ON SIMPLIFICATIONS OF SAMPLING DESIGN THROUGH REPLICATION WITH EQUAL PROBABILITIES AND WITHOUT STAGES

By

W. EDWARDS DEMING

Graduate School of Business Administration
New York University

Reprinted from

The Journal of the American Statistical Association,
Vol. 51, March 1956: pp. 24-53

ON SIMPLIFICATIONS OF SAMPLING DESIGN THROUGH REPLICATION WITH EQUAL PROBABILITIES AND WITHOUT STAGES

By

W. EDWARDS DEMING

GRADUATE SCHOOL OF BUSINESS ADMINISTRATION
NEW YORK UNIVERSITY

Reprinted from

THE JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION,
Vol. 51, March 1956: pp. 24-53

ON SIMPLIFICATIONS OF SAMPLING DESIGN THROUGH REPLICATION WITH EQUAL PROBABILITIES AND WITHOUT STAGES*

W. Edwards Deming
New York University

1. GENERAL DESCRIPTION

Purpose of This Paper. The purpose here is to explain a simplified procedure for the selection of a sample and for the numerical computation of the standard errors from the returns, with gains in overall statistical efficiency. There is no new theory here: instead, this is a synthesis and rearrangement of well-known principles and practices. The procedures to be described have been applied here and abroad in various kinds of social and economic studies, including estimates of acreage and of yield, in marketing research, and in studies of attitudes, in program-listening, in the appraisal of buildings and of other physical plants, in the testing of industrial materials, and in studies of accounting records. The main features are these:

1. Rapid, uniform, and valid computation of the estimates and of their standard errors through replication of the sampling procedure. Added simplicity, under certain conditions, by use of the range.

* The author expresses with pleasure his indebtedness to Leo P. Crespi and to Fred W. Trembour for the privilege of introducing in September, 1952, the plan of duplicate samples into the public opinion surveys carried out by the American Embassy in Germany (then the U. S. High Commission in Bonn). In the original experimental work, and in earlier and later drafts of the manuscript and of the instructions for the procedure, I had the continual assistance of Josephine D. Cunningham, without whom this publication would have suffered indefinite delay.

The illustrations and the tables from the sample of Cincinnati are presented through the courtesy of the firm O'Brien-Sherwood of New York. Helpful comments and criticisms of earlier manuscripts came from my friends and colleagues F. F. Stephan of Princeton, Tore Dalenius of the Central Bureau of Statistics in Stockholm, P. C. Tang now with the Food and Agriculture Organization of the United Nations, Octavio A. de Moraes of the Inter-American Statistical Institute, and Nathan Keyfitz of the Dominion Bureau of Statistics in Ottawa. I owe a special debt of gratitude to Max Bershad and William N. Hurwitz of the Bureau of the Census in Washington for their generous assistance, especially on the estimate of the variance of a ratio and on the variance of the general function. Finally, I owe much to D. B. Lahiri of the Indian Statistical Institute in Calcutta for several opinions, and for his paper cited in the first footnote.

2. Paper zones of equal size, which permit the use of equal probabilities and the theory of single-stage sampling.

3. The elimination of the complex formulas associated with multi-stage sampling and with unequal probabilities. The elimination of the further complexity of the formulas associated with 2 or more drawings from one primary unit.

4. Fine stratification by means of the paper zones.

5. The economy of multi-stage preparation of the sampling units, with simplifications and fewer mistakes in preparation.

6. The elimination of losses in efficiency and of special consideration of primary sampling units of extra-large size, or of extra-small size. The paper zones distribute the sample with the population—heavy where the population is dense, and light where the population is sparse.

7. Applicability with little modification (with reduced efficiency, of course) to conditions where there are not reliable census data.

8. Complete freedom in the basic design, such as in the size of the work-load, in the size of the segment, and in the modes of stratification. The basic design may include without complication (a) two-way stratification with forced selection of heterogeneous sampling units; (b) randomized segments for heavy pressure on a sample of nonresponses, or for the measurement of the differences between interviewers, different methods of training, alternative questions, etc.

9. A simplified field-procedure (optional), which permits the delineation of any economical or convenient size of segment, and if desired, immediate selection for coverage.

The design to be described in this paper is a combination of features all well known separately. For example, replicated samples are the foundation of the design of experiment. Replication in sampling has not hitherto been so common, yet it is the chief contributor to the simplicity of the Tukey plan (*vide infra*). Replication in sampling goes back in fact to the early work of Mahalanobis in 1936 in his surveys of jute in Bengal.[1] The key that opens the door to the advantages of replication and to equal probabilities, with the efficiency of strata in

[1] P. C. Mahalanobis, "On large-scale sample-surveys," *Philosophical Transactions of the Royal Society*, 231B (1944), 329–451; "Recent experiments in statistical sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society*, CIX (1946), 325–78. Recent papers by D. B. Lahiri relate more recent experience in the use of multiple drawings, and the elimination of primary units. The references are, "Technical paper No. 5 on the National Sample Survey" (The Department of Economic Affairs, Ministry of Finance, New Delhi, March, 1954); published also in *Sankhya*, 14 (1954), 264–316. His recommendations are in many ways parallel to those in this paper.

Hansen, Hurwitz, and Madow describe for an urban area a selection of sampling units that in preparation bears some features similar to the procedures in this paper, and which may therefore be usefully consulted in connection with this paper; see their book *Sample Survey Methods and Theory* (New York: John Wiley & Sons, 1953), vol. I, ch. 8, sec. 6.

fine classes, and to the theory of the single stage, is the paper zones that all contain the same number of work-loads. The combination of features to be described here brings within reach a wider use of probability sampling by research organizations and by government offices.

Although the possibility of showing a valid standard error for an estimate is by definition a feature of any probability sample,[2] it is a fact that results of probability samples have too often appeared in the past without standard errors because of the sheer labor of computation. This paper offers a solution of the difficulty.

The sampling unit. In the exposition presented here, the sample will be a survey of a human population. The elementary sampling unit will be a work-load, which will be transformed into a segment of area, or several segments of area, intended to supply to an interviewer an evening's work, or perhaps two evenings' work, or a week's work.

In other applications, not treated here, but involving no new principles, the sampling unit may be an acre, or a small plot in a field, a dollar of investment, a dollar of income or of expense, a card or 5 successive cards in a file, a line or 5 successive lines in a ledger, one employee on the roll, a group of employees, or a test-unit or a test-panel in a shipment of materials.

Measure of size; the Cdu. It is convenient now to invent the symbol Cdu's for the number of dwelling units that an area contained in the last census or according to any later information. Thus an area that contained 317 dwelling units at the last census and for which we have no new information contains now 317 Cdu's, even though the actual (unknown) number of dwelling units in the area today is different.

For illustration, a work-load will be 10 Cdu's. If the census or other information is not too far out of date, the work-loads will average about 10 dwelling units. If there has been a 10 per cent uniform growth, the work-loads will average about 11 dwelling units.

The size suggested here for a work-load only serves for illustration, although it is based on experience. Moreover, the segment to be assigned for canvass need not have the same size as a work-load. It may be half as big, or a third as big. In the section "Suggested field-procedure for the selection of segments," the intended size of the segment is, for illustration, 5 dwelling units, and the segments are scattered about the area that contains the random work-load. The size of the work-load and of the segment should be chosen in any proposed survey to achieve efficiency and smooth performance. This paper need not

[2] The term probability sample and its definition appeared first in the author's book *Some Theory of Sampling* (New York: John Wiley & Sons, 1950), p. 9.

treat the problem of what size of work-load is best for any given set of conditions, nor what size of segment is best, as these problems are already amply treated by theory in books.

The procedure will be illustrated first under the assumption that census figures exist for small areas. It will be illustrated later by a sample of Mexico where there are not such data.

The frame and the zones. We start with a list of census areas, which in urban areas may be sections, tracts, or blocks, and which in suburbs and in rural areas may be sections, enumeration districts, tracts, counties, or some other type of area. This list will be the frame.[3] The frame will show for each area listed the number of work-loads therein. The accumulated work-loads, area by area, will give a serial number to every work-load. Table 3 provides an example.

The frame assimilates a long string of beads, each bead being a work-load. Each bead bears a serial number. Markers, unequally spaced, show the end of one census area and the beginning of another. Other markers, uniformly spaced, will divide the entire string of beads into zones.

The number of work-loads (beads) in a zone will be determined by the average number of dwelling units or of people in a work-load, by the number of replications per zone, and by the total number of dwelling units or of people required in the sample. The symbol Z will denote the number of work-loads in a zone. Illustrations occur later.

The boundaries of a zone will cut across blocks, tracts, sections, counties, cities, and strata. The boundaries of a zone never show on a map. The zone is merely a convenience for assigning equal probabilities to the work-loads in a stratum.

Some remarks on the order of listing the census areas in the frame. The order in which the census areas appear in the frame should serve two requirements. First, a census area that is smaller than a zone (e.g., a block, tract, enumeration district) should if convenient go into the same zone with other areas that have similar characteristics (e.g., average income, occupation), in order to achieve possible gains from stratification. In other words, each zone should be as homogeneous as possible. Within a zone, however, the order in which the census areas appear is immaterial.

No problem of stratification arises for an area whose size is as big as a zone, as it will be in the sample no matter where it appears in the

[3] The frame was defined (but not named) by F. F. Stephan in his article, "Practical problems of sampling procedures," *American Sociological Review*, 1 (1936), 569–80, as a means of access to the universe. The term "frame" came from Frank Yates in a meeting of the United Nations' Sub-Commission on Statistical Sampling in 1947.

frame. Second, the order in which the zones appear in the frame should facilitate tabulation. The zone is the smallest possible building-block; any area of tabulation must be built up of zones.

For example, in a sample of a region, the frame might commence with the eastern portion (e.g., New England); 1st, the central portions of the big cities, in order by size; 2nd, the fringes of these cities; 3rd, the smaller cities, followed by the counties in the rural parts, in some significant order. This order would give strong geographic stratification and would permit separate tabulation for the central portions of all the big cities in the east, singly or by size-groups; likewise it would permit separate tabulations for the fringes of these cities; and for the rural part. It might not be the most convenient order for tabulation of an entire metropolitan district. Within a city, or within any other area, the requirements of stratification will usually be met if the order of listing the census areas (e.g., blocks or tracts) is geographic, or in the order of appearance in a census table (often geographic), or by rent-levels, or both.

Replicated drawings. The main feature of the plan is that there will be 2 random selections of work-loads from every zone. The 1st random number draws a work-load for Sample 1; repetition with a 2nd random number draws a work-load for Sample 2. If the sampling design calls for 1 segment to a work-load, then we draw the work-loads without replacement; we reject a duplicate random number within any zone. If the sampling design calls for the creation of 2 or more segments to a work-load, not all to be interviewed, then we accept a duplicate random number in the drawing of the work-loads, but we draw without replacement the segments for interview within the work-load (as we can not interview twice in the same households). Thus, the segments are drawn as a single-stage sample without replacement. The proportion of segments in the sample is $2/Z$, which explains the finite multiplier $1-2/Z$ in the variances later on.

The two sets of work-loads, when interviewed, will bring forth different results—different numbers of dwelling units and different numbers of males, females, and children. The procedure draws blocks, tracts, E. D.'s., and counties with probabilities in proportion to their sizes in Cdu's, yet because of the replication with equal probabilities within a zone, the theory for the estimates and for their standard errors is single stage.

More than 2 drawings per zone are permissible and will be treated later.

Suggested field-procedure for the selection of segments. A work-load once drawn into the sample will fall in a certain census area, perhaps a block, or a section, or other area. A field-worker thereupon goes to this area and creates segments, following any workable plan. She may use the half-open interval,[4] or she may indicate segments on the map furnished to her. She may find that the map requires considerable internal revision. The desired size of a segment may be 10 dwelling units, 5, or 4, or even 1, according to specifications. There are several methods in common use for creating segments, and there need be no elaboration here.[5] Detailed maps, directories, and aerial photographs are helpful where the delineation of the segments is to be done in the office.

Whatever the intended average size of a segment, there will usually be considerable variations in the sizes of the segments, because the primary necessity in delineation is clarity of boundary, not uniformity of size. Fortunately, segments need not contain equal numbers of people nor of dwelling units. All the dwelling units in a segment have the same probability as the segment of coming into the sample at this stage; likewise all the households therein, and all the people. Inequality and variability in the sizes of segments introduce no change in the probabilities of selection of the people or of the dwelling units therein, and no bias. Substantial inequalities in size will usually cause a negligible decrease in the precision of a ratio, and scarcely more in the precision of an estimate of a total. In any case, in a probability sample, the standard errors are calculable.

The field-worker may thus create segments one by one until she has exhausted the area assigned to her. She assigns to each segment a serial number. She may canvass a sample of segments at once if instructed to do so. Random numbers in a sealed envelope, which she will use only after she has defined and numbered the segments, will draw 1 random segment from each block of *c* segments. The number *c* is the segment-interval. Thus, if a block had a size of 9 work-loads, then the segment-interval would be 9, whatever be the sizes of the

[4] In the use of the half-open interval, a segment begins with and includes a certain address such as No. 46 Varick Street, and extends up to but does not include some other address such as No. 64, which address would form the commencement of the next segment. It is only necessary for the field-worker to list each address that will form the commencement of a segment, and to accompany the list occasionally with some brief directions or with a rough map so that the content of every segment will be unmistakable, now or later. The half-open interval was invented but not named by F. F. Stephan in the article cited in footnote 3. The name comes from Deming, *op. cit.*, p. 82, and independently from Frank Yates, *Sampling Methods for Censuses and Surveys* (New York: Charles Griffin and Company, 1949), p. 67.

[5] Cf. footnote 1, Mahalanobis and Hansen, Hurwitz, and Madow. Cf. also a summary by W. Edwards Deming, "On the possible types of sampling unit in the last stage of selection in a probability sample," New York: Advertising Research Foundation, 1955.

segments, and the supply of random numbers might 1, 14, 22, 29, 45, 52. Random numbers beyond the number of segments actually created are blanks and draw no sample. All the dwelling units and all the people in the block will thus have the probability 1/9 of coming into the sample, and no weighting factor will be required.

The instructions must contain a special provision to use when the field-worker encounters conditions that would cause delay and run up the costs excessively were she to create segments in the whole area. There, the first step would be to divide the area into portions, ascribe to each portion a size in work-loads, and to create segments in only one portion which the supervisor will select by the use of random numbers. The segment-interval for the selection of the segments for interview in this portion must then be reduced in the ratio of the number of work-loads in the selected portion to the number of work-loads in the whole area.

There should be an instruction to the field-worker to halt her proceedings and to call or write for special advice if she has more segments than random numbers: unless her segments are abnormally small, this condition may indicate abnormal growth, which will require special treatment.

This procedure has the following advantages: (1) The field-worker need not adjust the sizes of the segments in order to produce exactly c, $2c$, or any other special number of segments within an area. Instead, the intended size of the segments may be whatever size appears to be best from the standpoint of definition, efficiency, and completeness of coverage. (2) If the random numbers select more than one segment from an area, the segments selected will be scattered over the area, and there will thus be some small gain from the stratification so enforced. (3) The creation of the segments, the selection of the sample, and the interviewing, may take place in one visit, except for recalls where there was nonresponse.

The selection of the segments may take place in the office, after the field-worker turns in her identification of the segments. This, of course, must be the procedure for subsequent surveys.

This procedure goes smoothly in the field, with simplicity, economy, and statistical efficiency. It is nevertheless entirely optional, and is not an essential part of the methods of this paper.

Complete freedom in the basic design. One not only has complete liberty in the basic design, but must exercise it. Thus, one may use any mode of stratification that he deems to be efficient, by specifying the order for listing the census areas in the main frame. One may, if he wishes, use intermediate stratification of a preliminary sample, followed by Neyman allocation. For Neyman or disproportionate allocation, one merely (e.g.) halves the width of the zoning interval over any stratum to double the number of work-loads per 1000 Cdu's. One may meet subsequently the demand for some other size of sample for any district within the whole either by altering the zoning interval in that district, or by deleting a random work-load in a fixed number, or by drawing a supplemental sample. An additional random number per zone will produce a 50 per cent increase.

In a sample of New York, where the zoning interval was 2000 work-loads, in order to produce a 25 per cent supplementation in the boroughs of Brooklyn and Queens, I drew 2 supplemental random numbers between 0001 and 8000 for every 4 zones throughout the two boroughs. The interval 8000 was convenient because this number was to be the size of the thick zone for tabulation (*vide infra*).

The intended sizes of the work-loads, and of the segments as well, need not be constant within any zone nor even within any area. Thus, in areas that are difficult to carve, or which will be costly for the field-workers to reach, one may cut costs and increase the over-all efficiency by deliberately doubling the number of Cdu's in a work-load. The zoning interval will still cut off the same number of work-loads, and there will be no change in the probabilities, nor in the procedure subsequent.

One may feel free to choose any formula that appears to be efficient for the estimation of any total, proportion, or other characteristic of the frame. Whatever be the choice of the form of the estimate, the computation of the standard error thereof will be rapid and valid, as may be evident from formulas and procedures that appear later.

Triplicate and quadruplicate drawings from each zone may at times be desirable, or even 10 drawings per zone (the Tukey plan). Some recommendations on multiple drawings appear later.

For a national sample of several thousand work-loads, one may easily adapt this procedure so as to draw work-loads in clusters, separated but not too far. This modification should receive consideration in areas where travel is very costly, or where there is a desire to place several work-loads under one supervisor. It is only necessary to introduce a "local frame" of (e.g.) 1000 or 2000 work-loads as the unit of size (the bead) in the initial frame; then to draw (e.g.) 4 distinct work-loads from each local frame that falls into the sample. The 4 work-loads will seldom be separated more than some predictable distance. The 4 work-loads may be forced by stratification to come from 4 different quarters of the local frame. Replication is obtained by drawing a 2nd local frame with replacement from the same zone. A local frame may come into the sample twice. If it does, then a work-load of 2 or more segments, but not the segments selected therein, may also come into the sample twice.[6]

It is not difficult to lay out a plan by which to measure the variance between interviewers, or the difference between two methods of training, or between two questionnaires, or both, by balanced random assignments in successive zones, but this topic does not require elaboration here.

For extra pressure on (e.g.) one-third of the nonresponses, one may select at random 1 work-load from every successive 3 work-loads in Sample 1, and likewise for Sample 2; then weight these results by 3 and add them to the initial results. Or, one may attempt some other type of estimation.[7]

Heavy spotty growth that has taken place here and there since the last census will cause the same trouble in this sampling procedure that it causes in any other, and it can be handled in the same manner.[8] This problem, like many others, is not within the scope attempted here.

Two-way stratification with forced selection of heterogeneous sampling units. One important variation in design is a 2-way stratification which forces areas of unlike characteristics to fall together into the sample, a device that in some kinds of studies has shown notable increases in efficiency. Thus, one may force areas of heavy industry to fall into the sample along with areas of light industry, urban areas to fall with rural areas, center with fringe, high rent with low rent, etc. The forcing is accomplished by the use of two frames, where we used only one heretofore. The two frames will cover some region of the domain of study (a province, or a region, or a city), but they will commence with census areas that are opposite in character. Thus, one frame might commence with the central part of the biggest cities of predominantly heavy industry, and the other frame might commence with open country sparsely settled.

In studies of public opinion in Germany, a separation of communities into two lists on the basis of religion was effective. One list commenced

[6] The theory for the optimum number of segments in a work-load, and for the optimum number of work-loads in a cluster, is in Hansen, Hurwitz, and Madow, *op. cit.*, vol. I, p. 291.

[7] W. Edwards Deming, "On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponse," *Journal of the American Statistical Association* 48 (1953), 743–72.

[8] Cf. Hansen, Hurwitz, and Madow, *op. cit.*, vol. I, p. 351.

with communities that were practically 100 per cent Protestant; the other list commenced with communities that were practically 100 per cent Roman Catholic. Both lists merged at the bottom with communities that were about equally divided.

The work-loads in both lists will commence with serial number 1. The two zoning intervals will be equal, and each list will contain the same number of work-loads, with blanks at the end if necessary to fill out the last zone. However, the average sizes of the work-loads need not be equal in the two lists, nor the segments; and in fact their average sizes may vary from one zone to another within the same list to accommodate variable costs (as mentioned earlier).

One random number z draws work-load number z from one list; also work-load number z from the other list. Segments drawn as prescribed from each of these two work-loads form the 1st sample. Together they form one sampling unit, which consists of a dumbbell, with segments of opposite characteristics at the two ends. The sampling units are thus made heterogeneous, and there is the additional benefit of stratification.

A 2nd random number in the same zone draws 2 more work-loads, one from each list; and segments drawn from these work-loads form the 2nd sample, another dumbbell.

One may form a separate estimate for any region covered by either of the two frames alone, or for any region covered by both frames.

Further gains will accrue from the use of Masuyama's zigzag interval,[9] whereby random number z between 1 and Z' draws also work-load number $Z'-z$, where $Z'=2Z$.

Thickening the zones for economy and for assistance in computation. For speed and economy in the tabulation and computation, and to gain increased validity of some of the formulas that we shall use for the variances, we shall combine for tabulation several successive initial zones to form a "thick" zone. The thickened populations will go into the formulas ahead. The thickened populations will be bigger and less variable, relatively, than the populations of the initial zones, and their sampling distributions will be more nearly normal than the thinner populations of the initial zones. This approach to normality will im-

[9] Motosaburo Masuyama, "Recent advances in sampling surveys in Japan," *Bulletin of the International Statistical Institute*, xxxiii, part II (1951), 147–52, p. 149 in particular. The use of a heterogeneous primary sampling unit under one supervisor has been basic since 1940 in the sample for the Monthly Report on the Labor Force, designed by Morris H. Hansen and colleagues in the Census in 1940; cf. Hansen, Hurwitz, and Madow, *op. cit.*, chapter 12. Another plan for the use of heterogeneous sampling units was published by Roe Goodman and Leslie Kish, "Controlled selection, a technique in probability sampling," *Journal of the American Statistical Association*, 45 (1950), 350–72.

prove the estimate of the variance of a ratio, and will also validate the use of the range later on.

In thickening the zones, one should try to meet the following requirements. In a small survey, where there may be only 20 or 30 zones initially, the two requirements may be competitive.

1. The thickness of a zone should be enough to yield a minimum y-population of 10, although 5 can be tolerated. (The y-population is the number of dwelling units or of interviews or of some other basic population in the denominator of a ratio x/y. In practice, the y-population will almost always be far above the minimum.) Thickness that will yield populations of 10 or more will permit use of the range for the standard error of an estimate of this population.
2. The number of thick zones should be enough to yield a useful estimate of the sampling error, but no more. In this way one holds to a minimum the costs of tabulation and of computation. In duplicate drawings, each zone yields one degree of freedom in the estimate of a standard error. Triplicate or quadruplicate drawings will give more degrees of freedom per unit of tabulation: *vide infra* "Multiple drawings per zone".

It is interesting to note that the thickening process retains the statistical efficiency of the initial zones.

2. THEORY

Procedure for computing the estimates and their standard errors. For illustrative purposes we shall deal with 2 drawings per zone. If we use the subscripts 1 and 2 for the two samples in Zone i, then the results of the interviewers may be summarized as

x_{i1}, x_{i2} for the two x-populations (e.g., the x-population might be the number of packages of a certain item of food bought last week)

y_{i1}, y_{i2} for the two y-populations (e.g., the y-population might be the number of families that bought food of any kind).

Usually we need estimates of:

A the total x-population in the entire main frame
B the total y-population in the entire main frame
ϕ the ratio A/B

For example, A might be the total number of packages of a certain item of food that the families in a region purchased during the past two weeks, while B is the total number of families in the region that bought food of any kind. The symbol ϕ denotes the ratio A/B, the average number of packages purchased per family over this time-interval for this particular region.

The sample will provide estimates of the results that would have been obtained from a complete census over the region, with the same questionnaire as was used in the sample, by the same interviewers, working under the same instructions and supervisors, during the same period of time. For an estimate of the ratio ϕ we may take

$$f = \frac{\text{the } x\text{-population in the sample}}{\text{the } y\text{-population in the sample}} = \frac{x}{y} \tag{1}$$

in which x is the total x-population in the entire sample, both 1st and 2nd samples combined. The symbol y has a similar definition for the total y-population in the sample.

For an estimate of the variance of f we first define and calculate for (thick) Zone i,

$$\left.\begin{aligned} D_{xi} &= x_{i1} - x_{i2} \\ D_{yi} &= y_{i1} - y_{i2} \end{aligned}\right\} \tag{2}$$

then calculate

$$h_i = D_{xi} - fD_{yi} \tag{3}$$

and h_i^2, whereupon we may estimate

$$\operatorname{Var} f = \left(1 - \frac{2}{Z}\right)\frac{1}{y^2}\sum_{i=1}^{m} h_i^2 \tag{4}$$

The summation runs over the m thick zones. Z is the number of work-loads per zone, and the factor $1-2/Z$ is the usual finite multiplier for the reduction in variance owing to nonreplacement when 2 work-loads are drawn per zone. In practice, $1-2/Z$ is usually replaceable by 1.

Equation 4 is a simple adaptation of the usual approximate formula

$$C_f^2 = \left(1 - \frac{2}{Z}\right)\frac{1}{2m}(C_x^2 + C_y^2 - 2\rho C_x C_y) \tag{5}$$

for the square of the coefficient of variation of the ratio f. When there are 2 drawings per zone, we may estimate the quantity in parenthesis by the summation.[10]

$$\frac{1}{2m}\sum_{i=1}^{m}\left[\frac{x_{i1}-x_{i2}}{\bar{x}}-\frac{y_{i1}-y_{i2}}{\bar{y}}\right]^2=\frac{1}{2m\bar{x}^2}\sum_{1}^{m}(D_{xi}-fD_{yi})^2$$

$$=\frac{2m}{x^2}\sum_{1}^{m}h_i^2 \qquad (6)$$

Herein $\bar{x}$ and $\bar{y}$ are the average x- and y-populations per thick zone per sample; wherefore $x=2m\bar{x}$, and $f=\bar{x}/\bar{y}$. By definition

$$\hat{\sigma}_f=f\widehat{C}_f=(x/y)\widehat{C}_f \qquad (7)$$

whereupon (4) follows at once from (5) and (6).

Equation 4 achieves a drastic reduction in labor, when compared with any valid formula for a standard error in a multi-stage plan. The number of thick zones need not be large. In practice, from 10 to 20 serve well.

The above equations apply to any part of the frame over which the probability of selection remains constant. If the rate of sampling changes from one part of the frame to another, as in Neyman allocation of the sample, then each part of the frame requires a separate estimate, which the above equations will supply.

A simple numerical example of a ratio and its standard error. A survey of a small urban area gave the results[11] shown in Table 1. Five thick zones are of course too few to give a good estimate of a standard error, but they provide a simple illustration of the use of the formulas. Let the number of males be the x-population, and the males plus females the y-population. Then the over-all proportion male is

$$f=\frac{x}{y}=\frac{69+50}{127+115}=.49 \qquad (8)$$

The values of h_i^2 are in Table 2. In this particular survey, Z was 8; hence (4) gives

[10] Hansen, Hurwitz, and Madow, *op. cit.*, ch. 4C and p. 194. William G. Cochran, *Sampling Techniques* (New York: John Wiley and Sons, 1953), pp. 115–18. P. V. Sukhatme, *Sampling Theory of Surveys with Applications* (Ames: Iowa State College, and the Indian Society of Agricultural Statistics, 1954), pp. 139–46. Deming, *op. cit.*, ch. 5.

[11] I am indebted to Josephine D. Cunningham for the results of the survey (an experimental one conducted in one of my classes at New York University) and to Edith Del Peschio for compiling the populations by thick zone in Table 1.

$$\operatorname{Var} f = \left(1 - \frac{2}{8}\right)\frac{36.2}{(127 + 115)^2} \tag{9}$$
$$= .000464$$

TABLE 1

THE NUMBER OF MALES AND OF FEMALES BY THICK ZONE IN A SURVEY OF AN URBAN AREA

THICK ZONE	MALE AND FEMALE			MALE			FEMALE		
	Sample		D_i	Sample		D_i	Sample		D_i
	1	2		1	2		1	2	
$i=1$	21	29	− 8	13	13	0	8	16	−8
2	30	23	7	15	9	6	15	14	1
3	14	11	3	9	5	4	5	6	−1
4	48	26	22	25	12	13	23	14	9
5	14	26	−12	7	11	−4	7	15	−8
Sum	127	115	12	69	50	19	58	65	−7
*$\overline{D}$	*xxx*	*xxx*	10.4	*xxx*	*xxx*	5.4	*xxx*	*xxx*	5.4

* Computed without regard to the sign of D_i.

TABLE 2

CALCULATION OF h_i AND OF h_i^2

$f = x/y = (69+50)/(127+115) = .49$

THICK ZONE	FROM TABLE 1		$h_i = D_{xi} - fD_{yi}$	h_i^2
	D_{xi}	D_{yi}		
$i=1$	0	−8	3.9	15.4
2	6	7	2.5	6.2
3	4	3	2.5	6.2
4	13	22	2.2	4.8
5	−4	−12	1.9	3.6
Sum	*xxx*	*xxx*	*xxx*	36.2

whence

$$\hat{\sigma}_f = \sqrt{.000464} = .022 \tag{10}$$

Variance of a direct estimate. The sampling procedure furnishes a direct estimate, rarely useful, for the total of any population in the frame. Thus, one might wish to estimate the number of dwelling units in the entire frame, which might cover a city or the whole country. The direct estimate X of the total x-population A in the entire frame will be

$$X = \tfrac{1}{2}Zx \tag{11}$$

Z being, as before, the number of work-loads per zone. We may then use the estimate

$$\text{Var } X = \frac{1}{4}\left(1 - \frac{2}{Z}\right)Z^2 \sum_1^m D_{xi}^2 \tag{12}$$

or its equivalent

$$\hat{C}_X = \frac{\sqrt{\text{Var } X}}{X} = \sqrt{1 - \frac{2}{Z}}\,\frac{\sqrt{\sum_1^m D_{xi}^2}}{x} \tag{13}$$

This estimate is valid regardless of the sizes of the x-populations and regardless of their sampling-distributions. One could in fact use thin zones for this estimate, even if some of the x-populations were 0.

Var X is simple to calculate by the above equation, but we may, under suitable conditions, reduce the labor even further by use of the range, which does not require the sum of squares. We merely calculate $\overline{D}_x$, the average of the m values of D_{xi} taken without regard to sign, and substitute it into the formula

$$\hat{\sigma}_X = \sqrt{1 - \frac{2}{Z}}\sqrt{\frac{1}{2}\,m}\,\frac{Z\overline{D}_x}{1.13} \tag{14}$$

Expressed as a coefficient of variation this is

$$\hat{C}_X = \sqrt{1 - \frac{2}{Z}}\,\frac{\sqrt{2m}\;\overline{D}_x}{1.13x} \tag{15}$$

The factor 1.13 is the Tippett constant d_2 for samples of 2. One may calculate the numerical factors in these equations once and use them for all the standard errors in the survey: the factors are all constant except $\overline{D}_x$ and x. These equations will give almost the same result as the previous one for Var X, even when the sampling distribution of the x-population by thick zone departs considerably from the normal.

For a numerical illustration of a direct estimate and of its standard error we may turn back to Table 1 and compute from (11) that

$$X = \tfrac{1}{2}Zx = \tfrac{1}{2} \times 8(69 + 50) = 476 \tag{16}$$

for the number of males in the area. Table 2 shows the differences D_{xi}. Their sum of squares is 237, whence (13) gives

$$\widehat{C}_X = \sqrt{1 - \frac{2}{8}}\,\frac{\sqrt{237}}{69 + 50} = 11\% \tag{17}$$

for the coefficient of variation. Or, we may turn to Table 1 for $\overline{D}_x = 5.4$, with which (15) gives

$$\widehat{C}_X = \sqrt{1 - \frac{2}{8}}\,\frac{\sqrt{10} \times 5.4}{(69 + 50)1.13} = 11\% \tag{18}$$

A further illustration of a direct estimate and its standard error occurs later in the sample of Cincinnati.

Multiple drawings per zone. Triplicate drawings in each zone with an initial zone 50 per cent wider than the zone required for duplicate drawings will produce the same size of sample but will give a better estimate of the standard error, because there will be 33 per cent more degrees of freedom. Quadruple drawings with a zone twice as wide as the zone required for the duplicate drawings will yield 50 per cent more degrees of freedom (*vide* the table at the end of this section). However, the wider zones may under some conditions suffer a slight loss in the efficiency of the estimates of f and of X.

When there are k drawings per zone, instead of 2, equation (11) for an estimate of a total will contain the factor $1/k$ in place of the factor $\tfrac{1}{2}$. Equations 4 and 12 for the standard errors will also require modification. Thus, in place of (4) we now write

$$\operatorname{Var} f = \left(1 - \frac{k}{Z}\right)\frac{k}{y^2}\,\frac{1}{k - 1}\sum_{i=1}^{m}\sum_{j=1}^{k}\left[(x_{ij} - \bar{x}_i) - f(y_{ij} - \bar{y}_i)\right]^2 \tag{19}$$

wherein $\bar{x}$ and $\bar{y}$ have the same meanings as they did earlier, and $\bar{x}_i$ and $\bar{y}_i$ are the average populations per sample in thick Zone i.

In place of (12) we write

$$\text{Var } X = \left(1 - \frac{k}{Z}\right)\frac{Z^2}{k(k-1)}\sum_{i=1}^{m}\sum_{j=1}^{k}(x_{ij} - \bar{x}_i)^2 \tag{20}$$

These two equations take the simple forms of (4) and (12) when there are only 2 drawings per zone ($k=2$).

On the other hand, calculations with (14) or (15) go as rapidly as with 2 drawings per zone, as these equations require no modification at all except care to use D_{xi} as the range between all kx-populations in Zone i, and the appropriate change in the Tippett constant d_2, which will be 1.69 for a triple drawing, 2.06 for a quadruple drawing, and 3.08 for 10 drawings.[12]

A general plan for the variance of any estimate.[13] For an estimate other than a ratio (x/y) or a total (X) we need a more general theorem. If u_1 and u_2 are the two estimates of any characteristic, u_1 obtained from one random half of the full sample, u_2 from the remaining half, and u from both halves combined, then

$$\text{Var } u_1 = 2 \text{ Var } u \tag{21}$$

and it is a fact that

$$\text{Var } u = \text{Var }(u_1 - u) \tag{22}$$

very nearly. Hence, to estimate Var u one need only draw 1 of the 2 work-loads at random from every zone, and compute the estimate u_1 therefrom and $(u_1-u)^2$; then to repeat the procedure by drawing another random half, then another, and another. Sample 1 constitutes one random half; Sample 2 constitutes another, but yields no new information. Subsequent halves thus require random drawings from the full sample already at hand. The successive values of $(u_1-u)^2$ are correlated, but the cumulated average value of $(u_1-u)^2$ will soon settle down to some number which we may accept as an estimate of Var u. The number of degrees of freedom in this estimated variance will be m, the number of thick zones.

More generally, when there are k work-loads per zone, we may draw at random 1 work-load per zone and use the accumulated average of $(u_1-u)^2$ as $(k-1)$ Var u. The number of degrees of freedom will be $m(k-1)$.

[12] Cf. the table in Deming, *op. cit.*, p. 570.

[13] I am indebted to my colleagues William N. Hurwitz and Max Bershad of the Census in Washington for the privilege of publishing the method described in this section.

Use of the range for a general estimate. In place of the general method just described, one may often find a much simpler solution through use of the range. Let u_{i1} and u_{i2} be the two estimates per zone of some characteristic, and let u be the estimate obtained from the entire sample. Then the formula

$$\hat{C}_u = \frac{\overline{D}_u}{1.13u\sqrt{2m}} \tag{23}$$

will be satisfactory for the coefficient of variation of u, provided the (thickened) populations that enter into the calculation of u_{i1} and u_{i2} are all 10 or bigger. This formula will serve for a ratio, but the use of (4) is safer against small y-populations, and is easy to use.

The Tukey plan. There is a special convenience in some types of problems in the use of 10 drawings per zone, to form 10 interpenetrating subsamples, which extend over the entire area sampled, considered as one thick zone. Tabulations for each of the 10 subsamples will give 10 separate estimates of any characteristic, be it a ratio or a total or anything else. The 10 subsamples pooled give the final estimate u of the characteristic, and the variance between the 10 separate estimates $u_i (i=1, 2, \ldots, 10)$ furnishes an estimate of the variance of u, with 9 degrees of freedom. This plan goes under the name of the Tukey plan,[14] and it was used first in 1949.[15] Systematic subsamples are especially convenient in the Tukey plan, but it is better to draw fresh random numbers in each zone if there is any possibility of a loss in efficiency through serial correlation from zone to zone.

A quick mental calculation of the standard error in the Tukey plan with 10 subsamples may be had by using the following estimates:

$$\hat{\sigma}_u = \frac{D_u}{9.8} \doteq \frac{D_u}{10} \tag{24}$$

$$\hat{C}_u = \frac{D_u}{9.8u} \doteq \frac{D_u}{10u} \tag{25}$$

D_u is simply the range or the arithmetic difference between the highest and the lowest of the 10 estimates $u_i (i=1, 2, \cdots, 10)$. The factor 9.8 is the product of the Tippett constant $d_2=3.08$ for a sample of size 10, multiplied by $\sqrt{10}$. For a better estimate of the variance one calculates

[14] Described in Deming, *op. cit.*, pp. 99 and 353.

[15] W. Edwards Deming, "On the sampling of physical materials," *Revue de l'Institut International de Statistique*, 1950, 1–23.

$$\operatorname{Var} u = \frac{1}{10 \times 9} \sum_{1}^{10} (u_i - u)^2 \tag{26}$$

A still quicker solution to the margin of sampling error in u is the simple observation made by Mahalanobis that the probability is only $(\frac{1}{2})^{10} = 1/1024$ that the maximum of the 10 estimates u_i lies below the median of all the estimates that are possible by repetition of the same sampling procedure. There is an equal probability that the minimum of the 10 estimates lies above the median.

One may use one set of interviewers in a random 5 of the 10 subsamples, and another set of interviewers in the remaining 5 subsamples. There are then 8 degrees of freedom for the pure sampling variance, 1 degree of freedom between sets of interviewers, and 9 in the total variance. One may of course randomize the interviewers in all or in a part of the area covered by either of the 5 subsamples to obtain more detailed information on the variance between interviewers.

Lahiri in his paper cited in the first footnote gives arguments for the use of only 4 or 5 interpenetrating subsamples in a large-scale survey whose results will be of general interest, and he presents a further argument for presentation of the results by subsample. Briefly, (a) rare characteristic have a better chance to appear with sufficient frequency to provide a good estimate in each subsample if there are not too many subsamples; (b) the degrees of freedom furnished by the subsamples will be more effective if the distribution of the estimates by subsamples is nearly normal; (c) standard errors estimated from 4 or 5 interpenetrating subsamples will be sufficiently precise for most uses; (d) if the results of the survey are presented by subsample, the user may estimate the standard error or any function that he may wish to calculate from the populations so presented; (e) the results of 4 or 5 subsamples will be far less bulky in publication than the results of 10 subsamples. Lahiri's paper should be read for its illumination of a host of other practical problems in the preparation and use of sampling surveys.

One may compare in the table below the number of degrees of free-

Number of thick zones	Replications per zone	Number of tabulations required	Degrees of freedom (df)	df:tab
10	2	20	10	.50
7	3	21	14	.67
5	4	20	15	.75
1 (Tukey plan)	k	k	$k-1$	$1-1/k$

dom in the estimates of the standard errors against the number of separate tabulations required for different plans. Decision on the best plan will depend on convenience, and partly on this comparison, with consideration of possible losses from wide strata.

To gain experience with a new material, one may in the first survey use duplicate drawings, and compare the average variance between the two samples in adjacent thin zones with the average variance within thin zones. If this ratio lies between 1 and 1.05, there would be but little loss in doubling the size of the thin zones and using quadruple drawings in the next survey of this material.

I may say, however, that in my own practice, I have used either just 2 drawings per zone, or 10 with one thick zone for tabulation (the Tukey plan).

In connection with rare characteristics, it is interesting to note from some recent work by Jones[16] that the interpenetrating subsamples will still give valid estimates and valid standard errors so long as two or more of the subsamples pick up members of the universe that possess the specified characteristics. Of course, if some populations are rare or absent in one or more subsamples, the interpretation of the standard error must be made with the aid of the proper theory for skewed distributions.

3. ILLUSTRATIVE EXAMPLE: A SAMPLE OF A DISTRICT

A. A Sample of the Cincinnati Area

Preliminary calculations. The district for the first illustration will be the Cincinnati area, composed of the cities of Cincinnati and Covington, and the remainder of the counties Hamilton and Kenton, plus Campbell County adjacent. The aim of the study was to compare readers' opinions on two special features of a certain newspaper. The general plan outlined earlier was broken into two parts, the 1st part for the "block cities",[17] and the 2nd part for the remainder of the area. The reason for the split is that the firm that carried out the study could carry out the 1st part in their own office, as the only materials required therefor were the block statistics published by the Census

[16] Howard L. Jones, "On the lower moments of the mean of a Tukey sample," Chicago Illinois Bell Telephone Co., 1955. "Investigating the properties of a sample mean by employing random subsample means," this issue.

[17] A "block city" is a city for which the Bureau of the Census publishes "block statistics." To be eligible, the city must have contained 50,000 or more inhabitants in the preceding census. A block in America is the smallest area that is bounded by streets. The block statistics show for every block city the number of occupied dwelling units and many other useful figures for every block that had occupied dwelling units in the last census. In this illustration, the block cities in the sample were Cincinnati and Covington.

for the two block cities in the survey, plus maps or directories as further aids, all of which one can purchase readily. In contrast, the 2nd part makes use of figures and maps that only the Census possesses.

The area to be surveyed contained 276,600 Cdu's (1950), of which 179,139 Cdu's or 64.8 per cent were in the block cities Cincinnati and Covington. The number of dwelling units in the sample was to be about 900. A work-load was defined as 10 Cdu's, and 2 work-loads were to be drawn from each zone. Thus, 45 zones should yield a sample of 900 Cdu's; or, owing to growth since 1950, something over 900 occupied dwelling units today. To decide the zoning interval, we note that 276,000 Cdu's will give 27,600 work-loads, and that 27,600/45=613. The zoning interval actually adopted in the face of possible growth since the Census of 1950, was 630 work-loads.

Instructions for the part in the block cities.[18] The following instructions apply to the block cities Cincinnati and Covington.

1. Prepare a list of the tracts in the order shown in the Census statistics for tracts, and show for each tract the number of work-loads therein. Form the accumulated total work-loads tract by tract. The accumulated totals will ascribe a serial number to every work-load in the two block cities (Table 3).

2. Draw 2 random numbers for each of the 29 zones, one for the 1st sample, and one for the 2nd sample. The zoning interval will be 630 work-loads. Record these numbers in two columns in the order drawn (Table 4). Every random number will identify by serial number a certain work-load for the sample; also the tract and the block in which it lies.

3. Identify the blocks that contain the sample of work-loads. To this end, for every tract that was struck by the random numbers:

   a. Prepare a list of the blocks therein, in the order shown by the Census block-statistics. Show for each block the number of Cdu's therein and assign to each block a number of work-loads.

   b. First, tie any block of less than 20 Cdu's to another adjacent, and assign a size to the pair.[19] It may occasionally be necessary to tie 3 or more blocks together. Tie likewise blocks that had size 0, as they may now be occupied (*vide* Table 5 for an example).

   c. Force the total number of work-loads accumulated for the blocks within a tract to agree with the size ascribed to that tract in Step 1. Do the forcing by adding or subtracting a work-load from the biggest block or combination.

[18] These are the actual instructions to the firm. They are printed here only for illustration of the theory and principles, and not as patterns suitable without modification for other surveys.

[19] Step 3b provided enough segments for a second survey, to be taken a few months later. This step may be omitted if there will be no further surveys in the area.

TABLE 3

SERIAL NUMBERS OF THE WORK-LOADS BY TRACT IN THE BLOCK CITIES CINCINNATI AND COVINGTON

City	Tract	CDU's (1950)	Number of Work-Loads	Serial Numbers of the Work-Loads
Cincinnati	1	2019	202	1– 202
	2	2334	233	203– 435
	3	2729	273	436– 708
	4	2861	286	709– 994
	5	4577	456	995– 1450
	6	1461	146	1451– 1596
	7	1705	170	1597– 1766
	8	1473	147	1767– 1913
	9	2724	272	1914– 2185
	10	2221	222	2186– 2407
	.			
	.			
	.			
	107	750	75	15754–15828
	108	175	26	15829–15854
	109	87		
	110	486	49	15855–15903
	Ward			
Covington	1	2316	232	15904–16135
	2	1829	183	16136–16318
	3	2080	208	16319–16526
	4	2066	207	16527–16733
	5	6222	622	16734–17355
	6	5508	551	17356–17906
		*Blanks	364	17907–18270

* Blanks added to fill out the zone, to keep the zones for the block cities clean of the surrounding area, for ease in tabulation.

4. Draw a random number between 1 and the total number of work-loads in this tract, to locate the block or combination in which the work-load falls. Make a list of these blocks or combinations, and show how many work-loads were ascribed to each one in Step 3c.

In a big tract, it may save time to form groups of 5 successive blocks, and to show the detail block by block only for the group struck by the random number.

TABLE 4

THE SAMPLE OF LOCAL FRAMES IN THE BLOCK CITIES CINCINNATI AND COVINGTON, DRAWN BY USE OF KENDALL AND SMITH'S *Random Numbers*, 4TH THOUSAND, COLS. 5, 6, 7, LINE 15. EACH RANDOM NUMBER FROM THE BOOK LIES BETWEEN 000 AND 629, AND IS TRANSLATED BY ADDITION TO THE LEFT-HAND BOUNDARY OF THE ZONE

Zone (Thin)	Boundaries of the Zones	Random Drawings			
		From the book		Translated	
		Sample 1	Sample 2	Sample 1	Sample 2
1	00001–00630	401	211	402	212
2	00631–01260	122	087	753	718
3	01261–01890	222	559	1483	1820
4	01891–02520	457	424	2348	2315
5	02521–03150	211	012	2732	2533
etc.					

5. In each block or combination, create segments by the prescribed rules.[20]

6. Draw one segment at random from every successive *c* segments, and canvass it. (*c* is the number of work-loads in the block.) This the field-worker may do on the spot, immediately after she finishes the job of delineating the segments.

Tables 3 and 4 illustrate the above steps.

Instructions for the area outside the block cities. This part of the work was carried out by the Census, down through Step 10. The instructions to the Census follow (numbered continuously from the previous steps, for convenience in reference).

7. Prepare a list of the enumeration districts (hereafter E.D.'s.) in the three counties outside the block cities, in any order convenient. Show for each E.D. the number of work-loads therein, and form the accumulated totals. Tie an E.D. of less than 40 work-loads to an E.D. that is nearby on the Census list, and ascribe a number of work-loads to the combination. It may occasionally be necessary to tie 3 E.D.'s. in one combination. The accumulated totals will ascribe a serial number to every work-load.

8. In areas where there have been special censuses since 1950, without change of boundary, the Census will use the new figures.

[20] Cf. the section in Part I entitled, "Suggested field-procedure for the selection of segments."

TABLE 5

RANDOM NUMBER 1820 IN TABLE 4 STRUCK TRACT 8. THIS TABLE SHOWS THE BLOCKS OF TRACT 8, AND THE SERIAL NUMBERS OF THE 147 WORK-LOADS THEREIN. A RANDOM DRAWING BETWEEN 1 AND 147 SELECTED A WORK-LOAD FROM THE COMBINATION OF BLOCKS 5 AND 6 FOR THE SAMPLE

TRACT	BLOCK	CDU's	NUMBER OF WORK-LOADS	SERIAL NUMBERS OF THE WORK-LOADS
8	1	24	10	1767–1776
	2	78		
	3	101	10	1777–1786
	4	121	12	1787–1798
	5	77	9	1799–1807
	6	10		
	7	85	8	1808–1815
	8	118	12	1816–1827
	9	91	9	1828–1836
	10	23	6	1837–1842
	11	40		
	12	114	11	1843–1853
	13	55	6	1854–1859
	14	98	10	1860–1869
	15, 16	0		
	17	99	10	1870–1879
	18	61	6	1880–1885
	19	96	10	1886–1895
	20	0		
	21	26	5	1896–1900
	23	24		
	22	132	13	1901–1913
Total		1473	147	

9. Draw 2 random numbers for each of the 16 zones outside the block cities, one for the 1st sample, and one for the 2nd sample. The zoning interval will be 630 work-loads. Record these numbers in 2 columns in the order drawn. Every random number will identify by serial number a certain work-load; also the E.D. (or a combination of 2 or of 3 E.D.'s) in which this work load falls. The Census will furnish a map and a description of each of these E.D.'s. or combination, together with the figure that in Step 7 prescribed the number of work-loads therein.

10. As an alternative to Step 7 the Census may, for economy, use 2 stages. The first stage might be to list groups of 5 or 10 E.D's.

and to show for each group the number of work-loads. Within any group that is struck by a random number, it will then be necessary to ascribe a number of work-loads to every E.D. therein and then to force the total number of work-loads for the group to agree with the original measure of size assigned to it. Do the forcing by adding or subtracting a work-load from the biggest E.D. or combination.

11–12. The firm will now delineate segments in the E.D's. in which the work-loads fell, and will draw segments by random numbers for canvass. The work follows Steps 5 and 6 and need not be written out here.

B. Numerical Calculations for the Cincinnati Area

Some numerical results for Cincinnati. Table 3 shows the number of Cdu's in the tracts of Cincinnati and of Covington, the number of work-loads ascribed to each tract, and the serial numbers of these work-loads.

Table 4 shows the random numbers for these cities (Step 2). The work-loads that bear these serial numbers belong to the sample. There was a similar set of random numbers for the area outside the block cities (Step 9).

Comparison with Table 3 shows which tracts contained the work-loads in the sample. For example, work-load number 1820 lies in tract number 8. Next comes Step 3. One may turn to Hansen, Hurwitz, and Madow, *op. cit.*, pp. 248–52, to see an example that is similar; nevertheless, I give in Table 5 the detail in Tract 8, for the convenience of the reader. The sum of the work-loads block by block turned out to be 147, without any forcing. In Step 4 we draw a random number between 1 and 147: this random number turned out to be 33, which drew work-load number $1799=1766+33$, which draws the combination of blocks 5 and 6. Steps 5 and 6 need no description here.

The results for several characteristics for the two samples over the entire Cincinnati area are in Table 6. From this table we see that the total number of households in the sample is $(449+454)$, which substituted into (11) gives

$$X = \tfrac{1}{2} \times 630(449 + 454) = 284{,}445 \qquad (27)$$

for an estimate of the total number of households in the area in June 1954. This figure compares with the number 276,000 in the census of 1950. For the standard error of this estimate, we note that the average

TABLE 6

SUMMARY FOR THE 2 SAMPLES OVER THE ENTIRE CINCINNATI AREA

CHARACTERISTIC	SAMPLE 1	SAMPLE 2
Number of dwelling units encountered*	449	454
Number of individual persons		
Encountered†	452	460
Refused	41	42
On vacation	31	19
Not found;‡ deaf, sick	34	37
Households not qualified (do not receive the particular newspaper studied)	104	98
Households qualified	242	264

* This is the x-population used in (27).

† These are the interviews attempted. The rule was to interview all the people in families of 1 person; 1 at random from each family of 2 persons; 1 at random from every 3 persons in families of 3 or more persons. The final results for any household were then weighted inversely by the number of persons interviewed therein.

‡ Not found in 6 recalls.

difference $\overline{D}_x$ from Table 7 is 7.9, whence (15) gives

$$\hat{C}_X = \frac{\sqrt{2 \times 9} \times 7.9}{1.13(449 + 454)} = 3.3\% \tag{28}$$

for the coefficient of variation of the estimate X. Equation 13 gives 3.2 per cent. This agreement between (15) and (13) was predicted and is typical. The standard errors of any other characteristics are computable likewise, but I shall not show a further example here, except to add that the time required for the computation of 6 standard errors by (15) was about 15 minutes with a slide rule, once the figures by thick zones came to hand.

There was no randomization of interviewers here; hence (as in most surveys) the differences between the 1st and 2nd samples in Tables 6 and 7, and also the standard errors computed above, are not purely errors of sampling, but are a combination of the uncertainty introduced by sampling and of the differences between interviewers. If the same interviewer were assigned to both segments in any zone, then the standard errors as calculated would measure the sampling errors, plus the random component of response, not confounded with the differences between the interviewers. If 2 interviewers were assigned at

TABLE 7

TOTAL DWELLING UNITS BY THICK ZONES BY SUBSAMPLE, IN THE SURVEY OF CINCINNATI. THE DIFFERENCES AND THEIR SQUARES ARE FOR USE IN (15) AND (13) FOR THE STANDARD ERRORS

THICK ZONE	SAMPLE 1	SAMPLE 2	D_i	D_i^2
1– 5	51	45	6	36
6–10	55	40	15	225
11–15	50	47	3	9
16–20	37	56	−19	361
21–25	49	60	−11	121
26–30	40	42	−2	4
31–35	52	48	4	16
36–40	46	52	−6	36
41–45	69	64	5	25
Total	449	454	71*	833
Average per thick zone	49.9	50.4	7.9*	92.6

* Computed without regard to the sign of D_i.

random in 2 zones, it would be possible to compute the variance between interviewers as well as the random errors.[21]

Comparison with the Poisson variance. It is interesting to note that the precision delivered by the sampling plan here for the estimate of the number of dwelling units in the Cincinnati area is remarkably close to the precision that would have arisen from work-loads whose sizes followed a Poisson distribution. Thus, from Table 6, the average size of a segment in actual dwelling units encountered was $\frac{1}{2}(449+454)/45=10$. If the distribution of sizes were Poisson-like, the variance of this distribution would also be 10, and its coefficient of variation would be $1/\sqrt{10}$. The coefficient of variation of the mean of a sample of 90 segments would then be $1/\sqrt{900}$ or 3.3 per cent, in remarkable agreement with (28). The interpretation is that the sizes of the segments turned out to have a distribution approximately like a Poisson variate, partly as a result of the rule that clear and definite boundaries of a segment take precedence over equality of size, and partly as a result

[21] This is the fundamental idea in Mahalanobis's system of interpenetrating samples; see the references to his works in the first footnote. See also Hansen, Hurwitz, and Madow, *op. cit.*, vol. II, ch. 12, sec. 3. A completely orthogonal application occurs in the author's paper, "On the sampling of physical materials," *Revue de l'Institut International de Statistique*, 1950, pp. 1–23.

of uneven changes in population since the last census.[22] Under such conditions the equation

$$C_X = \frac{1}{\sqrt{x}} \tag{29}$$

sets a limit to the precision attainable in an estimate (X) of a total. For any characteristic that shows door-to-door correlation, the variance will of course be higher. With special effort to acquire recent figures on growth and changes in the populations of areas, it may be possible to create work-loads of more uniform size, and in recurring surveys it may pay to do so, especially if an estimate of a total is important. An estimate (f) of a proportion is usually not sensitive to the variation in the size of the work-load.

4. A NATIONAL SAMPLE

Preliminary calculations. In connection with another project, the number of work-loads in a national sample was to be about 500, and the number of zones about 250. The number of Cdu's, in the entire country from the Census of 1950 is about 43,000,000. A convenient zoning interval was 17,500, derived from 4,300,000/250 = 17,200. Because of growth since 1950, the yield in dwelling units of such a sample, with an average of 10 dwelling units per work-load, was well over 5000.

Instructions for the block cities.[23]

13. Prepare a list to show the block cities by geographic region (e.g., New England). List the cities in geographic order within each region, east to west, south to north within size-groups (over a million inhabitants, 250,000 to a million, under 250,000).[24] Opposite each city show the number of work-loads therein. The accumulated totals will give a serial number to every work-load. Continue the accumulation from one region to another.

14. Draw 2 random numbers in every zone. (The rest of this paragraph follows Step 2.)

15. It remains now to identify the sample blocks. To this end, for

[22] My friend F. F. Stephan of Princeton kindly pointed out this useful and interesting observation.

[23] For convenience in reference, the numbers assigned to the steps in this section are continuous with the previous steps.

[24] This order will give good statistical efficiency for most purposes, and it will facilitate tabulations by size of city. One may specify some other order for listing the cities if he prefers. The stratification, like many other parts of these instructions, should be altered to achieve the best statistical efficiency for the purposes at hand.

every city that was struck by the random numbers, identify the page in the statistics for tracts. This can be done by accumulating the number of work-loads for each page or for a whole group of pages for any city. First force the total number of work-loads assigned to this city in Step 13.

16. Identify the particular tract on the page that was identified in the preceding step. First force the total number of work-loads accumulated by tract to agree with the number assigned to this page.

17. Identify the blocks that contain the sample of work-loads, as in Step 4.

18 and 19. Same as Steps 5 and 6.

Instructions for the area outside the block cities. As in the sample for Cincinnati, this part was carried out by the Census. The strata were those defined by the Census for the Monthly Report on the Labor Force.[25] The steps were otherwise similar to Steps 7–12 for Cincinnati.

5. SAMPLING WITHOUT STATISTICS FOR SMALL AREAS

A sample of the Federal District of Mexico.[26] This area showed 604,000 dwelling units in the Census of 1950. There are totals for each section, but not for the blocks within the sections. The feature of special interest here, different from the previous illustration, is that within any section we assign to each block the average number of dwelling units.

Another feature is that the city has been growing at the rate of 6 per cent annually; hence, it seemed desirable to increase arbitrarily by 50 per cent the sizes of the blocks of some of the outlying sections wherein the growth appears to be most prominent. This decision illustrates the fact that the number of Cdu's in an area is not necessarily a figure published by the Census; it may instead represent one's best information.

Lack of census information for small areas does not alter the probability of selection nor introduce bias; it only subtracts from the efficiency that would be possible otherwise, and it adds a few problems to the fieldwork, owing to the fact that the work-loads assigned for interview will usually be more variable.

This study was to be an investigation on the cost of living. An interviewer could cover an average of 5 or 6 families in one day. The possible adjustments in the tasks of the interviewers that sometimes arise from

[25] Had there not been already in existence a suitable mode of stratification, it would have been necessary to prescribe one.

[26] I am indebted to Ana María Flores, Chief of the Department of Sampling in the Census of Mexico, for the data in Table 8 and for the privilege of working with her on this survey; likewise to José Nieto de Pascual in the same department.

inequalities in the sizes of the work-loads would not be serious here, because a small work-load in any area would likely be counterbalanced by a big one not far away. It appeared advisable, therefore, in the interest of simplicity of preparation and administration, to assign to each block within a section a designated number of work-loads, this number so chosen that it produced an average size of 5 dwelling units per work-load over the whole section. The plan for Mexico thus conforms otherwise to the plan for Cincinnati. Table 8 shows the start of the assignment of serial numbers to the work-loads in Estrata 1, a portion of Mexico City.

TABLE 8

THE COMMENCEMENT OF THE TABLE OF SERIAL NUMBERS FOR THE SAMPLE OF WORK-LOADS IN MEXICO CITY ESTRATA 1

Section	Number of blocks	Average number of Cdu's per block 1951	Number of work-loads		Serial numbers of these work-loads
			Per block	In the whole section	
1	4	140	28	112	0001–0112
2	11	127	25	275	0113–0387
3	16	126	25	400	0388–0787
4	17	100	20	340	0788–1127
5	8	145	29	232	1128–1359
.					.
.					.
.					.

One may also, in the absence of census data for small areas (such as by blocks), make a quick tour of a city or other area and estimate very roughly the number of dwelling units in each small area, then proceed according to the instructions for the survey of Cincinnati.