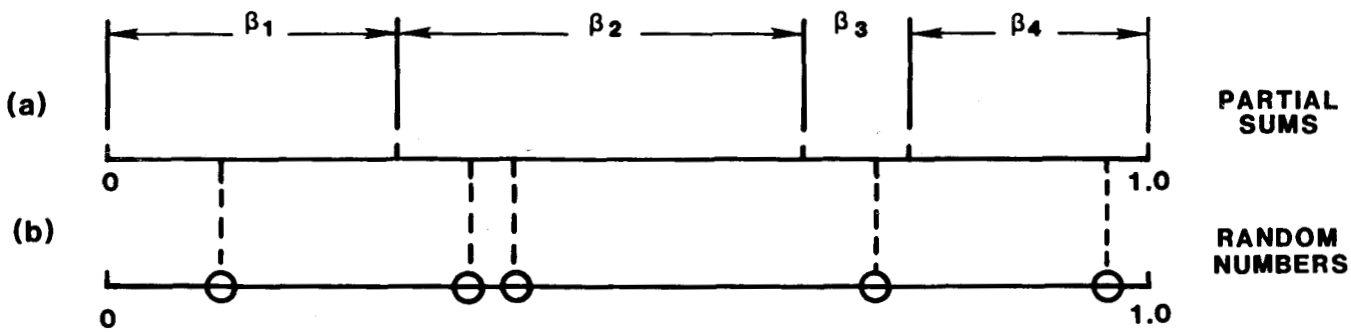


Figure 2. Actual Selection of Five Samples From Four Subclassifications

preselection scheme to assure that it does not contribute unnecessarily to the inaccuracy of the planning information.

The original scheme was weak because inherent in it was the need to arrange the subclassifications in some kind of order, thereby influencing results. One effect of such arrangement can be noticed in comparing Figure 1 with Figure 2, where the sampling from subclassification 3 changed, although its size had remained the same. Had subclassification 3 occurred first in the ordering, there would have been no change in the planning information.

The new scheme completely eliminates the influence that ordering has on the results. The steps in generating the planning information are identical to those of the original procedure (Figure 1). After that, however, we adjust the planning information by examining the subclassifications individually, after the random numbers are altered to show only their location in relation to their original subclassification. Specifically, a random number associated with subclassification s is adjusted by subtracting the partial sum,

$$\sum_{i=1}^{s-1} \alpha_i ,$$

from it, thus generating a new random number:

$$M_s = R - \sum_{i=1}^{s-1} \alpha_i ,$$

which measures the distance from the left boundary of the appropriate interval (Figure 1(a)) to the original number. From our example we might obtain the values in Table 1.

Table 1. Original Preselection Random Numbers Compared With New Scheme Random Numbers

Subclass. (s)	Preselection (α_s)	Selection (β_s)	Old Scheme Random Numbers (R)	New Scheme Random Numbers (M)
1	0.24	0.275	0.11	0.11
2	0.35	0.395	0.35, 0.39	0.11, 0.15
3	0.10	0.10	—	—
4	0.31	0.23	0.74, 0.96	0.05, 0.27

We will now discuss the algorithm for the new scheme. In the Appendix is proof that the desired statistical properties remain intact and that planning information has been perturbed as little as possible. Note that the algorithm deals with the fractions of the total population that are represented in each subclassification, not with the number of units.

A feature of the new scheme is that the planning information for subclassifications in which $\beta_i = \alpha_i$ will very rarely change. However, $\beta_i = \alpha_i$ does not imply that $A_i = B_i$, since the total population may change in size. In fact, if $A_i = B_i$, yet the total population has increased, say through more production, then $\alpha_i > \beta_i$. There is no guarantee, under such circumstances, that the planning information for subclassification i would not change, even though the subclassification remained the same size.

To determine how sample units move from subclassification to subclassification,* we begin by examining the changes in the relative sizes of the subclassifications. Specifically, we compute $\Delta_i = \alpha_i - \beta_i$ for all subclassifications i. Now, since

$$\sum \Delta_i = 0,$$

we know that any change in the size of one subclassification from preselection to selection must be compensated for by an opposite change in some other subclassification(s). We first consider all those subclassifications for which $\Delta_i > 0$ (e.g., Figure 1, subclassification 4). Carefully maintaining these subclassifications in the original order, we plot the partial sums

$$\sum_{i=1}^{s} \Delta_i$$

on a line to establish intervals for each $\Delta_i > 0$.

Next, on a separate line, we plot the partial sums

$$\sum_{i=1}^{s} |\Delta_i|$$

for those cases where $\Delta_i < 0$ (e.g., Figure 1, subclassifications 1 and 2), again being careful to maintain the original order.

*We assume that the total sample size does not change because of changes in population size. We would accommodate changes in sample size by choosing additional samples from the population at large, and randomly removing samples if too many were chosen at time of preselection.

We now return to those subclassifications for which $\Delta_i > 0$, and which, according to preselection, contained sample units. We determine whether each adjusted random number, M, associated with that subclassification, is greater than β_i. These are the sample units that will be moved;* for each such unit we determine the distance from the left side of the Δ interval (that is the distance, $M - \beta_i$). Using this distance, we then place sample units in the appropriate interval on the line showing the partial sums. This is shown in Figure 3(a) where the value of $M - \beta_i$ is for the sample unit from subclassification 4 in the example previously used. Directly transfering this point to the other line (Figure 3(b)) will identify both the subclassification and the distance, d_i, from the left side of the $|\Delta_i|$ interval which will accept the sample. In the example, $d_2 = 0.005$. This may now be translated back to the unit line showing the partial sums of the β_i (Figure 2(a)), where it will be located $\alpha_i + d_i$ from the left side of the ith interval. In the example, the sample will be located at a distance $\alpha_2 + d_2$ from the left side of the β_2 interval.

After all such sample units have been moved, the sample to be actually selected is represented by the adjusted random numbers shown on the plot of the partial sums of the β_i. Figure 4(a) is a repetition of Figure 1; Figure 4(b) a repetition of Figure 2; and, for comparison, Figure 4(c) shows the adjusted random numbers resulting from the new scheme.

In the new scheme, the distances from the left edge of the interval to the random numbers remain the same from preselection to actual selection, except when samples are transferred from one subclassification to another. This is noted in Figure 4 by distances a, b, and c. Both schemes resulted in the change of one sample unit (two pieces of planning information) in the example, which, because it was contrived to illustrate features of the schemes, does not illustrate

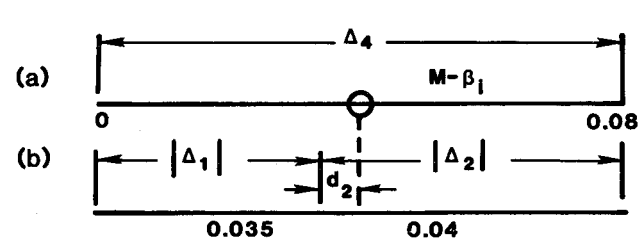


Figure 3. Moving a Sample From Subclassification 4 to Subclassification 2

*If there are none, the planning information has proved to be accurate and we may proceed to the final step of identifying a sequence number for each sample unit.

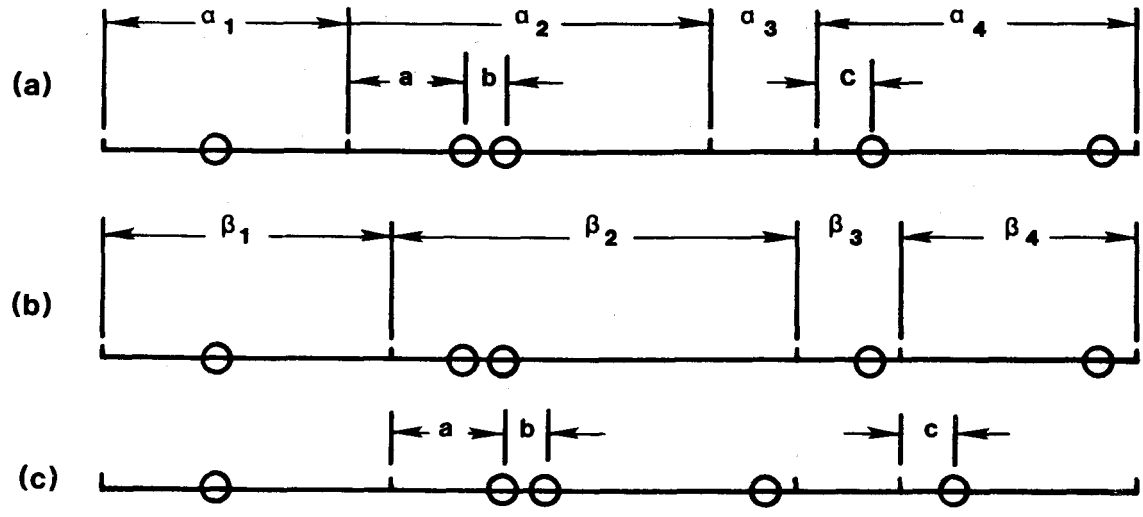


Figure 4. Comparison of New and Old Schemes

improvement in accuracy. However, we can compare accuracy by examining the probability that, before the fact, a particular sample unit will be associated with a random number that will place it where it will be moved. With the original scheme, this probability is measured by the amount of overlap between subclassifications (Figure 4(a), (b)). For our example this turns out to be 0.205, slightly better than one chance in five that a particular sample unit will be moved. For the new scheme, the probability is the sum of $\Delta_i > 0$, which, in our example, is 0.08, about one chance in twelve. Thus, in the example the new scheme would be expected to eliminate more than half the inaccuracies. Such improvement is not unusual; it depends a great deal on luck in arranging the subclassifications at the time of preselection.

Identifying Actual Sample Units

The final step in all schemes is to identify the units actually selected as samples. To do this, we identify the sample units in terms of a sequence number for the ordered units, where the units from subclassification 1 are associated with the first B_1 sequence numbers, the units from subclassification 2 are associated with the next B_2 sequence numbers, etc. Serial numbers can be determined by relating the sequence numbers to them as they occur in some ordering.

The sequence number is determined by multiplying the "adjusted" random number* by the number of units in the total population and discarding the non-integer part of the resulting number. This will provide a sequence number from 0 to N-1, where N is the number of units in the total population.

Unfortunately, converting a uniform [0,1) random variable to a random choice from $\{0, 1, ..., N-1\}$ creates a problem which we have not yet alluded to: the possibility that two separate random numbers could convert to the same sequence number. Because we are interested in sampling without replacement, the random numbers that convert to an already-identified sequence number will have to be replaced with a new random number. This seems to be a trivial problem because, at preselection, we can, and should determine whether more than one random number results in the same sequence number, and replace the extra numbers with new ones. However, the sequence number and conversion process depend upon the total number of units in the population, and any changes to the population from the time of preselecting to that of actually selecting the sample could affect both. In particular, random numbers that result in repeated sequence numbers in the preselection population may not do so when selecting samples, and vice versa.

To illustrate, suppose we had selected random numbers $R_1 = 0.6074$ and $R_2 = 0.6038$, and the population at preselection was 237. Both random numbers identify sequence number 143, so one would have to be discarded. Now, suppose that we learned (by ESP?) that the population would be 238 at actual selection; R_1 and R_2 would lead to different sequence numbers, 143 and 144, so both random numbers could be used. (Different sequence numbers would also occur if the population were reduced to 236!)

*The "adjusted" random number in the old scheme would be the distance from the left side of the subclassification interval in which the random number fell.

It is this feature that makes the sampling without replacement a little troublesome. If we replace a random number at preselection, which is desirable to assure that planning information does not change for unchanging populations, we must retain it for later use when the true population is known, at which time sequence numbers must be determined again. As illustrated above, population changes may require that we re-establish a discarded random number and drop the one that replaced it. Or perhaps, when the time comes to actually select a sample, we have to replace random numbers that survived in the preselection process.

Any replacement random numbers must be initially related to the population at large, and then related (by subtracting a partial sum) to the subclassification in which the rejection was made. For more details on this aspect, see the Appendix.

Uses of the Planning Information

The Appendix shows proof that, if planning information is not used to manipulate the population, the preselection schemes will preserve the statistical properties of sampling, assuring a sample randomly selected from the population existing at sample selection time. Unfortunately, there can be no guarantee that, if early planning information is given out, it will not be used to manipulate the population.

Manipulating the stockpile did not concern DOE in regard to the kind of planning information it desired because the Mods and yields were fundamental to the military capability inherent in the stockpile and could not be changed by DOE. It is difficult to change a weapon from one subclassification to another and, since DOE does not have custody of the weapons, it is fairly certain that planning information would not be used to modify the subclassifications or the weapons in them.

However, with regards to the information desired by DOD, the situation is much less clear. Because location is the basis for subclassification, simple movement of weapons is all that is needed to change subclassifications. And, because DOD personnel that desire the information are specifically concerned with minimizing "unnecessary" movement of weapons, it is not at all doubtful that such movement would indeed be affected by the planning information — if, by using it, movement could be reduced. That is why DOE does not want to provide actual serial numbers at preselection; it is quite probable that weapons preselected early for sampling would not be shipped to a forward site, only to be called back if selected. Rather, they would be held at the original site, hence would not experience the true stockpile handling and shipping environments, whose effects are of primary interest in the stockpile evaluation program.

Providing preselection information regarding sites will still permit some "game playing" with weapon movements, but, in this case, groups of weapons would be involved, and it is doubtful that restricting the movements of whole groups would be tolerated by military planners. However, since such movements, merely to censor part of the population from the sampling, are counter to the desire to minimize movement, they are not a real concern — so long as DOD desires to minimize weapon movements.

It is important to realize that any preselection information given out can be used to jeopardize the sampling process or to alter the true condition of the stockpile that is ultimately sampled. The more preselection information given out, the more likely that it would be used in a way that militates against the reasons for sampling. On the other hand, the less information given out, the more difficult and expensive it is to use it to impair the sampling program. We believe that providing site information will not harm the sampling program, but we also admit that the situation could be marginal and we strongly suggest that it be audited or reviewed to determine if the planning information is being used improperly.

Use of Two Classification Characteristics

In arriving at this point, we have merely identified subclassifications without concerning ourselves if they actually resulted from several characteristics or just one. However, we previously indicated that planning information was desired for several characteristics: Mods, yields, and locations, and that persons interested in Mods and yields are not interested in locations, and vice versa. Let us now consider planning information related to those groups of subclassifications discussed above.

Let us assume that a weapon is stored in three different locations and is either a Mod 0 or a Mod 1, resulting in six subclassifications (Table 2).

Table 2. Weapon Subclassifications When Mod and Location Are of Interest

Mod	Site A Subclass.	Site B Subclass.	Site C Subclass.
0	1	2	3
1	4	5	6

Let us now suppose that the accuracy of the planning information related to Mods is considered more important than that related to sites, and that we would like the planning information for the Mods to be the same as if we had only defined two subclassifications, namely the Mods.

To do this, we would first address two subproblems, one for Mod 0 and one for Mod 1, and would try to accommodate as many changes as possible within each Mod. Afterward, any change not accommodated would have to cross Mods and would provide a net change that would be equivalent to that which would have resulted had only two subclassifications of Mods been identified.

Because our priority was on information regarding Mods, it is possible that the information on sites may not be the same as it would have been had only sites been considered in establishing subclassifications. In establishing priorities, the accuracy of the forecast information available at preselection should be considered. For example, we should not give priority to sites when the forecasts about sites are very inaccurate. This would merely degrade planning information on Mods which, because such information can be forecast more accurately, would otherwise be quite accurate. A slight improvement in poor planning information is not worth the significant degradation of good planning information.

APPENDIX

Development of a Statistical Algorithm

The following is a development of an algorithm to modify a random sample that was originally chosen on erroneous subclassification population figures. At first, samples are identified only as belonging to certain subclassifications. The purpose of the algorithm is to preserve the statistical properties of a random sample drawn from the correct population figures, and to minimize the chances that the subclassification of the original sample will change when corrected.

Consider a number, k, of different subclassifications, and let A_i, B_i, $i = 1, 2, ..., k$ denote the population size of the ith subclassification in the original (possibly erroneous) and the actual (corrected) population figures, respectively. We assume that the subclassifications form a partition so that any element of the whole population is a member of one, and only one, subclassification. Let

$$\alpha_i = A_i / \sum A_i$$

and

$$\beta_i = B_i / \sum B_i$$

denote the proportion of the total population in the ith subclassification in each of the forecast and actual populations, respectively.

In discussing the problem we will first limit our attention to a single sample. We will then discuss modifications necessary for sample sizes larger than 1.

Single Sample

At the time of original selection, we select a uniform [0,1) random number, R, to determine from which subclassification the sample unit will be drawn.

Define

$$S_j = \sum_{i=1}^{j} \alpha_i, \; j = 1, 2, ..., k,$$

and let $j' = j'(R) = \min \{j : S_j > R\}$. Because $\{S_j\}$ is a nondecreasing sequence and $S_k = 1$, j' is properly defined. The subclassification indexed by j' is identified as the subclassification from which the sample is drawn when the original population figures are used.

To describe the algorithm used to possibly alter the sampling scheme and to prove that it has the proper statistical properties, we make the following definitions: Let Δ_i denote the change in relative size of subclassification i from the original to the corrected population figures, that is, $\Delta_i = \alpha_i - \beta_i$, $i = 1, 2, ..., k$. By definition,

$$\sum \alpha_i = \sum \beta_i = 1,$$

and therefore,

$$\sum \Delta_i = 0.$$

Let $J^+ = \{j : \Delta_j > 0\}$ and $J^- = \{j : \Delta_j \leq 0\}$. If $j \in J^+$, then the chance of the sample being drawn from subclassification j was too great by the amount Δ_j. Similarly, if $j \in J^-$, then the chance of subclassification j originally being sampled was smaller by $-\Delta_j$ than it should have been. The following quantities will be useful:

$$C_\ell = \sum_{i=1}^{\ell} (\Delta_i \vee 0),$$

$$D_\ell = \sum_{i=1}^{\ell} (-\Delta_i \vee 0), \; \ell = 1, 2, ..., k,$$

where $a \vee 0$ = maximum $\{a,0\}$. Note that C_ℓ is greater than $C_{\ell-1}$ if, and only if, $\ell \in J^+$. If D_ℓ is greater than $D_{\ell-1}$, then $\ell \in J^-$. It is also the case that

$$C_k = D_k = 1/2 \sum_{i=1}^{k} |\Delta_i| .$$

We determine the position of the random number R in the interval corresponding to j′ by forming the quantity $M_1 = R - S_{j'-1}$. It is this quantity that determines whether the subclassification from which the original sample was taken will be changed when the corrected population figures become known. Changes take place only to move samples that occur in the J^+ populations to the subclassifications in the J^- category.

Example: We have seven subclassifications with the forecast population data as (5, 11, 8, 7, 9, 7, 3). The actual population figures turn out to be (4, 13, 11, 7, 7, 6, 2). The above defined quantities are given in Table A1 for this example.

Table A1. Example With Seven Subclassifications, and R = 0.79 Resulting in Sample From Subclassification 5*

i	α_i	β_i	Δ_i	S_i	C_i	D_i
1	0.10	0.08	0.02	0.10	0.02	0.00
2	0.22	0.26	-0.04	0.32	0.02	0.04
3	0.16	0.22	-0.06	0.48	0.02	0.10
4	0.14	0.14	0.00	0.62	0.02	0.10
5	0.18	0.14	0.04	0.80	0.06	0.10
6	0.14	0.12	0.02	0.94	0.08	0.10
7	0.06	0.04	0.02	1.00	0.10	0.10

*$J^+ = \{1,5,6,7\}$, $J^- = \{2,3,4\}$
$j'(0.79) = 5$
$M_1 = 0.79 - S_4 = 0.79 - 0.62 = 0.17$

In describing the algorithm, we will use j′ to denote the index of the subclassification from which the sample is originally preselected, and j″ to denote the index of the subclassification from which the sample is actually taken. We now describe the algorithm assuming that the random number R has been picked and the original index j′ has been determined.

One of the following three conditions must hold:

- If $j' \in J^-$ then $j'' = j'$ and no change occurs
- If $j' \in J^+$ and $M_1 \leq \beta_{j'}$ then $j'' = j'$ and no change occurs
- If $j' \in J^+$ and $M_1 > \beta_{j'}$ then $j'' = \text{minimum } \{j{:}D_j > C_{j'-1} + M_1 - \beta_{j'}\}$

If there is a change, then the quantity, $D_{j''} - C_{j'-1} - M_1 + \beta_{j'}$ becomes the distance from the right side of the j″ interval associated with the sample placement. This quantity can be used in determining the actual sample unit to be taken from subclassification j″.

Example: Continuing the previous example we see that j′ = 5 and, thus, is in J^+, and that $M_1 = 0.17$, which is $> \beta_5$ (=0.14); thus a movement will occur. In this example:

$$C_{j'-1} + M_1 - \beta_{j'} = C_4 + M_1 - \beta_5 = 0.02 + 0.17 - 0.14 = 0.05 .$$

Thus, j″ = 3, since D_3 is the first value in the D sequence that exceeds 0.05. The random sample is placed at a distance $D_3 - 0.05$ (=0.05) from the right side of interval 3.

The above algorithm was given in a form such that computer implementation would follow easily. We comment that, graphically, the whole procedure is equivalent to subdividing the unit interval into portions with lengths given by the α_i's and, in each of the intervals associated with J^+, take the rightmost Δ portion and form another interval as the union of these intervals. This resultant interval is overlaid on an equivalent one formed from the subclassifications from J^-, and any sample unit from the J^+ interval is transferred to the J^- interval.

The crux of the algorithm is that R, a uniform [0,1) random number, when conditioned on lying in a subinterval contained within [0,1), has a uniform distribution on that subinterval. This being understood, we shall advance immediately to Proposition 1.

Proposition 1: In applying the above algorithm, the probability that the sample is ultimately drawn from subclassification j is β_j.

Proof: For all j, when using forecast population figures, the probability that the sample is from subclassification j is α_j. This follows immediately since R is a uniform random variable and j′ = j if, and only if, R lies in a subinterval of length α_j.

Consider $j \in J^+$, then the conditional probability that j″ = j given that j′ = j is the probability that the sample is not moved, given that j′ = j. This is $1 - \Delta_j/\alpha_j$; since Δ_j is the portion of the α_j interval that results in a move.

Thus, for $j \in J^+$,

$$\Pr\{j''=j\} = \Pr\{j''=j|j'=j\} \Pr\{j'=j\}$$

$$= (1 - \Delta_j/\alpha_j) \cdot \alpha_j = \alpha_j - \Delta_j = \beta_j .$$

For $j \in J^-$, we express the probability conditioned on the outcome of j'. Let η_i, $i \in J^+$ be the portion of the Δ_i interval in the $\{C_i\}$ sequence that overlaps the $-\Delta_j$ interval in the $\{D_i\}$ sequence. We then have,

$$\Pr\{j''=j\} = \sum_{i=1}^{k} \Pr\{j''=j|j'=i\} \Pr\{j'=i\}$$

$$= \sum_{i \in J^+} \Pr\{j''=j|j'=i\} \Pr\{j'=i\} + \Pr\{j'=j\}$$

$$= \sum_{i \in J^+} (\eta_i/\Delta_i)(\Delta_i/\alpha_i) \cdot \alpha_i + \alpha_j$$

$$= -\Delta_j + \alpha_j = \beta_j .$$

This establishes the proposition.

The above scheme is, in fact, the best that we can expect in terms of minimizing the probability of change and still maintain the probability, β_j, of ultimately being in subclassification j. We prove this in Proposition 2.

Proposition 2: Let j'' represent the final choice of subclassification for the sample, and j' be an initial choice, such that $\Pr\{j'=j\} = \alpha_j$ and $\Pr\{j''=j\} = \beta_j$. It follows that

$$\Pr\{j'' \neq j'\} \geq 1/2 \sum |\alpha_i - \beta_i| .$$

The algorithm described achieves the lower bound.

Proof: Define indicator functions for $j = 1, 2, \ldots, k$ as follows:

$$I_j = \begin{cases} 1 & \text{if } j'=j \\ 0 & \text{if } j' \neq j \end{cases}$$

$$I'_j = \begin{cases} 1 & \text{if } j''=j \\ 0 & \text{if } j'' \neq j \end{cases}$$

With the above definitions, and noting that the terms I_j (I'_j) will be 1 for a single j and 0 for all other j's, we note that

$$\sum_{j=1}^{k} |I_j - I'_j|$$

takes on the value of 0 if $j' = j''$ and the value 2 if $j' \neq j''$. Thus,

$$\Pr\{j' \neq j''\} = E\left[1/2 \sum_{j=1}^{k} |I_j - I'_j|\right]$$

$$= 1/2 \sum_{j=1}^{k} E|I_j - I'_j| ,$$

where E is the expectation operator. Since $|I_j - I'_j|$, for a given j, is itself an indicator function we can rewrite the above as follows:

$$\Pr\{j' \neq j''\} = 1/2 \sum_{j=1}^{k} E(I_j - I'_j)^2$$

$$= 1/2 \sum_{j=1}^{k} [E(I^2_j) + E(I'^2_j) - 2E(I_j I'_j)] .$$

Since $I_j \cdot I'_j \leq I_j$ and $I_j \cdot I'_j \leq I'_j$ we know that $E(I_j I'_j) \leq E(I_j) \wedge E(I'_j)$, where $a \wedge b = \min\{a,b\}$. Thus,

$$\Pr\{j' \neq j''\} \geq 1/2 \sum_{j=1}^{k} [\alpha_j + \beta_j - 2(\alpha_j \wedge \beta_j)]$$

$$= 1/2 \sum_{j=1}^{k} |\alpha_j - \beta_j| .$$

To demonstrate that the selection scheme described achieves this minimum, requires only that we show that $E(I_j I'_j) = E(I_j) \wedge E(I'_j)$ or, equivalently, that $\Pr\{j'=j, j''=j\} = \alpha_j \wedge \beta_j$ for $j = 1, 2, \ldots, k$. If $j \in J^-$ (i.e., $\alpha_j \leq \beta_j$), then by step 1 in the algorithm $j'=j$ implies that $j''=j$, thus, $\Pr\{j'=j, j''=j\} = \Pr\{j'=j\} = \alpha_j$. If $j \in J^+$ (i.e., $\alpha_j > \beta_j$), then $\Pr\{j'=j, j''=j\} = \Pr\{j''=j|j'=j\} \Pr\{j'=j\} = (1 - \Delta_j/\alpha_j)\alpha_j = \alpha_j - \Delta_j = \beta_j$. In both cases, $\Pr\{j'=j, j''=j\} = \alpha_j \wedge \beta_j$.

Note that this proposition relies only on the fact that the original distribution on j' was governed by the probabilities $\{\alpha_i\}$ and the final distribution on j'' was governed by the probabilities $\{\beta_i\}$. The form of any dependency between j' and j'' is inconsequential with respect to the lower bound given.

We have shown that the selection scheme picks a unit from the subclassifications in proportion to the ultimate population figures within the subclassifications and that it minimizes the movement from a preselection that had been made with possibly erroneous population figures. However, since the sampling is ultimately of individual units and not subclassifications, it is of individual units that we want

the random sample to consist of. To do this, any changes in subclassifications must be completely independent of the first choice, j'.

To illustrate this point more fully, consider a case in which units are identified in subclassifications according to physical location. Suppose also that there exists a unit in subclassification (location) 1 that we decide not to sample. If $j' = 1$ then we could exchange the unit with another in subclassification 2, making sure that $\Delta_1 = \Delta_2 = 0$; therefore, no movement would take place in these subclassifications and we would be assured that the unit of concern would not be sampled. Similarly, if $j' \neq 1$, then we assure that $\Delta_1 = 0$ and no movement at the second stage will take place into subclassification 1. Thus, armed with the knowledge of the outcome of j', we can, to some degree, manipulate the chances of drawing individual units in the sample. For this reason, it is essential that changes in subclassification identification of individual units occur completely independent of the first selection, j', if the random selection is to be based on individual units.

Multiple Samples and Modifications

The scheme described for drawing a single sample can be applied for drawing multiple samples. Each sample unit is associated with a random number, R, and the procedure is applied to each of them. Since Proposition 2 minimizes the probability of a change for a single sample; it applies equally to all samples. When drawing many samples, the expected number of moves is minimized. If no adjustment is made to the values of R, this procedure would be equivalent to sampling with replacement.

Theoretically, R is chosen from the continuum [0,1) and it is associated to physical units by sequencing the units, say 0 through N-1, and letting the random number, R, designate the [NR] unit ([X] is the greatest integer $\leq$ X). Therefore, if two random numbers, R_1 and R_2, are such that $[NR_1] = [NR_2]$, one of them is disallowed and another random number is drawn in its place. This amounts to sampling without replacement.

If the total population remained unchanged, i.e.,

$$\sum A_i = \sum B_i,$$

the possible reshuffling of units described in the algorithm would preserve the hypergeometric nature of the sample, because the Δ's would all be multiples of

$$1/\sum A$$

and the original process would insure that no two random numbers would corespond to the same unit. However, if

$$\sum A_i \neq \sum B_i ,$$

the original process restricts the resultant random numbers to lying in intervals of the form

$$\left[\ell/\sum A_i, (\ell + 1)/\sum A_i\right],$$

whereas the final draw should result in their being in intervals of the form

$$\left[\ell/\sum B_i, (\ell + 1)/\sum B_i\right].$$

As already noted, the essential properties of the algorithm come about because the restriction of the uniform random variable is again uniform. For this reason, the algorithm should be applied to the uniform [0,1) random variables rather than to the restricted random variables one gets when considering sampling without replacement.

Let $R_1, R_2, \ldots$ be a sequence of independent, identically distributed uniform [0,1) random variables, and let $A(R_1), A(R_2), \ldots$ be the result of applying the above described algorithm to each of the random variables. As long as the changes in the k classifications are independent of the sequence $\{R_i\}$, then $\{A(R_i)\}$ is also a sequence of independent, identically distributed uniform [0,1) random variables because A(R) is just a reshuffling of intervals on [0,1).

The complete algorithm to account for sampling without replacement would be given by the following steps: Assume that n samples are to be drawn. The preselections (possibly based on erroneous figures) are made by using the first n R's that have no repeats in the sequence numbers based on these figures. The actual selections will be made by using the first n A(R)'s that have no repeats in the sequence numbers based on the true population figures.

DISTRIBUTION:

US Department of Energy (2)
Albuquerque, Operations Office
P.O. Box 5400
Albuquerque, NM 87185
Attn: G. A. Hauquitz, QAD
V. L. Devers, QAD

1220 J. M. Wiesen
1222 G. T. Merren
1223 R. R. Prairie
1400 L. J. Heilman
1410 W. C. Kraft
1411 M. H. Thomas
1412 J. F. Calek
1414 W. E. Treibel
1415 J. T. Hillman
1417 F. W. Müller (20)
1417 E. E. Ard
1417 D. D. Spencer
1417 F. W. Spencer
1420 R. H. Schultz
1421 W. J. Denison
1422 D. E. Murphy
1423 A. C. Littleford
1424 L. G. Parsons
1426 A. J. Roth
4310 C. C. Burks
4320 G. J. Hildebrandt
4330 E. E. Ives
4340 H. W. Schmitt
4360 J. A. Hood
4370 B. E. Arthur, Jr.
8110 J. Barham
8150 J. L. Wirth
8160 R. D. Cozine
8320 R. L. Rinne
8329 O. H. Schreiber
8214 M. A. Pound
3141 L. J. Erickson (5)
3151 W. L. Garner (3)
For: DOE/TIC (Unlimited Release)
DOE/TIC (25)
(C. H. Dalin, 3154-3)