#108

Reprinted from Estadistica, Journal of the Inter-American Statistical Institute, Diciembre de 1959

SOME STRATIFIED SAMPLING PLANS IN REPLICATED DESIGNS

SOME STRATIFIED SAMPLING PLANS IN REPLICATED DESIGNS

W. Edwards Deming, U.S.A. *

RESUMEN

El propósito de este artículo es describir algunos de los posibles procedimientos de muestreo estratificado, y comparar las ganancias en la precisión que con cada uno de ellos se obtiene con los respectivos costos. Los procedimientos que el autor presenta siguen diseños de muestreo reiterado con aplicación especial de las muestras interpenetrantes de Mahalanobis.

La estratificación es una forma de usar la información estadística disponible para lograr una estimación mejor que la que sería posible obtener de otro modo. La información estadística que se usa en un diseño estratificado puede ya existir para todas las unidades de muestreo en el marco (como en el último censo), o puede exigir pruebas o entrevistas en una muestra preliminar. La estratificación tiene en realidad muchos significados y muchas formas de aplicación.

Una ventaja de la estratificación, además de la ganancia en la precisión, es la de disminuir el coeficiente del cuarto momento, con mejoramiento en la estimación de la varianza de $\bar{x}$

El autor describe varios planes de estratificación que distingue con las letras A, B, C, etc., y emplea el plan A como elemento principal de las comparaciones entre ellos.

Preliminary note

The purpose of stratification. Most frames for censuses, complete testing, or sampling, are already stratified into natural geographic zones. Material, as it comes to us in the frame, is never thoroughly mixed. This natural stratification is automatic and costs nothing. One example is census statistics, which are usually in some sort of geographic order. Another example is industrial product, each item of which emerges at the end of the production-line in the order in which it was made.

It sometimes pays to rearrange the sampling units into more homogeneous groups called strata. We sometimes rearrange all the sampling units in the frame before we draw the sample, and sometimes we rearrange only the sampling units in the sample, depending on costs. The aim of this paper is to describe some of the possible procedures of stratified sampling, and to compare the gains in precision with the costs. The procedures will follow replicated sampling designs as a special application of Mahalanobis's interpenetrating samples.

* Consultant in Statistical Surveys, 4924 Butterworth Place, Washington 16; Professor of Statistics, New York University. This article was prepared in December 1959.

Stratification is one way to use statistical information in our possession to acquire a better estimate than would be possible otherwise. The statistical information that we use in a stratified design may already exist for all the sampling units in the frame (as from the last census), or it may require tests or interviews of a preliminary sample. Stratification has in fact many meanings and many ways of application.

An advantage of stratification, aside from gain in precision, is a decrease in the fourth moment coefficient, with improvement in the estimate of Var $\bar{x}$.

Summary description of several plans.[1] I shall give at this point a brief summary of several stratified plans. I shall distinguish the various plans by the letters A, B, C, etc. Plan A is the basis for comparison, as it will be a replicated sample of the material just as it comes to us in the frame, without further stratification. We cannot be sure in the case of new material just how effective the natural stratification is, but Plan A, as we use it, will capture any possible gain from this source. We may think of Plan A as a proportionate stratified sample, where the strata are zones in the frame. Later sections show directions in more detail.

Plan A. Sample the frame as it is, with no rearrangement

THE PROPORTIONS P_i KNOWN

Classify the whole frame

Plan B. Classify all the sampling units of the frame. Then use proportionate allocation.

Plan C. Classify all the sampling units of the frame. Then use Neyman allocation.

Classify only the n sampling units of the sample

Plan D. Draw the sample as in Plan A, the size being determined exactly as for Plan B. Then classify the sampling units into the various strata. Use the entire sample; no thinning.

Plan E. Draw the sample as in Plan A, the size being determined as in Plan C. Then classify the sampling units into the various strata. Thin the strata to reach the Neyman ratios $n_i : n_j$. Choose the number n' so that with as little thinning as possible, the total sample will turn out to be about equal to the desired size n.

Classify only enough sampling units to fill the quotas n_i.

Plan F. Fix the sample sizes n_i by proportionate allocation, as in Plan B. Then draw sampling units one by one from the zones in the frame and classify them into the various strata until all the quotas n_i are full. Discard any sampling unit that belongs to a stratum whose quota is full. Form the estimates and the variances as in Plan B.

Plan G. Fix the sample-sizes n_i by Neyman allocation, as in Plan C. Proceed otherwise as in Plan F, but form the estimates and the variances as in Plan C.

---

[1] The theory and other illustrations of the various plans appear in Chapter 15 of my book *Sample Design in Business Research* (Wiley, 1960).

PROPORTIONS P_i NOT KNOWN

Classify only a preliminary sample of size N' to estimate the proportions P_i.

Plan H. Draw as in Plan A a preliminary sample of fixed size N'. Classify only these sampling units, not the whole frame. Then thin proportionately the samples from the various strata to reach a total specified size. Form the estimate X as directed later.

Plan I. Draw as in Plan A a preliminary sample of fixed size N', and classify only the sampling units in this preliminary sample, not the whole frame. Then thin the samples from the various strata by the use of ratios dictated by the Neyman allocation to reach a total specified size. Form the estimate X as directed later.

STRATIFICATION BEFORE SELECTION

Some notation and definitions in the frame. It will save time, before we go on, to introduce some definitions.

$$a = P_1 a_1 + P_2 a_2 + P_3 a_3 = \frac{A}{N}$$ the average population per sampling unit (1)

$$\bar{\sigma}_w = P_1 \sigma_1 + P_2 \sigma_2 + P_3 \sigma_3$$ the weighted average standard deviation within strata (2)

$$\sigma_w^2 = P_1 \sigma_1^2 + P_2 \sigma_2^2 + P_3 \sigma_3^2$$ the weighted average variance within strata (3)

$$\sigma_b^2 = P_1 (a_1 - a)^2 + P_2 (a_2 - a)^2 + P_3 (a_3 - a)^2$$
$$= P_1 a_1^2 + P_2 a_2^2 + P_3 a_3^2 - a^2$$ the variance between strata (4)

$$\sigma^2 = \sigma_b^2 + \sigma_w^2$$ the total variance (5)

$$\sigma_R^2 = Q_1 \sigma_1^2 + Q_2 \sigma_2^2 + Q_3 \sigma_3^2$$ the reverse internal variance $(P_i + Q_i = 1)$ (6)

$$\bar{\sigma}_R = Q_1 \sigma_1 + Q_2 \sigma_2 + Q_3 \sigma_3$$ the average reverse internal standard deviation (7)

It is helpful to see some of these definitions arrayed in Table 1.

More detailed description of Plan A. I now give fuller directions for the plans, step by step, starting with Plan A. The reader will in practice discover his own short-cuts, which will vary with local conditions and preferences. The notation will be obvious, I believe, by a glance at Table 2.

TABLE 1. NOTATION AND DEFINITIONS FOR THE FRAME
Extension to more strata will be obvious
M is the number of strata (here $M = 3$)

Stratum	Number of sampling units		Stratum's proportion of sampling units in the frame	Population		Between the populations of the sampling units within the stratum	
	In the frame	In the sample		Average per sampling unit in the stratum	Total in the stratum	Standard deviation	Variance*
1....................	N_1	n_1	$P_1 = N_1/N$	a_1	$A_1 = N_1 a_1$	σ_1	σ_1^2
2....................	N_2	n_2	$P_2 = N_2/N$	a_2	$A_2 = N_2 a_2$	σ_2	σ_2^2
3....................	N_3	n_3	$P_3 = N_3/N$	a_3	$A_3 = N_3 a_3$	σ_3	σ_3^2
Total for the frame...	N	n	1	—	A	—	—
Average per stratum..	$\bar{N} = \frac{N}{M}$	$\bar{n} = \frac{n}{M}$	$\frac{1}{M}$	—	$\bar{A} = \frac{A}{M}$	—	—
Average per sampling unit in the frame...	1	$\frac{n}{N}$	$\frac{1}{N}$	$a = \frac{A}{N}$	—	$\bar{\sigma}_w$	σ_w^2

* The weights are P_1, P_2, P_3.

PLAN A

1. Decide with the aid of the equation

$$Var\ \bar{x} = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2 \qquad \text{[Plan A]} \tag{8}$$

$$\doteq \frac{\quad}{n} \tag{9}$$

the size n of the sample required. Or, the equivalent form

$$\frac{N}{n} = N\left(\frac{\sigma_{\bar{x}}}{\sigma}\right)^2 + 1 \tag{10}$$

of Equation (8) may be handier, as it leads directly to Step 2. The symbol $\sigma_{\bar{x}}$ is the value desired for the standard error of $\bar{x}$. The symbol σ^2 is the average variance between sampling units within zones. In sampling new material, it is usually wise to be conservative and to assume that there will be little or no gain from the natural stratification, which means that we should at first set σ^2 equal to the predicted total variance between the sampling units in the frame. Later on, with experience, we may decrease σ^2 to allow for the gain from natural stratification in the frame.

2. Compute the zoning interval $Z = 2N/n$ for 2 subsamples, $10\ N/n$ for 10 subsamples. Form zones of Z consecutive sampling units in the frame.

3. Draw with random numbers between 1 and Z, 1 sampling unit from every zone. These drawings form Subsample 1. Draw another sampling unit from every zone for Subsample 2, and likewise for the other subsamples if any. Mark each sampling unit to show which subsample it belongs to, and which zone it came from.

TABLE 2. NOTATIONS AND DEFINITIONS FOR THE SAMPLE

Stratum	Population in the sample	Mean population per sampling unit	Estimated total population	Variance of this estimate
1......................	x_1 x-population in Stratum 1	$\bar{x}_1 = \frac{x_1}{n_1}$	$X_1 = N_1 \frac{x_1}{n_1}$	$Var\ X_1$
2......................	x_2 x-population in Stratum 2	$\bar{x}_2 = \frac{x_2}{n_2}$	$X_2 = N_2 \frac{x_2}{n_2}$	$Var\ X_2$
3......................	x_3 population in Stratum 3	$\bar{x}_3 = \frac{x_3}{n_3}$	$X_3 = N_3 \frac{x_3}{n_3}$	$Var\ X_3$
Sum....................	x	—	X	$Var\ X$*

* The variances are additive only if the N_i (or P_i) are known and used in the estimate X.

4. Carry out the interviews or the tests on the entire sample

5. Form the x–population by subsample.

6. Form from Subsample i the estimate.

$$X_{(i)} = Z_{x(i)} \tag{11}$$

where $^{(i)}$ is the x-population in Subsample i.

7. Form the final estimate

$$X = Z\bar{x} \tag{12}$$

from all subsamples combined. $\bar{x}$ is the average of the $x^{(i)}$, where $x^{(i)}$, is the x-population in Subsample i.

8. Estimate the precision attained. For example, if there were 10 subsamples and 1 thick zone,* we could use the range between the maximum and the minimum of the 10 values of $x^{(i)}$ to form the estimate

$$\hat{C}_X = \hat{C}_{\bar{x}} = \frac{1}{10\bar{x}}\left[\bar{x}^{(i)}_{max} - \bar{x}^{(i)}_{min}\right] \tag{13}$$

where $\hat{C}_{\bar{x}}$ denotes an estimate of the coefficient of variation of $\bar{x}$. One may of course estimate $C_{\bar{x}}$ by the sum of squares, if he prefers.

Ten subsamples and 1 thick zone give 9 degrees of freedom. One may acquire more degrees of freedom by tabulating the results in 2 thick zones. (It is a simple matter to form 2 thick zones so as to get almost exactly 18 degrees of freedom. For example, if the material is highly variable from 1 thin zone to another, one may allot odd thin zones to Thick Zone 1, and even thin zones to Thick Zone 2.)

If one is using 2 subsamples, it is advisable to tabulate the results in 8 or 10 or more thick zones to gain sufficient degrees of freedom, and to use the sum of squares in the estimate. The computation by which to estimate the precision $C_{\bar{x}}$ is in any case valid and simple with a replicated design.[2]

Proportionate allocation (Plan B). We proceed now to give fuller directions on Plan B. We assume in Plans B and C that all the sampling units in the frame have been classified into strata and counted, and that the sampling units in Stratum 1 have serial numbers 1 to N_1, in Stratum 2 from 1 to N_2, etc.

It is advisable, in all the plans, to maintain in each stratum the order in which the sampling units appeared in the original frame. By doing so, we capture the benefit of the natural stratification that already existed in the frame. This rule is not essential to the theory, but it is good for efficient design.

PLAN B

1. Decide with the help of the equation

$$n = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_w^2 \qquad \text{[Plan B]}$$

$$\doteq \frac{1}{n}\sigma_w^2$$

* *Editorial note*: See the article by Dr. Deming published in *Estadística* No. 55, under the title, "Simplificaciones en el Diseño del Muestreo Mediante Reiteración con Probabilidades Iguales y sin Etapas," p. 277-305.

[2] W. Edwards Deming, "On simplifications of sampling design through replication with equal probabilities and without stages," *Journal of the American Statistical Association*, volume 51, March 1956: pages 24-53.

2. Compute the zoning interval $Z = 2\,N/n$ for 2 subsamples, $10\,N/n$ for 10 subsamples. Form zones of Z sampling units in all strata.

3. Draw with random numbers between 1 and Z, 1 sampling unit from every zone, onward through all strata. These drawings form Subsample 1. Draw another sampling unit from every zone for Subsample 2, and likewise for the other subsamples if any. Mark each sampling unit to show which subsample it belongs to, and which stratum, and which zone it came from.

4. Carry out the interviews or the tests on the entire sample.

5. Form the x-populations by subsample: designate them $x^{(1)}$, $x^{(2)}$, etc., where $x^{(i)}$ is the x-population in Subsample i through all strata.

6. Form from Subsample i estimates as in Step 6 of Plan A.

7. Form the final estimate X as in Step 7 of Plan A.

8. Estimate the precision obtained, as in Plan A. No weighting is required, because all the sampling units in the frame, regardless of stratum, have in proportionate stratified sampling (Plan B) the same probability of selection. We may, of course, form separate estimates by stratum, if we need them. The estimates of X and of *Var* X are additive, as in Table 2. The degrees of freedom, however, are not additive.[3]

Neyman allocation (*Plan C*). One may be able in some problems to improve on proportionate sampling by altering n_1 and n_2 in proportion to σ_1 and σ_2. This is so when it is possible to form strata so that their variabilities (as measured by σ_1 and σ_2) are distinctly different. Such a plan was first put into practice by Neyman.[4] Plan C will denote Neyman allocation when we fix the sample-sizes in advance. The steps in Plan C follow

PLAN C

1. Decide with the aid of the equation

$$\sigma_{\bar{x}}^2 = \frac{(\bar{\sigma}_w)^2}{n} - \frac{\sigma_w^2}{N} \qquad \text{[Plan C]}$$

$$\doteq \frac{(\bar{\sigma}_w)^2}{n} \tag{15}$$

the size n of the sample required. Compute for Stratum i the sample-size,

$$n_i = \frac{P_i\,\sigma_i}{\bar{\sigma}_w} \tag{16}$$

[3] This formula in a more general form for any number of strata (or thick zones), and for any number of degrees of freedom in each stratum, was first given explicitly by F.E. Satterthwaite, "An approximate distribution of estimates of variance components," *Biometrics*, volume 2, 1946: pp. 110-114.

[4] J. Neyman, "On the two different aspects of the representative method," *Journal of the Royal Statistical Society*, volume XCVII, 1934: pages 558-606. The mathematical equations for Neyman allocation were nevertheless published earlier by A. Tschuprow, *Metron*, No. 3, 1923, page 672, but so far as I know, Tschuprow made no use of his formulas.

2. Compute the zoning intervals $Z_i = 2\,N_i/n_i$ for 2 subsamples, $10\,N_i/n_i$ for 10 subsamples. See Step 2 under Plan B. Form the zones in the various strata with the zoning intervals just calculated; see Step 2 under Plan B. The procedure is to make the selections from every stratum in Plan A.

3. Draw with random numbers between 1 and Z_1, 1 sampling unit from every zone of Stratum 1; with random numbers between 1 and Z_2, 1 sampling unit from each zone of Stratum 2; etc. The sample so drawn is Subsample 1. Draw another sampling unit from every zone for Subsample 2, and likewise for the other subsamples if any. Mark each sampling unit to show which subsample it belongs to, and which stratum, and which zone it came from.

4. Carry out the interviews or tests on the entire sample.

5. Form for Subsample i the x-populations $x_1^{(i)}$, $x_2^{(i)}$, $x_3^{(i)}$ stratum by stratum. Do this for every subsample.

6. Compute for Subsample i

$$
\begin{aligned}
X^{(i)} &= Z_1\, x_1^{(i)} + Z_2\, x_2^{(i)} + Z_3\, x_3^{(i)} \\
&= X_1^{(i)} + X_2^{(i)} + X_3^{(i)} \qquad (17)
\end{aligned}
$$

This is the estimate that Subsample i furnishes for the total x–population in the frame. The 3 terms, one by one, are estimates from Subsample i of the x–population stratum by stratum.

7. Calculate $\hat{C}_X = \hat{C}_{\bar{x}}$, following the advice in Plan A. For example

$$\hat{C}_X = \hat{C}_{\bar{x}} = \frac{1}{10\,X}\left[X^{(i)}_{max} - X^{(i)}_{min}\right] \qquad (18)$$

for 10 subsamples.

One may require separate estimates by stratum, in which case the procedure is to compute

$$X_1 = Z_1\, \bar{x}_1 \qquad \text{for Stratum 1} \qquad (19)$$

$$X_2 = Z_2\, \bar{x}_2 \qquad \text{for Stratum 2} \qquad (20)$$

Etc.

and

$$\hat{V}ar\, X_1 = Z_1^2 \frac{1}{k\,(k-1)} \sum \left(x_1^{(i)} - \bar{x}_1\right)^2 \qquad \text{for Stratum 1} \qquad (21)$$

$$\hat{V}ar\, X_2 = Z_2^2 \frac{1}{k\,(k-1)} \sum \left(x_2^{(i)} - \bar{x}_2\right)^2 \qquad \text{for Stratum 2} \qquad (22)$$

Etc.

for k subsamples. $\hat{Var}\, x$ signifies an estimate of $Var\ x$. $\bar{x}_1$ is the average of the k individual x-populations $\bar{x}_1^{(i)}$ in Stratum 1, $\bar{x}_2$ has a similar definition in Stratum 2. The estimates and their variances are additive, as in Step 8 under Plan B; hence we form

$$X = X_1 + X_2 + X_3 \tag{23}$$

$$\hat{Var}\, X = \hat{Var}\, X_1 + \hat{Var}\, X_2 + \hat{Var}\, X_3 \tag{24}$$

STRATIFICATION AFTER SELECTION WITH PROPORTIONS KNOWN

Plans D, E, F, G. These plans (like Plans B and C) all require advance knowledge of P_1, P_2, P_3.

PLAN D

1. Decide with the aid of the equation

$$Var\ \bar{x} = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\sigma_w^2 + \frac{1}{n}\,\sigma_R^2\right) \quad \text{[Plan D]}$$

$$\doteq \frac{1}{n}\left(\sigma_w^2 + \frac{\sigma_R^2}{n}\right) \tag{25}$$

the size n of the sample required. Proceed to draw this sample exactly as in Plan A.

2. Compute the zoning intervals $Z = 2N/n$ for 2 subsamples, $10\,N/n$ for 10 subsamples. Form zones of Z sampling units in the frame.

3. Draw with random numbers between 1 and Z, 1 sampling unit from every zone. These drawings form Subsample 1.

4. Classify each sampling unit into its proper stratum. Mark each sampling unit to show that it belongs to Subsample 1, which stratum, and which zone it came from.

5. Repeat Steps 3 and 4 for Subsample 2, and for the other subsamples if any. The sample-sizes $n_1^{(i)}$, $n_2^{(i)}$, $n_3^{(i)}$ thus drawn will be random variables. (The subscripts refer to the strata; the superscript is the subsample.)

6. Carry out the interviews or the tests on the entire sample.

7. Let $x_1^{(i)}$ denote as before the x-population in Subsample i from Stratum 1, with similar definitions for $x_2^{(i)}$, $x_3^{(i)}$. Form for Subsample i the estimate

$$X^{(i)} = N_1\,\frac{x_1^{(i)}}{n_1} + N_2\,\frac{x_2^{(i)}}{n_2} + N_3\,\frac{x_3^{(i)}}{n_3}$$

$$= X_1^{(i)} + X_2^{(i)} + X_3^{(i)} \tag{26}$$

of the total x-population. Do this for every sample. The 3 terms, one by one, are the estimates that Subsample i *furnishes* for the x-population in the 3 strata.

8. Form the final estimate

$$x = av.\ X^{(i)} \tag{27}$$

and if desired

$$\bar{x} = \frac{X}{N} \tag{28}$$

for the average x-population per sampling unit over all the whole frame.

9. Calculate $\hat{C}_X = \hat{C}_{\bar{x}}$, following the advice in Plan A. For example,

$$\hat{C}_X = \hat{C}_{\bar{x}} = \frac{1}{10\,X}\left[X^{(i)}_{max} - X^{(i)}_{min}\right] \tag{29}$$

for 10 subsamples.

Convenient modifications will occur to the user in practice. For example, it may be necessary to defer the classification of a sampling unit until after the final test or interview. This state of affairs introduces no complication. One simply proceeds into Step 7 after the final tests are complete.

Another point is that one may calculate the estimates X_1, X_2, X_3, for the strata separately, if there be need of them.

Then also, the user will in practice probably prefer, as I do, to make all his drawings from Zone 1 before he proceeds into Zone 2. The first sampling unit drawn from any zone belongs to Subsample 1; the second sampling unit belongs to Subsample 2; etc. The above description of the steps, as written, has the advantage, I hope, that it leaves no doubt about the independence of the subsamples, except for the fact that one will usually draw them without replacement.

PLAN E

1. Decide with the aid of the equation

$$Var\ \bar{x} = \overbrace{\frac{(\bar{\sigma}_w)^2}{n} - \frac{\sigma_w^2}{N}}^{\text{Plan C}} + \frac{1}{n}\left(\frac{1}{n'} - \frac{1}{N}\right)\bar{\sigma}_w\,\bar{\sigma}_R \qquad \text{[Plan E]}$$

$$\doteq \frac{1}{n}\left\{(\bar{\sigma}_w)^2 + \frac{1}{n'}\,\bar{\sigma}_w\,\bar{\sigma}_R\right\} \tag{30}$$

the size n of the sample required. Decide also with the help of the equation

$$n = n' \overline{T} = n' \sum P_i T_i \tag{31}$$

the size of the preliminary sample. Here T_i is the thinning ratio in Stratum i. For example, if we were to retain all the sampling units in Class 1 that appear in the preliminary sample, and half the sampling units in Class 2, then would

$$\overline{T} = P_1 + \frac{1}{2} P_2 \tag{32}$$

2. Compute the zoning interval $Z = 2 N/n'$ for 2 subsamples, $10 N/n'$ for 10 subsamples. Form the zones in the frame; see Step 2 under Plan B.

3, 4, 5. Proceed as in Steps 3, 4, 5 of Plan D. The sizes of the preliminary subsamples so drawn will be random variables in every stratum and in every subsample.

6. Decide on the most likely ratios $\sigma_1 : \sigma_2 : \sigma_3$ for the chief characteristic that the sample is to measure (hereafter, the x-population).

7. Fix for the next step the thinning ratios by the Neyman allocation

$$\frac{n_1}{n'_1} : \frac{n_2}{n'_2} : \frac{n_3}{n'_3} = \sigma_1 : \sigma_2 : \sigma_3 \tag{33}$$

in which n_1', n_2', etc., are the sizes of the preliminary sample in the several strata.

8. Thin the strata that require thinning.

The thinning ratios are relative, and the size n' of the preliminary sample was supposedly chosen so that the class with the biggest σ_1 will not take any thinning at all. Thus, if $\sigma_1 : \sigma_2 : \sigma_3$ were 1: 2: 4, we should leave Stratum 1 as it is; retain 1 sampling unit at random from every 2 of Stratum 2; retain 1 sampling unit from every 4 of Stratum 3. A convenient arrangement is to tie the beginning of Subsample 2 to the end of Subsample 1, to leave no gap. The final sample-sizes $n_1^{(i)}$, $n_2^{(i)}$, etc., in Subsample i will all be random variables.

9, 10, 11. Proceed as in Steps 7, 8, 9 of Plan D.

PLAN F

1. Decide the desired sample-size n as in Plan B. Compute also the sample-sizes $n_i = n P_i$ for Stratum i as in Plan B.

2. Compute the zoning intervals $Z_i = 2 N_i/n_i$ for 2 subsamples, $10 N_i/n_i$ for 10 subsamples.

3. Divide the frame into zones of Z_1 sampling units. Draw a sampling unit from Zone 1. If it belongs to Stratum 1, mark it so. If it belongs to some other stratum, ignore it and draw another and another until you get one for Stratum 1. Move into Zone 2 and repeat the procedure. Continue thus through all the zones. The units so obtained belong to Subsample 1 of Stratum 1.

4. Repeat the same procedure for Subsample 2 and for the other subsamples, if any, to complete the subsamples for Stratum 1.

5. Divide the frame into zones of Z_2 sampling units and repeat Steps 3 and 4 to obtain the sample for Stratum 2.

6. Use the same procedure for Stratum 3, and for the other strata if any.

7-11. Proceed as in Steps 4-8 of Plan B.

PLAN G

1. Decide the desired sample-size n, as in plan C. Compute also the sample-sizes $n_i = n \; P_i \sigma_i / \bar{\sigma}_w$ for Stratum i as in Plan C.

2. Compute the zoning intervals $Z_i = 2 N_i / n_i$ for 2 subsamples, $10 N_i / n_i$ for 10 subsamples.

3, 4, 5, 6. Proceed as in Steps 3, 4, 5, 6 in Plan F.

7-11. Proceed as in Steps 4-8 of Plan C.

In Plans F and G, the sample-sizes are fixed in advance: in Plans D and E they are not; they are random variables. That is why, as the student may have observed, in Plans F and G we may form the estimate X directly, whereas in Plans D and E we must form separate estimates by stratum and then add them. The variances of the plans with fixed sample-sizes are slightly smaller than the variances of the plans with variable sample-sizes for the same total number n of sampling units. The difference lies in the term σ_R^2/n. However, in Plans F and G with fixed sizes there is the additional cost of classifying some units, only to find that we cannot use them in the final sample because the quotas n_i are already filled.

Stratification after selection, continued; Plans H and I. In the previous plans, the proportions P_i were known, and we made use of them in fixing the sizes of the samples in the various strata, and in forming the estimates. We now encounter the problem were the proportions P_i are not known. What we do is to estimate the proportions P_i from classification of a preliminary sample of size N', which the reader will observe, serves as a new frame. Once we have this new frame, Plan H will resemble Plan B; Plan I will resemble Plan C.

STRATIFICATION AFTER SELECTION WITH PROPORTIONS NOT KNOWN

PLAN H

1. Decide with the aid of the equation

$$Var\ \bar{x} = \left(\frac{1}{n} - \frac{1}{N}\right) \sigma_w^2 + \frac{1}{n}\left(\frac{1}{N'} - \frac{1}{N}\right) \sigma_R^2 + \left(\frac{1}{N'} - \frac{1}{N}\right) \sigma_b^2 \qquad \text{[Plan H]}$$

$$\doteq \frac{\sigma_w^2}{n} + \frac{\sigma_b^2}{N'} \qquad \text{[If } N \text{ is large compared with } N' \text{ and with } n\text{]} \qquad (34)$$

the size n of the sample required.

2. Compute the optimum size N' of the preliminary sample. The equation for this is

$$\frac{n}{N'} = \frac{\sigma_w}{\sigma_b} \sqrt{\frac{c_1}{c_2}} \qquad (35)$$

c_1 is the cost of classifying a sampling unit in the preliminary sample, and c_2 is the cost of interviewing or testing a sampling unit in the final sample.

3. Compute the zoning interval $Z = 2N/N'$ for 2 subsamples, $10N/N'$ for 10 subsamples. Form zones of Z consecutive sampling units in the frame.

4. Draw from the frame, just as you would in Plan A, a preliminary sample of size N'. Mark each sampling unit to show which subsample it belongs to, and which zone it came from.

5. Carry out the preliminary tests or interviews on the entire preliminary sample to acquire information by which to classify every sampling unit therein. Mark each sampling unit to show which stratum it belongs to. Maintain in each stratum the order drawn.

6. Record for Subsample i the number of sampling units in each stratum, $N_1^{(i)}$, $N_2^{(i)}$, etc. These counts give the estimates

$$\hat{P}_1^{(i)} = (N_1' : N')^{(i)},\ \hat{P}_2^{(i)} = (N_2' : N')^{(i)},\ etc. \qquad (36)$$

of the proportions P_1, P_2, etc. Form these estimates for every subsample, and for all subsamples combined.

7. Reduce the number of sampling units proportionately in all strata and in all subsamples, to reach the required total sample of size n, decided in advance. For example, to retain 1/3 of the preliminary sample, select by random numbers, for the final sample, 1 out of every 3 consecutive sampling units in the preliminary sample. This reduction in size is called thinning.

8. Carry out on the final sample the main interviews or the main tests.

9. Calculate for Subsample i

$$\bar{x}^{(i)} = \sum P_1^{(i)} \frac{x_1^{(i)}}{n_1^{(i)}} \quad \text{[Sum over all strata]} \tag{37}$$

$$\bar{x} = av.\ \bar{x}^{(i)} \tag{38}$$

$x_1^{(i)}$ is the population in Subsample i from Stratum 1.

10. Calculate

$$\hat{C}_{\bar{x}} = \frac{1}{10\,\bar{x}} - \left[\bar{x}_{max}^{(i)} - \bar{x}_{min}^{(i)}\right] \tag{39}$$

for 10 subsamples, or use the sum of squares if you prefer. For 2 subsamples, form enough thick zones, as in Plan A.

PLAN I

1. Decide with the aid of the equation

$$Var\ \bar{x} = \frac{(\bar{\sigma}_w)^2}{n} - \frac{\sigma_w^2}{N} + \frac{1}{n}\left(\frac{1}{N'} - \frac{1}{N}\right)\bar{\sigma}_w\,\bar{\sigma}_R + \left(\frac{1}{N'} - \frac{1}{N}\right)\sigma_R^2 \quad \text{[Plan I]}$$

$$\doteq \frac{(\bar{\sigma}_w)^2}{n} + \frac{\sigma_b^2}{N'} \quad \text{[If } N \text{ is large compared with } N' \text{ and with } n\text{]} \tag{40}$$

the size n of the sample required.

2. Compute the optimum size N' of the preliminary sample. The equation for this is

$$\frac{n}{N'} = \frac{\bar{\sigma}_w}{\sigma}\sqrt{\frac{c_1}{c_2}} \quad \text{[Due to Neyman [5]]} \tag{41}$$

c_1 is the cost of classifying a sampling unit in the preliminary sample, and c_2 is the cost of interviewing or testing a sampling unit in the final sample.

3 and 4. Draw the preliminary sample N', as in Plan H.

5. Carry out the preliminary tests or interviews on the entire preliminary sample to acquire information by which to classify every sampling unit therein. Mark each sampling unit to show which stratum it belongs to. Maintain in each stratum the order drawn.

---

[5] J. Neyman, "Contributions to the theory of sampling human populations," *Journal of the American Statistical Association, volume* 33, 1938: pp. 101-116.

6. Decide on the most likely ratios $\sigma_1 : \sigma_2 : \sigma_3$ for the chief characteristic that the sample is expected to measure (hereafter, the x–population). Fix the thinning ratios $n_i : N'_i$ by the Neyman relations

$$\frac{n_1}{N'_1} : \frac{n_2}{N'_2} : \frac{n_3}{N'_3} = \sigma_1 : \sigma_2 : \sigma_3 \tag{42}$$

Comparison of the variances of Plans A, B, C. It is a simple matter to show that the following relations hold between the variances of Plans A, B, C. The letters A, B, C denote variances.

$$\frac{A - B}{A} = \left(\frac{\sigma_b}{\sigma}\right)^2 \tag{43}$$

$$B = A \left(\frac{\sigma_w}{\sigma}\right)^2 \tag{44}$$

$$\frac{B - C}{B} = 1 - \left(\frac{\bar{\sigma}_w}{\sigma_w}\right)^2 \tag{45}$$

$$C = B \left(\frac{\bar{\sigma}_w}{\sigma_w}\right)^2$$

$$= A \left(\frac{\bar{\sigma}_w}{\sigma}\right)^2 \tag{46}$$

The last 3 equations, involving Plan C, disregard $1/N$ in comparison with $1/n$.

EXAMPLE

An example of Plan I: A survey of mental retardation. This was a study of a sample of families in the State of Delaware to determine *inter alia* the prevalence of mental retardation. It turned out to be possible, by stratifying a preliminary sample on the basis of the relatively cheap Wexler-Bellevue test, to cut the cost of the field-work to about 46 percent of what its cost would have been without benefit of statistical theory. The calculations for the design of the sample follow. From a statistical standpoint there were two main aims:

a. To make an estimate of the prevalence of mental retardation (the average number of retarded persons per family);

b. To carry out certain psychological tests and measurements in both retarded and nonretarded families (a retarded family being one in which one member or more over 10 years of age is retarded).

Certain facts presented themselves: (1) The interviews and tests in aim *b* are long and expensive (hereafter this will be known as the L-test); (2) the L-test should therefore concentrate on the retarded families, and not dissipate itself on the much larger stratum of nonretarded families. (3) In aim *b* it was best, from the standpoint of statistical efficiency, that the number of tests should be about equal in the 2 groups.

Enquiry disclosed the fact that there is a simple, brief, and inexpensive test, called the Wexler-Bellevue test (hereafter the W-B test), which will classify a person above or below any designated point on the psychological scale, in almost exact agreement with the L-test (which, however, served many other purposes). Use of the W-B test would thus permit quick and inexpensive classification of the families in a preliminary sample into two classes:

Class 1: Families that are retarded, according to the W-B test;
Class 2: All other families.

Because of a fortunate relation between proportions and costs, this sample turned out to be efficient for both aims *a* and *b*.

The plan of sampling. Once the usefulness of the W-B test became clear, Plan I emerged as the best plan to use. Here is a brief description of the procedure:

1. Draw a master sample of about 2600 dwelling units in 10 subsamples, about 260 dwelling units in each subsample. The size of the master sample was chosen so that it would be big enough to furnish the preliminary sample N'.

2. Conduct a pilot study on one subsample (Subsample 10) to learn something about the proportions P_1 and P_2, and to get some experience and some figures on costs; to classify the families in the pilot study into Classes 1 and 2, on the basis of the W-B test.

3. Study by the L-test the families in Class 1, to estimate the distribution of retarded persons in Class 1.

4. Draw up the sampling plan and procedure. Decide on the thinning ratios and on the size of the final sample. Proceed with the field work and tabulations.

The pilot study would cost very little extra, as it was a portion of the main study, carried out in advance up to the point of thinning Class 2.

The results of the pilot study are in Table 3. The statistical characteristics of the 2 classes and for both classes combined appear below the table. The L-test conducted in Class 1 disclosed the fact that 1 family classed as retarded by the W-B test was actually not retarded. Thus, the W-B test made an excellent separation, though not perfect.

Of the 257 families in the pilot study, 57 went into Class 1 and 200 into Class 2. These figures gave the preliminary estimates of P_1 and P_2

that appear in the table. The estimates were good enough for planning, but not good enough to permit use of Plan E.

It was too expensive to conduct the L-test on the 200 families of Class 2, but it was possible by the aid of expert knowledge (not mine) to predict close enough the number of nonretarded families amongst the 200 families of Class 2, and thus to fix the sample-sizes and the thinning ratios.

This example provides an illustration of the sampling of new material. The pilot study gave the information that we needed about Class 1, and expert knowledge furnished all the information we needed about Class 2.

Optimum sizes of the sample. For the Neyman allocation of the final sample to the two strata we take

$$\begin{aligned} n_2 &= n_1 \frac{N_2 \sigma_2}{N_1 \sigma_1} \\ &= \frac{200 \times 0.14}{57 \times 0.69} n_1 = 0.71\, n_1 \end{aligned} \qquad (47)$$

whence

$$n = n_1 + n_2 = 1.7\, n_1 \qquad (48)$$

Considerations of cost and of the expected precision shown by Equation *infra* limited the final sample to about 400 families, which should be distributed as follows

$$\left.\begin{aligned} n_1 &= 400/1.7 = 235 \text{ in Class } 1 \\ n_2 &= 165 \text{ in Class } 2 \end{aligned}\right\} \qquad (49)$$

For the size N' of the preliminary sample we make use of Equation (41) in which c_1 is the average cost of using the W-B test in 1 family of the preliminary sample, and c_2 is the average cost of the L-test in 1 family of the final sample. The pilot study showed that

$$\left.\begin{aligned} c_1 &= \$\ 6 \\ c_2 &= \$\ 50 \end{aligned}\right\} \qquad (50)$$

whereupon

$$\frac{n}{N'} = \frac{\bar{\sigma}_w}{\sigma} \sqrt{\frac{c_1}{c_2}} = \frac{26}{57} \sqrt{\frac{6}{50}} = 0.16 \qquad (51)$$

If $n = 400$, then the optimum preliminary sample should be

$$N' = 400/0.16 = 2.500 \tag{52}$$

which will divide itself approximately into

$$\left.\begin{array}{ll} N'_1 = N' P_1 = \ \ 550 & \text{families in Class 1} \\ N'_2 = N' P_2 = 1950 & \text{families in Class 2} \end{array}\right\}$$

On this basis we calculate the thinning ratios

$$\frac{n_1}{N'_1} = \frac{235}{550}$$

$$\frac{n_2}{N'_2} = \frac{165}{1950}$$

practical approximations being

$$\frac{n_1}{N'_1} = \frac{1}{2} \tag{53}$$

$$\frac{n_2}{N'_2} = \frac{1}{12} \tag{54}$$

That is, we select 1 family at random from each successive 2 families in the preliminary sample of Class 1, and 1 family at random from each successive 12 in Class 2. These convenient ratios will give almost the same precision as the exact ratios. With these numbers the proposed sample should give

$$Var\ \bar{x} = \frac{(\bar{\sigma}_w)^2}{n} + \frac{\sigma_b^2}{N'} \tag{55}$$

$$= \frac{0.26^2}{400} + \frac{0.32}{2500}$$

$$= 0.00017 + 0.00013 = 0.00030$$

$$\sigma_{\bar{x}} = \sqrt{0.00030} = 0.017 \tag{56}$$

$$C_{\bar{x}} = \frac{\sigma_{\bar{x}}}{a} = \frac{0.017}{0.320} = \text{about } 5\% \tag{57}$$

TABLE 3. DISTRIBUTION OF RETARDATION

Class 1 (From the pilot study of 57 families)		Class 2 (By expert knowledge)
1 have 0 persons retarded 39 " 1 " " 11 " 2 " " 6 " 3 " " 0 " 4 " "		196 have 0 persons retarded 4 " 1 " " 0 " 2 or more persons retarded
$a_1 = 1.39$	Mean	$a_2 = \frac{4}{200} = 0.02$
$\sigma_1^2 = 0.48$	Variance	$\sigma_2^2 = \frac{4}{200}\ \frac{196}{200} = 0.02$
$\sigma_1 = 0.69$	Standard deviation	$\sigma_2 = \sqrt{\frac{4}{200} \times \frac{196}{200}} = 0.14$

Both classes combined

$$P_1 = 57/257 = 0.22$$

$$P_2 = 200/257 = 0.78$$

$$a = P_1\, a_1 + P_2\, a_2 = 32$$

$$\sigma_w^2 = P_1\, \sigma_1^2 + P_2\, \sigma_2^2 = 0.12$$

$$\sigma_b^2 = P_1\,(a_1 - a)^2 + P_2\,(a_2 - a)^2$$

$$= P_1\, P_2\,(a_2 - a_1)^2 = 0.32$$

$$\sigma_b = 0.57$$

$$\sigma^2 = \sigma_w^2 + \sigma_b^2 = 0.44$$

$$\bar{\sigma}_w = P_1\, \sigma_1 + P_2\, \sigma_2 = 0.26$$

This precision is sufficient, in view of the uncertainties of the tests, and of difficulties of definition of residence, nonresponse, and the like.

For reduction of the master sample from 2600 to the preliminary sample of size $N' = 2500$, it was convenient to delete 1 sampling unit in every successive 13 through all 10 subsamples.

Calculation of the expected saving over Plan A. Now let us see what size of sample this same precision would require if there were no stratification. Then would

$$Var\ \bar{x} = \frac{\sigma^2}{n} \qquad \text{[Plan A, no additional stratification]}$$

With $Var\ \bar{x} = 0.00030$, $\sigma^2 = 0.44$, we find that $n = 1500$.

Now how about the relative costs of Plans A and I? The computations are in Table 4, whence we see that Plan I will cost only about 45 percent as much as Plan A (no stratification).

TABLE 4. COMPARISON OF THE COSTS OF PLANS A AND I IN THE SURVEY OF MENTAL RETARDATION

Plan	Preliminary sample		Final sample		Total cost
	Number of families	Cost per family	Number of families	Cost per family	
A.........	—	—	1500	\$ 50	\$ 75,000
I.........	2400	\$ 6	400	50	34,400
Difference........					\$ 40,600
Saving effected by Plan 40,600/75,000........					54%

COMPARISON OF COSTS

Table of comparative costs. Table 5 shows how the average cost of any plan will compare with the average cost of Plan A, when both plans yield the same variance, viz., $(1 - \nu/N)\ \sigma^2/\nu$. This table is helpful in the choice of plan. The assumptions are in the heading of the table. There is the further assumption that the cost c_2 or test of an interview is the same in all strata.

There are occasional small intangible costs, besides those in the table, not easy to evaluate: these are mentioned in the notes at the end of the table: though small, they may serve to tip the balance one way or another in case the costs in the table for two plans turn out to be about the same. One may do well to reduce the size of the sample in a stratum where the cost of interviewing or of testing is excessive, and to build it up in other strata. The table of costs will then not be exactly applicable.

We may illustrate the use of the table with an example. Suppose that

$$N = 10{,}000$$

$$\nu = 500$$

$$c_1 = 20¢$$

$$c_2 = \$5.00$$

$$\sigma_w : \sigma = 0.8$$

$$K = \nu c_2 = \$2500, \text{ the cost of Plan A}$$

Then the cost of Plan B will be

$$Nc_1 + K\left(\frac{\sigma_w}{\sigma}\right)^2 = 10{,}000 \times 0.20 + 2500 \times 0.64$$

$$= \$2{,}000 + \$1{,}600 = \$3{,}600 \qquad (58)$$

which is more than the cost $K = \$2{,}500$ for Plan A.

Suppose that $\sigma_R : \sigma_w$ were 3. Then Plan D would cost

$$K\left(\frac{\sigma_w}{\sigma}\right)^2 + c_1\,\nu\left(\frac{\sigma_w}{\sigma}\right)^2 + (c_1 + c_2)\left(\frac{\sigma_R}{\sigma_w}\right)^2 = \$1{,}600 + 0.20 \times 500 \times 0.64$$

$$+ \$5.20 \times 3^2 = \$1{,}600 + \$64 + \$46.80 = \$1710.80$$

which is less than the cost of Plan A or of Plan B.

TABLE 5. COMPARATIVE COSTS OF THE VARIOUS PLANS OF STRATIFIED SAMPLING THAT WILL DELIVER THE SAME PRECISION AS PLAN A WILL DELIVER WITH THE SIZE OF SAMPLE EQUAL TO ν.

Assumptions: (1) The frame is not already stratified; (2) The unit costs of classification, and of interviewing or of testing, are the same in all strata. c_1 is the cost to classify one unit. In plans H and I the cost c_1 may include the cost of a preliminary test or short interview. c_2 is the cost to interview or to test one unit in the final sample for the main study.

Plan	Average cost	Remarks
A No stratification.................	$\nu c_2 = K$	This cost K and the sample-size ν furnish the bases for reference: $\sigma_{\bar{x}} = \sigma / \sqrt{\nu}$

THE PROPORTIONS P_i KNOWN
CLASSIFY ALL N SAMPLING UNITS IN THE FRAME

Plan	Average cost	Remarks
B Proportionate allocation; sample-sizes n_i fixed in advance	$Nc_1 + \nu c_2 \left(\frac{\sigma_w}{\sigma}\right)^2$ $= Nc_1 + K \left(\frac{\sigma_w}{\sigma}\right)^2$	Size of sample, $\nu \left(\frac{\sigma_w}{\sigma}\right)^2$
C Neyman allocation; sample-sizes n_i fixed in advance	$Nc_1 + K \left(\frac{\bar{\sigma}_w}{\sigma}\right)^2$	Size of sample, $n \doteq \nu \left(\frac{\bar{\sigma}_w}{\sigma}\right)^2$

THE PROPORTIONS P_i KNOWN
CLASSIFY ONLY THE n SAMPLING UNITS OF THE SAMPLE

Plan	Average cost	Remarks
D Draw and classify a sample of specified size n, which shall be also the final sample. The individual sample-sizes n_i are random variables.	$n(c_1 + c_2)$ $= K\left(\frac{\sigma_w}{\sigma}\right)^2 + \nu c_1 (\sigma_w/\sigma)^2$ $+ (c_1 + c_2)(\sigma_R/\sigma_w)^2$	Size of sample, $n \doteq \left(\frac{\sigma_w}{\sigma}\right)^2 + \left(\frac{\sigma_R}{\sigma_w}\right)^2$
E Draw and classify a specified number n' of sampling units. Thin the strata by use of the Neyman ratios. The total sample n and the individual sample-size n_i are all random variables.	$n'c_1 + nc_2$ $= n'c_1 + Kn/\nu = n'c_1$ $+ K\left(\frac{\bar{\sigma}_w}{\sigma}\right)^2 1 + \left\{\frac{\bar{T}\bar{\sigma}_R}{\bar{\sigma}_w}\right\}$	Average size of the final sample, $\bar{n} = \nu\left(\frac{\bar{\sigma}_w}{\sigma}\right)^2 + \left(\frac{\bar{T}\bar{\sigma}_R}{\bar{\sigma}_w}\right)$ $n' = n/\bar{T}$

THE PROPORTIONS P_i KNOWN
CLASSIFY ONLY ENOUGH SAMPLING UNITS TO FILL THE QUOTAS n_i

F	Fix the sample-sizes n_i in advance by proportionate allocation. Draw sampling units and classify them until all the quotas n_i are filled. n is fixed; also the n_i.	$n'c_1 + nc_2$ $= n'c_1 + K\left(\frac{\sigma_w}{\sigma}\right)^2$ Note that n' in this plan is not equal numerically to the n' in Plan E, nor to the n' in Plan G.	Total sample n and the quotas n_i as in Plan B. Variance the same as the variance of plan B. The number n' of sampling units that require classification will be, on the average, a bit bigger than n.
G	Fix the sample-sizes n_i in advance by the Neyman allocation. Draw sampling units and classify them until all the quotas n_i are filled. n is fixed.	$n'c_1 + nc_2$ $= n'c_1 + \nu(\bar{\sigma}_w/\sigma)^2 c_2$ $= n'c_1 + K(\bar{\sigma}_w/\sigma)^2$	Total sample n and the quotas n_i as in Plan C. Variance, the same as the variance of Plan C. The number n' of sampling units that require classification will be, on the average, a bit bigger than $n/\overline{T}$.

THE PROPORTIONS P_i NOT KNOWN IN ADVANCE
CLASSIFY A PRELIMINARY SAMPLE OF SIZE N' TO
ESTIMATE THE PROPORTIONS P_i

H	Classify the preliminary sample, and thin all classes proportionately to reach a specified final size n. The individual sample-sizes n_i are random variables.	$N'c_1 + nc_2$ $= N'c_1 + K\left(\frac{\sigma_w}{\sigma}\right)^2$	$Opt\ N' = \frac{\sigma_w}{\sigma_b}\sqrt{\frac{c_1}{c_2}}$ Size of final sample, $n \doteq \nu\left(\frac{\sigma_w}{\sigma}\right)^2\left\{1 + \frac{1}{N'}\left[\nu\left(\frac{\sigma_b}{\sigma}\right)^2 + \left(\frac{\sigma_R}{\sigma_w}\right)^2\right]\right\}$
I	Classify the preliminary sample, and thin the classes by the Neyman ratios to reach a specified final size n. The individual sample-sizes are random variables.	$N'c_1 + nc_2$ $= N'c_1 + K\left(\frac{\bar{\sigma}_w}{\sigma}\right)^2$	$Opt\ N' = \frac{\bar{\sigma}_w}{\sigma_b}\sqrt{\frac{c_1}{c_2}}$ Size of final sample, $n \doteq \nu\left(\frac{\bar{\sigma}_w}{\sigma}\right)^2\left\{1 + \frac{1}{N'}\left[\nu\left(\frac{\sigma_b}{\sigma}\right)^2 + \left(\frac{\sigma_R}{\bar{\sigma}_w}\right)^2\right]\right\}$

* * *