25 March 1947

(42) Allocation in stratified sampling

By W. Edwards Deming

How the problem arises. Thus far we have learned how to compute the variance of a sun or mean under certain sampling conddions? and emphasis has been place on the importance of either attaining a prescribed variance at minimum cost, or, alternatively, of attaining the smallest possible variance for a given allowable budget. However, so far, no indication has been given concerning the most efficient distribution of a sampling a mongest the several classes into which a universe may have been divided, in order to achieve such desirable aims. It would be possible by cut and try methods to increase the size of sampling in one class and dearease it in another, each time recomputing the total ~~s~~ cost and expected variance, and thus enventually to arrive at a satisfactory allocation of the sample amongst the classes, but this is tedious and it would be much better to have an indication fefor ehand of just what the desired aims depend on, and how to achine them directly. The solution of this problem, the proper allocation amongest the several classes, is ~~the~~ aim.

Formulation of the problem. The symbol E will be used hers to denote the sun of some particulor characteristic in the universe.

For example, E might be the total inventory of nails or flour, the total annual sales of nails or flour or of all commodities combined; or it might be the total unemployed, or employed at so many hours par week, or the number of nales or females or both under 18 and in school.

The universe of elements to be sampled might consist of a list of all the dealers of some specific or general type (all hardware dealers, all establishments selling flour at retail) within a given region. Or it might consist of a lest of all the dwelling units, or of all the small areas into which the region could be divided for purposes of sampling. The dealers, dwelling units, or areas, are supposed to have been chassified or stratified possibly on the basis of location and likely also on some other significant characteristic such as inventory, sales, or number of employees or number of occupied dewlling units reported at the last census. In each class at the time the sample is to be taken there will be an (unknow) average inventory or sales per dealer or per area; or an average number of people employed, unemployed, or in school, either per household or per area. If the Greek letter $\mu 1, \mu. \cdots$ denote these actual averages today, supposedly theoretically obtainable if a complete count were taken, then

$$\xi = N_1\mu_1 + N_2\mu_2 + \cdots \text{ through all classes} \qquad (1)$$

wherein N_1 is the actual (supposed known) number of number in Clase I. Bat, of course, we can only provide estimates of the Greek symbols, and if x, m_1, m_2, $\cdots$ denote the sample estimates of ξ, μ_1, μ_2, $\cdots$, the equation of estimate to be used here is

$$X = N_1 m_1 + N_2 m_2 + \cdots \qquad (2)$$

$m_1, m_2, \cdots$ are here the average inventories, sales, employed, etc. obtained from the samples in the various classes. There estmates are subject to sampling erros; in fact if m_1 is obtained by averaging the inventories of n_1 sample dealers. drawn at random from class I, then

$$\mathrm{Var}\ m_1 \doteq \frac{N_1 - N}{N_1 - 1}\ \frac{\sigma_1^2}{n_1} \qquad [p. \qquad] \quad (3)$$

There will be a like contribution of variance from each class, wherefore

$$\mathrm{Var}\ x = N_1^2\ \mathrm{Var}\ m_1 + N_2^2\ \mathrm{Var}\ m_2 + \cdots$$

$$= N_1^2\,\frac{N_1 - n_1}{N_1 - 1}\,\frac{\sigma_1^2}{n_1} + N_2^2\,\frac{N_2 - n_2}{N_2 - 1}\,\frac{\sigma_2^2}{n_2} + \cdots \quad (4$$

It should be noted that the problem to be solved here is associated with a par-ticular kind of sampling and estimation. A random sample of n_1 dealers, house-holds, or areas is to be tahne from Class 1, n_2 from Class 2, etc. Moreover, and very importance, the estimate x in to be formed frist by straight averaging to get $m_1, m_2, \cdots,$ followed then by multiplication by $N_1, \mu_1, \mu_2, \cdots$ as Eq. 2 indicates. In symbols, if x_{ij} denote the inventary of Dealer j in Class i, then

$$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$ [The n_i dealeas or dwelling units are supposedly numbered $j = 1, 2, \cdots, n_i$ after the sample is drawn] ⑤

In practice, instead of forming these average one by one, it is usuall more erpedition to form the term of Eq. 2 directly by noting that

$N_i m_i = \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}$ [Note that $\sum_{j=1}^{n_i} x_{ij}$ is merely the sum of the inventories of the sample dealers in Class i (6)

Here it is the sampling interval N_i/n_i, or the ratio of universe to sample size in Class i, that is used as a miltiplier of the sample inventory $\sum_{j=1}^{n_i} x_{ij}$.

If some other procedure of estimation were introduced, for example, the ratio of the inventories of the sample-dealers of Class i at two different dates (the Census date and the sample date), then Eqs. 2 and 3 would be replaced by corresponding but quite different equations, and all that fallowed would be affected. That is to say, optimum allocation ~~alloca~~ under one procedure of sampling and estimating will ~~be~~ not be optimum under another procedure. Thus, although in this chapter we shall solve a very importan~~ce~~t problem in optimum allocation, ~~under one~~ the solution should be regarded as applicable only in the specific procedure described. It should be said, though, that most of the remarks made here will hold good in other procedures as well.

A more complicated problem, frist sampling areal units and then subsampling dwelling units from them, is treated in a book by Hansen, Hurwitz, and Deming entitled A chapter in Poulation Sampling, Bureau of the Census, 1947.

The cost of conducting the survey will be

$$C = n_1 c_2 + n_2 c_2 + \cdots$$

wherein c_i is the cost of a questionnaire in class i (including callection, follow-up, and tabulation, with a pro-rated share of overhead).

The problems to be salved can be stated as follows:

Problem I Find what sampling intervals N_I/n_I, N_2/n_2, etc. will produce minimum Var x for a fixed allowable budget C.

Problem II. Find what sampling intervals N_I/n_I, N_2/n_2, etc. will produce a prescribed Var X at minimum cost

Fortunatly, the two problems almost completely overlap, and we shall set off to salve the frist

Solution of the frist problem; cost presoribed. The solution, when finally arrived at, will determine a set of values of n_I, n_2, --- which constute the best sizes of sample in the several strata (i.e., best in the sense of giving minimum Var X for a given budget C). These will be termed the "equilibrium" values from the analogy with problems of equilibrium in mechanics

Let δ denote a variation from an equilibrium value. Then if Var x is at a proper minium, its variation will be zero to within higher power of δn_i: i.e.,

$$\delta \operatorname{Var} x = \frac{\partial \operatorname{Var} x}{\partial n_1} \partial n_1 + \frac{\partial \operatorname{Var} x}{\partial n_2} \partial n_2 + --- = 0 \quad (8)$$

~~It~~ will be observed that ~~the~~ ~~procedure~~.

It will be observed that the procedure is to perform the differentiations exactly as if $n_1 \, n_2 \, \cdots$ were continuous variables, when in fact they can proceed only by steps of unity. This procedure can ~~proceed only~~ be relied upon to give results for $n_1 \, n_2 , \cdots$ correct with in a unit; in fact the results (Eq. 13 ahead) are probably exact

Now differentiation of the first terms on the right-hand side of Eq. 4. gives

$$\frac{\partial}{\partial n_1} N_1^2 \frac{N_1 - n_1}{N_1 - 1} \frac{\sigma_1^2}{n_1} = \frac{\partial}{\partial n_1} \frac{N_1^3 \sigma_1^2}{(N_1 - 1) n_1}$$

$$= \frac{\partial}{\partial n_1} \frac{N_1^2 \sigma_1^2}{N_1 - 1} = - \frac{N_1^3 \sigma_1^2}{(N_1 - 1) n_1^2} - 0 \quad (9)$$

The second term on the right is 0 because n_1 is not contained in the term directly above it.

Every term on the right of Eq. 4 will contribute a derivative like the one just written; hence from Eq. 8 it follows that

$$\frac{N_2^3 \; \sigma_1^2}{(N_2-1)n_1^2}\,\delta n_1 + \frac{N_2^3 \; \sigma_2^2}{(N_2-1)n_2^2}\,\delta n_2 + \cdots = 0 \quad (10)$$

Now differentiate Eq. 7 to get

$$c_1 \delta n_1 + c_2 \delta n_2 + \cdots = 0 \text{-----} (11)$$

The C on the right expresses the fact that however the sample sizes are permitted to vary, the total budget C is to be kept fixed.

This equation thus imposes a condition on the variations $\delta n_1, \delta n_2, \cdots$; they can not all be independent; one of them is fixed by Eq. 11

Next, multiply Eq. 11 through by $-\lambda^2$ and add it to Eq. 10, collecting the coefficients of $\delta n_1, \delta n_2, \cdots$ The result is

$$\left[\frac{N_2^3 \, \sigma_2^2}{(N_2-1)n_1^2} - \lambda^2 c_2\right]\delta n_2 + \left[\frac{N_2^3 \, \sigma_2^2}{(N_2-1)n_2^2} - \lambda^2 c_2\right]$$

$$\delta n_2 + \cdots = 0 \text{-----} (12)$$

Now suppose I let $\delta n_2, \delta n_3 \cdots$ take on any values whatever, forcing Eq. 11 to hold by a proper choice of δn_1, at the same time choosing λ so as to make the coefficient of δn_1 vanish in Eq. 12; the frist term is then 0 no matter what δn_1 may be. But this requires that the sun of all the remaining terms be 0 for any values whatever of $\delta n_2, \delta n_3, ---,$ which can be so only if each coefficient is separately equal to 0. And thus every coefficient vanishes, giving

$$\lambda^2 c_1 = \frac{N_i^2}{n_i^2}\sigma_1^2 \quad \frac{N_i}{N_i - 1} \text{ for all } i \text{ --(13)}$$

$$= \frac{N_1^2}{n_i^2}\sigma_i^2 \text{ ------------ (14)}$$

[The last from arises by neglecting the =1 compared with N_1, an approximation that will be assumed in what follows and for which no apology need be made when N_I is large.]

Interpretation of the result. The equations just written are extremely important; they express the optimum allooation of the sample. For turned around, Eq. 14 can written in the form

$$\frac{n_1}{N_1} = \frac{\sigma_1}{\lambda\sqrt{c_i}} \text{ ---------- (15)}$$

This equation tells us that if Var x is to be a minimum, the sampling-size in any class should not only be in proportion to the size of that class, but also in should not only

but also in proportion to the standard deviation of the inventories in that class. Morever, if the unit cost C_I of collecting information within a particular class is relatively very great, the sample in that class should be cut down because of $\sqrt{C_i}$ in the demominator. However, in practice, one must not forget that differential sampling ratios may introduce complications into the field and office procedures. There is simplicity in strict proportionality, wherein n_i/N_I is a constant for all classes, and the statistician should therefore not hastily prescribe differential sampling ratios in accordance with Eq. 15 without some preliminary calculations on gains and costs.

It is usually possible to determine in advance what the costs and sampling errors will be under various plans, and to make a rational choice of sampling ratios By evaluating λ it is possible to assign an absolute numerical magnitude to n_I. To do this, go back to Eq. 14 and multiply each side by n_i, then add for all classes, using the summation sign Σ to indicato the operation of summing. The result is

$$\lambda^2 \Sigma n_I C_I = \frac{\Sigma N_i^2}{n_i} \sigma^2$$

But by Eqs. 7 and 15 this gives

$$\lambda^2 C = \lambda \Sigma N_I \sigma_I \sqrt{C_I}$$

whence

$$\lambda = \frac{\Sigma N_i \sigma_I \sqrt{C_i}}{C} \text{ - - - - - - - - - - - -(1b)}$$

Everything on the right is supposed known or approximately determenable beforehand, wherefore a value for λ can be calculate and used in Eq. 15 to settle in advance the sample size n_I to be down from the various classes. It is possible to make a direct calculation of the minimized value of the Var X. This is done by going back to Eq. 4 and reiwriting it in the form.

$$\text{Var } X = N_1\left(\frac{N_1}{n_1}-1\right)\sigma_1^2 + N_2\left(\frac{N_2}{n_2}-1\right)\sigma_2^2 + \cdots \qquad (17)$$

This is obtained from Eq. 4 by dropping the -1 in each denominator. As remarked earlier, this approximation is of negligible importance, and the work is greatly simplified.

Going on, the equilibrium value of N_1/n_1 is introduced from Eq. 15 with the result that the minimized

$$\text{Var } X = \left(\frac{\lambda_1 \nu C_1}{\sigma_1}-1\right)N_1\sigma_1^2 + \left(\frac{\lambda \nu C_2}{\sigma_2}-1\right)N_2\sigma_2^2 + \cdots$$

$$= \lambda \sum N_i\sigma_i \nu C_i - \sum N_i\sigma_i^2 \qquad (18)$$

Again, λ is to be determined from Eq. 16 then used here. And thus the Var x to be expected from the wisest expenditure of the budget C can be calculated pretty closely beforehand, the only limitation being the prior determinations of $N_1, N_2, \cdots\cdots, \sigma_1, \sigma_2, \cdots\cdots$

The on the right arises from the finite multipliers $(N_i - n_i)/(N_i-1)$ and can be neglected if the sample sizes n_i/N_i are all small.

A simplifying assumption—equal unit costs. The variation in unit costs may be considerable, as when one class of household in the sample is urban and the cost of traveling from one household to another is small while another class is rural in a sparsely populated area and the cost of traveling from one household to another is considerable. Again, one class might be handled by mail at low cost, while another class requires personal interviews at high cost. However, in many enquiries, such as mail reports from business firms, the unit costs of collection and tabulation are nearly equal in all classes. and as cost enters only through the factor vc_1 in the denominator of Eq. 15 anyhow, it is then convenient and satisfactory under such circumstances to put

$$C_1 = C_3 = \text{-----} = C \qquad (19)$$

and

$$\lambda v_C = K \qquad (20)$$

whereupon

$$\frac{n_1}{N_1} = \frac{\sigma_1}{K} \qquad (21)$$

so that

$$\frac{n_1}{N_2} ? \frac{n_2}{N_3} = \sigma_1 \, 8 \, \sigma_2 \qquad (22)$$

Thus, if $\sigma_1 = 2\sigma_3$, then the sampling interval N_1/n_1 in class 1 is only half that of N_2/n_2 in Class 2.

This interpretation is introduced here for the very practical peason that a systematic selection within classes is permissible and even advantageous in place of a random selection in many types of sampling problems.

If the unit coasts are so nearly equal that Eq.

19 can be assumed, then by solving Eq. 21 for n_1 and summing over all class it follows that

$$K = \frac{\Sigma N_i \sigma_i}{n} \qquad (23)$$

whence upon returning to Eq. 21, we see that it can be written

$$\frac{n_1}{N_1} = \frac{n \ \sigma_i}{\Sigma N_j \sigma_i} \qquad (24)$$

wherein n, the total sample, must satisfy the relations

$$n = n_1 + n_2 + ----- = \frac{C}{c} \qquad (25)$$

Under the assumption of equal unit costs (Eq. 19), Eq. 18 for the minimised Var x reduces to

$$\mathrm{Var}\, x = \left(\frac{K}{\sigma_1} - 1\right) N_1 \sigma_1^2 + \left(\frac{K}{\sigma_2} - 1\right) N_2 \sigma_2^2$$

$$+ \ ---------$$

$$= K \Sigma N_i \sigma_i - \Sigma N_i \sigma_i^2$$

$$= \frac{(\Sigma N_i \sigma_i)^2}{n} - \Sigma N_i \sigma_i^2 \qquad (26)$$

Along with Eqs. 15 or 21, this equation is one of the most important onece in the chapter. It is very handy for calculation, for is evaluates at ones the minimised Var x without the intermediate step of first finding n_1, n_2, ----- and substituting them into Eq. 4. As a matter of fact, if it is known that the sampling fraction n_i/N_i are all going to be small, the second term in the last equation may be neglected. at least for a ready evalnation of the expected precision, for which the from

$$C.V.x = \frac{1}{\sqrt{n}} \; \frac{\Sigma N_1\sigma_1}{\xi} \qquad (27)$$

may be preferred. This form has the drawback of requiring an advance estimate of the total inventory ξ, which is just the characteristic that it is desired to measure, but this is usually not aerious because in practice C.V.x is not aseded with extreme accuracy.

The right-hand fraction in the east equation can be written

$$\frac{\Sigma N_i\sigma_i}{\xi} = \frac{N_1\sigma_1 + N_2\sigma_2 + ----}{N_1\mu_1 + N_2\mu_2 + ----} \qquad (28)$$

which can be regarded as a generalized coefficient of variation of the atratified universe. Eq.27 then appears as

$$c.V.x = \frac{C.V.\ universe}{\sqrt{n}}$$

exactly as was encountered in an earlier chapter (p.) in consideration of sampling from an unstratified universe, the finite miltiplier being placed equal to unity.

If there is only one class in the universe, the right-hand side of Eq.28 reduces to $\sigma : \mu$ (dropping the subscript), and this is the ordinary definition of the coefficient of variation of a distribution.

Historical note the advantages of allocating the sample according to Eq.21 (or its equivalent, Eq.27) were first noted by Neyman. The reference is to p.580 of his article entitled, on the two different aspects of the representative method, journal of the Royal Statistical Society, Vol. XCV11, 1934; PP. 558-606.

Remark 1. In sample-design on is often confronted with the decisiton between two plans—

Plan A: Sample-sizes proportional to N_i
Plan B: Sample-sizes proportional to $N_i\sigma_i$
(Eq. 21 or 24).

One should not too hastily specify Plan B. It will always show an apparent gain (*vide infra*) but it is essential to bear in mind that the variable sampling retios of Plan B also require variable weighting factors, and that the increased costs and complications in the field-work and tabulation may wipe out the apparent savings. The expected savings are not difficult to compute, and do not require accurate advance estimates of the standard deviations $\sigma_1, \sigma_2, ----$ within classes. Preliminary calculations should always be carried out before making a decision. Simple examples are given further on.

It is often possible to classify the elements of the universe into strata in such manner that the standard deviations σ_i are almost sure to be approximately equal.

Plans A and B are then identical, and maximum efficiency and simplicity are attained. This is often accomplished well enough by grouping the elements on equal class-intervals as of a provious consus.

Remark 2. Every survey is multipurpose, and sometimes two or more characteristics are competitive— that is, the sample design that is adequate for attaining the desired reliability in one important characteristic is not adequate for attaining the required reliability in another important characteristic. Under such conditions the best procedure seems to be strata of approximately equal range (as of the last census), when ordinarily the standard deviations of the competing characteristics will be approximately equal. Proportionate sampling (Plan A) then attains maximum simplicity and approximately maximum efficiency for all the competing characteristics.*

A further simplifying assumption, often usful. In a wide variety of sampling problems it can be assumed that when the elements (dealers, areas) of the universe are grouped according to same measure of size (inventory, sales, number of employees, number of dwelling units) as of a past consus, the standard deviations in the various classes at a later state will be nearly proportional to their sizes. In symbols,

$$\frac{\sigma_1}{\mu_1} = h, \text{ a constant} \quad \left[\begin{array}{l}\text{First exploited by} \\ \text{Hansen and Hurwitz}\end{array}\right] \qquad (29)$$

This assumption is often addbittedly rough, yet extremely useful. Substitution into Eq. 26 gives

$$\mathrm{Var}\, x = h^2\left(\frac{\xi^2}{n} - \sum N_1 \mu_1^{\,2}\right) \qquad (30)$$

in which approximations to h and ξ give a ready value of Var x.

If the finite multipliers in Eq. 4 are all nearly unity, the second term in parenthesis will be small compared with the first and will serve its purpose even though approximated with still less relative accuracy than ξ. In fact, if the second term is neglected altogether there is left the very simple relation

$$C.V.\, x = \frac{h}{\sqrt{n}} \qquad (31)$$

wherein n, the total sample, comes from Eq. 25.

* Jessen, Blythe, Kempthorne, and Deming, "On a population sample for Greece," Journal of the American Statistical Association, vol 42, 1947.

By introducing Eq. 29 when applicable, the sampling fractions n_i/N_i may be found quickly by noting that

$$\frac{n_i}{N_i} = \frac{n\sigma_i}{\Sigma N_i \sigma_i} \qquad [\text{Eq. } 24]$$

$$= \frac{n}{\xi}\mu_i$$

h has disappeared and it is only necessary to assume a mean μ_i for each class in order to achieve the allocation demanded by Eq. 21. These means (μ_i), should preferably be advance estimates referring to today, rather than to the last census; then today's estimate of ξ is simply $\Sigma N_i \mu_1$ (see the numerical illustrations further on), and σ_i/μ_1 in Eq. 29 represents today's coefficient of variation estimated in advance for Class 1.

Actually, in Eqs. 30 and 31 it is only necessary that h be an average value of $\sigma_i 8 \mu_1$. In using Eq. 32 it is possible to make allowances for departures from the average (see Step iv in the next section).

The assumption contained in Eq. 29 is the constancy of the coefficient of variation of the distribution of inventories, sales, or employed, from one class to another. It is a reasonable one in saying that the variability amongst big dealers is bigger than the absolute variability amongest small once. The assumption of strict proportionality, however, can not be expected to hold very closely, but fortunately the usefulness of the assumption is not seriously impaired if some classes depart only moderately from strict proportionality, provided we have some knowledge of there departures and are able to make corrections (again see step iv in the next section). For example, it is usual to find considerably greater values of $\sigma_i : \mu_i$ in the smaller classes, a good example being the blocks occupied by none or only 1 or 2 families at the time of a census: a few year later there may be many new houses on these blocks where there were only vacant lote at the time of the census. It might be well to double the sample of this class of block, over the sample-size calculated from Eq. 32. If there is any danger of a huge housing project being cuilt meanwhile, any such areas should be thrown out for a separate sampling-procedure (cf. the book by Hansen, Hurwitz, and Deming mentioned in footnote ef).

<u>Steps in the use of the above results in sample-design</u>. The assumption will be made here that the unit costs are the same in all classes, as specified by Eq. 19. Then if the sample is to be allocated according to Eq. 21. the procedure is as follows:

<u>Step i.</u> Prescribe an allowable coast C and calculate the total allowable sample size $n = C/c$. c is assumed known

<u>Step ii.</u> Arrive at suitable advance

approximations for ξ, h, and μ_i if Eq. 29 is useable; otherwise arrive at suitable approximations for N_i and σ_i. This step requires considerable knowledge of the universe, as might have been acquired in provious surveys or perhaps in a pilot study, as is often a wiss plan.

<u>Step iii.</u> Calculate the minimized expected Var X or C.V. X from Eqs. 26, 27, 30, or 31 (whichever is applicable).

<u>Step iV</u> If the expected precision, just calculated, appears to be good enough, proceed to calculate the sampling fractions n_i/N_i by Eqs. 15, 21, or 32, the latter if applicable, in which case reise or lower sample-sizes by appropriate amounts in those classes wherein the value h assumed for $\sigma_i \div \mu_i$ is known to be too low or too high.

If, on the other hand, the expected Var x as computed in Step iii is not considered to be small enough, then it will be necessary to increase the allotted budget from c to some new value C', and recompute. If additional funds are not to be had, it will be necessary either to accept a lower standard of precision (greater Var x) or to abadon the survey as not being worth while for the allowable funds.

If the assumption of equal unit is not justifiable, these stens require some obvious modifications.

Solution of the second problem: Var x prescribed. Here, c is to be made a minimum, the n_i to be of such magnitude that Var x has a prescribed value (such as C.V. x = 2 percent, 5 percent, 25 percent, etc.)

The mathmatical solution is indentical with the first problem down to the point where λ is determined. It is now to be found from Eq. 18. which gives

$$\lambda = \frac{Var\, x + \Sigma N_i \sigma_i^2}{\Sigma N_i \sigma_i \sqrt{c_i}} \qquad (33)$$

Everything on the right is supposed known so λ can be calculated ‡ Then from Eq. 16 the cost is seen to be.

$$c = \frac{1}{\lambda} \Sigma N_i \sigma_i \sqrt{c_i} \qquad (34)$$

The required sample sizes n_i would be computed from Eq. 15 at the same time. The total sample size will of course be

~~n=ΣΣ~~ $n = \Sigma n_i$ [As in Eq. 25]

Simplifications may be introduced under the same conditions as those mentioned in earlier paragraphs. In the frist place, if the finite multipliers in Eq. 4 can all be replace by unity, the second term in the numerator of Eq 33 can be thrown away in the calculation of λ Second, if the unit cost are the same in all classes, so that Eq 19 can be used, then in ~~all~~ place of Eqs 33 and 34 it would be simpler to write.

$$K=\lambda\sqrt{c} = \frac{Var\ x + \Sum N_i \sigma_i^2}{\Sigma N_I \sigma_i} \qquad (35)$$

(The second term, arising from the finite multipliers, can sometimes be neglected, as mentioned)

$$\text{and} \quad c = \frac{C}{K} \Sigma N_1 \sigma_i \qquad (36)$$

The sample sizes (n_i) would be compated from Eq 21 at the same time. The total sample will be

$$n = \Sigma n_i = \frac{C}{c} = \frac{\Sigma N_i \sigma_i}{K} \qquad (37)$$

If in addition, the constancy of the ratio σ_i/μ, may be usefully assamed, as expressed by Eq. 29, then Eq. 35 gives

$$K = \lambda\sqrt{\ }_0 \frac{Var\ x + h^2 \Sigma N_i \mu_i^2}{h\,\Xi} \qquad (38)$$

(The second term, arising from the finite multipliers, can sometimes be neglected, as mentioned)

Whereupon the cost simplifies to

$$c = \frac{c}{K} \Sigma N_i \sigma_i \quad (Eq\ 36)$$

$$= \frac{ch\xi}{K} \qquad (39).$$

$$= \frac{ch^2 \xi^2}{Var\ x} = \frac{ch^2}{(C.V.X)^2}$$

(Neglecting the second term in Eq 38)

The simple sizes would then be found from Eq 32

When the variance is prescribed, and the unite costs c_i are all assumed to be equal, the steps to be applied in allocating the sample according to Eq 21 can be outlined as follows.

Step I Decide on the prescribed value of Var x, or perhaps, if easier, think in terms of C.V. x (2 percent 5 percent, 25 percent, etc)

Step II Arrive at suitable advance approximations for ξ, h, and μ, if Eq 29 is useable; otherwise arrive at sluitable approximations for N_i and σ_i This step requires considerable knowledge of the universe, as might have been acquired in previous surveys or perhaps in a

. . study, as is often a wise plan.

Step III. Calculate λ from Eq. 33 and then the minimum cost C from Eq. 34. or preferably from Eq. 36 or 39. if applicable.

Step IV. If the cost is apparently not going to be too heavy, proced to allocate the sample by Eqs. 15, 21, or 32, the latter if applicable, in which case raise or lower the sample-sizes by appropriate amounts in those classes wherein the value 17 assumed for $\sigma_1 : \mu_1$ is known to be too low or too high.

If, on the other hand, the expected minimum cost C as computed in Step III is too high, it will be necessary to accept a lower level of precisin than initially prescribed, and to recompute the cost. If this plain is ~~not~~ acceptable, the survey must be abandoned, as being too costly.

If the assumption of equal unit costs is not justifiable these steps require some obvious modifications.

Gains complaret for different allocation.

Two simple illustrations will help. Let the problem be the determination of the total inventories of 3000 dealers, which have been divided (for simplocity) into just two classes, as of some previous census. σ_1 and μ_1 represent advance estimates, as of today. The data are shown in Table 1. It will be assumed that the

Table 1

class	N_1	μ_1	$N_1\mu_1$	σ_1	Simple sizes, n_1	
					Plan A n_1 propoartional to N_1	Plan B n_1 proportional to $N_1\sigma_1$ (Eq. 32)
1	1000	100	1000 00	90	40	60
2	2000	50	100 000	45	80	60
Total	3000	XXX	Σ=200 000	XXX	120	120

unite ~~cons~~ costs are the same in both classes, wherefore the total costs of both plans are equal because the total sizes are equal (120). Eq. 4 will give the variance of the total estimated inventory, but for simplicity the finite multipliers will be neglected. Then

$$\text{Var } x = N_1^2 \frac{\sigma_1^2}{n_1} + N_2^2 \frac{\sigma_2^2}{n_2} \quad \text{(Approximation to Eq. 4)}$$

Under Plan A this gives

$$\text{Var } x = 1000^2 \times \frac{90^2}{40} + 2000^2 \times \frac{45^2}{80}$$

$$= \frac{1000^2 \times 45^2}{40}(4+2) = \frac{1000^2 \times 45^2}{120} \times 18$$

Under Plan B the same formula gives

$$\text{Var } x = 1000^2 \times \frac{90^2}{60} + 2000^2 \times \frac{45^2}{60}$$

$$= \frac{1000^2 \times 45^2}{60}(4+4) = \frac{1000^2 \times 45^2}{120} \times 16$$

The loss in using Plan A is thus (18-16)/16 or one-eighth. That is, Plan B will produce an eighth less variance at the same cost, or will cost an eighth less for the same variance. Under other conditions (different N_1 and σ_1) the gains arising from Plan B may be much greater, but here the gain is not much more than enogh to offset the additional costs and complications in the field and tabulations arising from the introduction of different sampling ratios in the two classes.

This illustration is introduced to show the danger of jumping too rapidly into one plan or the other without some preliminary computations. Even rough guesses at the standard deviations σ_1 and σ_2 may be sufficient to indicate which type allocation will best serve the purpose.

[The student should satisfy himself that Eq 26 with the second terms neglected gives the same result for Plan B. Thus, Indent Red.

$$\text{Var } x = \frac{(\Sigma N_i \sigma_i)^2}{n}$$

[Eq.26, second term neglected] (41)

$$= \frac{(1000 \times 90 + 2000 \times 45)^2}{120}$$

$$= \frac{1000^2 \times 45^2}{120} \times 16 \text{ as before}$$

the r Comparison with the result for Var X using no stratification at all is interesting. The total variance of the universe is

$$\sigma^2 = P\sigma_1^2 + Q\sigma_2^2 + PQ(\mu_2 - \mu_1)^2 \text{ [F.] } (42)$$

$$= \frac{1}{3} \times 90^2 + \frac{2}{3} \times 45^2 + \frac{1}{3}\ \frac{2}{3}(100-50)^2$$

$$= \frac{41,450}{9}$$

P and Q have denote the proportions in the two class; $P = N_1/N$ and $Q = N_2/N$ where $N = N_2 + N_2$.

An unstratified random sample of 120 will give

$$\text{Var } x = N^2 \frac{\sigma^2}{n} \text{ [The finite multiplier neglected}$$

$$= 3000^2 \times \frac{41,450}{120 \times 9} = \frac{1000^2 \times 45^2}{120} \times 20.5 \ (43)$$

It is now possible to compare the variance arising from all three sampling proced~~ory~~ures.

Table 2 illustrates a further important point, viz.,

Table 2

Procedure	Var x
Unstratified random sampling	$(1000^2 \times 45^2/120)$ 20.5 by Eq 43
Plan A: sample-sizes proportional to N_i	$(1000^2 \times 45^2/120)$ 18 by Eq.40
Plan B: sample-sizes proportional to $N_i \sigma_i$	" 16 by Eq.40 or 41.

That under certain conditions the gains arising from stratified sampling, even when the sample-sizes are proportional to N_i or $N_i \sigma_i$, will be disappointing.

The conditions under which stratification pays off are difficult to describe in general terms, but it is probably safe to say that no striking gains will be made unless the stratification produces widely different means or variance. In the latter case the advantage of using sample-sizes proportional to $N_i \sigma_i$ may be very much worth while. Another illustration (Table 3), being but a simple alteration of the foregoing one, may make this point clear. The only essential difference between this and the forgoing illustration is that the ratio $\sigma_1 : \sigma_2$ is now 4 instead of ~~to~~2, yet the results, shown in Table 4, are obviously considerably different from those in Table 2. In fact, Plan B in Table 4 is seen to require only two-thirds the sample-size of Plan A, which in turn requires only three-quarters the sample size in.

Table 3

Class	N_i	μ_i	$N_i\mu_i$	σ_i	Sample sizes, n_i Plan A n_i proportional to N_i	Plan B n_i proportional to $N_i\sigma_i$ (Eq.32)
1	1000	200	200 000	2000	80	160
2	2000	50	100 000	50	160	80
Total	3000	XXX	Σ=300 000	XXX	240	240

Table 4

Procedure	Var X
Unstratified random sample	$(1000^2 \times 50^2/80)$ 24 by Eq.42
Plan A: sample-sizes proportional to N_i	" " 18 by Eq.40
Plan B: sample-sizes proportional to $N_i\sigma_i$	" 12 by Eq.40 or 41.

unstratified random sampling. In this example, samples proportional to $N_i\sigma_i$ would most certainly show a substantial saving over and above the extra costs of differential sampling and weighting. (The fact that the total sample in this second example is 240 in ~~240 in~~ place of 120 in the first one is of no consequence whatever.) The calculations are left as an exercise.

As a further exercise the student might calculate Var x using still other sample allocations, such as $n_1 = 200$, $n_2 = 40$; also the reverse. The student should also calculate C.V. x for Plan B in both illustrations by using Eq. 31, then verifying the results by calculating $\sqrt{\text{Var}\, x/\mathcal{E}^2}$, taking Var x and $\mathcal{E}$ from the table.

Note in regard to advance estimates of the variance. The variance $\sigma_1^2, \sigma_2^2, \cdots$ within the several classes at the time the sample is taken are assumed to be known in advance, approximately at least, as has been explained. Oftentimes considerable ingenuity is required to evaluate these variances without actually carrying out a full-scale survey. One should make use of census information and other previous studies, consultations with experts in the subject-matter, and sometimes a pilot study for estimating some of the variance as well as for testing the questionnaire and instructions. There always seems to be a way out if an honest effort is made, particularly in view of certain favorable features of the problem.

For instance, it is importan~~ce~~t to note that successful allocation does not require accurate knowledge of the variance involved. This is indeed fortunate.

It is so because Var x does not increase rapidly from its minimum value at N_1, N_2, --- are moved only small distance $\delta n_1, \delta n_2$ --- away from their equilibrium values: this can be seen from Eq. 8 in which the increase in Var X (319頁へ續く)。

arising from the <u>first</u> powers of $\delta n_1, \delta n_2$, ------ has been placed equal to zero (the condition for a proper minimun), wherefore Var x is affected only through the squares and higher powers of $\delta n_1, \delta n_2$, ---- . Now if δn_1 is small, its square is still smaller, and the effect on Var x may be negligible. Hence it is not necessary to come out with exactly the equilibrium values of n_1, n_2, ---- ; approximations will still serve the purpose tolerably well, which means that only reasonably good values of σ_1^2, σ_2^2, ----- are required.

[This point is brought out in the first numerical illustration, wherein two different allocations of the sample did not produce greatly different values of Var x. In the second illustration the two allocations of the sample were sufficiently different to produce distinctly different values of Var x.]

Another noteworthy feature is that optimum allocation is obtained if only the advance estimates of σ_1, σ_2, ---- are in the <u>right</u> proportions. If all are estimated too high by (e.g.) 20 percent, as when h in Eq. 29 is set too high, the sampling fractions n_i/N_i are unaffected, as is easy to see form Eqs. 24 or 32. Damage is done, however, to the advance estimate of the Var x or the estimated cost. For instance, if the variances were all initially over-estimated by 20 percent, then for a prescribed cost C (Problem I) the advance calculation of the C.V. x will also be too high by about 10 percent, as will be discovered when the variances are evaluated from the returns (cf. the section, "A word on the final determination of the precision reached"). In other words, the error band will turn out to be 10 percent smaller than expected. The same thing will happen when a particular C.V. x has been prescribed (Problem II); the cost will then be about 20 percent greater than was necessary to reach the prescribed precision. The additional expediture goes for unneeded precision, purchased interestingly enough at the cheapest possible rate (optimum allocation).

Exact formulation of the gains arising from stratification. In the first place, this title implicitly bears with it the thought of allocating the sample in one way or another to the various classes. It is not enough to specify the method of stratification; one must also specify the manner of drawing the sample before he can calculate the variance of the result.

Here we shall compare thee plance of allocation, the same three in fact that were just illustrated with numerical examples. Now, however, we shall seek an exact formulation of the comparisons. The variances associated with the three different plans are those shown below.

Plan A: in which n_1 is proportional to N_i. Here

$$\text{Var } x = \Sigma N_i \, (Ƀ-1)\sigma_1^2 \qquad [\text{From Eq. 17}] \quad (44)$$

wherein for all classes

$$\frac{N_i}{n_i} = Ƀ \; [\text{the sampling interval for all classes}] \quad (45)$$

Plan B: in which n_1 is proportional to $N_i\sigma_i$. Here

$$\text{Var } x = \frac{(\Sigma N_i \sigma_i)^2}{n} - \Sigma N_i \, \sigma_1^2 \; [\text{Eq. 26}].$$

Plan C: in which there is no stratification at all. Here

$$\text{Var } x = N^2 \frac{N-n}{N-1} \, \frac{\sigma^2}{n} = N(Ƀ-1)\sigma^2 \qquad [N-1 \text{ replaced by } N] \quad (46)$$

wherein

$$\text{Ъ} = N/n$$

as in Eq. 45

Let these three variances be designated by A, B, C. Then

$$\begin{aligned} C - A &= (\text{Ъ}-1)\left[N\sigma^2 - \Sigma N_i\sigma_i^2\right] \\ &= (\text{Ъ}-1)\,\Sigma N_i(\sigma^2-\sigma_i^2) \\ &= N(\text{Ъ}-1)\,\Sigma P_i(\sigma^2-\sigma_i^2) \\ &= N(\text{Ъ}-1)(\sigma^2-\Sigma P_i\sigma_i^2) \\ &= N(\text{Ъ}-1)\,\Sigma P_i(\mu_i-\mu)^2 \qquad (47) \end{aligned}$$

Herein

$$P_i = \frac{N_i}{N}, \text{ the proportion of the universe in Class } i \qquad (48)$$

and the line follows from the preceding one from the fact that

$$\sigma^2 = \Sigma P_i\sigma_i^2 + \Sigma P_i(\mu_i-\mu)^2 \qquad (49)$$

The relative gain of Plan A over Plan C will be

$$\frac{C-A}{C} = \frac{\Sigma P_i(\mu_i-\mu)^2}{\sigma^2} \qquad (50)$$

and this relative gain is the same whether it is the total inventory, total sales, total unemployed this is being estimated, or the average inventory or sales per dealer, or the average unemployed per family or per area. The result shows that stratification with size of sample proportional to N_i (Plan A) removes the variance between classes.

obviously. this will be appreciably better than straight random unstratified sampling only if one or more of the large strata have widely different means, so that at least one term $N_i(\mu_i-\mu)^2$ will contribute safficiently to the difference A - C. Thus Plans A and C are compered.

Now compare Planes A and B.

$$A - B = \frac{N}{n}\sum N_i\sigma_i^2 - \frac{(\sum N_i\sigma_i)^2}{n}$$

$$= \frac{1}{n}\left[\sum N N_i\sigma_i^2 - (\sum N_i\sigma_i)^2\right]$$

$$= \frac{1}{n}\left[\sum N N_i\sigma_i^2 - \sum N_i^2\sigma_i^2 - \sum\sum N_i N_j\sigma_i\sigma_i\right] \quad i \neq j$$

$$= \frac{1}{n}\left[\sum N_i\sigma_i^2(N-N_i) - \sum\sum N_i N_j\sigma_i\sigma_j\right].$$

$$= \frac{1}{n}\left[\sum N_i^2\sigma_i^2 + \sum\sum N_i N_j\sigma_i^2 - \sum N_i^2\sigma_i^2 - \sum\sum N_i N_j\sigma_i\sigma_j\right]$$

$$= \frac{1}{n} \sum \sum N_i N_j (\sigma_i - \sigma_j)^2 \qquad i < j$$

$$= BN \sum \sum P_i P_j (\sigma_i - \sigma_j)^2 \qquad (51)$$

Again, it is probably best to compute $\frac{A-B}{B}$ for the relative grin of Plan B over A. The gain will be worth while only if two or more of the σ_i are greatly different, and only then provided neither class is very small. For although $(\sigma_i - \sigma_j)^2$ be large for one pair of classes, the factor $P_i P_j$ may reduce the term to some negligible amount if either Class i or Class j is relatively very small, and the extra cost of using sample sizes proportional to $N_i \sigma_i$ rather than the uniform density of Plan A might not be worth while.

In particular, if two classes have equal variance they contribute nothing at all to the advantage of Plan B over Plan A. To go further, if all classes have to same variance, Plans A and B are identical.

The student should return to the numerical illustration of the last section and test out Eqs. 50 and 51 for the differences between C − A and A − B.

A word on the final determination of the precision reached. At the risk of repetition, a word is offerd here on the importance of evaluating the precision actually reached, for at least some of the characteristics computed from the sample.* What has gone before in this chapter is advice on sample-design, which in the modern sense implies <u>advance</u> calculation of the expected variance for some critical characteristic to be obtained from the sample (total inventory, sales, or employment). This advance calculation can only be made on the basis of incompl knowledge — the best obtainable at the time, but incomplete nvertheless. For example, one never knows what snags the field-work will run into. but these sangs affect the variances σ_1^2, σ_2^2, -, ---. A subsample take from the returns will give estimates of these variances as they actually occurred in the sampling. Substitution into Eq. 4 then gives a close estimate of Var x.

It is usually not permissible to use Eq. 26 for evaluating Var x after the survey is completed because oftentimes the actual field procedure and sample-sizes differ considerably from those intended, owing to special and unforeseen circumstances encountered, as

---

Examples: i. Deming and Simmons, "On the design of a sample for dealer "inventories," Journal of the American Statistical Assoc., Vol. 41, 1946: PP. 16-33. ii, Hansen, Hurwitz, and Deming, <u>A Chapter in Population Sampling</u> (Bureau of the Census, 1947). iii. Jessen, Blythe, and Deming, "On a population sample for Grace", Journal of the Americal Statistical Assoc., Vol., 1947: PP. —

well as sometimes to misunderstandings or ignoring the instructions. It is therefore better to use Eq. 4 with the actual samples $n_1, n_2, ---$ and the new estimates of the variances σ_1^2, σ_2^2 ----

The proper size of subsample to take for estimating the variances from the returns has been discussed elsewhere (pp. __ - ___).

This last stage -- the final evaluation of Var $\bar{x}$ -- be it noted, is independent of the assumptions that went into the design, however crude, and hence impendent of whether the design wass really a good one.

This last stage is the guarantee of precision, and a good sample deserves it. The cost is trifling.

A further argument lies in the experience to be gained by computing some actual variances under the conditions met, in order to have them available for ~~future~~ sample designs.

Exercises

1. In sampling the dwelling units of a city for estimating the proportion colored, it is possible to classify the dwelling units into two groups, in one of which the proportion colored is .1 and in the other of which it is .9. Prove that sample sizes proportional to to N_i are identical with sample sizes proportional to $N_i \sigma_i$; in other words, Plans A and B are endentical. (Hint: $\sigma_1^2 = P_1 Q_1 = P_2 Q_2 = \sigma_2^2$)

2. If the two groups of dwelling units of the preceding exercise are of equal size, then statefied sampling with a uniform sampling ~~ratio~~ ratio will show a gain of 64 percent over no stratification at all. This is, only 36 percent as many dwelling units would be required on Plans A or B as on Plan C, to reach a prescribed precision

3. (a) If the proportions colored in the two classes are .25 and .75, then the gains of Plans A and B drop to 25 percent. (b) If the proportions are .6 and .4, the gains in efficiency are only 4 percent; that is, the gains accomplished by ~~st~~ stratification are now almost negligible

此処に発表するDr W.E.Demingの論文は昨年同博士及米国統計使節団の一人として来朝されたとき統計数理研究所で講演されたものでその節同博士講究録へ発表を承諾されたもので帰国後手を入れて送付されたものである。（編輯者 小川所員）