Sampling Aspects of a Consumer Survey of Milk Products

By Eugene E. Hixson

An accumulation of knowledge of sampling errors is valuable for purposes of improving the design of samples and the interpretation of results. This article contributes to such knowledge a comparative analysis of alternative sampling methods, based on data from a recent marketing survey.

DATA from a consumer survey of milk products[1] were analyzed to obtain estimates of random sampling errors for selected items, the effects of restrictions on the geographical dispersion of the sample, and the effects of clustering (that is, the selection of groups of households rather than individual households). A more thorough understanding of these effects should enable one to do a better job of designing samples and to see whether the well known formula $\sqrt{\frac{pq}{n}}$ can be used to provide rough approximations of the standard error of p.

This survey was conducted in October and November of 1952 in Memphis, Tenn. Its purpose was to discover and analyze factors that were influencing the decline in sales of fresh fluid milk and the upward trend in sales of nonfat dry milk solids in consumer-sized packages in the Memphis market. It was also designed to disclose consumer purchase patterns and habits in use of milk products in the homes.

The statistical population was defined as all private households residing within the city limits of Memphis, Tenn., at the time of the survey. The 1950 Census block statistics for Memphis provided information on the number of households in each block, which was used as a basis for sampling.

Each block was assigned one sampling unit if it contained from 5 to 50 occupied dwelling units. Blocks with fewer than 5 dwellings were combined with adjacent blocks, and blocks with more than 50 were assigned one sampling unit for every 50 occupied dwelling units. Then a sample of blocks was selected by choosing the block in which every nth sampling unit fell. This was done by proceeding through the bulletin on block statistics in the order that the blocks were listed.

This gave a sample of blocks with probabilities of selection proportional to the number of sampling units and with "forced" geographical dispersion[2] as the blocks were listed in a geographical order. If a selected sample block contained more than one sampling unit it was divided by means of Sanborn maps into the appropriate number of sampling units from which one was selected at random. The sampling rate was .0179, giving a total of 69 sampling units (mostly whole blocks) whose boundaries were rigidly defined.

Interviewers were instructed to make contact with all private households in selected sampling units. One set of questions was asked of all users of nonfat dry milk solids. Another set was asked of a subsample of nonusers of nonfat dry milk solids. Thus one sample of 306 users of nonfat dry milk solids and another of 308 nonusers were obtained. At least two callbacks were required at a sample household in which no one was found at home, one callback to be made after 6 p. m.

Table 1 shows results of an analysis of selected items. For each item, estimates of sampling errors were computed for three different sample designs in order to compare their efficiencies. The first method corresponds to the design actually used in the survey. The second assumes an unrestricted random sample of sampling units, that is, no restriction on the geographical dispersion of the sample. As cluster sizes are variable, estimated percentages from the two methods are ratio estimates. The third method assumes an unrestricted random sample of individual households,

[1] Dwoskin, Philip B. MILK PRODUCTS: CONSUMER PURCHASE PATTERNS AND USE, MEMPHIS, TENN. U. S. Dept. Agr. Marketing Research Report No. 39, 68 pp. illus. May 1953.

[2] "Dispersion" is used instead of "stratification" because the sample is not a stratified sample in the usual sense.

which is impossible unless a list is available. The estimated standard errors for these three methods are designated s_1, s_2, and s_3 respectively. The methods of estimating the sampling error for each of these methods is shown in the footnotes to table 1.

TABLE 1.—*Analysis of the sample data*

Item	Number of households[1]	Percent of households	Estimated sampling errors: Cluster sampling: With geographical dispersion[2] s_1	Estimated sampling errors: Cluster sampling: Without geographical dispersion[3] s_2	Estimated sampling errors: Unrestricted random sampling[4] s_3	$\frac{s_1}{s_3}$
		Percent	*Percent*	*Percent*	*Percent*	
Percent of:						
All households using NFDMS[5]	1, 649	18. 6	1. 1	1. 1	1. 0	1. 10
All NFDMS households using buttermilk	306	68. 6	3. 8	3. 7	2. 7	1. 41
High income[6] NFDMS households using buttermilk	105	72. 4	4. 2	4. 7	4. 4	. 95
Negro NFDMS households using buttermilk	79	86. 1	4. 7	4. 5	3. 9	1. 21
All non-NFDMS households using buttermilk	308	64. 9	2. 7	3. 0	2. 7	1. 00
High income[6] non-NFDMS households using buttermilk	98	63. 3	5. 2	5. 2	4. 9	1. 06
Negro non-NFDMS households using buttermilk	75	86. 7	3. 1	3. 5	3. 9	. 79
All NFDMS households using evaporated milk	306	79. 7	3. 6	3. 5	2. 3	1. 57
High income[6] NFDMS households using evaporated milk	105	75. 2	6. 4	6. 1	4. 2	1. 52
Negro NFDMS households using evaporated milk	79	91. 1	5. 9	5. 8	3. 2	1. 84
All NFDMS households using NFDMS less than 6 months	306	21. 9	3. 1	2. 9	2. 4	1. 29
All NFDMS households who drink buttermilk made from NFDMS	306	33. 7	2. 9	2. 9	2. 7	1. 07
Cream using, non-NFDMS households using cream bought only in stores	110	53. 6	6. 1	5. 7	4. 8	1. 27
Homemakers having heard of or used NFDMS more than 6 months, who think: NFDMS has less food value than fresh milk	292	23. 9	2. 8	2. 8	2. 5	1. 12
Fresh milk using, NFDMS households who only buy home delivered milk	287	24. 0	3. 0	3. 2	2. 5	1. 20
Fresh milk using, middle income[7] NFDMS households who only buy home delivered milk	93	33. 3	4. 4	4. 5	4. 9	. 90
NFDMS households having fresh milk home delivered because it is: convenient, handy, or less trouble	138	49. 3	4. 5	4. 4	4. 3	1. 05
NFDMS households who do not have home delivery now but have had in the past, because: Use less, family needs change	73	21. 9	5. 4	5. 1	4. 8	1. 13
Fresh milk using, NFDMS households who prefer cartons to bottles	287	20. 9	2. 6	2. 4	2. 4	1. 08
Fresh milk using, NFDMS households who prefer bottles to cartons, because: Don't like taste, flavor or smell of cartons	193	21. 8	2. 8	3. 1	3. 0	. 93

[1] This column contains the number of households upon which the percents in the next column are based.

[2] In order to estimate sampling errors for this design, a procedure analogous to the "collapsed strata" technique was used to account for the geographical dispersion. (That is, the 69 sampling units were defined as 23 sets, each containing 3 sampling units, where the assignment of sampling units to a set coincided with the order of sample selection.) The estimated standard error of a percent for this design is given approximately by the formula:

$$s_1=\sqrt{\frac{\sum_{i=1}^{c}\sum_{j=1}^{m_i} X^2_{ij}-\sum_{i=1}^{c}\frac{X^2_i}{m_i}}{\bar{n}^2 m(m-c)}}$$

where k_{ij}=total number of homemakers responding yes to item k in the jth sampling unit and the ith set.

n_{ij}=total number of homemakers who were asked item k in the jth sampling unit and the ith set.

m_i=number of sampling units in the ith set.

$m=\sum_{i=1}^{c} m_i$, total number of sampling units in the sample.

$\bar{n}=\sum_{i=1}^{c}\sum_{j=1}^{m_i} n_{ij}\Big/ m$, average number of homemakers per sampling unit who were asked item k.

c=number of sample sets.

Footnotes continued on p. 52.

A comparison of s_1 and s_2, item by item, reveals that no appreciable gain was accomplished by geographical dispersion. In order to compare the given design with unrestricted random sampling, $\frac{s_1}{s_3}$ was computed. This comparison measures the combined effects of clustering and geographical dispersion. It also shows how well $\sqrt{\frac{pq}{n}}$ approximates the standard error of the design used. The data show that $\frac{s_1}{s_3}$ varies considerably for the different items. Since the effect of geographical dispersion was generally small, the variation in $\frac{s_1}{s_3}$ from item to item must be due primarily to differences in the intrasampling unit correlation or clustering effect. The greater the value of $\frac{s_1}{s_3}$, the greater the tendency for households within a sampling unit to be alike, and the greater the loss in statistical efficiency from use of large sampling units rather than small sampling units.

Clustering effects appear to be greatest for those users of nonfat dry milk solids who use evaporated milk. But a study of the data for individual blocks revealed that most of the effect of clustering is attributable to one or two blocks. The ratio of $\frac{s_1}{s_3}$ exceeded 1.25 for six items and was 1.00 or less for five items. Unpublished results from some investigations have indicated a smaller range in the values of $\frac{s_1}{s_3}$ from item to item and that $k\sqrt{\frac{pq}{n}}$ could be used as an expedient for getting a rough approximation of the sampling standard error of a percentage, where k is a constant equal to about 1.25. For this particular study, use of the "expedient" would be questionable. However, the s_1's are subject to rather large sampling errors so the "expedient" might have some value, particularly if the alternative is to have no information on sampling error.

In surveys of this type, where many estimates are to be made, one tries to design a sample that is near optimum for the more important items or for the majority of items. Thus, if one has some knowledge of the average cluster effect for the items of most interest, he can estimate the required sample size for an unrestricted random sample by using $\frac{pq}{n}$, and then multiply this sample size by the factor $\left(\frac{s_1}{s_3}\right)^2$ to obtain a rough estimate of the sample size necessary for a cluster sample.

Footnotes continued from p. 51.

$$X_{ij}=k_{ij}-p\ (n_{ij})$$

$$p=\sum_{i=1}^{c}\sum_{j=1}^{m_i} k_{ij} \Big/ \sum_{i=1}^{c}\sum_{j=1}^{m_i} n_{ij}$$

$$X_i=\sum_{j=1}^{m_i} X_{ij}, \text{ total of } X_{ij} \text{ for the ith set.}$$

[3] The estimated standard error of a percent for this design is given approximately by the formula:

$$s_2=\sqrt{\sum_{i=1}^{c}\sum_{j=1}^{m_i} X^2{}_{ij}/\bar{n}^2 m(m-1)}$$

[4] The estimated standard error of a percent for this design is given approximately by the formula:

$$s_3=\sqrt{pq/n}$$

where: $q=1-p$

$$n=\sum_{i=1}^{c}\sum_{j=1}^{m_i} n_{ij}$$

[5] NFDMS is used as abbreviation for nonfat dry milk solids.

[6] A weekly gross family income of $100 and over.

[7] A weekly gross family income of $50–$99.