DESIGN EFFECTS FOR ALPHABETIC CLUSTER SAMPLES

Lanny L. Piper and James R. Chromy, Research Triangle Institute

INTRODUCTION

In a number of sampling applications, involving selection of individuals, alphabetized lists of the individuals in the target population are available or obtainable. Standard methods employed to sample such lists generally involve serial numbering and selection of the sample using simple random sampling or systematic sampling with a random start point. Another alternative within the probability sampling context, is to select a sample of alphabetic segment clusters with each cluster defined in terms of a portion of an alphabetized list. Words in dictionaries are naturally clustered on pages; these clusters are defined unambiguously in terms of the first word appearing on a page and continuing up to but not including the first word appearing on the next page. This same type of definition could be used to identify clusters of words on any alphabetized list if proper provision were made for repeated words. The first sample word in a selected cluster would be either the first word defining the cluster or if the defining first word were not in the sampling frame, then the first sample word would be the word in the frame immediately following the defining first word in alphabetical order. The cluster then contains words in the ordered list up to, but not including the first word of the next cluster.

If an alphabetized sampling frame consisting of the names of members of some target population exists, it is possible to select a probability sample of alphabetic segment clusters without having immediate access to the list. Unambiguous instructions can be written for the data collection staff for obtaining the list and specifying the sample. The principal disadvantages of the technique are some loss of control over the size of the total sample and possible increased variance of estimates due to homogeneity of clusters and variability in cluster size.

APPLICATIONS AND PRIOR RESULTS

Alphabetic clusters can be used as first-stage sampling units if the entire target population is listed on an alphabetic file. Examples of such target populations are association memberships or registries. The technique was applied to first stage units defined on the National Register of Scientific and Technical Personnel [4]. That methodological study investigated the potential increases in variances due to the homogeneity of alphabetic segment clusters with cluster sizes 215 and 430. For almost all statistics studied the increase in variance due to clustering was less than 20 percent.

Possibly, a greater potential exists using alphabetic segment clusters in the selection of second-stage samples where the first-stage units are schools, hospitals, doctors' offices, or other units that maintain alphabetized files on students, patients, or other individuals. Some examples of applications at the second stage of sampling are discussed below.

The technique was applied to select a sample of high school seniors to determine their responses to a job skills screening questionnaire [2]; further analysis of data from this study is presented later in this paper. The procedure is also being employed in another study as a means of reducing the workload required at the individual school level to identify lists of graduates. These lists are screened to identify those students who graduated from high school before attaining the age of sixteen and one-half years old [3].

Another application of alphabetic segment clusters was the selection of a sample of patient medical records from doctors' office and nursing home files. These clusters were designed to contain approximately an equal number of records. These equal-sized clusters were then combined so that approximately 200 records could be identified in the sampled alphabetic segments. From this sample, 50 records were selected for the purpose of abstracting medical history information.

ALPHABETIC SEGMENT DEFINITIONS

Various sources containing names of persons are available to construct any desired number of equal sized or unequal-sized alphabetic segments. Examples of such sources are telephone directories, membership directories of professional organizations, employee listings, and various types of computerized listings. Such sources can be used either individually or in combination.

Probably the most readily available source is the telephone directory. Table 1 shows the accumulated percentages for each letter of the alphabet based on the names listed in five different telephone directories:

1. Rochester, New York;
2. Raleigh, North Carolina;
3. Lincoln, Nebraska;
4. St. Paul, Minnesota;
5. Phoenix, Arizona.

The percentages shown in this table were computed from the number of pages having names beginning with the appropriate letter (approximated to the nearest one-tenth of a page). Table 2 shows further use of the telephone directory to define varying numbers of approximately equal sized alphabetic segments. Once the 36 segments were defined, different combinations of these segments were applied to form 18, 12, 9, 6, and 4 approximately equal-sized segments respectively.

The alphabetic segments shown in table 3 were constructed from a computerized list of participants in the federally sponsored Upward Bound program during the fall of 1973. This list contained over 6,100 students in grades 10 through 12 who were from a low socioeconomic background and considered to be academic risks. Both the set of 35 segments and the set of 25 segments were constructed so that approximately equal proportions of students were included in each segment.

Table 4 shows 28 approximately equal-sized alphabetic segments defined using the 1970 American Statistical Association directory [4]. The segments in this table were also constructed to contain approximately equal proportions of members listed in the directory.

Table 1. Accumulated percentage of number of pages in the specified telephone directory having names beginning with the designated letter

	St. Paul, Minn.	Lincoln, Neb.	Phoenix, Ariz.	Rochester, N.Y.	Raleigh, N.C.	Average
			(cumulative percentages)			
A	4.1	3.6	4.8	3.5	3.9	4.0
B	12.1	12.1	13.5	12.0	13.3	12.6
C	17.9	18.2	21.0	19.8	21.2	19.6
D	21.8	26.2	25.1	25.2	25.1	24.7
E	24.2	28.4	27.3	27.1	27.5	26.9
F	27.8	31.9	31.0	31.2	30.6	30.5
G	32.4	35.9	35.8	36.2	34.8	35.0
H	39.4	43.7	42.9	42.4	43.4	42.4
I	40.0	44.2	43.4	43.1	43.8	42.9
J	43.4	46.9	46.0	44.8	47.8	45.3
K	48.6	51.5	49.5	49.1	50.3	49.8
L	53.9	56.2	53.8	54.0	54.1	54.4
M	62.9	64.3	62.9	63.6	63.2	63.4
N	65.7	66.9	64.8	65.8	65.4	65.7
O	67.8	68.4	66.2	67.2	66.3	67.2
P	72.6	72.6	71.2	72.3	72.1	72.2
Q	72.8	72.8	71.4	72.6	72.3	72.3
R	77.6	77.4	76.5	77.6	76.6	77.1
S	88.7	88.2	87.1	88.3	86.2	87.7
T	92.0	90.9	90.6	91.4	90.0	91.0
U	92.5	91.3	91.2	91.7	91.0	91.5
V	93.8	92.6	92.8	93.6	91.5	92.9
W	98.9	98.5	98.2	98.8	99.1	98.7
X	98.9	98.5	99.0	98.8	99.1	98.9
Y	99.3	99.1	99.5	99.3	99.8	99.4
Z	100.0	100.0	100.0	100.0	100.0	100.0

ANALYSIS OF THE JOB SKILLS DATA

The technique of alphabetic segment cluster sampling was applied in a mail survey of high school seniors to determine how many would claim to possess certain specified job skills. The data were to be used to develop efficient screening procedures for a national job skills survey for the National Assessment of Educational Progress. Seniors were asked to write yes after each of thirty different job skills listed that fit the category "can do now".

The sample was selected as a subsample of schools participating in a previous study. Two or three alphabetic segment clusters were randomly selected and specified in those schools where sampling of students was required to obtain a sample of approximately 50 seniors. Table 5 shows the sampling rates as determined by each school's estimated senior class enrollment. The segments specified for selection are shown in table 2.

Since time constraints did not allow a pretest of the sample selection procedures, a decision was reached to employ expanded segment definitions to make sampling instructions easier to follow for the school personnel. As a result specified segments were enlarged; e.g., if the segment AAA-BAR was selected, field instructions were to include all seniors with last names beginning with A or B. This modification affected all alphabetic segments except the 1/4 segments. Ratio estimates at the school were used to partially adjust for any bias introduced by this modification and for bias due to student nonresponse.

Since the school selection was through a complex two-phase sample, the total design effect would reflect a number of factors other than alphabetic cluster sampling. The within-school sample selection was, however, relatively straightforward. Design effects for estimated proportions of students able to do seven selected job skills were computed for a sample of 57 schools where student sampling was required. Within-school design effects estimates were obtained by computing the variance of the ratio estimate for cluster sampling and dividing by the simple random sampling variances based on the estimated proportion and a sample size equal to the combined size of the selected clusters.

Averages of the within-school design effects are shown in table 6 . The overall average increase in within-scnool variance due to clustering was 19 percent. No perceivable relation of average design effect to school size is evident from these estimates. Table 7 shows the job skills "can do" estimates in decreasing order of magnitude. "Stenographer" has the lowest estimated proportion and the lowest average within-school design effect. The remaining design effects do not exhibit any strong relation to the level of estimate.

A comparison of the expected sample size and the reported sample size using alphabetic segment clusters in 78 schools where sampling was required for the Job Skills study is shown in table 8 . The

Table 2. Approximately equal-sized alphabetic segments constructed from five telephone directories

Alphabetic segment number	Approximate proportion of names contained in each segment					
	1/36	1/18	1/12	1/9	1/6	1/4
1	AAA-ARM					
2	ARN-BAR	AAA-BAR				
3	BAS-BLZ		AAA-BLZ			
4	BMA-BRO	BAS-BRO		AAA-BRO		
5	BRP-CAQ					
6	CAR-CNZ	BRP-CNZ	BMA-CNZ		AAA-CNZ	
7	COA-CRD					
8	CRE-DED	COA-DED		BRP-DED		
9	DEE-DZZ		COA-DZZ			AAA-DZZ
10	EAA-FEZ	DEE-FEZ				
11	FFA-GEN					
12	GEO-GZZ	FFA-GZZ	EAA-GZZ	DEE-GZZ	COA-GZZ	
13	HAA-HAX					
14	HAY-HOK	HAA-HOK				
15	HOL-HZZ		HAA-HZZ			
16	IAA-JOH	HOL-JOH		HAA-JOH		
17	JOI-KEK					
18	KEL-KZZ	JOI-KZZ	IAA-KZZ		HAA-KZZ	EAA-KZZ
19	LAA-LIS					
20	LIT-MAR	LAA-MAR		JOI-MAR		
21	MAS-MDZ		LAA-MDZ			
22	MEA-MON	MAS-MON				
23	MOO-NAX					
24	NAY-OZZ	MEA-OZZ	MEA-OZZ	MAS-OZZ	LAA-OZZ	
25	PAA-PIN					
26	PIO-RAX	PAA-RAX				
27	RAY-RZZ		PAA-RZZ			LAA-RZZ
28	SAA-SEA	RAY-SEA		PAA-SEA		
29	SEB-SIQ					
30	SIR-SNZ	SEB-SNZ	SAA-SNZ		PAA-SNZ	
31	SOA-STQ					
32	STR-THN	SOA-THN		SEB-THN		
33	THO-UZZ		SOA-UZZ			
34	VAA-WER	THO-WER				
35	WES-WIL					
36	WIM-ZZZ	WES-ZZZ	VAA-ZZZ	THO-ZZZ	SOA-ZZZ	SAA-ZZZ

expected values were computed using the "Average" column of table 1 for the appropriately selected alphabetic segments in each school and multiplying this proportion by the estimated total number of seniors in the school.

CONCLUSION

The general methodology of using alphabetic segment clusters for selecting samples appears useful for many applications. This study confirms the results of an earlier study that only moderate increases in variance are realized due to the alphabetic segment clustering. The decision to apply the procedure in specific cases must, of course, be based on the relative costs of applying alternative sample selection procedures, as well as on the relative magnitudes of the variance of key estimates.

REFERENCES

[1] Edward G. Bryant, Nancy W. Caldwell, Morris H. Hansen, and Vernon E. Palmour, A Study of the Use of Sampling in Surveys of Scientific and Technical Personnel, Westat Inc., Report to National Science Foundation, June 1970.

[2] James R. Chromy and Lanny L. Piper, A Survey of High School Seniors to Determine Responses to the Job Skills Sampling Questionnaires, Final Report to the National Assessment of Educational Progress, Research Triangle Institute, October 1973.

[3] Bruce L. Jones, NAEP Year 05 Supplementary Frame Sampling, Weighting, and Survey Results, NAEP Working Paper No. 10, Research Triangle Institute, August 1975.

[4] 1970 Statisticians and Others in Allied Professions, The American Statistical Association.

Table 3. Approximately equal-sized alphabetic segments constructed from a computerized roster of Upward Bound participants from the fall of 1973

Segment number	Segment definition	Segment number	Segment definition
1	Aaa-Arr	1	Aaa-Bal
2	Ars-Bea	2	Bam-Blu
3	Beb-Boo	3	Blv-Bry
4	Bop-Bro	4	Brz-Cha
5	Brp-Can	5	Chb-Cri
6	Cao-Cla	6	Crj-Doo
7	Clb-Cri	7	Dop-Fek
8	Crj-Den	8	Fel-Gar
9	Deo-Edw	9	Gas-Gri
10	Edx-Fit	10	Grj-Hen
11	Fiu-Gar	11	Heo-Hyl
12	Gas-Gor	12	Hym-Jon
13	Gos-Ham	13	Joo-Lem
14	Han-Hen	14	Len-Mar
15	Heo-How	15	Mas-Met
16	Hox-Jel	16	Meu-Nat
17	Jem-Kan	17	Nau-Peo
18	Kao-Lea	18	Pep-Ray
19	Leb-Lug	19	Raz-Roo
20	Luh-May	20	Rop-Sha
21	Maz-Met	21	Shb-Spe
22	Meu-Mor	22	Spf-Tho
23	Mos-Ort	23	Thp-Vil
24	Oru-Pet	24	Vim-Wig
25	Peu-Ram	25	Wih-Zzz
26	Ran-Rob		
27	Roc-Rui		
28	Ruj-Sha		
29	Shb-Smi		
30	Smj-Sto		
31	Stp-Tho		
32	Thp-Var		
33	Vas-Wea		
34	Web-Wil		
35	Wim-Zzz		

Table 4. Approximately equal-sized alphabetic segments constructed from the American Statistical Association directory

Segment number	Segment definition	Segment number	Segment definition
1	Aaa-Bah	15	Lad-Lin
2	Bai-Bil	16	Lio-Mar
3	Bim-Bru	17	Mas-Mey
4	Brv-Che	18	Mez-Nad
5	Chd-Cra	19	Nae-Oub
6	Crb-Div	20	Ouc-Pos
7	Diw-Eva	21	Pot-Rob
8	Evb-Fri	22	Roc-Sch
9	Frj-Gol	23	Sci-Sie
10	Gom-Han	24	Sif-Sta
11	Hao-Hir	25	Stb-Ter
12	His-Jam	26	Tes-Ver
13	Jan-Ker	27	Ves-Whi
14	Kes-Lac	28	Whj-Zzz

Table 5. Planned student sampling rates and expected sample sizes

School senior enrollment		Prescribed sampling	Expected sample size	
From	To	rate	From	To
1	100	all	1	100
101	120	2/4	50	60
121	180	2/6	40	60
181	270	2/9	40	60
271	360	2/12	45	60
361	540	2/18	40	60
541	720	3/36	45	60
721	1,080	2/36	40	60

Table 6. Average within-school design effects by size of school and job skill

			Average estimated within-school design effects							
Size range	Target sampling rate	No. of schools (n)	Waiter-waitress	Stenog-rapher	Typist	Retail sales person	Carpenter	Service station attendant	Tractor operator	Average
101- 120	2/4	2	.07	.54	.36	.45	1.58	2.03	1.56	.94
121- 180	2/6	13	1.63	1.18	1.00	.41	.97	.88	.89	1.00
181- 270	2/9	12	1.04	1.05	.94	1.39	.63	.85	.72	.95
271- 360	2/12	7	1.50	.94	1.91	1.89	.62	1.55	1.63	1.43
361- 540	2/18	9	1.00	.41	1.27	1.16	1.21	.89	2.58	1.22
541- 720	3/36	11	2.38	.55	.90	2.52	2.90	1.43	1.24	1.70
721-1080	2/36	3	.63	.05	.22	1.85	.13	.47	.63	.63
Average		57	1.43	.80	1.06	1.40	1.24	1.08	1.29	1.19

Table 7. Average within-school design effects by level of estimate

Job skill	Estimated proportion that "can do"	Average design effect
Waiter-waitress	.61	1.43
Service station attendant	.42	1.08
Typist	.38	1.06
Retail sales	.31	1.40
Farm tractor operator	.19	1.29
Carpenter	.19	1.24
Stenographer	.12	.80

Table 8. Expected and reported sample sizes using alphabetic segment clusters in the job skills study by size of school

Size range	Expected no. of students	Reported no. of students	Number of sample schools
101- 120	121.7	84	2
121- 180	812.0	793	15
181- 270	909.7	982	14
271- 360	605.7	599	8
361- 540	1983.8	1790	21
541- 720	1079.1	1125	11
721-1080	698.3	659	7
Total	6210.3	6032	78