An Examination of Optimization in the Missouri Master Sample

Bernard Lazerwitz, University of Missouri

1. Optimization

The specific sample examined was selected from the Missouri master sample design at an overall sampling fraction of 1 in 1250. Within that portion of the sample selected from segments and chunks, there was an average of 5.7 sample hu's per chunk, 4.1 sample hu's per segment, 8.3 sample hu's per secondary selection; and 20.6 sample hu's per county. Within that portion of the sample selected from city directories (and block supplements),[1] there was 2.7 sample hu's per city directory cluster. Within sample hu's, one adult respondent was selected by means of an adult selection table technique.[2]

Kish (2) gives the following two equations to use in determining optimum occupied sample hu size per primary sampling unit.

1. $\text{deff} = [1 + \text{roh}\,(b-1)]$

where:

a) deff = cluster sample design efficiency. For the entire sample, deff. is 1.56. For that portion of the sample selected through chunks and segments, deff. is 1.96; for that portion of the sample selected from city directories (and block supplements), deff is 1.44.

b) roh is the intraclass correlation coefficient.

c) b represents the average number of occupied sample housing units per primary sampling unit.

2. optimum $b = \left[\dfrac{C_a\,(1-\text{roh})}{(c)\quad(\text{roh})}\right]^{\frac{1}{2}}$

where:

C_a = the average cost per primary sampling unit of training, planning, travel time, listing, mileage, and miscellaneous expenses.

C = the average cost per occupied sample housing unit of actual interviewing, other field time, and editing.

Applying these equations, in turn, to the entire sample; sample hu's from the non-metropolitan areas (primarily from chunks and segments); and sample hu's from the metropolitan areas (St. Louis and Kansas City) --almost exclusively from city directories and block supplements--gives the information of Table 1.

The design effect is somewhat larger in the non-metro areas than in the metro areas. This is reasonable since the metro area sample hu's were primarily selected in small city directory clusters while the non-metro area sample hu's were in the larger clusters of a county-town-chunk design. For the same design reasons, the non-metro area primary sampling units have larger b's than the metro areas. The roh factor is almost twice as large in the metro areas as in the non-metro areas. The C_a costs per primary sampling unit are considerably larger in the non-metro areas which make up 47% of the sample. Again, this is to be expected because of the greater travel distances in these small towns and rural areas. The average costs per occupied sample hu of interviewing, other field time, and editing does not vary too much between the two parts of the sample. Note that the optimum b figures are consistently larger than the actual b's for both parts of the sample. The metro area sample design is a one stage selection of city directory clusters (apart from the small block supplement). Hence the optimum b for metro areas not only refers to the desired clustering per primary sampling unit, but also gives the optimum level for the actual final stage selection clusters.

The data of Table 1 indicate that it should be possible to introduce additional field work savings by increasing sample cluster sizes. This can readily be done by selecting larger clusters of city directory sample lines from the master sample city directory clusters. It can be done in the chunk-segment sample portion of the master sample by selecting clusters of segments for any particular survey instead of individual setments. For the next statewide survey, we shall double previous city directory and segment clustering. This would raise directory selections from clusters of five lines to clusters of ten lines. We shall double the within-chunk rate and then select segments in clusters of two.

1. OPTIMIZATION FACTORS FOR PROJECT 030, Missouri Master Sample, 1971

Sample Category	deff	b	roh	C_a	C	optimum b
Entire Sample	1.56	6.4	0.100	$32.47	$2.92	10.0
Non-Metro Areas	1.96	14.0	.074	$70.54	$3.03	17.1
Metro Areas	1.44	4.2	1.1375	$21.05	$3.53	6.1

II. Sampling Errors and Statistical Inference on Project 030

2. GENERALIZED SAMPLING ERROR OF PERCENTAGES[a] - PROJECT 030, 1971.

(in percentages)

Reported Percentages	Number of Interviews								
	100	200	300	400	500	600	700	800	900
50	10.0-12.5	7.1-8.9	5.8-7.2	5.0-6.2	4.5-5.6	4.1-5.1	3.8-4.7	3.5-4.4	3.3-4.1
30 or 70	9.2-11.5	6.5-8.1	5.3-6.6	4.6-5.7	4.1-5.1	3.7-4.6	3.5-4.4	3.2-4.0	3.1-3.9
20 or 80	8.0-10.0	5.7-7.1	4.6-5.7	4.0-5.0	3.6-4.5	3.3-4.1	3.0-3.7	2.8-3.5	2.7-3.4
10 or 90	6.0-7.5	4.2-5.2	3.5-4.4	3.0-3.7	2.7-3.4	2.4-3.0	2.3-2.9	2.1-2.6	2.0-2.5

[a]The figures in this table represent two standard errors. Hence, for most items the chances are 95 in 100 that the value being estimated lies within a range equal to the reported percentages, plus or minus the sampling error.

In order to enable survey users to employ correct statistical inference procedures with these multi-stage sample survey data, we have developed generalized sampling error tables for individual percentages and for the difference between two percentages for varying numbers of interviewers. Here, I shall present just Table 2 for individual percentages. In Table 2 the low level estimates found in the cells give the 95 per cent confidence limits based upon the usual simple random sample formula. The high level estimates take into consideration the additional amount of variance derived from the use of a clustered sample. The procedures and statistical formulas used to obtain these sampling errors can be found in Kish (2) or Lazerwitz (3). The necessary computer program has been obtained from the Sampling Section of the Survey Research Center of the University of Michigan.

To illustrate the use of the table, let us find the sampling error for that 29% of the women of the survey who feel that "professors who advocate controversial ideas have no place in a state supported university." Since the total number of female interviews is 502, we enter the column of Table 2 headed "500" and the row headed "30 or 70". This tells us that chances are 95 out of 100 that this 29 per cent is subject to a sampling error of plus or minus 5.1 per cent (using the high level estimate).

Frequently, the difference between two percentages of the data of the statewide survey exceeds their proper high level estimate of sampling error. Hence two such percentages can be considered significantly different at a 95 per cent confidence level. Occasionally, some of the survey data are based upon percentages whose differences do not exceed their low level estimates. In all such cases, the percentages cannot be considered significantly different. When the difference between two percentages falls between their low and high level estimates of sampling error, the question of significance is considered unresolved. In such situations, it would be best to compute the specific sampling error of the involved difference rather than try to work with generalized tables.

III. Yield and Coverage Expectations

How well did this new sample design turn out with regard to actual sample hu coverage? On the whole, there is a good match between an expected yield of 1328 sample hu's and an actual yield of 1357 sample hu's. Here the excess of 29 sample housing units are primarily a result of the block supplement sample yield in St. Louis City. The very nature of the block supplement sample exposes one to the risk of encountering large clusters of new construction or of unlisted housing units in older structures missed by city directories. It would take extensive field work to avoid such situations which can be better handled by allowing more sample size variation and the technique of a "surprise stratum" (which was utilized for the St. Louis supplement sample).

Footnotes

[1]The block supplement yield on this survey was just 65 hu's, many of which were vacant.

[2]See Kish (1) for these selection tables.

References

(1) Kish, Leslie, "A Procedure for Objective Respondent Selection Within the Household," Journal of the American Statistical Association, 44 (September 1949), 380-87.

(2) _____, Survey Sampling, New York: John Wiley, 1965, 206-17, 268-70, 282-99.

(3) Lazerwitz, Bernard, "Sampling Theory and Procedures," in Methodology in Social Research, (edited by H. Blalock), New York: McGraw-Hill, 1968, 298-313.