NOTES

This section is devoted to brief research and expository articles, notes on methodology and other short items.

THE ALLOCATION OF SAMPLINGS AMONG SEVERAL STRATA

By J. Stevens Stock and Lester R. Frankel

1. **Introduction.** The problem of selecting a random sample so as to obtain optimum precision in making estimates has been the subject of inquiries by Bowley,[1] Neyman,[2] Sukhatme[3] and others. In estimating an average value of a variate in a population it is often profitable to stratify the universe into several homogeneous parts and sample at random within each of these parts. In order to obtain maximum efficiency for a given size of sample it appears that the number of samplings from each stratum should be proportional to the standard deviation of the characteristic under consideration and to the total number of units within the stratum. By distributing the sample in such a manner optimum precision will be obtained in estimating a general average.

However, it often happens that it is not the purpose of an investigation to study the aggregate of the universe. Evaluations and interrelations of characteristics in different groups or strata within the universe may be of importance. Thus, in cost-of-living surveys in a number of urban centers the object is to compare costs among the cities of different backgrounds. In such cases it is desirable for each city to have equal reliability so that each one may be treated as a unit. There are many other situations in the social sciences where analyses of this type are of importance.

2. **The Problem.** In general, the sampling problem is: Given several well defined areas of study and a fixed number of observations with which to make the survey, how best to distribute the observations such that each area will be represented with equal precision.

There are n observations to be distributed among m areas or strata. In the

[1] A. L. Bowley, "Measurement of the precision attained in sampling," *Bulletin de l'Institute International de Statistique* 1926 Rome, Tome XXII, 1-ere Livraison, 3-eme partie, pp. 1-62 (supplement).

[2] J. Neyman, "On the two different aspects of the representative method," *Journal of the Royal Statistical Society*, 1934, pp. 558-625.

[3] P. V. Sukhatme, "Contribution to the theory of the representative method," Supplement to the *Journal of the Royal Statistical Society*, Vol. II, 1935, pp. 253-268.

i-th stratum, if N_i is the total number of units, S_i^2 the variance of the characteristic to be measured, and n_i the size of the sample, the sampling error of the arithmetic mean is

$$\sigma_i = \sqrt{\frac{S_i^2}{n_i}\frac{(N_i - n_i)}{(N_i - 1)}}. \tag{1}$$

The problem then is, given N_i, numbers proportional to S_i^2 and n, to determine n_i such that

$$\sigma_1 = \sigma_2 = \cdots = \sigma_m.$$

3. **First Solution.** If we assume that the variances S_i^2 are all equal and that for $N_i - 1$ we may substitute N_j, we have

$$\frac{N_1 - n_1}{n_1 N_1} = \frac{N_2 - n_2}{n_2 N_2} = \cdots = \frac{N_m - n_m}{n_m N_m}. \tag{2}$$

From the total amount of money available and the cost per sampling unit we can determine the total number of observations to be apportioned among the m populations

$$n = \sum_1^m n_i. \tag{3}$$

We are able then to write m equations in m unknowns:
From (2) we may write $m - 1$ equations

$$\frac{1}{n_1} - \frac{1}{N_1} = \frac{1}{n_j} - \frac{1}{N_j} \qquad (j = 2, 3, \cdots m) \tag{4}$$

and from (3) we may write one equation.

$$n_1 + n_2 + \cdots + n_m = n. \tag{5}$$

But equations (4) are not easily soluble in their present form; they can be made linear by writing the approximation

$$\frac{1}{n_i} \equiv \frac{1}{L_i(1 + \alpha_i)} \doteq \frac{1 - \alpha_i}{i_i}.$$

Where L_i is some reasonable approximation of n_i chosen such that

$$\sum_1^m L_i =' \sum_1^m n_i$$

and α_i is some small correction for L_i to be determined. We have then approximately,

$$\frac{1 - \alpha_1}{L_1} - \frac{1}{N_1} = \frac{1 - \alpha_j}{L_j} - \frac{1}{N_j} \qquad (j = 2, 3, \cdots m) \tag{6}$$

and from equation (5) we get

$$\alpha_1 L_1 + \alpha_2 L_2 + \cdots + \alpha_m L_m = 0. \tag{7}$$

If we write

$$\phi_i \equiv L_1 L_i\left(\frac{1}{N_1} - \frac{1}{N_i}\right) + L_1 - L_i$$

we may write (6) and (7) in the following form:

$$(8)\qquad \begin{array}{lllll} -L_2\alpha_1 + L_1\alpha_2 & & & & = \phi_1 \\ -L_3\alpha_1 & + L_1\alpha_3 & & & = \phi_3 \\ \cdot\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot \\ -L_m\alpha_1 & & & + L_m\alpha_1 & = \phi_m \\ L_1\alpha_1 + L_2\alpha_2 & + L_3\alpha_3 & + \cdots & + L_m\alpha_m & = 0 \end{array}$$

The matrix of the coefficients is

$$(9)\qquad \begin{Vmatrix} -L_2 & L_1 & 0 \cdots & 0 \\ -L_3 & 0 & L_1 \cdots & 0 \\ \cdot & \cdot & \cdot \cdots & \cdot \\ -L_m & \cdot & \cdot \cdots & L_1 \\ L_1 & L_2 & \cdot \cdots & L_m \end{Vmatrix}$$

From this matrix we find that

$$(10)\qquad \alpha_1 = \frac{-\sum_2^m \phi_i L_i}{\sum_2^m L_i^2}$$

and from the general form of equation (8) we have

$$(11)\qquad \alpha_i = \frac{\phi_i + L_i\alpha_1}{L_1}.$$

These two equations (10) and (11) give us all the α_i. It is then only necessary to compute the second approximations of n_i by

$$(12)\qquad L_i' = L_i(1 + \alpha_i) \doteq n_i.$$

Closer approximations, though perhaps unnecessary, can be made by repeating the computation with the next approximations. The final approximations may be checked by substituting them in equations (4).

4. **Second Solution.** Sometimes the numbers S_i^2 are known or at least proportionate numbers can be estimated with a fair degree of accuracy for each area. We shall call these proportionate number ξ_i^2. We now have the conditions

$$(13)\qquad \xi_1^2\,\frac{N_1 - n_1}{n_1 N_1} = \xi_2^2\,\frac{N_2 - n_2}{n_2 N_2} = \cdots = \xi_m^2\,\frac{N_m - n_m}{n_m N_m}$$

and as before

$$\sum_1^m n_i = n\,.$$

We may write m equations in m unknowns, α_i, using the approximations L_i as before:[4]

$$(14)\qquad \begin{array}{l} -\,S_1^2L_2\alpha_1 + S_2^2L_1\alpha_2 \qquad\qquad\qquad = \theta_2 \\ -\,S_1^2L_2\alpha_1 \cdot \qquad + S_3^2L_1\alpha_3 \qquad\quad = \theta_3 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ -\,S_1^2L_m\alpha_1 \qquad\qquad + S_m^2L_1\alpha_m = \theta_m \\ \quad L_1\alpha_1 + L_2\alpha_2 + \cdots + L_m\alpha_m \quad = 0 \end{array}$$

Where

$$(15)\qquad \theta_i = L_1S_i^2 - L_iS_i^2 + L_iL_1\left(\frac{S_1^2}{N_1} - \frac{S_i^2}{N_i}\right).$$

Solving these m linear equations for α_i we get

$$\alpha_1 = \frac{-\sum_2^m \theta_i\,L_i/S_i^2}{S_1^2\sum_2^m L_i^2/S_i^2}$$

and from the general form of equations (14) we have

$$\alpha_i = \frac{\theta_i + S_1^2L_i\alpha_1}{S_i^2L_1}\,.$$

These α_i may be applied as before to the approximations L_i for new approximations L_i' of the numbers n_i.

5. **Remarks.** (i) In either case the applications of the corrections to the approximation L_i may be applied in two different ways:

$$(16)\qquad L_i' = L_i(1 + \alpha_i)$$

$$(17)\qquad L_i' = \frac{L_i}{1 - \alpha_i}.$$

When the corrections are applied according to (16) the sum of the new approximations adds up correctly to the total n, and no further adjustment need be made in the L_i' either for repeating the operation again for nearer approximations or for final use. If, however, the corrections are relatively large, say

[4] The numbers S_i^2 and ξ_i^2 may be used interchangeably since they are by hypothesis proportional.

greater than .10, there seems to be better convergence with the second approximations if formula (17) is used and the resulting L'_i adjusted proportionately such that they add up to n. These numbers then can again be adjusted with new α_i for final approximations.

(ii) The numbers S_i^2 or ξ_i^2 are not always estimable. If they are not known at all or are known to be all nearly equal the first solution is perhaps the more useful. If these numbers are known, and known to be different, the second solution is necessary. However, some saving in computation by the second method may be effected if the approximations L_i are first adjusted by the first solution before being entered into the computation of the second solution.

(iii) Further accuracy, though perhaps unnecessary, may be attained in the second solution by substituting throughout $S_i'^2$ for S_i^2 where

$$S_i'^2 = \frac{N_i}{N_i - 1} S_i^2$$

This substitution eliminates any slight inaccuracies caused by substituting N_i for $N_i - 1$.

(iv) The initial approximations L_i may in almost every case be gotten from the following formula:

$$L_i = \frac{n}{m} - \left(\frac{n}{m}\right)^2 \left(\frac{1}{N_i} - \frac{1}{m} \sum_1^m \frac{1}{N_i}\right).$$

(v) In all that has been presented above it has been assumed that the sample has been drawn without replacements from a finite universe. Whether or not this assumption is tenable depends upon the particular object of the research.

6. Example. In the Survey of Youth in the Labor Market conducted by the Division of Research in the Works Progress Administration youth who completed the eighth grade in the school years 1928–1929, 1930–1931 and 1932–1933 were studied. In six cities, Duluth, Denver, Birmingham, Seattle, San Francisco, and St. Louis random samples from school records were selected. Funds permitted the use of 40,000 schedules.

From school records it was possible to determine the total number of eighth grade graduates in each city for the years in question. The problem arose then as to what would be the most efficient method of distributing these 40,000 schedules among the six cities in order to compare the problems of youth.

Assuming equal variances within cities, quotas were computed for each of the cities. From Table 1, summarizing the computations, it can be seen that the quotas fall somewhere between proportionate and equal frequencies. This last result would be expected if samplings had been made from infinite universes.

7. Note. In the social sciences interest centers in deriving relationships among the various strata where each stratum is considered as a single unit. In such cases equal precision is desired. However, if the object of research is

TABLE 1

City	8th grade graduates	Initial approximation	First correction term	First approximation	Second correction term	Quotas	Percent sampled
Duluth, Minn.	5,500	4,000	−.02968	3,881	−.00077	3,878	70.51
Birmingham, Ala.	9,000	5,500	+.06641	6,399−	+.00148	5,343	59.37
Denver, Colo.	12,500	6,000	−.02690	5,352	−.00164	6,409	51.27
Seattle, Wash.	15,000	6,500	+.07525	6,989	+.00257	7,007	46.71
San Francisco, Cal.	21,000	8,000	+.01425	8,114	−.00341	8,086	38.50
St. Louis, Mo.	31,000	10,000	−.07349	9,265	+.00129	9,277	29.93
Total	94,000	40,000		40,000		40,000	

simply to draw contrasts between any two strata we would seek to minimize the standard error of the difference,

$$\sigma_{\Delta_{jk}} = \sqrt{S_j'^2\left(\frac{1}{n_j} - \frac{1}{N_j}\right) + S_k'^2\left(\frac{1}{n_k} - \frac{1}{N_k}\right)}$$

subject to the condition,

$$\sum_1^m n_i = n\,.$$

This leads to the result

$$\frac{S_j'}{n_j} = \frac{S_k'}{n_k}\,.$$

Thus, the number of samplings from each stratum is, for all practical purposes, proportional to the standard deviations, irrespective of the size of the various strata.

WASHINGTON, D. C.

ON THE COEFFICIENTS OF THE EXPANSION OF $X^{(n)}$

BY J. A. JOSEPH

Let us construct the following triangular arrangement of numbers:

$$\begin{array}{c}
1 \\
1 \qquad 1 \\
1 \qquad 3 \qquad 2 \\
1 \qquad 6 \qquad 11 \qquad 6 \\
1 \qquad 10 \qquad 35 \qquad 50 \qquad 24 \\
\cdot \qquad \cdot \qquad \cdot \qquad \cdot \qquad \cdot \qquad \cdot \\
1 \quad f_1(n-1) \quad f_2(n-1) \quad \cdot \quad \cdot \quad \cdot \quad f_{n-2}(n-1) \quad f_{n-1}(n-1) \\
1 \quad f_1(n) \quad f_2(n) \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad f_{n-1}(n) \quad f_n(n)
\end{array}$$