A Review of Recent Statistical Developments in Sampling and Sampling Surveys

By F. Yates, Sc.D.

[Read before the Royal Statistical Society, January 23rd, 1946, the President, the Rt. Hon. Lord Woolton, P.C., C.H., in the Chair]

1. *Introduction*

Sampling is not a new subject. My object in bringing it up for discussion to-night is twofold. Firstly, with the increase in economic planning, and the development of the social sciences, the need for economic and social censuses and surveys has greatly increased. Such surveys generally require some form of sampling for their efficient and speedy execution. Secondly, development of new statistical methods in agriculture and biology has led to developments in sampling theory which are relevant to all branches of sampling. While it is broadly true that no really new methods of selecting representative samples have been introduced in recent years, the theory underlying the various methods is now much better understood, and practical procedures are available for estimating the sampling errors of complicated as well as simple sampling methods.

Although I have devoted considerable space to discussion of methods of estimating sampling errors, I do not wish to imply that such estimates of error are always required before a critical analysis of the results of a survey can be undertaken—we may often be satisfied, from previous experience, or from the general behaviour of the results themselves, that adequate accuracy has been achieved. On the other hand, comparison of the efficiency of different sampling methods, and intelligent planning of future surveys, can only be made by detailed analysis of the various components of variation to which the material is subject, and a study of their effects on sampling errors.

I have not attempted to discuss this latter problem in detail in the present paper. What has been attempted is a comprehensive summary of the various sampling methods that are commonly employed, and their inter-relations, together with an outline of the appropriate methods of arriving at estimates of the quantities under survey, and of determining the errors to which these estimates are subject.

While the parts of the paper dealing with the estimation of sampling errors are necessarily somewhat technical, the remainder of the paper has, I hope, been written in reasonably non-technical language. Sampling, after all, is largely a matter of common-sense, and the common-sense approach has often resulted in more rapid progress in technique than has the more purely mathematical approach. What has been lacking in the purely common-sense approach is a method of estimating the efficiency of different sampling procedures. A synthesis is therefore required. It is this synthesis I have attempted in the present paper.

2. *Position in* 1924

In May 1924 the Bureau of the International Institute of Statistics appointed a commission for the purpose of studying the application of the representative method in statistics. The members of the commission were Professor Arthur Bowley, Professor Corrado Gini, Mr. Adolf Jensen, M. Lucien March, Professor Verrijn Stuart, and Professor Frantz Žižek. The commission presented its report in 1925 (Jensen, 1926). Appended to the report were a paper by Jensen on " The Representative Method in Practice," and a paper by Bowley on " The Measurement of the Precision Attained in Sampling," together with the two shorter papers by Stuart and March. The report and the attached papers give a good picture of the sampling methods then in use in economic and social surveys and in the analysis of census material, and of the statistical theory on which these methods were based.

The conclusions of the report are embodied in a set of resolutions which were adopted by the International Institute of Statistics. These are as follows:

The International Institute of Statistics

Considering that it is necessary in many cases to draw general conclusions based upon partial investigations owing to the impossibility of procuring a complete statistical material;

Considering that even in such cases where complete material is available it may be sufficient to work up a portion of the material, provided that this working up is done in a rational manner; and

Considering that the saving of labour, time and money which is possible by limiting the investigation to a portion of the material will often make it possible to make a much more extensive use of the information at hand and to enter far more deeply into the subject than is possible by a working up of the whole material;

I. With reference to the Resolution passed at the Session at Berlin in 1903, again calls attention to the very considerable advantages which can be obtained by applying the Representative Method under the following conditions:—

The results of a partial investigation should only be generalized provided that the sample used is in its nature sufficiently representative of the totality. In such respects the sample may be selected in different ways; the following two main cases, however, are to be distinguished:

(*A*) *Random Selection.* A number of units are selected in such a way that exact equality of chance of inclusion is the dominant rule. Then precision is related to the number included which should be large enough to render insignificant accidental deviations;

(*B*) *Purposive Selection.* A number of groups of units are selected which together yield nearly the same characteristics as the totality. In order to have any knowledge of the precision of the estimate it is necessary that sufficient groups should be included to allow the variation between the characteristics of the groups to be measured. But since the precision often depends to a great extent on the discretion used in making the selection, the following controls are recommended:—

1. The selection on the same principle should be made twice or more; after their comparison, the samples can be merged. (This recommendation is also applicable to the Random Selection);
2. In repeated observations, the relation between the part and the whole should from time to time be examined more minutely;

II. Recommends that the investigation should be so arranged wherever possible, as to allow of a mathematical statement of the precision of the results, and that with these results should be given an indication of the extent of the error to which they are liable;

III. Repeats the wish expressed in the Resolution of 1903, that in the report on the results of every representative investigation an explicit account in detail of the method of selecting the sample adopted should be given.

Two main features will strike the present-day reader: first, the considerable prominence given to the method of purposive selection, and second, the lack of any very clear conception of the possibility, except by the selection of units wholly at random, or by the inadequate procedure of sub-dividing the sample into two or more parts, of so designing sampling enquiries that the sampling errors should be capable of exact estimation from the results of the enquiry itself.

3. *Developments in biological and agricultural research*

At about this time a number of new developments were taking place in the statistical methods used in biological and agricultural research. Those connected with the design and analysis of agricultural field experiments have been of particular importance to sampling theory. In such experiments each treatment is repeated only a few times, and consequently the then customary procedure of estimating the error of each treatment mean from deviations of the yields of the individual plots from that mean led to very imprecise results. Moreover, such estimates were in any case usually invalidated by the fact that each replicate of the experiment was arranged in a compact block of plots on the ground, so as to eliminate fertility differences as far as possible.

The introduction of the analysis of variance technique by R. A. Fisher gave a convenient arithmetical procedure for pooling estimates of error from different treatments, and simultaneously eliminating variation due to blocks, or other features of the layout. Consideration of the theoretical basis of the analysis of variance led in turn to the introduction of the principle of randomization, thereby ensuring that the estimates of error should be valid whatever the nature of the variations in fertility.

The first published application of the analysis of variance (Fisher and Mackenzie, 1923) is of considerable interest, as it shows how the ideas developed. The potato experiment there analysed contained several systematic features of arrangement, which were not discussed or fully taken account of in the analysis.

Fuller consideration of the problems raised by the analysis of this and similar experiments led to a rapid refinement of the technique, and by 1926 the essentials of good experimental design and analysis were fully realized (Fisher, 1926). The necessity of certain elements of randomization had been recognized, and the analysis of variance, as then developed, provided a method of pooling of the estimates of error from the different treatments, of eliminating components of variability which did not affect the treatment comparisons, and of furnishing separate estimates of error for treatment comparisons which, because of features of the design, were of differing accuracy. Parallel development of the t and z tests enabled these estimates of error to be used as a basis for exact tests of significance.

The first steps in the application of the new technique of analysis to sampling problems were taken by Clapham (1929). Clapham was concerned with the problem of estimating the yields of experimental plots of cereals from a number of small sampling units cut from each plot. He used the analysis of variance to calculate the sampling errors to which sampling units of various types were subject. The method was subsequently tried out in practice by Clapham (1931), and its efficiency was further examined by the present author (Yates and Zacopanay, 1935). A similar technique was developed for potatoes by Wishart and Clapham (1929).

Clapham's work led the way in the development of sampling techniques applicable to many agricultural and biological problems, such as the estimation of the growth rate and chemical composition of crops, the degree of infestation of the soil and crops with diseases and insect pests, the bacterial content of soils and liquids, the sampling of heterogeneous materials for chemical analysis. In these investigations the sampling problems were approached from a new angle, and for the first time those concerned with drawing up sampling schemes made a general practice of estimating sampling errors from the results of their observations, both to ascertain whether the sampling actually undertaken was adequate for the purpose in hand, and to increase the efficiency of future sampling of the same type of material. Trial sampling schemes of various kinds were also tested out, to determine the most suitable type of sampling unit.

The principles on which this work was based are very simple. They can be summarized in the following terms (Yates, 1935):

(1) If bias is to be avoided, the selection of the samples must be determined by some process uninfluenced by the qualities of the objects sampled and free from any element of choice on the part of the observer.

(2) If a valid estimate of sampling error is to be available, each batch of material must be so sampled that two or more sampling units are obtained from it. These sampling units must be a random selection from the whole aggregate of sampling units that can be taken from the batch of material, and all the sampling units in the aggregate must be of approximately the same size and pattern, and must together comprise the whole of the batch of material.

The importance of the first condition was well known to those who drew up the report to the International Institute of Statistics in 1924. It was the methodology implicit in the second condition which constituted the new advance. Realization of the functions of strict processes of randomization in agricultural field experiments had led to a corresponding realization of its importance in providing a valid estimate of error in sampling. " At random " no longer meant " haphazard." Again, the analysis of variance, by making possible the pooling of estimates of error and the separation of components of error which were not homogeneous, enabled the number of independent sampling units taken from each batch of the material to be reduced to a small number, and so permitted the use of relatively complicated sampling schemes, often involving sampling in two or more stages.

In most of these early applications the material sampled was of such a nature that it could be divided into sampling units of approximately the same size and shape, and consequently formally simple sampling schemes could be adopted. When, however, the methods began to be applied to such problems as crop estimation, the estimation of timber resources, and surveys of economic conditions and practices on farms, where the natural units of the population under survey are of widely differing size, new difficulties arose, which were by no means fully resolved at the outbreak of the war (see, for example, Cochran (1939)). In particular, methods were required for dealing with differences in variability of different parts of the material, and of handling sampling errors of material in which the sampling units were of widely differing size.

During the war intensive use of sampling in economic and social surveys and in problems arising in operational research, such as bomb distributions, amount of damage in blitzed towns, etc., has led to a further study of method both here and in America, and many of the above difficulties have been resolved, though there has, as yet, been no proper codification of the new methods, similar to that which has existed for some years for methods suitable for agricultural and biological material. I have therefore attempted, in the succeeding sections of this paper, to give an outline of these developments and their applications to the problems arising in area sampling, and in social and economic surveys.

4. *Summary of sampling methods which provide estimates of sampling errors*

It will first be profitable to make a short list of the various sampling methods which are in common use in one or other of the various fields to which sampling is applied, and which are capable, in virtue of the random elements in the selection process, of furnishing exact estimates of sampling error.

All forms of sampling involve some choice of sampling unit. The sampling units chosen may be natural units of the material to be sampled, such as the individuals of a human population, or natural aggregates of such individuals, such as households, or they may be formed by arbitrary sub-division of the material, such as the areas formed by grid squares on a map. Sampling units need not necessarily all be of the same size, though if there is marked variation in size, estimates based on ratios or percentages are usually required, and this complicates the estimation of sampling errors.

(*a*) *Random sampling* (*without restrictions*)

This is the simplest form of sampling, often referred to, but rarely used in practice, since its place is taken by some form of quasi-random sampling (see Section 10).

In random sampling selection from the whole population of sampling units into which the material is divided is made by some strictly random process, such as numbering all the units and selecting the requisite number of numbers at random by drawing lots or by the aid of a table of random numbers.

(*b*) *Stratified sampling* (*random sampling from groups*)

In stratified sampling the whole of the material to be sampled is divided into groups or strata. The same proportion is then selected from each stratum by some process of strict random selection within each stratum.*

If the strata are so chosen that each forms a relatively homogeneous group, the accuracy of the sample will be considerably increased, since each stratum is represented in the correct proportion in the sample.

It is worth noting that if the total number of units falling in each stratum is already known, as is frequently the case from previous census material, there is no need to divide the whole population into strata before selecting the sample. A stratified sample can be constructed by selecting at random from the whole population, classifying the sampling units into the strata as they are selected, and rejecting any further units falling in a given stratum as soon as the quota for that stratum is obtained.

(*c*) *Sub-sampling*

Sampling may be performed in two or more stages, the whole population being divided into a number of large sampling units, each of which contains a number of smaller sampling units. A

* The term " stratified sampling " is also used to cover the case in which different proportions are taken from the various strata. This case is described in Section 9, under the term " variable sampling fraction." (See Mr. Kendall's contribution to the discussion.)

sample of the large sampling units may then be taken, and from each of the large sampling units so chosen a proportion of the smaller sampling units may be selected. Thus, for example, if a sample of all the inhabitants of a country is required, the country may first be divided into large sampling units consisting of towns and rural areas. A random selection from these towns and rural areas may then be made, and from each of the towns and rural areas selected a further random selection of individuals may be made.

Stratification and other sampling devices may be and often are used in conjunction with sub-sampling at any or all of the stages of the sampling.

(*d*) *Stratification for two or more factors*

Stratification can be carried out simultaneously for two or more factors. Thus, we might classify a population according to income group and also according to age and sex. The number of sub-strata will then be equal to the number of cells in the two or more way table, corresponding to the factors for which the stratification is carried out. A random selection of equal proportions from each of these sub-strata will be exactly equivalent to ordinary stratified sampling with the sub-stratum as the unit of stratification.

If the number of cells is large, the sampling procedure is likely to become very involved. Moreover, it frequently happens that although the marginal totals of the numbers of sampling units in the two or more way table of the factors are known from previous census material, the numbers in the separate cells are unknown. These difficulties can be overcome by constructing a sample which is so adjusted that the proportions in the sample agree with the marginal totals of each factor separately. Such a sample can be called a sample stratified for two or more factors without control of sub-strata. The same process of selection can be used as was suggested for constructing an ordinary stratified sample without actually sub-dividing the population. The number of rejections at the end of the process is larger the larger the number of factors for which stratification is required.

(*e*) *Balancing*

Control of a quantitative character—*e.g.*, income—may be obtained by stratification in groups corresponding to ranges of values of the quantitative character, but if the number of units to be included in the sample is small, and if control of other (non-quantitative) characters is also desired, it may be simpler to select a sample which is balanced for the quantitative factor. In a balanced sample the mean value of the balanced factor in the sample is equal to the mean of the factor in the whole population.

It is important, in selecting a balanced sample, that the process of selection is such that, apart from the restriction imposed by the balancing requirements (and any other restrictions due to stratification, etc., for other factors), the selection is equivalent to random selection.* This is the essential difference between a balanced sample and a purposively selected sample.

A procedure analogous to thaf suggested for the construction of a stratified sample without actually dividing the whole population into strata will effect this. A random sample is first selected (stratified for other factors if required). Further members are then selected by the same random process, the first member being compared with the first member of the original sample, the second with the second member and so on, the new member being substituted for the original member if balance is thereby improved.

Such a procedure is apt to become tedious if balance for a number of factors is attempted. In such cases the simultaneous choice of a number of alternative members from pairs selected as above will facilitate the process, though the exact conditions required for strict equivalence to a random sample have not yet been specified.

5. *Methods of forming estimates and calculating sampling errors*

All the sampling methods set out in Section 4 are governed by the condition that the chance of inclusion of any sampling unit or sub-unit in the sample is the same. Consequently, the

* There does not appear to be any procedure which will give exact equivalence to a restricted random sample whatever the form of the parent distribution: even if such a sample could be obtained estimates derived from it would in fact be subject to some elements of bias, though such bias is not likely to be of importance in practice.

estimate of the mean of any quantity (or proportion of any characteristic) in the parent population may be obtained directly by taking the mean of the corresponding quantity in the sample data. Similarly, estimates of the totals of the parent population may be formed by multiplying the sample totals by the reciprocal of the sampling fraction.

The estimation of the sampling errors to which these estimates are subject depends on the sampling procedure which has been followed. The estimation of the errors of qualitative data differs from that of quantitative data in that the variability of the proportion possessing a given characteristic in a random sample follows the binomial law of error, and therefore depends only on the proportion in the universe and the number in the sample. Consequently the standard deviation of a sampling unit does not require estimation. If, however, the sampling units are made up of numbers of individuals, and the proportion of individuals possessing a given characteristic is of interest, correlation between individuals in the same sampling unit will invalidate estimates of sampling error derived from the binomial distribution, and the standard deviation of a sampling unit must be estimated in the same manner as for quantitative data, scoring the individuals 1 if they possess the characteristic and 0 if they do not.

(*a*) *Random sampling* (*no restrictions*)

The standard deviation of a single sampling unit can be calculated in the ordinary manner and used to estimate the sampling error of the mean or total as required.

If y_r is the value of a variate y in unit r of the sample, $\bar{y}$ the mean and $S(y)$ the sum of y for all units of the sample, $\bar{Y}$ and $\Sigma(Y)$ the estimates of the mean and the total of the population, n the number in the sample, N the number in the population, f the sampling fraction, s the estimated standard deviation of a sampling unit, and S.E. denotes the estimated sampling standard error, we have

$$f = n/N$$
$$\bar{Y} = \bar{y}$$
$$\Sigma(Y) = \frac{1}{f}S(y)$$
$$s^2 = \frac{S(y_r - \bar{y})^2}{n-1} = \frac{S(y^2) - \bar{y}S(y)}{n-1}$$
$$\text{S.E. of } \bar{Y} = s\sqrt{\frac{1-f}{n}}$$
$$\text{S.E. of } \Sigma(Y) = \frac{ns}{f}\sqrt{\frac{1-f}{n}}$$

The factor $\sqrt{1-f}$ occurs in the estimates of the sampling standard errors because the population sampled is finite. This factor should not be included when testing the difference of the means of two sampled populations to see whether, for example, they are subject to different causal agents.

(*b*) *Stratified sampling*

In estimating the sampling error of a stratified sample the variability between the different strata must be eliminated from the estimate of the variance of a single sampling unit. As pointed out above, this can be effected by the analysis of variance procedure. If t is the number of strata, the analysis of variance can be set out as follows:

	Degrees of freedom	Sum of squares	Mean square
Between strata	$t-1$	A	
Within strata	$n-t$	B	s_t^2
Total	$n-1$	C	s^2

If S, S' and S'' indicate summation over the whole sample, a single (s'th) stratum, and the whole group of strata respectively, the between strata and total sums of squares are given by the formulæ:

$$A = S''\{\bar{y}_s S'(y_s)\} - \bar{y}S(y)$$
$$C = S(y^2) - \bar{y}S(y)$$

The sum of squares B within strata is then obtained by subtraction. It will be seen that the sum of squares within strata is the sum of the squares of the deviations from the strata means, and the total sum of squares is the sum of the squares of the deviations from the general mean.

The mean squares can then be obtained by division of the sums of squares by the corresponding degrees of freedom. The formulæ given for the standard errors of a random sample without restrictions hold with the substitution of s_i for s. The increase in efficiency due to the use of a stratified sample instead of a fully random sample is given by the ratio s^2/s_i^2.

If the variability of the different strata is very different, it is best to calculate the standard deviation of a single sampling unit, and thence the standard errors of the strata means and totals, for each stratum separately. The standard error of the estimate of the total of the whole population will then be given by the square root of the sum of the squares of the standard errors of the estimates of the totals of the separate strata. The differences between the means of different strata can likewise be tested by means of their separate standard errors (omitting, if inappropriate, the factor $\sqrt{1-f}$).

If the numbers of sampling units in some or all of the strata are too small for estimation of separate standard errors, the pooled estimate given by the ordinary analysis of variance will still give an estimate of error applicable to the mean or total of the whole population, even when the variability of the different strata is different. If the numbers in the different strata are unequal, some small adjustment is theoretically necessary, as can be seen by making separate estimates, but this can usually be ignored.

A useful method of dealing with variation in sampling error when the number of units sampled in each stratum is small, is the construction of an error graph. In such a graph the estimated standard deviation (of a single sampling unit) is plotted against the estimated mean of the corresponding stratum, or other suitable statistic. A curve can be fitted to the points so obtained (either graphically or by some more exact method) so as to give an improved estimate of the standard deviation associated with any value of the mean. This procedure is also of value when sampling from batches of material, if the number of sampling units taken from each batch is too small for precise determination of the sampling error. It was followed, for example, in the Wireworm Survey (Yates and Finney, 1942) in which 20 sampling units were taken from each field, and in which there was strong association between the sampling variability and the degree of infestation.

(*c*) *Sub-sampling*

In the case of sub-sampling there will be a separate sampling error corresponding to each stage of the sampling. As a simple example we may take the case of a two-stage sampling process, stratified in the first stage, in which the sampling units of the first stage (from which the sub-samples are drawn) are all of the same size.

If there are h sub-units in each main unit of which k are included in the sub-sample, the analysis of variance will be as follows:

		Degrees of freedom	Sum of squares	Mean square
Between main sampling unit means	Between strata	$t-1$	A	
	Within strata	$n-t$	B	s_i^2
	Total	$n-1$	C	
Sub-units	Between main units	$n-1$	kC	
	Within main units	$n(k-1)$	D	s_u^2
	Total	$nk-1$	E	

A complication arises in that each main unit contains k sampled sub-units. The numerical quantities entering into the first part of the analysis will therefore consist of the means (or totals) of k sub-units. Totals are generally more convenient for actual analysis, but the analysis of variance is best set out in terms of means, in which case all sums of squares in the first part derived from the totals must be divided by k^2. The total sum of squares of the first part of the analysis will enter into the second part as " between main units," but must here be multiplied by k to make it comparable with the other sums of squares in this part. (An alternative procedure is to divide all sums of squares of totals in the first part by k instead of k^2 so that both parts are directly comparable.)

The variability contributed by the second stage of sampling will be included in the overall estimate provided by the first part of the analysis. Consequently if the sampling fraction at the first stage is small, the sampling error can be derived directly from s_i^2 as in ordinary stratified sampling, and the second part of the analysis will then only be required if the efficiency of the sampling procedure requires review. The full expression for the sampling standard error of the estimate of the mean of all sub-units in the population is

$$\sqrt{\frac{s_i^2}{n}(1-f_1)+\frac{s_u^2}{nk}f_1(1-f_2)}$$

where f_1 and f_2 are the sampling fractions of the first and second stage respectively.

Knowing the values of s_i^2 and s_u^2, the effect on the sampling error of changes in the intensity of sampling at each stage can be evaluated. If the number of sub-sampling units taken from each main unit is altered from k to k', s_u^2 will remain unaltered, but s_i^2 will asume a new value given (except for errors of estimation) by

$$s_i'^2 = s_i^2 + \left(\frac{1}{k'}-\frac{1}{k}\right)s_u^2$$

The effect of varying the number of sub-sampling units in each main unit, or the number of main units in the whole sample, or combinations of the two, on the accuracy of the sampling process can thus be evaluated.

(*d*) *Stratification for two or more factors*

The analysis of variance procedure used in a stratified sample can be followed. If the sample is stratified for two or more factors without control of sub-strata, the variability due to the separate factors can only be eliminated by fitting constants for these factors in the manner developed for the analysis of variance of multiple classifications with unequal numbers in the different classes (Yates, 1934).*

If there is no control of sub-strata it is difficult to estimate differences in variability between the different strata, especially if there are changes in variability in the various strata of more than one factor. Nevertheless, an overall analysis of variance will give an estimate of error which can be applied to the means and totals for the whole population.

(*e*) *Balancing*

In order to make a proper estimate of the sampling error of a balanced sample, it is necessary to use the analytical procedure known as the analysis of covariance. This is exactly analogous to the analysis of variance procedure which would be used if the sample were not balanced.

The first step is to calculate the sums of squares for the control variate as well as for the main variate. At the same time a third column is inserted in the analysis of variance table in which are entered the sums of products of the control and main variate which correspond to the sums of squares of the control and main variate respectively in the other two columns. Thus, the total sum of products will be given by the formula $S(x-\bar{x})(y-\bar{y})$, corresponding to the two sums of squares $S(x-\bar{x})^2$ and $S(y-\bar{y})^2$. In the formulæ given above squares are therefore replaced by products, and products of means and corresponding totals by products of means of the control variate and totals of the main variate (or vice versa).

Thus in a balanced stratified sample the sums of squares and products would appear as follows:

	Degrees of freedom	Sx^2	Sxy	Sy^2
Between strata	$t-1$	A'	A''	A
Within strata	$n-t$	B'	B''	B
Total	$n-1$	C'	C''	C

where, for example, $A'' = S''\{\bar{x}_s S'(y_s)\} - xS(y)$.

A corrected within strata mean square is now obtained from the formula

$$s_i'^2 = \frac{B - B''^2/B'}{n-t-1}$$

All the sampling errors are calculated using this corrected mean square.

* As I indicated at the meeting, this problem requires further investigation. I hope to publish a note on the matter shortly.

It may be noted that $B - B''^2/B'$ or $B - bB''$ is the sum of squares of deviations about the regression line (or in the case of a stratified sample a series of parallel regression lines) of the main variate on the control variate, the regression coefficient being given by

$$b = B''/B'$$

In the calculation of s_i^2 the number of degrees of freedom is reduced by one to allow for the estimation of the regression coefficient from the data. The reason why it is unnecessary to estimate the regression coefficient when estimating the population mean or total, is because balance automatically ensures that the adjustment for the regression is zero, whatever the value of the coefficient.

When two or more variates are balanced, the same general procedure holds, but it is now necessary to estimate the sum of squares of deviations from the values given by a partial regression equation. This involves the solution of simultaneous linear equations, the procedure for which is described by Fisher in *Statistical Methods for Research Workers* and elsewhere.

6. *Methods of improving the accuracy of a sample enquiry by adjustment of the results*

In all the methods of sampling outlined above the estimates of the population means and totals are obtained directly from the sample means and totals. In certain circumstances, however, when additional facts about the whole population are known, it is possible to provide more accurate estimates by adjustment of the sample means in the light of these facts. In most types of sampling enquiry such adjustment would not be justified: it is usually better to use the additional information when taking the sample so that the above simple rules of estimation can be followed, rather than to make adjustments involving additional calculations after taking the sample. Nevertheless cases do arise where such adjustments are of value, either because the imposition of additional restrictions would be impossible, or would be more trouble than making the adjustments, or because more accurate estimates are required from material already collected. Furthermore, study of the principles involved helps in the appreciation of the underlying principles governing all sampling.

(*a*) *Adjustment of a random sample to give the effect of a stratified sample*

If the members of a random sample are classified into different strata and the proportions $p_1, p_2, p_3, \ldots$ of the whole population falling in these strata are known, the population mean can be estimated from the means $\bar{y}_1, \bar{y}_2, \bar{y}_3, \ldots$ of the different strata derived from the sample by the formula

$$\bar{Y} = p_1\bar{y}_1 + p_2\bar{y}_2 + p_3\bar{y}_3 + \ldots$$

It can be shown that, provided the number of units falling in each stratum is not too small, the difference in accuracy between such an adjusted estimate and an estimate obtained from a stratified sample containing the same total number of sampling units will be very small. The sampling error appropriate to such an adjusted estimate can be calculated as if the sample were a stratified sample. To avoid minor complications it is best to estimate the error of the mean of each stratum separately, and calculate the error of the final estimate from these errors by the formula

$$\sqrt{p_1^2 s_1^2 \frac{1-f_1}{n_1} + p_2^2 s_2^2 \frac{1-f_2}{n_2} + \ldots}$$

where the symbols are those used in Section 5(*a*) and the suffixes refer to the different strata.

(*b*) *Use of covariance to give the effect of a balanced sample*

The estimate of the population mean can be adjusted for one or more control factors which have not been balanced in the sample, by the use of a regression. The regression can be calculated in the manner outlined above for balanced samples. In the case of a single control factor the adjusted estimate of the mean will be

$$\bar{Y} = \bar{y} + b(\bar{X} - x)$$

when $\bar{X}$ is the population mean of the control variate.

The error of this adjusted estimate will be the same as the error of the estimate from a balanced sample of the same size and type, except for a slight increase due to error in the estimation of b, which, however, is usually small enough to be ignored.

7. *Use of ratios or percentages*

It frequently happens that the value of more than one quantitative variate is recorded for each sampling unit. Often, also, the ratio between a pair of such variates is considerably less variable than are the separate variates. Such ratios may be of interest in themselves, or their average values may provide a basis for the estimation of the population total of one of the variates, when the total of the other variate is known accurately.

Thus, for example, the yield per acre of a field is likely to be less variable than either the total yield or the acreage of a field, since fields vary considerably in size. If therefore in a crop estimation scheme the yields of a random sample of fields are determined, and the total acreage under the crop is known from previous returns, the total yield can best be estimated from the formula:

$$\text{total yield} = \text{estimated mean yield per acre} \times \text{total acreage.}$$

The method of estimating the mean yield per acre requires some consideration. If the mean of the yields per acre of all the sample fields is taken, the estimate of the total yield will be subject to bias when yield per acre is associated with size of field. This bias can be avoided by weighting the estimates of the mean yields per acre by the acreages of the fields from which they were obtained. Such a weighted estimate is exactly equivalent to sum of the yields of all the sampled fields divided by the sum of the acreages of all the sampled fields.

Although the estimates based on the weighted means of the yields per acre are free from bias, they are not necessarily the most accurate that can be obtained. If the yields per acre, as determined by the sampling procedure, are equally variable for small and large fields, the unweighted mean will give the most accurate estimate of the mean yield per acre, though one that will be biased if there is association between yield per acre and size of field.

In practical sampling procedures we are therefore frequently confronted with the possibility of using various alternative estimates, some of which will be certainly unbiased, but which may be of lower accuracy than other estimates, which will themselves only be unbiased if certain conditions hold. Whether these conditions do in fact hold can usually be tested by means of the sampling data. In certain cases we may even be prepared to accept some risk of bias in return for reduction in the variability of the results. This is especially the case when the sampling is repeated at intervals and changes are of more importance than absolute magnitudes.

In the Survey of Fertilizer Practice (Yates, Boyd and Mathison, 1944), for example, information was obtained from a sample of farms on the amount and composition of fertilizer applied per acre to different crops, one field being selected at random on each farm from all the fields under each given crop. In the analysis these rates of application were weighted by the total area of the given crop on the farm. This procedure eliminates all bias due to farmers who grow large acreages of the crop applying fertilizers at different rates (on the average) from farmers who grow small acreages—such differences certainly exist—but it will not eliminate bias arising from differences of treatment of small and large fields on the same farm. To do this each rate should be weighted by a quantity nx, where x is the acreage of the sampled field and n is the number of fields on the farm.

The estimation of the sampling error of an estimate based on percentages or ratios presents certain special problems. In the case of the unweighted mean of the ratios

$$\bar{p} = \frac{1}{n} S\left(\frac{y}{x}\right)$$

the variability of the mean may be calculated in the ordinary manner by carrying out the appropriate analysis of variance on the individual ratios, though even in this case complications are apt to arise, as for instance when a stratified sample of fields has been taken but the area figures are only available for the whole of the country.

The calculation of the sampling error of a weighted mean of the ratios

$$\bar{p}_w = \frac{S(y)}{S(x)}$$

is more difficult. It might at first sight be thought that the estimate of error could be calculated by the ordinary rules applicable to weighted observations. This procedure, however, will be found to give an estimate of error which may be seriously biased if the weights x are not inversely proportional to the sampling variances of the separate ratios.

An estimate of error which is unbiased whatever the law of variation in the sampling variances * can be obtained from the unweighted sum of squares of the deviations from the line $y = \bar{p}_w x$. This sum of squares is given by

$$Q = S(y - \bar{p}_w x)^2 = S(y^2) - 2\bar{p}_w S(xy) + \bar{p}_w{}^2 S(x^2).$$

Since the fitting is not "least square," the rule as to degrees of freedom does not hold exactly, but for practical purposes Q may be taken as based on $n - 1$ degrees of freedom. The standard error of the ratio $\bar{p}_w$ is therefore

$$\frac{1}{S(x)}\sqrt{\frac{nQ(1-f)}{n-1}}.$$

It will be recognized that utilization of percentages in the manner outlined above is only appropriate when additional information is available on one of the variates, or when the value of the ratio itself is of interest. If the population values of both variates are estimated from the sample, then each variate should be treated separately. If the acreage as well as the yield per acre is estimated from the same sample of fields, for example, the total yield of the crop should be estimated directly from the total yield of all the sampled fields. Use of a weighted estimate of the yield per acre is then in fact exactly equivalent to the direct estimate based on the total yield of all the sampled fields, since

$$\bar{p}_w \frac{1}{f} S(x) = \frac{S(y)}{S(x)} \frac{1}{f} S(x) = \frac{1}{f} S(y)$$

It should be further recognized that adjustment of the sample mean of y to allow for differences between the sample and population means of x can be carried out by means of a regression of y on x in the manner of Section 6(*b*). This will prove a more accurate method of adjustment if the mean ratio of y to x does not remain substantially constant over the whole range—*i.e.*, if the regression line does not pass near the origin.

8. *Treatment of sampling units containing different numbers of individuals*

A commonly occurring problem, which throws further light on sampling problems arising in the use of ratios or percentages, is the treatment of surveys in which the sampling units contain different numbers of similar individuals. Surveys of human populations in which households are taken as the sampling units are a case of this type. In such surveys percentages of the population falling in given categories, or quantitative estimates per head of population, are usually required, rather than estimates of totals for the whole population. The problem is therefore essentially one of ratios.

As an example we may take the case of a nutrition survey covering a sample of households in which some quantitative assessment of the degree of adequacy of the nutrition of each individual in the household is made. The results for all the members of one household are clearly likely to be correlated, and account must therefore be taken of the fact that the sampling is by households and not by individuals.

Considering, for simplicity, households of one, two and three members, and denoting some quantitative measure for members of such households by y_{1r}, y_{2rs}, y_{3rs}, where r denotes the household and s the member, we may write

$$\begin{aligned} y_{1r} &= a_1 + u_{1r} + v_{1r} \\ y_{2rs} &= a_2 + u_{2r} + v_{2rs} \\ y_{3rs} &= a_3 + u_{3r} + v_{3rs} \end{aligned}$$

where a_1, a_2, a_3 are the means for families of one, two and three in the population (assumed large), u_{1r}, u_{2r} and u_{3r}, are quantities whose means are zero and which vary from household to household, and v_{1r}, v_{2rs}, v_{3rs} are similar quantities which vary from individual to individual. (In the case of

* Cochran (1942) recognized the existence, but did not bring out the importance, of an unbiased estimate of error of this type in discussing the adjustment of samples by linear regression. He did not give an unbiased estimate applicable to ratios.

families of one, u and v cannot be separated: both have been included to preserve formal symmetry.)

Let the variance of u_1 be U_1, etc., and let the numbers of households of one, two and three in the sample be n_1, n_2, n_3, with $n = n_1 + 2n_2 + 3n_3$. The analysis of variance will be

		Degrees of freedom	Expectation of mean square
Between means of households	Between households of different sizes	2	—
	Between households of one	$n_1 - 1$	$U_1 + V_1$
	,, ,, two	$n_2 - 1$	$U_2 + \frac{1}{2}V_2$
	,, ,, three...	$n_3 - 1$	$U_3 + \frac{1}{3}V_3$
Within households	Within households of two	n_2	V_2
	,, ,, three	$2n_3$	V_3

This analysis is similar to that obtained in two stage sampling (Section 5 (*c*)).

The estimates of the population means for the different household sizes will be given by the corresponding sample means, $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_3$, which will have sampling standard errors of

$$\sqrt{\frac{1}{n_1}(U_1 + V_1)},\ \sqrt{\frac{1}{n_2}(U_2 + \tfrac{1}{2}V_2)},\ \sqrt{\frac{1}{n_3}(U_3 + \tfrac{1}{3}V_3)}$$

If the proportions p_1, p_2, p_3 of the different households in the population are known, the estimate of the general population mean for all households will be

$$\frac{p_1\bar{y}_1 + 2p_2\bar{y}_2 + 3p_3\bar{y}_3}{p_1 + 2p_2 + 3p_3}$$

The sampling standard error of this quantity is

$$\frac{1}{p_1 + 2p_2 + 3p_3}\sqrt{\frac{p_1^2}{n_1}(U_1 + V_1) + \frac{4p_2^2}{n_2}(U_2 + \tfrac{1}{2}V_2) + \frac{9p_3^2}{n_3}(U_3 + \tfrac{1}{3}V_3)}$$

or since $n_1 : n_2 : n_3 = p_1 : p_2 : p_3$ approximately, the standard error is approximately

$$\sqrt{\frac{p_1(U_1 + V_1) + 4p_2(U_2 + \frac{1}{2}V_2) + 9p_3(U_3 + \frac{1}{3}V_3)}{n(p_1 + 2p_2 + 3p_3)}}$$

If a random sample of n individuals had been taken the sampling standard error would have been

$$\sqrt{\frac{p_1(U_1 + V_1) + 2p_2(U_2 + V_2) + 3p_3(U_3 + V_3)}{n(p_1 + 2p_2 + 3p_3)}}$$

Comparison of these two standard errors will give the loss of accuracy due to taking households as sampling units.

If the proportions of the different households in the population are not known, the sampling standard error given above, with n_1, n_2, n_3, substituted for p_1, p_2, p_3, is still applicable if the estimate is regarded as the estimate of the mean of a set of households containing the same proportions of households of different sizes as the actual sample. If it is regarded as an estimate of the mean of the population from which the sample is drawn, allowance for errors due to differences between the proportions in the sample and in the population must be made. For the determination of this latter error the formula already given in Section 7 is appropriate. If $\bar{y}_{2r}$ is the mean of the rth family of two, etc., and $\bar{Y}$ is the estimate of the mean of the population, so that

$$\bar{Y} = \frac{n_1\bar{y}_1 + 2n_2\bar{y}_2 + 3n_3\bar{y}_3}{n_1 + 2n_2 + 3n_3}$$

the expression for Q becomes

$$Q = S_1(y_{1r} - \bar{y}_1)^2 + 4S_2(\bar{y}_{2r} - \bar{y}_2)^2 + 9S_3(\bar{y}_{3r} - \bar{y}_3)^2 + n_1(\bar{y}_1 - \bar{Y})^2 + 4n_2(\bar{y}_2 - \bar{Y})^2 + 9n_3(\bar{y}_3 - \bar{Y})^2$$

If the adequacy of nutrition varies considerably between households of different size, the distinction between these two errors is of importance. If we wish to compare the adequacy of nutrition of two towns, for example, we can either compare the actual means, recognizing that part (perhaps the major part) of the observed difference will be due to differences in the proportions of the different sizes of household in the two towns, or we can standardize the proportions for the two towns, calculating adjusted means for these standard proportions. In the latter case the first of the above errors is appropriate, in the former the second.

The contrast between the overall sum of squares between households derived from the standard analysis of variance on individuals and the quantity Q is worth noting. In the standard analysis of variance the sum of squares derived from the means of households of size 2, for example, is multiplied by 2, and therefore the contributions to the total sum of squares are weighted in proportion to the size of household. In the quantity Q, on the other hand, they are weighted in proportion to the square of the size of household. The standard analysis will give the most accurate estimate of the sampling error if the variance of the household means is inversely proportional to size of household, but the estimate will be seriously biased if some other variance relationship holds. The quantity Q will give an estimate of sampling error which is unbiased, whatever the variance relationship.

9. *Use of a variable sampling fraction*

In economic material, as mentioned in Section 3, the units often vary very greatly in size, the variability amongst units of similar size being usually closely related to size. In such cases, for a given amount of sampling, a considerably more accurate estimate will in general be obtained if a greater proportion of the more variable units is taken. Thus, for example, if we wish to ascertain by sampling methods the number of workers employed at a given time in an industry, it will pay to obtain returns from a greater proportion of the big firms than of the small firms. The practical method of doing this is to stratify the firms into size groups, using larger sampling fractions for the groups containing the bigger firms.

In principle this idea is not new. It was recognized, for example, by Neyman (1934), but Neyman recommended that a preliminary sample should be taken in order to determine the variability of the different groups, whereas in practice the variation in the variability from group to group can often be roughly foretold from the nature of the material, so that reasonably efficient sampling fractions can be chosen in advance of any sampling.

To obtain the greatest precision with a given total number of sampling units, the sampling fractions must be proportional to the standard deviations within the different groups. Thus, if the standard deviations are $\sigma_1, \sigma_2, \sigma_3, \ldots$ the sampling fractions $f_1, f_2, f_3, \ldots$ will be given by

$$f_1/\sigma_1 = f_2/\sigma_2 = f_3/\sigma_3 \ldots$$

If in a finite population, this equation results in any sampling fraction having a value greater than unity, then the whole of that group must be included in the sample.

No great accuracy in choice of sampling fractions is required. If the sampling fractions actually used are roughly proportional to the ideal values, almost full efficiency will be attained. In the common case in which the standard deviations are approximately proportional to the means of the different size groups, the sampling fractions will be given by the formula

$$f_1/m_1 = f_2/m_2 = f_3/m_3.$$

where $m_1, m_2, m_3, \ldots$ are prior estimates of these means.

If a variable sampling fraction is used, it is of course necessary to weight the sample totals for the different groups in proportion to the reciprocals of the sampling fractions. In other words, the total of the population must be estimated from the sample totals of the different groups by the formula

$$\Sigma(Y) = S(y_1)/f_1 + S(y_2)/f_2 + S(y_3)/f_3 + \ldots$$

The mean of the population can then be estimated from the estimated total $\Sigma(Y)$.

Where the material to be sampled consists of a small number of very large units with high variability, together with a much larger number of small units subject to considerably lesser variability, use of a variable sampling fraction will give very considerable gains in precision. The results of the National Farm Survey are an example of material of this type. The survey covered all holdings in England and Wales over five acres. In order to obtain a summary of the results over the whole country (with sub-divisions for counties, types of farming, etc.), the holdings were divided into size groups, and sampling fractions were chosen as follows:

Size group (acres)	Average size (acres)	No. of holdings	Sampling fraction (per cent.)	No. of holdings in sample
5– 25	12	101,450	5	5,072
25–100	55	111,360	10	11,136
100–300	165	65,210	25	16,302
300–700	413	11,150	50	5,575
over 700	1,035	1,430	100	1,430
		290,600	(13·6)	39,515

To obtain adequate representation of the smaller size groups for purposes where contrasts between farms of different sizes were required, and to give additional accuracy to estimates involving numbers of farms (regardless of size), the sampling fractions for these groups were relatively larger than those which would give the most accurate estimates of the areas of land falling in different categories.

Some calculations of the gain in efficiency resulting from the use of a variable sampling fraction in the National Farm Survey showed that the sampling variance for measures of the percentage of farm land possessing given attributes was one third that which would have been obtained from a sample containing the same total number of farms with equal proportions taken from each of the size groups. On the other hand, stratification into size groups would not in itself have resulted in any great reduction in error, since the proportions of the different size groups possessing such attributes did not differ substantially.

Another case in which the use of a variable sampling fraction results in large gains of efficiency is where some such quantity as crop acreage is being estimated, and the crop in question occupies very different proportions of the total land area in different districts. Here more intense sampling of the areas with high proportions of the crop will be advantageous. The large gain in efficiency obtained by Mahalanobis (1944) in the estimation of the area of jute as a result of minimising a cost function appears to have been mainly due to the fact that the cost function was such that a variable sampling fraction was determined.

10. *Quasi-random sampling*

In practice, at least in economic and social studies, it is rare for a sample to be selected by strict random selection. Generally some form of selection from a list, such as taking every tenth name in the list, or other form of systematic selection, is used. If the list is arranged in substantially random order—*e.g.*, if it is alphabetical—a quasi-random sample can be treated as if it were a random sample. In general, however, quasi-random sampling automatically results in some form of stratification. This, if every tenth house is taken from an electoral register which is arranged by wards and streets, one tenth of all the houses in each ward will automatically be taken. Consequently, the sample will be stratified by wards. Further elements of stratification will also occur to a greater or less extent. For example, in the long streets the number of houses taken will be very approximately one-tenth of the whole.

In calculating the sampling errors of material of this type it is usually sufficient to eliminate the effects of the more important divisions, such as wards, ignoring the effects of the smaller sub-divisions, such as streets, it being recognized that the sampling error so obtained is likely to be a slight over-estimate.

Attempts are sometimes made to overcome difficulties of this kind by taking two random starting points for each street and then selecting houses at equal intervals. Thus, instead of taking every tenth house, we might select two numbers at random between 1 and 20, say 12 and 15, and take two groups of houses from the street, namely 12, 32, 52, etc., . . . and 15, 35, 55, etc. . . Differences between the means of the groups will provide a formally correct estimate of the sampling error. The additional trouble, both of selection and of computation, is such, however, that this procedure is not of great practical value.

11. *Systematic sampling*

Area sampling presents special problems. The classical case is the sampling of a field under an agricultural crop to determine the yield or other characteristics of the crop. Inasmuch as the fertility may vary over different parts of the field, more accurate results will clearly be obtained if the units of the sample are distributed more or less evenly over the field. The simplest way of ensuring this, while satisfying the conditions required for a valid estimate of error, is to divide the field into blocks, sub-divide each block into small areas, and to select two or more small areas at random from each block. This constitutes an ordinary stratified sample, with the blocks as strata and the small areas as sampling units; fertility differences between the different blocks will be eliminated from the sampling error. A modification of this procedure is to use complex sampling units, each unit being made up of a set of small areas arranged in some pattern which ensures a reasonably even spread over the block. This will eliminate a good deal of the variation in fertility within blocks. Provided condition (2) of Section 3 is satisfied a valid estimate of sampling error is still possible.

In many types of area sampling the location of the sampled areas on parallel lines is desirable. In sampling agricultural crops, for instance, the rows may have to be followed; in sampling a forest area it may be convenient to follow a fixed compass bearing from a point located on a base line. Sampling is then often two-stage, the chosen lines constituting the first stage, and the sampling of the lines the second.

Considering the first stage, the area may be divided into a set of blocks bounded by lines parallel to the direction of sampling, and two lines may be selected at random from each block. This, however, will not lead to regular spacing of the lines, and in extreme cases four contiguous lines may be selected. It is clear that an even spacing of the lines will in general give a more accurate representation of the area. Even spacing has the additional advantages that location of the lines is simpler, and that the construction of an approximate map of the area, if this is required, is facilitated, since there will be no blank patches such as tend to occur with random location. Such even spacing may be termed *systematic sampling*. There are certain possibilities of bias which must not be overlooked, but which are usually of little practical importance in the types of area sampling for which systematic sampling is most suitable.

The same considerations apply to the second stage, the sampling of the selected lines. If evenly spaced units are taken on these lines, and the lines themselves are evenly spaced, a systematic grid pattern of sampling units will result.

The relative merits of systematic and random sampling of the grid and line type have been the subject of lengthy controversy, particularly in the United States, in connection with the sampling of forest areas. I have recently been making an investigation of the problem designed to ascertain:

(1) The gain in accuracy obtained by the use of systematic samples instead of samples randomly located in pairs within blocks.

(2) What methods, if any, are available to make an approximate estimate of the sampling errors involved in systematic sampling.

This investigation is not fully complete, but certain tentative conclusions may be set out here. First, as regards gain in accuracy, this will naturally depend very much on the nature of the material. If the chief source of variation is of a random nature, then clearly the gain in accuracy of systematic over random samples will be small. On the other hand, if the variation is of a continuous type, the gain may be very considerable.

The gain in accuracy due to systematic sampling with lines spaced at a distance A compared with random sampling with pairs of lines randomly located in blocks of width $2A$ can be looked on as made up of two parts, (*a*) that due to a reduction in block size from width $2A$ to A, (*b*) that due to the location of the line at the centre of a block of width A instead of at random in such a block. The gain in accuracy due to (*a*) in any particular type of material can therefore be assessed if the variances within blocks of width A and $2A$ are known. The assessment of the part (*b*) of the gain is much more difficult, since we become involved in certain properties in the variation of the sampled material which are very difficult to determine.

The variance in blocks of width A can with a certain amount of ingenuity be determined from the results of random sampling in which random pairs of samples are taken from blocks of width $2A$, provided the locations of the sampled lines are known. Thus it is possible, in any particular case of random sampling, to make an estimate of the part (*a*) of the gain that would have resulted from the use of systematic sampling.

If a systematic sample has been taken, on the other hand, we cannot, from the internal evidence of the sample itself, determine either the gain in accuracy over a random sample or the actual accuracy of the systematic sample. Nevertheless an upper limit to the sampling error can always be obtained from the differences of consecutive pairs of systematically located units, but such an estimate, with continuously varying material, may be much above the true value.

To make an estimate of the degree to which the continuous features in the variation of the material reduce the sampling error, some supplementary observations are necessary. One method which appears to give satisfactory results is to take systematic samples at four times the normal density—*i.e.*, at a spacing of $\frac{1}{4}A$—over part of the material. From these observations what may be called " systematic sub-samples with spacing A, adjusted for end conditions," covering blocks of moderate width—say $4A$—can be constructed.* There will be four such sub-samples per block, and the variances within blocks of sub-samples separated by $\frac{1}{2}A$ and of those separated by $\frac{1}{4}A$ can be calculated. These variances can then be used as the basis of an estimate of the sampling variance of systematic samples with spacing A.

An alternative method of which the possibilities have not yet been fully investigated is to take a second systematic sample which has a slightly different spacing to the original sample, so that information is available on the differences between lines at all spacings.

These methods, however, require further investigation before they can be confidently advocated for use in practice. All that can be said at present is that although a systematic sample will unquestionably give more accurate information than a random sample on material which varies in a more or less continuous manner, its accuracy cannot be assessed without supplementary sampling: from the internal evidence of the sample itself only an upper limit can be assigned to the sampling error, and the greater the advantage of systematic over random sampling, the more widely will this upper limit be separated from the true value. Nevertheless, I feel certain that in extensive area surveys systematic line or grid sampling is preferable to random sampling. The additional work required to investigate the accuracy attained will be more than compensated for by the saving in work in the survey itself.

12. *Applications in social surveys*

With the development of the social sciences and the adoption of a planned economy, surveys should become of increasing importance. The use of sampling in social surveys is not new, and in some respects the methods are relatively simple. For most types of survey covering a single town or rural area the population can easily be divided into relatively small sampling units, either separate individuals or households, and there is usually no difficulty in obtaining a quasi-random sample of such a population. The only points of difficulty in the estimation of sampling errors arise from variation in size of sampling unit when households are used, and the quasi-random nature of most samples.

If a small sample of the population of a whole country is required, a further difficulty is encountered. A sample selected in equal proportions from all the towns and rural areas in the country would be so scattered that the amount of travelling involved in order to carry out interviews would be quite excessive. Frequently, therefore, sampling has to be limited to certain towns and rural areas, even when a sample representative of the whole country is required.

For certain purposes the selection of towns in which opportunities of survey are particularly favourable more than outweighs the objection that such a sample cannot be regarded as fully representative of the whole country. Provided this lack of representativeness is clearly recognized, no harm will ensue. Moreover, if surveys have been carried out in a number of towns and rural areas which are widely contrasted, and there is found to be little difference between the results of the different towns and areas, it may be assumed that the results will be reasonably representative for the whole country.

The simplest method of dealing with the problem of securing a fully representative sample is to take a random sample of towns and rural areas, stratifying as far as possible into geographical regions, and by such simple characteristics as size and degree of industrialization. There does, however, seem to be a strong case for attempting the construction of some form of balanced

* If the observed values are numbered consecutively, the systematic sub-sample totals will be of the form: $\frac{1}{4}y_1 + y_5 + y_9 + y_{13} + \frac{3}{4}y_{17}$, $\frac{1}{2}y_2 + y_6 + y_{10} + y_{14} + \frac{1}{2}y_{18}$, $\frac{3}{4}y_3 + y_7 + y_{11} + y_{15} + \frac{1}{4}y_{19}$, $y_4 + y_8 + y_{12} + y_{16}$.

sample in which various control factors are used. As far as I am aware, the possibilities of effecting this have never been thoroughly investigated. In this connection the ingenious method followed in constructing a sample of farming districts for economic surveys in the United States is worth consideration (Hagood and Bernert, 1945). In order to obtain a sample which was reasonably balanced for a number of control factors, while avoiding the disadvantages of purposive selection, an index consisting of a linear function of these control factors was formed, and the districts were then stratified on the basis of the values of this index, random selection of one district from each stratum being made. In certain cases an auxiliary index was also used in forming the strata.

Unfortunately, emphasis on economy of interviewers' time has only too often led to the abandonment of the principle of random or quasi-random selection. In surveys of public opinion the procedure known as " quota " sampling in the United States, whereby a sample is made up by selection of the requisite quota of individuals in various income groups, etc., by the interviewers themselves, has become popular.

If such selection were otherwise random, it would of course be equivalent to a stratified sample, and would be quite satisfactory, but the actual methods of selection adopted are by no means random. Thus Box and Thomas (1944) give the following description of one of the methods followed by the Wartime Social Survey:

> " Where income group is used to control the sample, information is sought from the police, or other officials with local knowledge, as to the districts in which families belonging to different income groups are most frequently to be found. In towns where there are regional investigators this information has already been collected and areas marked off on street maps. The investigator then goes to the different types of district and selects streets where she expects to find families in the different groups. The houses to be visited in these streets are selected according to some previously determined plan—*e.g.*, every tenth house or from lists of random numbers. If the household does not belong to the expected income group, it is classified in the group to which it does belong, and further calls are made in other streets until the required quota of each group is completed."

It is clear that such a method may very easily introduce serious biases. It is also quite impossible to assess the accuracy of the results from the internal evidence of the observations themselves.

On this point Deming, in the course of a valuable review of the principles governing the conduct of sampling surveys (Deming, 1945), has written:

> " Unless biases can be removed satisfactorily a method of collection that appears to be cheap is too often cheap only in the sense of providing a lot of schedules per dollar, but may actually be very costly when measured in the amount of useful information per dollar or the damage done through misinformation."

That the quota method, in spite of objections that can be raised to it, has given satisfaction for certain types of work is due primarily, I think, to the fact that it has been mainly used for opinion surveys where there is in general little check on the accuracy of the results actually obtained and where, in any case, no very precise results are required. Many opinion surveys are also repeated from time to time, and changes of opinion are the points of chief interest. If the sample is selected in the same manner on each occasion, any bias will affect the results on all occasions more or less equally, and consequently trends of opinion will be truly reflected in the results.

When, however, quota sampling is used for more serious work on which administrative decisions have to be based, and which require proper quantitative estimates of the various characteristics under survey, its weakness becomes apparent. There is also the further important objection that the quota method involves a considerable loosening of control over the interviewers. This often has a bad effect on the selection of the sample—for example, interviewers may tend to choose houses or districts from which it is easy to obtain answers.

Even apart from questions of bias, it is doubtful whether the use of an elaborate quota system is likely to produce results which are appreciably more accurate than some simple form of quasi-random selection. Particularly in the case of qualitative characters, the gain due to stratification is much less than is commonly supposed. If, for example, a population is divided into five equal groups of which 70, 60, 50, 40 and 30 per cent. respectively give a positive answer to a certain question, a fully random sample will only need to be 8 per cent. larger (for the same accuracy) than a stratified sample. Compared with a quasi-random sample from an electoral register arranged by wards, which will automatically be partially stratified by social class and income group, the gain of a stratified sample is likely to be much less. If all the percentages are small, or in the neighbourhood of 100 per cent., the gain is in any case quite trivial: with percentages of 10, $7\frac{1}{2}$, 5, $2\frac{1}{2}$ and 0 per cent. for example, the gain of a stratified over a fully random sample is 0·7 per cent.

The doubts and difficulties to which the use of the quota method by the Wartime Social Survey has given rise are well illustrated by the lengthy discussion on this point in the paper referred to above (Box and Thomas, 1945). It is there stated:

> " In inquiries relating to the whole adult civilian population, the population is stratified by sex and by occupation, as well as by region and by urban and rural areas. Stratification by sex presents no difficulty, but the division of the population into occupation groups and allotting the appropriate number of interviews to each group is a matter of some concern. The lack of any up-to-date information on the proportions of the population following different occupations is a serious drawback to the Survey. Also such classifications as have been made at different times are not altogether satisfactory for social survey purposes."

The authors do not appear to recognize that this particular problem would not have arisen at all had they been content to use some method of random sampling. While a quasi-random sample over the whole population of a town is undoubtedly the simplest, a sample in which certain districts only are taken, these being a sample of all districts, with a sub-sample of houses in each selected district, would probably be nearly as accurate, and would substantially cut down travelling time, which is the main argument in favour of the method used at present.

It was these considerations that led to the decision, in the social surveys undertaken by the Ministry of Home Security, to use a quasi-random sample over the whole of each surveyed town, with scrupulous attention to such details as " calling back." These surveys were carried out in order to determine the reactions of populations of raided towns to air raids, and involved the evaluation of such quantities as amount of time lost from work, the amount of evacuation, etc., and it was of the utmost importance that the reliability of the results should not be called in question on statistical grounds.

In conclusion I would like to emphasize that in many social surveys the sampling problems involved constitute only a minor difficulty, compared with the problem of obtaining reliable information on the subject under enquiry. This involves careful drafting of the questionnaire, the inclusion only of questions for which it is reasonably certain that the questionees know the relevant facts, and are prepared to give truthful answers, and a high degree of skill and tact on the part of the field workers. These points were well brought out in the paper on the Wartime Social Survey and the subsequent discussion.

13. *The planning of sampling enquiries*

The efficient planning of a sampling enquiry requires knowledge of the different components of variability of the material, and also a knowledge of the relative costs of collecting the information with different types of sampling unit.

The ideal method of studying variability is to have available complete information on an adequate and representative body of material. We can then make a thorough analysis of the different components of variability, and from the results of this analysis calculate the sampling errors to be expected in different sampling procedures. Alternatively, trial sampling schemes of various kinds can be tried out, and the errors to which they are subject estimated in the ordinary manner. If this is done, however, some supplementary study of the variability will usually throw light on the reasons why one method is more accurate than another, and may suggest further improvements.

The importance of having available an adequate and representative body of material for study must, however, be emphasized. Individual batches of material often possess features of variability which specially favour one type of sampling unit, but which are not repeated in other batches of similar material. Thus the procedure of harvesting part of a single field under an agricultural crop in small units, and reaching conclusions as to the best type of sampling unit from the performance of various units on this one small area, cannot be regarded as sound.

In practice, however, complete information of this kind is frequently not available. Often it would be quite impracticable to collect it. It is not sufficiently realized that critical analysis of the results of any properly executed sampling enquiry will not only enable the accuracy of the survey itself to be determined, but will also usually throw considerable light on whether future enquiries on the same type of material can be more efficiently planned. To give a few simple examples: the gain due to stratification can be evaluated, whether the original enquiry was itself stratified or not, provided the individual units can be assigned to the strata under consideration; the gain due to the use of a variable sampling fraction can be similarly estimated; the best ratio between intensity of sampling at two different stages can be assessed in the light of the sampling errors at the two stages and the relative costs of sampling at the two stages.

It is not possible, however, to assess the effect of all possible changes in sampling procedure from the results of a single enquiry: thus it is impossible to determine the variability within households from a random sample of individuals unless the sampling is so intense that an adequate number of pairs of individuals from within households are obtained. Though frequently a small amount of supplementary sampling, which can be carried out at the same time as the main enquiry, will provide vital information on variability that would otherwise be lacking.

The need for thorough studies of the efficiency of different sampling procedures on different types of material is great, and I hope that many more workers will be persuaded to undertake them in future. All such studies demand a considerable amount of numerical work, but the burden of computation can be lightened by the use of mechanical aids, in particular punched-card methods.

One reason why such studies are often neglected is that once a survey has been completed the question of whether it could have been carried out more efficiently is historical as far as that survey is concerned. Often, too, the person in charge of a series of surveys makes a sufficient study of the results of the first one or two surveys to satisfy himself of the adequacy of the technique, and to introduce any necessary improvements, but he fails to publish his conclusions, and other workers in the same field have to go over the same ground again, or follow blindly in his footsteps. Is it too much to hope that the results of such enquiries will be published more frequently in future?

The full value of any study of efficiency, and indeed of the magnitude of the sampling errors, is realized when a new survey on the same type of material is undertaken. Had it not been that a fairly thorough investigation of the sampling errors of the 1938–9 Census of Woodlands had been made,* it would have been impossible to plan the 1942 Census with any confidence. (For a brief description of this Census see Yates, (1943).) As it was, it was possible to give a firm assurance that the survey, if properly carried out in the manner planned, would give results of the required accuracy.

References

Bowley, A. L. (1926), *Bull. Int. Inst. Stat.*, **22**, 440.
Box, H., and Thomas, G. (1944), *J. Roy. Stat. Soc.*, **107**, 151.
Clapham, A. R. (1929), *J. Agric. Sci.*, **19**, 214.
Clapham, A. R. (1931), *ibid.*, **21**, 366.
Cochran, W. G. (1939), *J. Amer. Stat. Ass.*, **34**, 492.
Cochran, W. G. (1942), *ibid.*, **37**, 199.
Deming, W. E. (1945), *ibid.*, **40**, 307.
Fisher, R. A., and Mackenzie, W. A. (1923), *J. Agric. Sci.*, **13**, 311.
Fisher, R. A. (1926), *J. Min. Agric.*, **33**, 503.
Hagood, M. J., and Bernert, E. J. (1945), *J. Amer. Stat. Ass.*, **40**, 330.
Jensen, A. (1926), *Bull. Int. Inst. Stat.*, **22**, 359, 381.
Mahalanobis, P. C. (1944), *Phil. Trans. Roy. Soc. Lond.*, **231**, 329.
March, L. (1926), *Bull. Int. Inst. Stat.*, **22**, 440.
Neyman, J. (1934), *J. Roy. Stat. Soc.*, **97**, 558.
Stuart, C. A. V. (1926), *Bull. Int. Inst. Stat.*, **22**, 440.
Wishart, J., and Clapham, A. R. (1929), *J. Agric. Sci.*, **19**, 600.
Yates, F., and Zacopanay, I. (1935), *ibid.*, **25**, 545.
Yates, F. (1934), *J. Amer. Stat. Ass.*, **29**, 51.
Yates, F. (1935), *Ann. Eug.*, **6**, 202.
Yates, F., and Finney, D. J. (1942), *Ann. App. Biol.*, **29**, 156.
Yates, F. (1943), *Roy. Soc. Arts*, 1.
Yates, F., Boyd, D. A., and Mathison, I. (1944), *Emp. J. Exp. Agric.*, **12**, 163.

* But, I have to confess, not published!