1. INTRODUCTION

In sample surveys the need often arises to select all or part of the sample by the cutoff method. A sample (or portion of a sample) selected by the cutoff method consists of all sampling units with values above a given level on some available measure of size (MOS). Units lying below the cutoff are either subjected to further sampling or omitted altogether.

Pure cutoff samples are nonprobability samples because, given the MOS information, all units are selected with certainty. As nonprobability samples, cutoff samples have been criticized by probability sampling theorists as lacking a verifiable assumption that the relationship between large and small units is relatively constant (Raj 1972; Hansen et al. 1953). However, cutoff samples have been tolerated in establishment surveys for two main reasons:

1. Many variables of interest in establishment surveys (e.g., sales volumes) are characterized by population distributions which are highly skewed. A small proportion of the population accounts for the majority of the item. The cost of adding a sample of smaller establishments would not be justified.

2. The larger units are more able to cooperate with the data needs of the surveys, due to their larger scale of operation.

Hansen et al. (1953) accept cutoff samples, with the provision that the units in the sample account for at least 90 to 95 percent of the aggregate being estimated, and the assumption of an essentially constant relationship between large and small units is tenable.

Given a single item of interest, a single-item cutoff sample is easy to draw. Either all units with MOS above a certain value could be selected (e.g., all establishments employing more than 1,000 persons) or a coverage goal could be set. In the latter case, the sample would be drawn by (1) forming the total of the MOS over all units; (2) ranking the units in descending order of the MOS, and (3) selecting all units until the cumulative MOS sum exceeds a desired proportion (e.g., 50 percent) of the total MOS.

However, this paper is concerned with cases in which there is more than one variable of interest, and more than one MOS. Extending single-item sampling procedures to the multiple-item case is never easy for any type of sample selection procedure. For example, Cochran (1977, pp. 113-123) presents several methods of sample allocation for multiple-item probability samples, none of which are as strongly supported by theory as the single-item methods.

Multiple-item cutoff samples could be chosen by sampling each variable separately, and then merging the resulting samples into one. However, such a procedure would ignore the relationships among the MOS. When adding a unit to the sample based on its MOS for one item, the unit's other MOS also contribute to the sample coverage, regardless of whether the unit would qualify for the sample based on these other MOS. In extreme cases (e.g., Table 1 for a simple example) a procedure merging single-item 50 percent samples could lead to a census of the population, with 100 percent coverage on all MOS variables.

Table 1. Selection of a Merged Single-Item 50 Percent Cutoff Sample from a Population with 4 Sampling Units and Two Measures of Size (MOS)

<table>
<thead>
<tr>
<th>Sampling Unit</th>
<th>First MOS Value</th>
<th>MOS Top 50 Percent</th>
<th>Second MOS Value</th>
<th>MOS Top 50 Percent</th>
<th>In Merged Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>No</td>
<td>4</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>No</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>Yes</td>
<td>2</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>Yes</td>
<td>3</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

As a solution to this problem, consider that, in a simple single-MOS cutoff sample, the selected sample does not just cover p percent of the MOS; it is also the smallest possible sample which covers p percent of the MOS. This observation suggests that, in the development of multiple-item cutoff sampling algorithms, a search be made for the smallest possible sample which meets all the cutoff criteria. This requirement demands that the algorithm be able to consider all MOS at once, and thus be able to take into account the effect that the selection of one unit for one MOS has on sample coverage for the other MOS.

Reconsidering the 50 percent sampling problem posed in Table 1, suppose that the sample were to be initialized by placing unit number 4 (the largest unit overall) in the sample. Examining the other 3 units, it is apparent that adding only one more unit, number 1, would satisfy the 50 percent coverage criteria for both MOS simultaneously. (In this example, a sample of number 2 and 3 would also satisfy.) The multiple-item viewpoint has therefore resulted in a 50 percent smaller sample than that produced by the
merger of separately selected single-item samples. The next section describes how this insight was developed into a formal procedure.

2. THE MULTIPLE-ITEM CUTOFF SAMPLING PROCEDURE

The multiple-item cutoff sampling program is listed in full in Appendix A. Before using the program, data preparation consists of the creation of a data set in which there is one observation for each sampling unit. Each observation contains an MOS variable for each item that is to be estimated from the sample survey data. Each sampling unit is also identified by some form of identification number. If some sampling units are to be included in the sample a priori, the observations corresponding to these units are put into a data set named CERTAIN, otherwise the observations are put into a data set named NONCERTAIN.

In lines 2-9 of the procedure, the observations from NONCERTAIN are used to initialize a matrix called NONSAMPLE with the MOS information and a column vector called SAM_ID with the corresponding identification numbers of the nonsampled units. If there are any observations in CERTAIN, these are (1) counted and (2) used to initialize the SAMPLE matrix and its corresponding SAM_ID vector (lines 10-16). If there are no observations in CERTAIN, SAMPLE, SAM_ID, and the counter are initialized with zeroes (lines 17-19).

After initializing the data matrices, a vector of totals for each MOS over all sampling units is calculated (line 20), and from this total the coverage goal is calculated (line 21). In the Appendix A listing, the goal is taken to be 90 percent of the total for all items. However, the scalar quantity .90, could be replaced by a vector specifying different coverage goals for each MOS.

Once the goal vector has been defined, sample selection can begin. The selection algorithm has a forward mode (lines 25-53) and a reverse mode (lines 54-78). In the forward mode, the general strategy is to reach the goal in as few steps (added units) as possible. This can be done by selecting at each stage the unit which can make the greatest overall contribution towards reaching the goal. The reverse mode ensures that there are no units in the tentatively selected sample (after the forward mode has terminated) which are not needed. The general strategy in reverse mode is to reduce the sampled total back down to the goal in as many steps (deleted units) as possible. This can be done by first eliminating the unit with the smallest contribution towards overfulfillment of the goal.

To determine whether it is necessary to enter forward selection mode, the program computes a DEFICIT (line 22). The DEFICIT is a vector defined as

<table>
<thead>
<tr>
<th>GOAL•SAMPLE</th>
<th>DEFICIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOAL•SAMPLE</td>
<td>0</td>
</tr>
<tr>
<td>GOAL•SAMPLE</td>
<td>0</td>
</tr>
</tbody>
</table>

If DEFICIT has any nonzero elements, the forward mode is entered. First, the amount that each nonsampled unit can contribute towards meeting the goal for each MOS is computed (line 26). The contributions are contained in a matrix, CONTRIB, whose elements are defined as

<table>
<thead>
<tr>
<th>NONSAMPLE</th>
<th>CONTRIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONSAMPLE</td>
<td>DEFICIT</td>
</tr>
</tbody>
</table>

Defining the contribution in this manner ensures that units only receive "credit" for the amount that they can contribute towards meeting the goal. Any amount a unit has which would cause the sample to exceed the goal is discarded, and only the amount which would meet the goal is counted. The unit chosen for the sample is that which would make the greatest contribution, summed over all MOS (line 27). If there are no ties for the greatest contributor, the unit is added to the SAMPLE (lines 43-54), its MOS are set to zero in NONSAMPLE, and the DEFICIT is recomputed. In case of ties, three tie-breakers are used. As the first tie-breaker, the unit which has the greatest amount (disregarding the goal) of the deficient items is identified (lines 30-32). If there is still a tie, the second tie-breaker identifies the largest unit (summed over all MOS) (lines 35-37). If the tie persists, it is finally broken using random numbers (line 39).

Once the nonzero DEFICITS have been eliminated, forward selection is terminated. At this point in the process, the SAMPLE matrix can be partitioned as

SAMPLE |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CERTAIN</td>
</tr>
</tbody>
</table>

where the upper submatrix reflects the SAMPLE initialization, and the lower submatrix consists of the units added during forward selection. To complete forward selection, the initial zero vector (if any) is eliminated from SAMPLE, and the number of units in the sample is counted (lines 48-53).

To determine whether to enter reverse mode, the program first computes a vector called the SURPLUS (line 54). The SURPLUS is defined as

<table>
<thead>
<tr>
<th>SAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CERTAIN</td>
</tr>
</tbody>
</table>

selected units from NONSAMPLE |
When sampling has been completed (indicated by the remaining observation), the program (line 55) identifies all of the units which are larger than the SURPLUS for one or more MOS. Such units are too big to be dropped from the sample. However, if the program identifies some units which are smaller than the SURPLUS for all MOS, and these units were originally in the NOTCERTN set (i.e., not required a priori for the sample), then these units become candidates for deletion in the reverse mode (lines 58-78). Thus, only units from SAMPLE’s lower submatrix can be affected by the reverse mode operations.

In reverse mode the smallest of the candidate for deletion is identified, and, if there are no ties, the unit is deleted from the sample (lines 61-74). If there is a tie for the smallest unit, the tie is broken with random numbers (lines 63-60).

After deleting the smallest candidate, the SURPLUS is recomputed, and the SAMPLE is again searched for deletion candidates (lines 76-77). Reverse mode continues until no more deletion candidates can be found.

When sampling has been completed (indicated by the message in line 79), the list of identification numbers and the selection order is output into a data set for whatever further processing is desired (e.g., a listing of the sample results merged with the MOS data).

3. THE PROCEDURE EXAMINED IN DETAIL

A Detailed Example of Multiple-Item Selection

To illustrate the operation of the sampling procedure presented in the previous section, the program (Appendix A) was run on a test data set (Appendix B) using the PRINT option in line 1 to produce detailed output for program monitoring. There were no a priori certainty units in the test data set, and each of the 122 observations was represented by eight MOS. There were four distinct items in the MOS data set: Item 1, Item 2, Item 3, and Item 4. However, in this example, estimates were required for three regions (Regions A, B, and C) for Items 1 and 4. Each observation therefore contained MOS for these two items at the regional level and for the remaining two items at the overall level. The procedure chose a sample for all of these MOS simultaneously.

Figure 1 depicts the initial stage of sample selection. The MOS had been input and the 80 percent coverage goals (indicated by a horizontal reference line) had been determined, but no units had yet been added to the sample.

Since there was an initial deficit (the initial DEFICIT equals the GOAL when there are no certain units), forward selection began. The first unit selected was number 82 (refer to observation numbers in Appendix B). Number 82 was also the largest unit in terms of the sum of the MOS. The next unit selected was number 109, followed by 39, 3, 67, 121, 4, 44, 32, and 28, in that order. After these first ten units were added to the SAMPLE, the state of the sample selection was as depicted in Figure 2. All items were partially filled, some much more than others.

For the first 10 selections, the results were identical to those which would have been obtained by selecting based on the sum of the MOS. The reason for this situation was that the SAMPLE total was not close enough to the GOAL for the type of truncation implied by the definition of CONTRIB to occur. Truncation was not observed until the seventeenth iteration, at which point the contribution of number 1 was affected due to a surplus of Item 3. Number 1 eventually was the twentieth unit selected and, as a result of its selection, Item 3 became the first item to reach full coverage. After 25 units had been selected, sample coverage reached the state shown in Figure 3.

The SAMPLE may have seemed almost complete after 25 selections. However, an additional 23 selections were made before the forward selection mode terminated. Along the way, the CONTRIB matrix continually reflected the presence of truncation. The tie-breaker procedures were invoked three times: once for a tie between numbers 9 and 88 (broken by random numbers), once for a tie between numbers 9 and 117 (broken by choosing 9), and once for a tie between numbers 9 and 8 (broken by choosing 9, which had the larger MOS for the needed item), and once (at the next-to-last iteration) to break a five-way tie between numbers 2, 38, 73, and 75 (decided in 38’s favor based on having the largest MOS for the needed item).

After forward selection terminated, the SURPLUS was calculated and compared with the MOS for the units tentatively selected for the sample. One candidate for deletion was identified, number 22 (selected at the 57th iteration). Number 22 was deleted, yielding the final 80 percent cutoff sample given in Figure 4.

Comparing Multiple- with Single-Item Selection

The multiple-item procedure illustrated in the preceding section required a good deal of work. Was that work worth it? How much different were the results from those which would have been obtained by selecting single-item cutoff samples separately and merging the results? The answer depends in part on the data and in part on the importance attached to obtaining the smallest possible sample size.
Regarding the data, the two approaches will give similar results if there are few MOS and the inter-MOS correlations are high. In an extreme case, results will be identical whenever there is just one MOS. The example in Table 2 represented another extreme case, in which the correlation between the two MOS was -1.00. However, the MOS in Appendix B are not atypical of those encountered in many establishment surveys, in which the MOS are highly skewed and are weakly correlated.

To compare the multiple- and single-item approaches empirically, cutoff samples with coverage goals of 50, 60, 70, 80, and 90 percent were selected using both procedures on the Appendix B data set. Results are presented in Figure 5 and Table 2.

Table 2. Comparison of Single- and Multiple-Item Cutoff Sample Sizes and Item Coverage Ranges for Selected Cutoff Coverage Goals

<table>
<thead>
<tr>
<th>Percent Coverage</th>
<th>Single-Item Sample</th>
<th>Multiple-Item Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>Size</td>
<td>Range (1)</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
<td>50.0-45.5</td>
</tr>
<tr>
<td>60</td>
<td>24</td>
<td>60.0-52.2</td>
</tr>
<tr>
<td>70</td>
<td>29</td>
<td>71.0-65.0</td>
</tr>
<tr>
<td>80</td>
<td>36</td>
<td>82.0-76.5</td>
</tr>
<tr>
<td>90</td>
<td>52</td>
<td>92.0-85.7</td>
</tr>
</tbody>
</table>

In each case, the multiple-item algorithm chose the smaller sample with the greatest reductions occurring for the lower coverage goals. These are fairly substantial reductions, and could well justify the additional computation time if smaller sample sizes were important. As shown by the coverage ranges, the merged single-item samples lose efficiency by not taking full advantage of all the MOS information provided on the sampling units.

4. SUMMARY AND DISCUSSION

This paper has presented a procedure for selecting multiple-item cutoff samples which takes advantage of the multivariate nature of the data. The algorithm aims at producing the smallest possible sample which satisfies all cutoff criteria. Compared with samples selected by merging separately drawn single-item cutoff samples, the samples selected by this algorithm are significantly smaller.

Some issues remain to be discussed. In the example given in the previous section, the MOS variables were taken to represent the level of the series, e.g., annual sales in thousands of dollars. With PROH MATRIX, it would be a trivial problem to select samples based on each unit's proportion of the total. Furthermore, since the goal is expressed in terms of percent coverage it would seem reasonable to work directly on the proportions. To test this conjecture, five samples were drawn using different cutoff goals, and the results of the multiple-item results presented in the previous section. Results were identical for the 50, 60, and 90 percent cutoff samples. The 70 percent sample was the same size, although two units were different. The 60 percent sample size proportions included two units not in the original sample, but excluded three other units, for a net drop of one unit.

While the preceding results suggest that there is no significant gain to be achieved by sampling proportions, the results do raise two interesting questions: (1) are the multiple-item samples unique? and (2) could the algorithm guarantee the smallest possible sample satisfying all coverage goals? Regarding uniqueness, both the example in Table 1 and the results of the 70 percent sample comparison demonstrate that the multiple-item procedure, unlike the single-item procedure, does not yield unique samples. More importantly, the case of the 60 percent sample comparison reveals that samples arrived at by the algorithm are not necessarily the smallest possible samples satisfying all the cutoff criteria.

The reason that samples chosen by the algorithm may be larger than the absolute minimum is that the algorithm only deals with one sampling unit at a time. The resulting sample is one which could not be improved (i.e., made smaller) by removing any one of its members. However, there might be possibilities of improvements by deleting n units from the sample and replacing them with m different units, where n > m. In the 60 percent sample comparison, the proportion-based sample replaced three units from the original sample with two other units. Unfortunately, a practical way of identifying these m and n units in general remains elusive.

Finally, different starting rules could be formulated to speed up the algorithm. If there are many units, with small values for the MOS, the algorithm will require a good deal of computation. Since the selection algorithm operates as a select-the-largest-overall algorithm until the table is approached for one MOS, the sample could be initialized with a single-item sample on the total of all MOS variables. The algorithm described in this paper could then be used to fine-adjust the sample by adding units to satisfy any remaining deficits and by eliminating any surplus units. Computer time was not critical for the sample selection example given in the previous section.
5. REFERENCES


Appendix A

A Listing of the Cutoff Sampling Program

1. READ DATA APPENDIX.
2. FETCH RANDOM.
3. CALL MCHL.
4. IF (ITEM 14, ITEM 18, ITEM 2, ITEM 8) ITEM 14, ITEM 15, ITEM 16, ITEM 17, ITEM 18.
5. FETCH ITEM 19.5.
6. RXN RESS (ITEM 10).
7. CALL MCHL (RXN).
8. CALL MCHL.
9. CALL MCHL.
10. CALL MCHL.
11. CALL MCHL.
12. CALL MCHL.
13. CALL MCHL.
14. CALL MCHL.
15. CALL MCHL.
16. CALL MCHL.
17. CALL MCHL.
18. CALL MCHL.
19. CALL MCHL.
20. DO WHILE (ANY (DEFINIT)).
21. CALL MCHL (ITEM 14, ITEM 15, ITEM 16, ITEM 17, ITEM 18).
22. CALL MCHL (ITEM 19).
23. CALL MCHL (ITEM 20).
24. CALL MCHL (ITEM 21).
25. CALL MCHL (ITEM 22).
26. CALL MCHL.
27. CALL MCHL.
28. CALL MCHL.
29. CALL MCHL.
30. CALL MCHL.
31. CALL MCHL.
32. CALL MCHL.
33. CALL MCHL.
34. CALL MCHL.
35. CALL MCHL.
36. CALL MCHL.
37. CALL MCHL.
38. CALL MCHL.
39. CALL MCHL.
40. CALL MCHL.
41. CALL MCHL.
42. CALL MCHL.
43. CALL MCHL.
44. CALL MCHL.
45. CALL MCHL.
46. CALL MCHL.
47. CALL MCHL.
48. CALL MCHL.
49. CALL MCHL.
50. CALL MCHL.
51. CALL MCHL.
52. CALL MCHL.
53. CALL MCHL.
54. CALL MCHL.
55. CALL MCHL.
56. CALL MCHL.
57. CALL MCHL.
58. CALL MCHL.
59. CALL MCHL.
60. CALL MCHL.
61. CALL MCHL.
62. CALL MCHL.
63. CALL MCHL.
64. CALL MCHL.
65. CALL MCHL.
66. CALL MCHL.
67. CALL MCHL.
68. CALL MCHL.
69. CALL MCHL.
70. CALL MCHL.

Appendix B

Test Data Used for the Sample Selection Example
Figure 1. Sample Coverage at Time of Initialization

Figure 2. Sample Coverage After Selection of First 10 Units

Figure 3. Sample Coverage After Selection of First 25 Units

Figure 4. Sample Coverage at the Completion of Sample Selection

Figure 5. Comparison of Single and Multiple Item Cutoff Sample Sizes for Selected Cutoff Coverage Goals