THE ALPHABETIC ARRAY SAMPLING TECHNIQUE

Abraham Frankel and Patricia Wright

U. S. Office of Education

In many of the surveys conducted by the U. S. Office of Education, it is often necessary to select samples of elementary reporting units such as students and teachers. Because listings of these units are usually not available, a two-stage sample design has been used in such situations. In order to facilitate the selection of samples of these elementary units, an alphabetic array sampling technique was designed.[1] Informal studies have shown that the first letter of the surnames of the faculty and students in educational institutions is generally distributed in a non-random fashion; that is, some of the letters appear more often than others. On this basis, in the alphabetic array technique the letters of the alphabet may be grouped so that selections may be made with known probabilities. Instructions are then issued to respondents in the first stage of the sample design to select for the sample of elementary units a group or a combination of groups of letters. For example, some university deans might be asked to select in their sample all faculty members whose surname began with the letters A, P, R, or T. Other deans might be requested to select other combinations of letters in order to provide the required sample size.

The sampling design of the alphabetic array is based on the fact that the relative frequency of the first letters of the surnames is quite stable for large populations. Hence, the technique involves arranging the 26 letters of the alphabet into groups of letters representing the first letter of the surname of individuals in such a way that each group of letters includes about the same proportion of all names of the large population. That is, for an n line alphabetic array,

$$\Sigma p_{i1} = \Sigma p_{i2} = \cdots = \Sigma p_{ij} = \cdots = \Sigma p_{in}$$

and

$$\Sigma p_{i1} + \Sigma p_{i2} + \Sigma p_{i3} + \cdots + \Sigma p_{in} = 1.00$$

where

$j = 1, 2, 3, \ldots, n$
n = the number of lines in the array
p_{ij} = the probability of the occurence of the i-th letter which is grouped in the j-th line
Σp_{ij} = the probability of the occurence of all the letters which are grouped in the j-th line

The relative frequency of the occurence of the letters that is used in this technique is based on the surnames in the Social Security records which are fairly precise and surely representative of all surnames in the United States. The percentage distribution of surnames by initial letter from the Social Security records and from other listings are shown in Table 1.

Five and six line alphabetic arrays were developed from this distribution. In the five line array the sum of the probabilities of the occurence of the letters in each line should equal .20 or 1/5th of the population; in the six line alphabetic array the sum of the probabilities should equal .1667 for each line. The groupings of the letters to form these two arrays are shown below.

5 Line Alphabetic Array

Line	Letters	Σp_{ij}
1	J K O S V	.1980
2	B G L U Y	.1992
3	I M T W Z	.2011
4	C E F H X	.2023
5	A D N P Q R	.1994
Total		1.0000

6 Line Alphabetic Array

Line	Letters	Σp_{ij}
1	F J S X	.1678
2	B E L U Y	.1670
3	D M N Z	.1657
4	C G I K	.1670
5	H O Q V W	.1662
6	A P R T	.1663
Total		1.0000

The alphabetic array sampling technique was used in the sample selection procedure for the survey of the "Status and Career Orientation of College Faculty (COLFACS)" which was conducted by the U. S. Office of Education in the spring of 1963. The purpose of this study was to obtain basic data (personal characteristics, position and assignment, educational background, work experience, economic status and occupational plans) about college faculty in the aggregate United States. The universe of inquiry was defined as the full-time faculty with the rank of instructor or above, who taught at least one degree credit course in the spring of 1963 in universities, liberal arts colleges, teachers colleges or independent technological schools.

The sample design provided for a two-stage stratified sample (see Table 2). In the first stage, the universe of institutions was stratified by public or private control, by type of institution, and by size of faculty. The institutions that were included in the sample were selected with varying probabilities of 1:1, 1:2 or 1:5 depending upon the number of institutions in the stratum universe. In the second stage, the number of faculty that were sampled was selected in such a manner that the overall sampling fraction in each stratum was 1:10. In other words, where the institution

[1] Philip Desind, now with the Post Office Dept., developed the alphabetic array sampling technique in 1962 when he was with the U] S] Office of Education.

TABLE 1--PERCENTAGE DISTRIBUTION OF SURNAMES BY INITIAL LETTER FROM VARIOUS LISTINGS

Initial Letter	Social Security[1]	Who's Who in America (Vol. 30)	Who's Who in American Education (19th Ed.)	Washington,D.C. Metropolitan Telephone Dir. (Fall 1961)	ASA Directory 1961
A	3.051	3.223	3.536	3.327	3.310
B	9.357	9.947	9.491	9.620	9.172
C	7.267	7.406	7.072	7.594	7.162
D	4.783	4.803	4.901	4.846	4.279
E	1.888	2.138	2.171	1.917	1.978
F	3.622	3.936	4.032	3.653	4.081
G	5.103	4.648	5.025	4.991	5.372
H	7.440	8.119	8.561	7.594	6.839
I	.387	.434	.558	.362	.552
J	2.954	2.107	2.543	2.821	1.853
K	3.938	3.874	4.280	3.616	5.236
L	4.664	4.741	4.839	4.484	5.070
M	9.448	9.482	8.933	9.475	9.109
N	1.785	1.735	2.047	1.700	2.238
O	1.436	1.394	1.365	1.302	1.426
P	4.887	4.307	4.032	4.774	4.310
Q	.175	.155	.124	.217	.167
R	5.257	4.896	4.529	4.991	4.914
S	10.194	10.381	10.361	10.306	10.639
T	3.450	3.223	2.854	3.544	3.310
U	.238	.279	.248	.325	.354
V	1.279	1.209	1.365	1.085	1.249
W	6.287	6.694	6.017	6.582	5.944
X	.003	.019	.000	.007	.000
Y	.555	.465	.620	.506	.666
Z	.552	.403	.496	.361	.770
	100.0	100.0	100.0	100.0	100.0
Total Names	117,358,888	3,227	1,612	690,000[2]	9,606

[1] In 1957 the social security record keeping operations were converted to an electronic data processing system. At that time the Division of Accounting Operations, BOASI, SSA, HEW, compiled the material shown in this column. It represents all account numbers issued from the beginning of the social security program in 1936 to the middle of 1956.

[2] This represents the total listings and not the individual names in the telephone directory.

sampling fraction was 1:1, the faculty sampling fraction was 1:10; where it was 1:2, the faculty sampling fraction was 1:5; and where it was 1:5, the faculty sampling fraction was 1:2.

The overall 1:10 sampling rate was obtained by the use of the alphabetic array sampling technique in the following manner. If the institution in the sample was selected with a sampling fraction of 1:1, a line was selected at random from the five line alphabetic array; for example, line #3. Then that institution was told to include in their sample all faculty members meeting the specifications stated above whose surname began with the letters I, M, T, W or Z. Since this provided a 20 percent (1/5th) sample of the faculty under study, it was necessary to select, after a random start, every other name in order to obtain the desired 1:10 sample.

For those institutions that were selected with a 1:2 sampling fraction, a line was selected at random from the five line alphabetic array and the institution was asked to include in their sample those faculty members whose surname began with the letters of the selected line. No subsamples were taken as the required sample take of 1:5 was obtained.

For those institutions that were selected with a 1:5 sampling fraction, a sample of 1:2 of the faculty was required. This was obtained by taking three lines at random from the six line alphabetic array and forwarding these letters to the institution to be used in selecting their faculty to be included in the sample.

The sample selection plan for the COLFACS survey may be summarized as follows:

	Sampling Fraction		
First stage:			
Select institutions	1:1	1:2	1:5
Second stage:			
Select faculty:			
a. Use 1 line of five line array -	1:5	1:5	-
b. Use 3 lines of six line array -	-	-	1:2
c. Subsample -	1:2	-	-
Overall sampling fraction -	1:10	1:10	1:10

TABLE 2--SAMPLE DESIGN USED IN THE SURVEY OF STATUS AND CAREER ORIENTATION OF COLLEGE FACULTY, SPRING 1963

Type of		Public Control				Private Control			
Institution	Faculty Size	M	m/M	m	n/N	M	m/M	m	n/N
All Inst.		360		289		782		304	
University	Total	82		82		58		58	
	1,000 & over	15	1:1	15	1:10	4	1:1	4	1:10
	500--999	24	1:1	24	1:10	22	1:1	22	1:10
	Under 500	43	1:1	43	1:10	32	1:1	32	1:10
Liberal Arts	Total	85		85		668		190	
	150 & over	31	1:1	31	1:10	27	1:1	27	1:10
	75--149	27	1:1	27	1:10	115	1:2	58	1:5
	Under 75	27	1:1	27	1:10	526	1:5	105	1:2
Teachers College	Total	167		96		31		31	
	150 & over	23	1:1	23	1:10	0		0	
	75--149	69	1:2	35	1:5	1	1:1	1	1:10
	Under 75	75	1:2	38	1:5	30	1:1	30	1:10
Technological Schools	Total	26		26		25		25	
	150 & over	7	1:1	7	1:10	6	1:1	6	1:10
	75--149	9	1:1	9	1:10	6	1:1	6	1:10
	Under 75	10	1:1	10	1:10	13	1:1	13	1:10

M = number of institutions in the universe (1,142)
m = number of institutions in sample (593)
N = number of faculty in the universe
n = number of faculty in sample
m/M = sampling fraction for selecting institutions
n/N = sampling fraction for selecting faculty members

All institutions that were selected in the sample responded. A total of 15,494 names was received and questionnaires were mailed to these individuals. The response breakdown is shown below.

Number mailed out	15,494
Number of useable forms received	13,017
Number out of scope	1,694 1/
Refusals	56
Nonresponse	727

The number of faculty in 1962-63 meeting the survey specifications was estimated from the preliminary reports on "Faculty and Other Professional Staff in Institutions of Higher Education, 1961-62" (OE: 53014-62, Nov. 1963) and for 1963 - 64. These reports provide a count of the faculty as of the first semester of the 1961-62 and the 1963-64 academic years, respectively, classified by types of duty. The estimated number in 1961 was 129,700 and in 1963, 148,000. The number in 1962 is estimated at 138,500. The sample take, using the alphabetic array technique was 9.94 percent of the estimated universe.

II

In the belief that it may be of interest to others we discuss in this section of the paper a method of selecting elementary sampling units through the use of an n x 1 matrix where, by use of combinatorial methods, it is possible to include a fixed number of elementary units in the sample with a lesser or a larger number of primary sampling units (psu's) as desired. A larger number of psu's usually has the advantage of reducing the variance of estimates if the between psu variance is large. Conversely, when the within psu variance is large, a smaller number of psu's with an increased number of elementary units will tend to reduce the overall variance.

Let us consider a universe of M primary sampling units that has a total of T elementary units. A sample of t elementary units is to be selected from this universe. The following method may be used for selecting a sample of the primary sampling units from the universe and for determining the sampling fractions to be used in selecting the elementary units for a fixed sample size.

If the k-th row of an n x 1 matrix is arbitrarily selected to divide the n rows of a matrix into two groups, then we define those rows in the matrix from 1 through k as letter rows and those rows from k + 1 to n as number rows. This is shown in the

1/ This includes members of the faculty who were not employed full-time or who were serving in administration, personnel services, research, etc., and were not teaching.

figure below where

k = the number of letter rows
ϕ = the number of number rows
$n = k + \phi$

$$\left[\begin{array}{c} 1\\ 2\\ 3\\ \cdot\\ \cdot\\ k\\ k+1\\ k+2\\ \cdot\\ \cdot\\ n \end{array}\right] \begin{array}{l} \left.\begin{array}{c} a_1\\ a_2\\ a_3\\ \cdot\\ \cdot\\ a_k \end{array}\right] \text{--- letter rows}\\ \left.\begin{array}{c} 1\\ 2\\ \cdot\\ \cdot\\ \phi \end{array}\right] \text{--- number rows} \end{array}$$

The n rows of the matrix are then combined p rows at a time. The total number of combinations is

$$\binom{n}{p} = \frac{n!}{p!(n-p)!} \qquad (1)$$

This equation may be rewritten as

$$\binom{n}{p} = \binom{k}{p}\binom{\phi}{0} + \binom{k}{p-1}\binom{\phi}{1} + \ldots + \binom{k}{p-(p-j)}\binom{\phi}{p-j} + \ldots + \binom{k}{0}\binom{\phi}{p} \qquad (2)$$

where $k \geq p$ and $\phi \geq p$. For clarity, we call that part of the combination that has the letter k in it the letter row component of the combination and the part that has ϕ in it the number row component.

Equation (1) provides the number of combinations into which the universe (M) has been divided and equation (2) the composition of the combinations according to the number of letter and number rows.

As the total number of psu's in the universe is M, the average number of psu's in each combination is

$$\bar{M} = \frac{M}{\binom{n}{p}} \qquad (3)$$

and the total number of psu's associated with the different combinations in equation (2) is

$$M = \bar{M}\binom{k}{p}\binom{\phi}{0} + \bar{M}\binom{k}{p-1}\binom{\phi}{1} + \ldots \bar{M}\binom{k}{0}\binom{\phi}{p} \qquad (4)$$

In order to determine the number of elementary sampling units that is to be taken from the psu's in the sample a weighting factor (w) is used in each of the different combinations. The value of the weighting factor w_i for the i-th combination is the value of p in the letter row component of the combination. For example, in equation (4), w equals p in the first combination; w equals (p-1) in the second combination; and w equals 0 in the last combination.

For the desired sample size (t) the following equation is set up and evaluated for x.

$$t = \bar{M}\binom{k}{p}\binom{\phi}{0}w_p x + \bar{M}\binom{k}{p-1}\binom{\phi}{1}w_{p-1}x + \ldots + \bar{M}\binom{k}{0}\binom{\phi}{p}w_0 x \qquad (5)$$

Then

$$\frac{w_i x}{\frac{T}{M}} \qquad (6)$$

is the sampling fraction to be used to select the elementary sampling units from those psu's that fall in the combination having the value of p_i in the letter row component of the combination.

An illustration of the above method is shown in the following example. A sample of 10,000 (t) is to be selected from a universe of 50,000 (T) teachers. The number of schools in the universe is 1,000 (M). It is believed, from other sources, that the between psu variance for the characteristic under study is high so that we would want a large number of schools in our sample. As our first trial we set up a 10 x 1 matrix with 5 letter rows (k) and 5 number rows (ϕ) and take the 10 rows in combination 2 at a time (p).

Using equations (1) and (2), we get

$$\binom{10}{2} = \binom{5}{2}\binom{5}{0} + \binom{5}{1}\binom{5}{1} + \binom{5}{0}\binom{5}{2}$$

$$45 = 10 + 25 + 10$$

This means that the 45 combinations are identified in the following manner: 10 combinations having 2 letter rows; 25 combinations having 1 letter and 1 number row; and 10 combinations having 2 number rows.

Note 1. Table 3 shows the number of combinations that result from various n x 1 matrices and row combinations.

The average number of schools in each combination is, by equation (3),

TABLE 3--NUMBER AND COMPOSITION OF COMBINATIONS FOR N x 1 MATRICES WITH 5 LETTER ROWS AND DIFFERENT ROW COMBINATIONS

Total no. of rows (N)	5	6	7	8	9	10	15	20	25
No. of letter rows	5	5	5	5	5	5	5	5	5
No. of number rows	0	1	2	3	4	5	10	15	20
2 row combinations									
Total combinations	10	15	21	28	36	45	105	190	300
2 letters (ij)	10	10	10	10	10	10	10	10	10
1 letter (i0)	0	5	10	15	20	25	50	75	100
0 letters (00)	0	0	1	3	6	10	45	105	190
3 row combinations									
Total combinations	10	20	35	56	84	120	455	1140	2300
3 letters (ijk)	10	10	10	10	10	10	10	10	10
2 letters (ij0)	0	10	20	30	40	50	100	150	200
1 letter (i00)	0	0	5	15	30	50	225	525	950
0 letters (000)	0	0	0	1	4	10	120	455	1140
4 row combinations									
Total combinations	5	15	35	70	126	210	1365	4845	12650
4 letters (ijkl)	5	5	5	5	5	5	5	5	5
3 letters (ijk0)	0	10	20	30	40	50	100	150	200
2 letters (ij00)	0	0	10	30	60	100	450	1050	1900
1 letter (i000)	0	0	0	5	20	50	600	2275	5700
0 letters (0000)	0	0	0	0	1	5	210	1365	4845

$$\bar{M} = \frac{M}{\binom{n}{p}} = \frac{1,000}{45} = 22.2$$

Substituting in equation (5)

$$10,000 = 22.2\binom{5}{2}\binom{5}{0}2x + 22.2\binom{5}{1}\binom{5}{1}1x + 22.2\binom{5}{0}\binom{5}{2}0x$$

and solving for x, we find that x = 10.

As T/M = 50,000/1,000 = 50, the sampling fraction to be used in those schools associated with the first type of combination (2 letter rows) is, by equation (6), (2 x 10)/50 = 2/5; for the second type (1 letter row and 1 number row), (1 x 10)/50 = 1/5; and for the third type (2 number rows), (0 x 10)/50 = 0; that is, schools in the combination identified by 2 number rows will not be included in the sample.

The sample selection procedure then involves the selection of 222 schools at random from the 1,000 schools in the universe and denoting them as 2 letter row schools and selecting 556 schools at random and denoting them as 1 letter and 1 number row schools. A 40 percent sample (2:5) of the teachers will be selected in the former schools and a 20 percent sample (1:5) of the teachers in the latter schools.

The results of the above example along with those obtained by using other n x 1 matrices are shown in Table 4. In all of these matrices the number of letter lines was equal to 5 and the rows were taken in combination 2 at a time.

TABLE 4 - RESULTS OBTAINED BY USING DIFFERENT n x 1 MATRICES

	n x 1 matrix			
	n = 5	n = 10	n = 20	n = 25
Number of combinations	10	45	190	300
Number of schools in sample	1,000	778	448	366
Number of schools (2 letter rows)	1,000	222	53	33
Sampling fraction for teachers	1:5	2:5	4:5	1:1
Number of schools (1 letter row and 1 number row)	0	556	395	333
Sampling fraction for teachers	-	1:5	2:5	1:2