Sample Surveys

By

W. Edwards Deming

International Encyclopedia of STATISTICS

Edited by

WILLIAM H. KRUSKAL and JUDITH M. TANUR
University of Chicago *State University of New York at Stony Brook*

SAMPLE SURVEYS

There is hardly any part of statistics that does not interact in some way with the theory or the practice of sample surveys. The differences between the study of sample surveys and the study of other statistical topics are primarily matters of emphasis.

The field of survey research is closely related to the statistical study of sample surveys [*see* SURVEY ANALYSIS]. *Survey research is more concerned with highly multivariate data and complex measures of relationship; the study of sample surveys has emphasized sampling distributions and efficient design of surveys.*

I
THE FIELD

The theory of sample surveys is mathematical and constitutes a part of theoretical statistics. The practice of sample surveys, however, involves an intimate mixture of subject matter (such as demography, psychology, consumer research, medicine, engineering) with theory. The germ of a study lies in the subject matter. Translation of a substantive question into a stimulus (question or test) enables man to inquire of nature and to quantify the result in terms of estimates of what the same inquiry would produce were it to cover every unit of the population.

Sampling, properly applied, does more. It furnishes, along with an estimate, an index of the precision thereof—that is, a margin of the uncertainty, for a stated probability, that one may reasonably ascribe to accidental variations of all kinds, such as variability between units (that is

between households, blocks, patients), variability of the interviewer or test from day to day or hour to hour, variations in coding, and small, independent, accidental errors in transcription and card punching.

The techniques of sampling also enable one to test the performance of the questionnaire and of the investigators and to test for differences between alternative forms of the questionnaire. They enable one to measure the extent of under-coverage or over-coverage of the prescribed units selected and also to measure the possible effects of differences between investigators and of departures from prescribed rules of interviewing and coding.

This article describes *probability sampling*, with special reference to studies of human populations, although the same theory and methods apply to studies of physical materials, to accounting, and to a variety of other fields. The main characteristic of probability sampling is its use of the theory of probability to maximize the yield of information for an allowable expenditure of skills and funds. Moreover, as noted above, the same theory enables one to estimate, from the results themselves, margins of uncertainty that may reasonably be attributed to small, accidental, independent sources of variation. The theory and practice of probability sampling are closely allied to the design of experiments.

The principal alternatives to probability sampling are judgment sampling and convenience sampling [*see* SAMPLE SURVEYS, *article on* NONPROBABILITY SAMPLING].

Uses of sampling. Probability sampling is used in a wide variety of studies of many different kinds of populations. Governments collect and publish monthly or quarterly current information in such areas as employment, unemployment, expenditures and prices paid by families for food and other necessaries, and condition and yield of crops.

In modern censuses only basic questions are asked of every person, and most census information is elicited for only a sample of the people, such as every fourth household or every twentieth. Moreover, a large part of the tabulation program is carried out only on a sample of the information elicited from everyone.

Sampling is the chief tool in consumer research. Samples of records, often supplemented by other information, furnish a basis on which to predict the impact that changes in economic conditions and changes in competitive products will have on a business.

Sampling is an important tool in supervision and is helpful in many other administrative areas, such as studies of use of books in a library to improve service and to make the best use of facilities.

Sampling—what is it? Everyone acquires information almost daily from incomplete evidence. One decides on the basis of the top layer of apples in a container at the fruit vendor's whether to buy the whole container. The top layer is a good sample of the whole if the apples are pretty well mixed; it is a bad sample and may lead to a regrettable purchase if the grocer has put the best ones on top.

The statistician engaged in probability sampling takes no chances on inferences drawn exclusively from the top layer or from any other single layer. He uses random numbers to achieve a standard degree of mixing, thereby dispersing the sample throughout the container and giving to every sampling unit in the frame an ascertainable probability of selection [*see* RANDOM NUMBERS]. He may use powerful techniques of stratification, ratio estimation, etc., to increase accuracy and to decrease costs. For instance, in one type of stratified sampling he in effect divides the container of apples into layers, mixes the apples in each layer, and then takes a sample from each layer.

Some history of sampling. Sir Frederick Morton Eden estimated the number of inhabitants of Great Britain in 1800 at nine million, using data on the average number of people per house in selected districts and on the total number of houses on the tax-rolls, with an allowance for houses not reported for taxation. The first census of Great Britain, in 1801, confirmed his estimate. Messance in 1765 and Moheau in 1778 obtained estimates of the population of France by multiplying the ratio of the number of inhabitants in a sample of districts to the number of births and deaths therein by the number of births and deaths reported for the whole country. Laplace introduced refinements in 1786 and calculated that 500,000 was the outside margin of error in his estimate of the population of France, with odds of 1,161 : 1. His estimate and its precision were more successful than those of the complete census of France that was attempted at the same time. [*See* LAPLACE.]

A. N. Kiaer used systematic selection in a survey of Norwegian workers in 1895, as well as in special tabulations from the census of Norway in 1900 and from the census of Denmark in 1901 and in a study of housing in Oslo in 1913.

Bowley in 1913 used a systematic selection of every twentieth household of working-class people in Reading (England) and computed standard errors of the results.

Tabulation of the census of Japan in 1921, brought to a halt by the earthquake of 1923, went forward with a sample consisting of the records of every thousandth household. The results agreed with the full tabulation, which was carried out much later. The Swedish extraordinary census of 1935 provides a good example of the use of sampling in connection with total registrations.

One strong influence on American practice came in the 1930s from Margaret H. Hogg, who had worked under Bowley. Another came when controversies over the amount of unemployment during the depressions of 1921 and 1929 called for improved methods of study—Hansen's sample of postal routes for estimates of the amount of unemployment in 1936 gained recognition for improved methods; without it the attempt at complete registration of unemployed in the United States at the same time would have been useless.

Mahalanobis commenced in 1932 to measure the yield of jute in Bengal and soon extended his surveys to yields of rice and of other crops. In 1952 all of India came under the national surveys, the scope of which included social studies and studies of family budgets, sickness, births, and deaths. Meanwhile, the efforts of statisticians, mainly in India and England, had brought advances in methodology for estimation of yield per acre by random selection of small plots to be cut and harvested.

A quarterly survey of unemployment in the United States, conducted through interviews in a sample of households within a sample of counties, was begun in 1937. It was soon made monthly, and in 1942 it was remodeled much along its present lines (Hansen et al. 1953, vol. 1, chapter 9).

Sampling was used in the census of the United States in 1940 to extend coverage and to broaden the program of tabulation and publication. Tabulation of the census of India in 1941 was carried out by a 2 per cent sample. Subsequent censuses in various parts of the world have placed even greater dependence on sampling, not only for speed and economy in collection and tabulation but also for improved reliability. The census of France used sampling as a control to determine whether the complete Census of Commerce of 1946 was sufficiently reliable to warrant publication; the decision was negative (Chevry 1949). [*For further history, see* Stephan 1948. *Some special references to history are contained in* Deming (1943) 1964, p. 142. *See also* STATISTICS, *article on* THE HISTORY OF STATISTICAL METHOD.]

Misconceptions about sampling. Sampling, of course, possesses some disadvantages: it does not furnish detailed information concerning every individual person, account, or firm; furthermore, error of sampling in very small areas and subclasses may be large. Many doubts about the value of sampling, however, are based on misconceptions. Some of the more common misconceptions will now be listed and their fallacies pointed out.

It is ridiculous to think that one can determine anything about a population of 180 million people, or even 1 million people, from a sample of a few thousand. The number of people in the country bears almost no relation to the size of the sample required to reach a prescribed precision. As an analogy (suggested by Tukey), consider a basket of black and white beans. If the beans are really mixed, a cupful would determine pretty accurately the proportion of beans that are black. The cupful would still suffice and would give the same precision for a whole carload of beans, provided the beans in the carload were thoroughly mixed. The problem lies in mixing the beans. As has already been noted, the statistician accomplishes mixing by the use of random numbers.

Errors of sampling are a hazard because they are ungovernable and unknown. Reliability of a sample is a matter of luck. Quality and reliability of data are built in through proper design and supervision, with aid from the theory of probability. Uncertainty resulting from small, independent, accidental errors of a canceling nature and variation resulting from the use of sampling are in any case determinable afterward from the results themselves.

Errors of sampling are the only danger that one has to worry about in data. Uncertainty in statistical studies may arise from many sources. Sampling is but one source of error. [*See below, and see also* ERRORS, *article on* NONSAMPLING ERRORS].

Electronic data-processing machines, able to store and retrieve information on millions of items with great speed, eliminate any need of sampling. This is a fanciful hope. The inherent accuracy of original records as edited and coded is the limitation to the accuracy that a machine can turn out. Often, complete records are flagrantly in error or fail to contain the information that is needed. Moreover, machine-time is expensive; sampling reduces cost by reducing machine-time.

A "complete" study is more reliable than a sample. Data are the end product of preparation and of a long series of procedures—interviewing, coding, editing, punching, tabulation. Thus, error of sampling is but one source of uncertainty. Poor workmanship and structural limitations in the method of test or in the questionnaire affect a

complete count as much as they do a sample. It is often preferable to use funds for improving the questionnaire and tests rather than for increasing the size of the sample.

Statistical parts of sampling procedure. A sampling procedure consists of ten parts. In the following list, *M* will denote those parts that are the responsibility of the expert on the subject matter, and *S* will denote those that are the responsibility of the statistician. (The technical terms used will be defined below.)

(*a*) Formulation of the problem in statistical terms (probability model) so that data will be meaningful (*M*, *S*). A problem is generated by the subject matter, not by statistical theory.

(*b*) Decision on the universe (*M*). The universe follows at once from a careful statement of the problem.

(*c*) Decision on the frame (*M*, *S*). Decision on the type and size of sampling units that constitute the frame (*S*).

(*d*) Procedure for the selection of the sample (*S*).

(*e*) Procedure for the calculation of estimates of the characteristics desired (averages, totals, proportions, etc.) (*S*).

(*f*) Procedure for the calculation of standard errors (*S*).

(*g*) Design of statistical controls, to permit detection of the existence and extent of various nonsampling errors (*S*).

(*h*) Editing, coding, tabulation (*M*, *S*).

(*i*) Evaluation of the statistical reliability of the results (*S*).

(*j*) Uses of the data (*M*).

Definitions of terms

The technical terms that have been used above and that will be needed for further discussion will now be defined.

Universe of study. The universe consists of all the people, firms, material, conditions, units, etc., that one wishes to study, whether accessible or not. The universe for any study becomes clear from a careful statement of the problem and of the uses intended for the data. Tabulation plans disclose the content of the universe and of the information desired. Examples of universes are (*i*) the housewives aged 20–29 that will live in the Metropolitan Area of Detroit next year, (*ii*) all the school children in a defined area, (*iii*) all the pigs in a country, both in rural areas and in towns.

Frame. The frame is a means of access to the universe (Stephan 1936) or to enough of the universe to make a study worthwhile. A frame is composed of sampling units. A sampling unit commonly used in house-to-house interviewing is a compact group or segment of perhaps five consecutive housing units. A frame is often a map, divided up—either explicitly or tacitly—into labeled areas. In a study concerned with professional men, for example, the frame might be the roster of membership of a professional society, with pages and lines numbered. The sampling unit might be one line on the roster or five consecutive lines.

Without a frame probability sampling encounters numerous operational difficulties and inflated variances (see, for example, the section "Sampling moving populations," below).

In the types of problems to be considered here (with the exception of those treated in the section "Sampling moving populations," below) there will be a frame, and every person, or every housing unit, will belong to one sampling unit, or will have an ascertainable probability of belonging to it. In the sampling of stationary populations, a sampling procedure prescribes rules by which it is possible to give a serial number to any sampling unit, such as a small area. A random number will then select a definite sampling unit from the frame and will lead to investigation of all or a subsample of the material therein that belongs to the universe.

Selection of persons within a dwelling unit. Some surveys require information concerning individuals, and in such cases it may be desirable, for various reasons (contagion, fatigue, and so on), to interview only one eligible person in a dwelling unit that lies in a selected segment. In such surveys, the interviewer may make a list of the eligible people in each dwelling unit that falls in the sample and may select therefrom, on the spot, by a scheme based on random numbers, one person to interview. Appropriate weights are applied in tabulation (Deming 1960, p. 240).

Nominal frame and actual frame. One must often work with a frame that fails to include certain areas or classes that belong to the universe. A list of areas that contain normal families may not lead to all the consumers of a product, as some consumers may live in quasi-normal quarters, such as trailers and dormitories. Extension of the sampling procedure into these quarters may present problems. Fortunately, the proportion of people in quasi-normal households is usually small (mostly 1 per cent to 3 per cent in American cities), and one may therefore elect to omit them.

A frame may be seriously impaired if it omits too much of certain important classes that by definition belong to the nominal frame. It is substantive judgment, aided by calculation, that must decide whether a proposed frame is satisfactory.

Sampling from an incomplete frame. Almost every frame is in some respects out of date at the time of use. It is often possible, however, to use an obsolete or incomplete frame in a way that will erase the defects in the areas that fall into the sample. One may, for example, construct rules by which to select large sampling units from an incomplete frame and then to amend those units, by local inquiry, in order to bring them up to date. Selection of small areas within the larger area, with the appropriate probability, will maintain the prescribed over-all probability of selection.

Sampling for rare characteristics. One sometimes wishes to study a rare class of people when there is no reliable list of that class. One way to accomplish this is to carry out a cheap, rapid test in order to separate a sample of households into two groups (strata)—one group almost free of the rare characteristic, the other heavily populated with it—and then to investigate a sample drawn from each group. Optimum sampling fractions and weights for consolidation may be calculated by the theory of stratified sampling (discussed below; see also Kish in Symposium . . ., 1965).

Equal complete coverage of a frame. The equal complete coverage of a frame is by definition the result that would be obtained from an investigation of all sampling units in a given frame, carried out by the same field workers or inspectors, using the same definitions and procedures, and exercising the same care as they exercised on the sample, and at about the same period of time. The adjective "equal" signifies that the same methods must be used for the equal complete coverage as for the sample.

Some operational definitions. *Sampling error.* Suppose that for a given frame, sampling units bear the serial numbers 1, 2, 3, and on to N. However it be carried out, and whatever be the rules for coding and for adjustment for nonresponse, a complete coverage of the N sampling units of the frame would yield the numerical values

$$a_1, a_2, a_3, \cdots, a_N \text{ for } x,$$
$$b_1, b_2, b_3, \cdots, b_N \text{ for } y.$$

In a survey of unemployment, for example, the x-characteristic of a person might be the property of being unemployed and his y-characteristic the property of belonging to the labor force. Then a_1, the x-population of sampling unit No. 1 (which might consist of five successive households), would be the count of people that have the x-characteristic in that sampling unit. That is, a_1 would be the count of unemployed persons in the five households. Similarly, b_1, the y-population, would be the count of people in the labor force in those same households. Then a_1/b_1 would be the proportion unemployed in the sampling unit of five households.

Again, x might refer to expenditure for bread and y to expenditure for all food. Then a_1/b_1 would be the proportion of money that goes for bread in sampling unit No. 1, expenditure for all food being the base.

Here, the people with the x-characteristic form a subclass of those with the y-characteristic, but this may not be so in other surveys. Thus, the x-characteristic and the y-characteristic might form a dichotomy, such as passed and rejected or male and female. One often deals with multiple characteristics, but two will suffice here.

Denote the sum of the x-values and of the y-values in the N sampling units by

$$A = a_1 + a_2 + a_3 + \cdots + a_N = Na = x\text{-total},$$
$$B = b_1 + b_2 + b_3 + \cdots + b_N = Nb = y\text{-total},$$

which makes a and b the average x-value and the average y-value per sampling unit in the frame, as in Table 1. For example, A might be the total number unemployed in the whole frame and B the total number of people in the labor force. Then $\phi = A/B$ would be the proportion of people in the labor force that are unemployed.

An operational definition of the sampling process and of the consequent error of sampling is contained in the following experiment.

(a) Take for the frame N cards, numbered serially 1 to N. Card i shows a_i and b_i for the values of the x-characteristic and y-characteristic.

(b) Draw a sample of n cards, following the specified sampling procedure (which will invariably require selection by random numbers).

Table 1 illustrates the notation for the frame and for the results of a sample. The serial numbers on the cards in the sample are not their serial numbers in the frame but denote instead the ordinal number as drawn by random numbers. Sample card No. 1 could be any card from 1 to N in the frame. In general, another sample would be composed of different cards, as the drawings are random.

(c) Form estimates by the formulas specified in the sampling plan. For illustration, one may form, from the sample, estimators like

$$(1) \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$(2) \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$(3) \qquad X = N\bar{x},$$

$$(4) \qquad Y = N\bar{y},$$

$$(5) \qquad f = \bar{x}/\bar{y}.$$

Table 1 — Some notation for frame and sample

	FRAME		SAMPLE		
	Serial number of sampling unit	x-value	y-value	Serial number in order drawn in sample	x-value	y-value
	1	a_1	b_1	1	x_1	y_1
	2	a_2	b_2	2	x_2	y_2
		⋮	⋮		⋮	⋮
	N	a_N	b_N	n	x_n	y_n
Total		A	B		x	y
Average per sampling unit		$a = A/N$	$b = B/N$		$\bar{x} = x/n$	$\bar{y} = y/n$
Variance*		$\sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i - a)^2$	$\sigma_b^2 = \frac{1}{N}\sum_{i=1}^{N}(b_i - b)^2$		$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$
Standard deviation		σ_a	σ_b		s_x	s_y

* Some authors define variances by means of $N-1$ and $n-1$ rather than N and n.

If (1), (2), (3), (4), and (5) are used as estimators of a, b, A, B, and ϕ, respectively, and if the results of the complete coverage were known, then one could, for any experiment, compute errors of sampling, such as

$$\Delta\bar{x} = \bar{x} - a, \tag{6}$$

$$\Delta\bar{y} = \bar{y} - b, \tag{7}$$

$$\Delta f = f - \phi. \tag{8}$$

It is an exciting fact that a single sample—provided that it is big enough (usually 25, 30, or more sampling units), and provided that it is designed properly and skillfully in view of possible statistical peculiarities of the frame and is carried out in reasonable conformance with the specifications—will make possible an estimate, based on theory to follow later, of the important characteristics of the distribution of sampling variation of all possible samples that could be drawn from the given equal complete coverage and processed by the specified sampling procedure.

Standard error and mathematical bias. We continue our conceptual experiment.

(*d*) Return the sample of n cards to the frame, and repeat steps (*b*) and (*c*) by the same sampling procedure, to form a new sample and new estimates $\bar{x}$, $\bar{y}$, f. Repeat these steps again and again, 10,000 times or more.

Explicit statements will now be confined to $\bar{x}$. The 10,000 experiments give an empirical distribution for $\bar{x}$, by which one may count the number of samples for which $\bar{x}$ lies between, for example, 100 and 109. We visualize an underlying theoretical distribution of $\bar{x}$, which the empirical distribution approaches closer and closer as the number of repetitions increases.

We are typically concerned with relationships between (*i*) the empirical distribution of $\bar{x}$ and (*ii*) the theoretical distribution of $\bar{x}$, for the given sampling procedure. Study of these relationships helps in the use of sampling for purposes of making estimates of characteristics of the frame.

Let a be the characteristic of the complete coverage that the generic symbol $\bar{x}$ estimates. Then if

$$E\bar{x} = a, \tag{9}$$

the sampling procedure is said to be unbiased. (The symbol E denotes *expectation*, the mean of the theoretical distribution of $\bar{x}$.) But if

$$E\bar{x} = a + C, \qquad C \neq 0, \tag{10}$$

the sampling procedure has the mathematical bias C. In any case, the variance of the distribution of $\bar{x}$ is

$$\sigma_{\bar{x}}^2 = E(\bar{x} - E\bar{x})^2, \tag{11}$$

and its square root, $\sigma_{\bar{x}}$, is the standard error of the sampling procedure for the estimator $\bar{x}$. Thus, a sampling procedure has, for any estimator, an expected value, a standard error, and possibly a mathematical bias (see the section "Possible bias in ratio estimators," below).

Uncertainty from accidental variation. Under the conditions stated above, the margin of uncertainty in the estimator $\bar{x}$ that is attributable to sampling and to small, independent, accidental variations, including random error of measurement (Type III in the next section), may be estimated, for a specified probability, as $t\hat{\sigma}_{\bar{x}}$, where $\hat{\sigma}_{\bar{x}}$ is an

estimator of $\sigma_{\bar{x}}$. The factor t depends on the probability level selected for the margin of uncertainty (which will in turn depend on the risks involved) and also on the number of degrees of freedom in the estimator $\hat{\sigma}_{\bar{x}}$. In large samples the distributions of most estimators are nearly normal, except for frames that exhibit very unusual statistical characteristics. The standard deviation, $\sigma_{\bar{x}}$, then contains nearly all the information regarding the margin of uncertainty of $\bar{x}$ that is attributable to accidental variation. Presentation of the results of a survey requires careful consideration when there is reason to question the approximate normality of estimators (Fisher 1956, p. 152; Shewhart 1939, p. 106).

Random selection. It is never safe to assume, in statistical work, that the sampling units in a frame are already mixed. A frame comes in layers that are different, owing to geographic origin or to order of production. Even blood, for example, has different properties in different parts of the body.

A random variable is the result of a random operation. A system of selection that depends on the use, in a standard manner, of an acceptable table of random numbers is acceptable as a random selection. Methods of selection that depend on physical mixing, shuffling, drawing numbers out of a hat, throwing dice, are not acceptable as random, because they have no predictable behavior. Neither are schemes that merely remove the choice of sampling units from the judgment of the interviewer. Pseudo-random numbers, generated under competent skill, are well suited to certain types of statistical investigation [*see* RANDOM NUMBERS].

Types of uncertainty in statistical data

All data, whether obtained by a complete census or by a sample, are subject to various types of uncertainty. One may reduce uncertainties in data by recognizing their existence and taking steps for improvement in future surveys. Sample design is an attempt to strike an economic balance between the different kinds of uncertainty. There is no point, for example, in reducing sampling error far below the level of other uncertainties.

Three types of uncertainty. The following discussion will differentiate three main types of uncertainty.

Type I. Uncertainty of Type I comprises built-in deficiencies, or structural limitations, of the frame, questionnaire, or method of test.

Any reply to a question, or any record made by an instrument, is only a response to a stimulus. What stimulus to apply is a matter of judgment. Deficiencies in the questionnaire or in the method of test may therefore arise from incomplete understanding of the problem or from unsuitable methods of investigation. Structural limitations are independent of the size or kind of sample. They are built in: a recanvass will not discover them, nor will calculation of standard errors or other statistical calculations detect them.

Some illustrations of uncertainty of Type I are the following:

(*a*) The frame may omit certain important segments of the universe.

(*b*) The questionnaire or method of test may fail to elicit certain information that is later found to be needed. The questionnaire may contain inept definitions, questions, and sequences. Detailed accounting will give results different from those given by mere inquiry about total expenditure of a family for some commodity; date of birth gives a different age from that given in answer to the simple question, How old are you? There may be differential effects of interviews depending on such variables as sex and race of the interviewer.

(*c*) Use of telephone or mail may yield results different from those obtained by personal interview.

(*d*) Judgments of coders or of experts in the subject matter may differ.

(*e*) The date of the survey has an important effect on some answers.

Type II. Uncertainty of Type II includes operational blemishes and blunders—for example:

(*f*) One must presume the existence of errors of a noncanceling nature (persistent omission of sampling units designated, persistent inclusion of sampling units not designated, persistent favor in recording results).

(*g*) One must presume the existence of bias from nonresponse.

(*h*) Information supplied by coders for missing or illegible entries may favor high or low values.

(*i*) There may be a large error, such as a unique blunder.

Type III. Uncertainty of Type III is caused by random variation. Repeated random samples drawn from the same frame will give different results. Besides, there are inherent uncorrelated, nonpersistent, accidental variations of a canceling nature that arise from inherent variability of investigators, supervisors, editors, coders, punchers, and other workers and from random error of measurement.

Standard error of an estimator. The standard error of a result includes the combined effects of all kinds of random variation, including differences within and between investigators, supervisors, coders, etc. By proper design, however, it is possible to get separate estimates of some of these differences.

A small standard error of a result signifies

(*i*) that the variation between repeated samples will be small and (*ii*) that the result of the sample agrees well with the equal complete coverage of the same frame. It usually tells little about uncertainties of Type II and never anything about uncertainties of Type I.

Limitations of statistical inference. Statistical inference (estimates, standard errors, statistical tests) refers only to the frame that was sampled and investigated. No statistical calculation can, by itself, detect or measure nonsampling errors, although side experiments or surveys may be helpful. No statistical calculation can detect defects in the frame. No statistical calculation can bridge the gap between the frame covered and the universe. This is as true of probability sampling as it is of judgment sampling, and it is true for a complete census of the frame as well.

Comparison of surveys. Substantial differences in results may come from what appear to be inconsequential differences in questionnaires or in methods of hiring, training, and supervision of interviewers and coders or in dates of interviewing. The sampling error in a sample is thus not established by comparison against a complete census unless the complete census is the equal complete coverage for the sample.

Recalls on people not at home. Many characteristics of people that are not at home at first call, or that are reluctant to respond, may be very different from the average. What is needed is response from everyone selected, including those that are hard to get. To increase the initial size of the sample is no solution. Calculations that cover a wide variety of circumstances show that the amount of information per dollar expended on a survey increases with the number of recalls, the only practicable limit being the time for the completion of the survey. Good sample design therefore specifies that four to six well-timed recalls be made or specifies that recalls continue until the level of response reaches a prescribed proportion. Special procedures, such as intensive subsampling of those not at home on the first or second call, have been proposed (see Leven 1932; Hansen & Hurwitz 1946; Deming 1960).

Surveys by post. One can often effect important economies by starting with a mail survey of a fairly large sample properly drawn from a given frame, then finishing with a final determined effort in the form of personal interviews on all or a fraction (one in two or one in three) of the people that failed to reply (Leven 1932). Mail surveys require a frame, in the form of a list of names with reasonably accurate addresses, and provision for keeping records of mailings and of returns. They are therefore especially adaptable to surveys of members of a professional society, subscribers to a journal, or subscribers to a service. [*For further discussion of mail surveys, see* ERRORS, *article on* NONSAMPLING ERRORS.]

Simple designs for enumerative purposes

The aim in an *enumerative* study is to count the number of people in an area that have certain characteristics or to estimate a quantity, perhaps their annual income, regardless of how they acquired these characteristics. The aim in an *analytic* study is to detect differences between classes or to measure the effects of different treatments.

For illustration consider a study of schizophrenics. One enumerative aim might be to estimate the number of children born to schizophrenic parents before onset of the disease or before the first admission of one of the parents to a hospital for mental diseases. Further aims of the same study might be analytic, such as to discover differences in fertility or in duration of hospitalization caused by different treatments, differences between communities, or differences between time periods.

The finite multiplier typified by $1/n - 1/N$ (to be seen later) appears in estimators for enumerative purposes. It has no place in estimators for analytic purposes.

Optimum allocation of effort for an enumerative aim may not be optimum for an analytic aim. Moreover, what is optimum for one enumerative characteristic may not be optimum for another. Hence, it will usually be necessary to compromise between competitive aims.

Enumerative aims will occupy most of the remaining space in this article.

The theory presented in this section is for the design commonly called *simple random sampling*. This is often a practicable design, and the theory forms a base for more complex designs.

A simple procedure of selection and some simple estimators. Definitions of "frame," "sample," and other terms were introduced above. In addition, it will be convenient to define the *coefficient of variation*. For the x-population and y-population of the frame, the coefficients of variation are defined as

$$(12) \qquad C_a = \frac{\sigma_a}{a}, \qquad C_b = \frac{\sigma_b}{b}, \qquad a > 0, b > 0.$$

In like manner, the symbol C_x denotes the coefficient of variation of the empirical or theoretical distribution of the random variable x. The square, C_x^2, of the coefficient of variation C_x is called the *rel-variance* of x. The coefficient of variation is especially useful for characteristics (such as height)

that are positive. It is often helpful to remember, for example, that $C_{\bar{x}} = C_x = C_{\bar{x}}$ because x, $\bar{x}$, and X are constant multiples of each other.

The procedure of selection specified earlier gives every member of the frame the same probability of selection as every other member, wherefore

$$E\bar{x} = a, \quad E\bar{y} = b. \tag{13}$$

That is, $\bar{x}$ and $\bar{y}$ are unbiased estimators of a and b, respectively. Moreover,

$$X = N\bar{x}, \quad Y = N\bar{y} \tag{14}$$

are unbiased estimators of A and B.

Often, a ratio such as

$$\phi = \frac{A}{B} = \frac{a}{b} \tag{15}$$

is of special interest. The sample gives

$$f = \frac{X}{Y} = \frac{x}{y} = \frac{\bar{x}}{\bar{y}} \tag{16}$$

as an estimator of ϕ. If the total y-population, B, is known from another source, such as a census, A may be estimated by the formula

$$X' = Bf. \tag{17}$$

This estimator X' is called a *ratio estimator*. It will be more precise than the estimator $X = N\bar{x}$ in (14) if the correlation between x_i and y_i is high. Other estimators will be discussed later (for example, regression estimators). Theory provides a basis for the choice of estimator.

Possible bias in ratio estimators. Necessary and sufficient conditions for there to be no bias in f as an estimator of ϕ are that $Ey \neq 0$ and that x_i/y_i and y_i be uncorrelated—that is, that $E[(x/y)y] = E(x/y)Ey$. In practice, if bias exists at all, it is usually negligible when the sample contains more than three or four sampling units.

Sampling with and without replacement. Usually, in the sampling of finite populations, one permits a sampling unit to come into the sample only once. In statistical language, this is sampling *without replacement*. Tests of physical materials are sometimes destructive, and a second test would be impossible. To draw without replacement, one simply disregards a random number that appears a second time (or uses tables of so-called random permutations).

There are circumstances, however, in which one accepts the random numbers as they come and permits a sampling unit to come into the sample more than once. This is sampling *with replacement*.

Hereafter, most equations will be written for sampling without replacement. It is a simple matter to drop the fraction $1/N$ from any formula to get the corresponding formula for sampling with replacement. Actually, in practice, samples are usually such a small part of the frame that the fraction $1/N$ is ignored, even though the sampling be done without replacement.

Variances. The variances of the estimator $\bar{x}$ derived from the sampling procedure described earlier are

$$\begin{cases} \text{without replacement} & \sigma_{\bar{x}}^2 = \dfrac{N-n}{N-1}\dfrac{\sigma^2}{n} \cong \left(\dfrac{1}{n} - \dfrac{1}{N}\right)\sigma^2, \\ \text{with replacement} & \sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}. \end{cases} \tag{18}$$

(The sign $\cong$ indicates an approximation that is sufficiently close in most practice.)

Similar expressions hold for $\bar{y}$. For the ratio $f = \bar{x}/\bar{y}$, the approximation

$$C_f^2 \cong \left(\frac{1}{n} - \frac{1}{N}\right)(C_a^2 + C_b^2 - 2abC_{ab}) \tag{19}$$

is useful if n be not too small. Here

$$C_{ab} = \frac{1}{Nab}\sum (a_i - a)(b_i - b) \tag{20}$$

is the rel-covariance of the x-population and y-population per sampling unit in the frame.

When the ratio estimator of the total x-population is derived as in eq. (17), eq. (19) gives the same approximation for $C_{\bar{x}}^2$.

Estimate of aggregate characteristic—number of units in class unknown. It often happens in practice that one wishes to estimate the aggregate value of some characteristic of a subclass of a group when the total number of units in the subclass is unknown. For example, one might wish to estimate the aggregate income of women aged 15 or over that live in a certain district, are gainfully employed, and have at least one child under 12 years old at home (this specification defines the universe). The number of women that meet this specification is not known. An estimate of the average income per woman of this specification, prepared from a sample, suffers very little from this gap in available knowledge, but an estimate of the total income of all such women is not so fortunate.

As an illustration, suppose that the frame is a serialized list of N women aged 15 or over and that the sample is a simple random sample of n of these women, drawn with replacement by reading out n random numbers between 1 and N. Informa-

tion on the n women is collected, and it is noted which ones belong to the specified subclass—that is, which ones live in a certain district, are gainfully employed, and have at least one child under 12 years old at home. Suppose that this number is n_s and that the average income of the n_s women is $\bar{x}_s$. Of course, n_s is a random variable with a binomial distribution.

What is the rel-variance of $\bar{x}_s$? Let C_s^2 be the rel-variance between incomes of the women in the frame that belong to the subclass. It is a fact that the conditional rel-variance of $\bar{x}_s$, for samples of size n_s of the specified subclass, will be C_s^2/n_s, just as if the women of this subclass had been set off beforehand into a separate stratum and a sample of size n_s had been drawn from it.

The conditional expected value of $\bar{x}_s$ over all samples of fixed size n_s in the subclass has moreover the convenient property of being the average income of all the women in the frame that belong to this subclass. It is for this reason that the conditional rel-variance of $\bar{x}_s$ is useful for assessing the precision of a sample at hand. For purposes of design, one uses the rel-variance of $\bar{x}_s$ over all samples of size n, which is $C_s^2E(1/n_s)$, or very nearly $C_s^2[1 + Q/nP]/nP$, where P is the proportion of all women 15 or over that meet the specification of the subclass, and $P + Q = 1$.

In contrast, any estimator, X_s, of the aggregate income of all the women in the specified subclass will not have such convenient properties as $\bar{x}_s$. The conditional expectation of X_s, for samples of size n_s, is not equal to the aggregate income of all the women in the frame that belong to the subclass. The conditional rel-variance of X_s for a sample of size n_s at hand, although equal to the conditional rel-variance of $\bar{x}_s$, therefore requires careful interpretation. Instead of attempting to interpret the conditional rel-variance of X_s, one may elect to deal with the variance of X_s in all possible samples of size n. Thus, if X_s is set equal to $(N/n)n_s\bar{x}_s$ (here N/n is used as an expansion factor equal to the reciprocal of the probability of selection), it is a fact that the rel-variance of X_s over all samples of size n will be $(C_s^2 + Q)/nP$ (see Deming 1960, p. 129).

The problem with X_s arises from the assumption that N_s, the number of women in the frame that meet the specification of the subclass, is unknown. If N_s were known, one could form the estimator $X_s = N_s\bar{x}_s$, which would have all the desirable properties of $\bar{x}_s$.

One way to reduce the variance of the total income, X_s, of the specified class is (1) to select from the frame a large preliminary sample, (2) by an inexpensive investigation to classify the units of the preliminary sample into two classes, those that belong to the specified class and those that do not, (3) to investigate a sample of the units that fell into the specified class, to acquire information on income. The preliminary sample provides an estimate of N_s, and the final sample provides an estimate of $\bar{x}_s$. The product gives the estimate $X_s = N_s\bar{x}_s$ for the total income in the specified class. (For the variance of X_s and for optimum sizes of samples, see Hansen, Hurwitz, & Madow, 1953, vol. 1, pp. 65 and 259.)

If, further, N were not known and only the probability, π, of selection, to be applied to every sampling unit in the frame, were known, both n and n_s will be random variables, and there will be a further inflation of the rel-variance of any estimator of the aggregate income of all the women in the specified subclass. Thus, if X_s be set equal to $n_s\bar{x}_s/\pi$ for such an estimator, then the unconditional rel-variance of X_s will be $(C_s^2 + 1)/nP$. The conditional rel-variance of $\bar{x}_s$, however, is still C_s^2/n_s.

It may be noted that for a small subclass there is little difference between $C_s^2 + Q$ and $C_s^2 + 1$.

Examples are common. Thus, one might read out a two-digit random number for each line of a register, following the rule that the item listed on a line will be drawn into the sample if the random number is 01. If counts from outside sources are not at hand or are not used, then the rel-variance of an estimator, X_s, of the total number or total value of any subclass of items on the register contains the factor $C_s^2 + 1$.

Use of thinning digits. Reduction of the probability of selection of units of specified characteristics (such as items of low value) through the use of thinning digits may produce either the factor $C_s^2 + Q$ or the factor $C_s^2 + 1$ in the rel-variance of an estimator of an aggregate, depending on the mode of selecting the units.

Estimates of variances. Estimates of variances are supplied by the sample itself, under proper conditions, as was discussed above. Some of the more important estimators follow, denoted by a circumflex ($\hat{}$). For the variance of $\bar{x}$,

$$(21) \qquad \hat{\sigma}_{\bar{x}}^2 = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum (x_i - \bar{x})^2,$$

with a similar expression for $\hat{\sigma}_{\bar{y}}^2$. For the covariance,

$$(22) \qquad \hat{\sigma}_{\bar{x}\bar{y}} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}).$$

Eqs. (21) and (22), with N infinite, were developed by Gauss (1823). These estimators are un-

biased; $\sqrt{\hat{\sigma}_{\bar{x}}^2}$ is a slightly biased estimator of $\sigma_{\bar{x}}$, but the bias is negligible for n moderate or large.

Under almost all conditions met in practice, one may set

$$t = \frac{\bar{x} - E\bar{x}}{\hat{\sigma}_{\bar{x}}} \tag{23}$$

and compare this quantity with tabulated values of t to find the margin of uncertainty in $\bar{x}$ for any specified probability. Such calculations give excellent approximations unless the distribution of sampling units in the frame is highly skewed. Extreme skewness may often be avoided by stratification (discussed below).

A useful approximate estimator for the rel-variance of $f = X/Y = \bar{x}/\bar{y}$ is

$$\hat{C}_f^2 \cong \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{(n-1)\bar{x}^2}\sum (x_i - fy_i)^2. \tag{24}$$

This formula is derived by combination of eqs. (19), (21), and (22). In accordance with a previous remark, one may take $\hat{C}_{X'} = \hat{C}_f$, where X' is the ratio-estimator of A as given by (17).

Size of frame usually not important. Because of the way in which N enters the variances, the size of the frame has little influence on the size of sample required for a prescribed precision, unless the sample is 20 per cent or more of the total frame. For instance, the sample required to reach a specified precision would be the same for the continental United States as for the Boston Metropolitan Area, on the assumption that the underlying variances encountered are about the same for the entire United States as for Boston.

Special form for attributes (0, 1 variate). In many studies a sampling unit gives only one of two possible observations, such as *yes* or *no*, *male* or *female*, *heads* or *tails*. The above equations then assume a simple form.

If each person in a frame is a sampling unit, and if $a_i = 1$ for *yes*, $a_i = 0$ for *no*, then the total x-population, A, in the frame is the total number of *yes* observations that would be recorded in the equal complete coverage, and a is the proportion *yes*, commonly denoted by p. The variance between the a_i in the frame is

$$\sigma_a^2 = pq, \tag{25}$$

where $p + q = 1$.

The random variate, x_i, will take the value 0 or 1;

$$x = \sum x_i \tag{26}$$

will be the number of *yes* observations in the sample, and

$$\hat{p} = x/n \tag{27}$$

will be the proportion *yes* in the sample. Replacement of $\bar{x}$ by $\hat{p}$ in previous equations shows that $\hat{p}$ is an unbiased estimator of the proportion *yes* in the frame and that

$$\sigma_{\hat{p}}^2 = \left(\frac{1}{n} - \frac{1}{N}\right)pq. \tag{28}$$

It is important to note that this variance is valid only if each sampling unit produces the value 0 or 1. It is not valid, for instance, for a sample of segments of area if there is more than one person per segment, or if the segments are clustered (as discussed below).

For an estimate of the variance of $\hat{p}$ (provided the sampling procedure meets the conditions stated) one may use

$$\hat{\sigma}_{\hat{p}}^2 = \left(\frac{1}{n} - \frac{1}{N}\right)\hat{p}\hat{q}, \tag{29}$$

where $\hat{p} + \hat{q} = 1$.

How good is an estimator of a variance? The variance of the estimator $\hat{\sigma}_{\bar{x}}^2$ in eq. (21) depends on the standardized fourth moment, β_2, of the frame and on the number of degrees of freedom for the estimator. Thus, if one defines

$$\beta_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{a_i - a}{\sigma}\right)^4, \tag{30}$$

then the rel-variance of the estimator $\hat{\sigma}_{\bar{x}}^2$ of eq. (21) will be $(\beta_2 - 1)/n$, which diminishes with n.

Systematic selection. A simple and popular way to spread the sample over the frame is to select every Zth unit, with a random start between 1 and Z, where $Z = N/n$. This is called systematic sampling with a single random start, and it is one form of patterned sampling. In certain kinds of materials, specifically those in which nearby sampling units are, on the average, more similar than units separated by a longer interval, systematic sampling will be slightly more efficient than stratified random sampling (Cochran 1946).

A disadvantage of systematic sampling with a single random start is that there is no strictly valid way to make a statistical estimate of the standard error of a result so obtained. This is because the single start is equivalent to the selection of only one sampling unit from the Z possible sampling units that could be formed. One may nevertheless, under proper conditions, get a useful approximation to the rel-variance by using the sum of squares of successive pairs. Eq. (21) with $n = 2$ and $N = Z$ gives the estimator

$$\hat{C}_{\bar{x}}^2 = \left(1 - \frac{2}{Z}\right)\frac{\sum (x_{i1} - x_{i2})^2}{[\sum (x_{i1} + x_{i2})]^2}, \tag{31}$$

where the summation runs over all pairs.

Hidden and unsuspected periodicities often turn up, and in such cases the above formula may give a severe underestimate or overestimate of the variance. For example, every nth household might be nearly in phase with the natural periodicity of income, rent, size of family, and other characteristics associated with corners and with the configuration of dwelling units within areas and within apartment houses. Systematic sampling of physical elements or of time intervals can lead to disaster.

A statistician will therefore justify a single random start and use of eq. (31) only if he has had long experience with a body of material.

Instead of a single random start between 1 and N/n, one may take two random starts between 1 and $2N/n$ and every $(2N/n)$th sampling unit thereafter. Extension to multiple random starts is obvious. Two or more random starts give a valid estimate of the variance. Fresh random starts every six or eight zones will usually reap any possible advantage of systematic sampling and will avoid uncertainty in estimation of the variance.

Efficiency of design. The relative efficiency of two sampling procedures, I and II, that give normally distributed estimators of some characteristic are by definition the ratio of the inverses of the variances of these estimators for the same size, n, of sample. In symbols (E denotes efficiency),

$$(32) \qquad \frac{E_{\text{I}}}{E_{\text{II}}} = \frac{\sigma_{\text{II}}^2}{\sigma_{\text{I}}^2}, \qquad n_{\text{I}} = n_{\text{II}}.$$

This concept of efficiency is due to Fisher (1922).

Comparison of costs is usually more important than comparison of numbers of cases. Let the costs be c_{I} and c_{II} for equal variances. Then

$$(33) \qquad \frac{E_{\text{I}}}{E_{\text{II}}} = \frac{c_{\text{II}}}{c_{\text{I}}}, \qquad \sigma_{\text{I}} = \sigma_{\text{II}}.$$

Comparison of efficiencies of estimators whose distributions depart appreciably from normality require special consideration.

Sampling moving populations. A possible procedure in sampling moving populations is to count and tag all the people visible from a number of enumerators' posts through a period of a day or a week (the first round) and then to repeat the count from the same or different posts some time later (the second round). The n_1 people counted and tagged in the first round constitute a mobile frame for the second round. If the number of people counted in the first round is n_1, and if the number counted in the second round is n_2, with an intersection of n_{12} for people counted in both rounds, then an estimator of the total number of mobile inhabitants in the whole area is $\hat{N} = n_1 n_2 / n_{12}$ (Yates 1949, p. 43; Deming & Keyfitz 1967).

More complex designs

Considerations of cost—clustering. The total cost of a survey includes cost of preparing the frames and cost of travel to the units selected. In some surveys it may be possible to get more information per unit cost by enlarging the sampling unit, a procedure commonly called *clustering*. One may, for example, define a sampling unit as comprising all the dwellings in a compact segment of area. Further, one may, with experience and care, subsample dwelling units from a selected cluster or select one member of a family where two or more members qualify for the universe. Again, in a national survey, one may restrict the sample to a certain number of counties that will come into the sample by a random process. Or, in a survey of a city, one may restrict the sample to a random selection of blocks.

Any such plan reduces the interviewer's expenses for travel and reduces the cost of preparing the frame. However, restriction of the sample usually also increases variances, unless the total number of households in the sample be increased as compensation. It should be remembered, though, that the actual precision obtained by the use of cluster sampling may be nearly as good as that obtained by an unrestricted random selection of the same number of dwelling units with no clustering.

Theory indicates the optimum balance between enlargement of the sampling unit and the number of sampling units to include in the sample. Obviously, the theory is more complex than that discussed in the last section. Stratification, ratio estimators, and regression estimators are additional techniques that, under certain conditions, yield further increases in efficiency (see below).

An example. The following illustration refers to a sample of a city: (*i*) Suppose that it has been determined in advance that for the main purposes of a survey the optimum size of areal unit is a compact group of five dwelling units, called a segment. (*ii*) A sampling unit within the city will consist of $\bar{n}$ segments from a larger number of segments contained in a block. The $\bar{n}$ segments of a sampling unit (if $\bar{n} > 1$) should be scattered over the block. A good way to effect this scatter is by a systematic selection. (*iii*) The m sampling units in the city will be selected by random numbers. For simplicity, assume that all blocks in the city contain an equal number, $\bar{N}$, of segments. Suppose that there are

M blocks in the whole city,
$\bar{N}$ segments in a block,
$N = M\bar{N}$ segments in the whole city,

$\bar{n}$ segments in a sampling unit,
$\bar{N}/\bar{n}$ sampling units in a block,
$M\bar{N}/\bar{n}$ or $N/\bar{n}$ sampling units in the whole city,
m sampling units in the sample.

Then if

$$\bar{x} = x/m\bar{n}, \tag{34}$$

one may take

$$X = N\bar{x} \tag{35}$$

for an estimator of the x-population in the whole city. For this estimator,

$$C_X = C_{\bar{x}}, \tag{36}$$

and

$$\text{var}\,\bar{x} = \frac{N - m\bar{n}}{N - \bar{n}}\left(\frac{\sigma_b^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N} - 1}\frac{\sigma_{iw}^2}{m\bar{n}}\right). \tag{37}$$

If m is small compared with M,

$$\text{var}\,\bar{x} \cong \frac{\sigma_b^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N} - 1}\frac{\sigma_{iw}^2}{m\bar{n}}. \tag{38}$$

If, also, $\bar{n}$ is small compared with $\bar{N}$,

$$\text{var}\,\bar{x} \cong \frac{\sigma_b^2}{m} + \frac{\sigma_{iw}^2}{m\bar{n}}. \tag{39}$$

Here σ_b^2 is the variance between blocks of the mean x-population per sampling unit, and σ_{iw}^2 is the average variance between sampling units within blocks.

Important principle in size of secondary unit. Suppose that the cost of adding one more block to the sample is c_1 (cost of maps, preparation, delineation of segments, travel) and that the cost of an interview in an additional sampling unit is c_2. Then the total cost of the survey will be

$$K = mc_1 + m\bar{n}c_2. \tag{40}$$

In eq. (37) var $\bar{x}$ will be at its minimum for a fixed cost K if

$$\bar{n} = \frac{\sigma_{iw}}{\sigma_b}\sqrt{\frac{c_1}{c_2}}, \qquad \text{optimum } \bar{n}. \tag{41}$$

This equation was derived by both L. H. C. Tippett and Shewhart, independently, in 1931.

Note that m does not appear in this equation. That is, the optimum value of $\bar{n}$ on the basis of the cost function (40) is independent of m, the number of sampling units in the sample (and very nearly independent of the number of blocks in the sample).

The optimum m is found by substituting the optimum $\bar{n}$ from eq. (41) into eq. (40) and solving for m. Of course, it is necessary to assume values for σ_{iw}/σ_b and for $\sqrt{c_1/c_2}$ to do this. (Because each sampling unit will usually fall in a different block, m will usually be exactly or nearly as large as the number of blocks in the sample.) Usual numerical values of $\sigma_{iw} : \sigma_b$ and of $c_1 : c_2$ lead to small values of $\bar{n}$ and to large values of m. Efficient design therefore usually requires a small sample from a block and dispersion of the sample into a large number of blocks.

Extension of this theory to a national sample, and to stratified designs and ratio estimators, leads to the same principle.

Variation in size of segment will increase var $\bar{x}$ by the factor $1 + C_u^2/n$, where C_u^2 is the rel-variance of the distribution of the number of dwelling units per segment. A similar factor, $1 + C_{N_i}^2/m$, measures the increase in var $\bar{x}$ from variation in the number of segments per block.

Replicated designs for ease in estimation of variance. Replication of a sample in two or more interpenetrating networks of samples will provide a basis for rapid calculation of a valid estimate of the standard error of any result, regardless of the complexity of the procedure of selection and of the formulas for the formation of estimates [Mahalanobis 1944; Deming 1950; 1960; *see also* INDEX NUMBERS, *article on* SAMPLING].

Stratified sampling

The primary aim of stratified sampling is to increase the amount of information per unit of cost. A further aim may be to obtain adequate information about certain strata of special interest.

One way to carry out stratification is to rearrange the sampling units in the frame so as to separate them into classes, or strata, and then to draw sampling units from each class. The goal should be to make each stratum as homogeneous as possible, within limitations of time and cost. Stratification is equivalent to blocking in the design of an experiment. It is often a good plan (*i*) to draw a preliminary sample from the frame without stratification; (*ii*) to classify into strata the units in the preliminary sample, and (*iii*) to draw, for the final sample, a prescribed number of sampling units from each stratum so formed. Step (*i*) will sometimes require an inexpensive investigation or test of every sampling unit in the preliminary sample to determine which stratum it belongs to.

Stratification is one way to make use of existing information concerning the frame other than the information obtained from investigating the sampling units in the final sample itself. Other ways to use existing information are through ratio estimators and regression estimators (see below).

In practice a frame is to some extent naturally

Table 2 — Notation and definitions for the frame ($M = 2$ strata)

STRATUM	NUMBER OF SAMPLING UNITS: In the frame	NUMBER OF SAMPLING UNITS: In the sample	STRATUM'S PROPORTION OF SAMPLING UNITS IN THE FRAME	POPULATION: Average per sampling unit in the stratum	POPULATION: Total in the stratum	BETWEEN THE POPULATIONS OF THE SAMPLING UNITS WITHIN THE STRATUM: Standard deviation	BETWEEN THE POPULATIONS OF THE SAMPLING UNITS WITHIN THE STRATUM: Variance
1	N_1	n_1	$P_1 = \frac{N_1}{N}$	a_1	$A_1 = N_1 a_1$	σ_1	σ_1^2
2	N_2	n_2	$P_2 = \frac{N_2}{N}$	a_2	$A_2 = N_2 a_2$	σ_2	σ_2^2
Total for the frame	N	n	1	—	A	—	—
Unweighted average per stratum	$\bar{N} = \frac{N}{M}$	$\bar{n} = \frac{n}{M}$	$\frac{1}{M}$	—	$\bar{A} = \frac{A}{M}$		—
Weighted average per sampling unit	—	—	—	$a = \frac{A}{N}$	—	$\bar{\sigma}_w$	σ_w^2

Source: Deming 1960, p. 286.

stratified to begin with. Thus, areas in geographic order usually are already pretty well stratified in respect to income, occupation, density of population, tastes of the consumer, and other characteristics. No frame arrives thoroughly mixed, and any plan of sampling should be applied by zones, so as to capture the natural stratification. Theory serves as a guide to determine whether further stratification would be profitable.

Plans of stratification for enumerative studies. Several plans of stratified sampling for enumerative studies will now be described.

The notation and definitions to be used in this discussion are given in tables 2 and 3. (Note that $\bar{N}$ and $\bar{n}$ are defined differently here than they were earlier.) These tables are presented in terms of two strata ($M = 2$), but extension to a greater number of strata follows obviously. The following additional definitions are needed:

$$(42) \qquad \sigma_R^2 = Q_1\sigma_1^2 + Q_2\sigma_2^2 + Q_3\sigma_3^2,$$

the average reverse variance between sampling units within strata, and

$$(43) \qquad \bar{\sigma}_R = Q_1\sigma_1 + Q_2\sigma_2 + Q_3\sigma_3,$$

the average reverse standard deviation between sampling units within strata, where $Q_i + P_i = 1$.

Plan A (no stratification): The scheme of sampling described above will be designated plan A. It is needed here for comparison, and also because it constitutes the basis for selection from any stratum.

Note that in plan A, as in plans B, D, F, and H, below, all the sampling units in the frame have equal probability of selection, namely n/N, wherefore $E\bar{x} = a$ and $EX = A$.

P_i known—whole frame classified. Two sampling plans for which the proportions in each stratum are known (or ascertainable) and the the whole frame is classified will now be described.

Plan B (proportionate sampling): Decide with the help of eq. (47) the size, n, of the sample required. Compute next

$$(44) \qquad n_i = nN_i/N = nP_i.$$

Draw by random numbers, as in plan A, a sample of size n_i from stratum i. Investigate every member of the sample, and calculate

$$(45) \qquad X_1 = N_1x_1/n_1, \qquad X_2 = N_2x_2/n_2, \qquad \text{etc.}$$

Table 3 — Notation and definitions for the sample

Stratum	Population in the sample	Mean population per sampling unit	Estimated total population	Variance of this estimator*
1	x_1 x-population in stratum 1	$\bar{x}_1 = \frac{x_1}{n_1}$	$X_1 = N_1 \frac{x_1}{n_1}$	$\text{var } X_1$
2	x_2 x-population in stratum 2	$\bar{x}_2 = \frac{x_2}{n_2}$	$X_2 = N_2 \frac{x_2}{n_2}$	$\text{var } X_2$
Sum	x	—	X	$\text{var } X$

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

Source: Deming 1960, p. 287.

(For simplicity, most formulas will henceforth be written for two strata, in conformance with tables 2 and 3. Extension to more strata is obvious.) Here, $n_i/N_i = n/N$, wherefore

$$(46) \qquad \begin{aligned} X &= X_1 + X_2 \\ &= \frac{N}{n}(x_1 + x_2) \\ &= N\frac{x}{n} = N\bar{x} \end{aligned}$$

and

$$(47) \qquad \operatorname{var}\bar{x} = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_w^2, \qquad \text{plan B.}$$

The n_i of eq. (44) and later expressions will not in general be integers. In practice one uses the closest integer; the effects on variance formulas are usually completely negligible.

Plan C (Neyman sampling): Decide with the help of eq. (49) the size, n, of the sample required. Compute next the Neyman allocation (Neyman 1934),

$$(48) \qquad n_i = nP_i\sigma_i/\bar{\sigma}_w.$$

Draw by random numbers, as in plan A, a sample of size n_i from stratum i. Investigate every member of the sample. Form estimators X_1, X_2, and $X = X_1 + X_2$. Form $\bar{x} = X/N$ for an unbiased estimator of a. Here

$$(49) \qquad \operatorname{var}\bar{x} = \frac{(\bar{\sigma}_w)^2}{n} - \frac{\sigma_w^2}{N}, \qquad \text{plan C.}$$

The Neyman allocations are the optimal n_i for minimizing $\operatorname{var}\bar{x}$ when the P_i are known.

P_i known—only a sample classified. One may, in appropriate circumstances, require only the classification of a preliminary sample drawn from the frame. The decision hinges on the costs of classification and the expected variances of the plans under consideration.

Plan D: Decide with the help of eq. (50) the size, n, of the sample required. Draw the sample as in plan A. Classify the sampling units into strata. The number, n_i, of sampling units drawn from stratum i will be a random variable. Carry out the investigation of every unit of the sample. Form X_1, X_2, X, and $\bar{x}$ as in plan B. Then

$$(50) \qquad \operatorname{var}\bar{x} = \overbrace{\left(\frac{1}{n} - \frac{1}{N}\right)\left(\sigma_w^2 + \frac{1}{n}\sigma_B^2\right)}^{\text{plan B}}, \qquad \text{plan D.}$$

Plan E: Decide with the help of eq. (52) the size, n, of the final sample. Draw by random numbers a preliminary sample of size n'. Thin (reduce) by random numbers the strata of the preliminary sample to reach the Neyman ratios

$$(51) \qquad \frac{n_1}{n_1'} : \frac{n_2}{n_2'} : \cdots = \sigma_1 : \sigma_2 : \cdots$$

and simultaneously the total sample, n. Here n_1', n_2', etc., are the sizes of the preliminary sample in the several strata, and n_1, n_2, etc., are the sizes of the final sample. For greatest economy, choose n' so that one stratum will require no thinning. Carry out the investigation of every unit of the final sample. Form the estimators X_1, X_2, and $X = X_1 + X_2$. Then $\bar{x} = X/N$ will again be an unbiased estimator of a, but now

$$(52) \qquad \begin{aligned} \operatorname{var}\bar{x} &= \overbrace{\frac{(\bar{\sigma}_w)^2}{n} - \frac{\sigma_w^2}{N}}^{\text{plan C}} + \frac{1}{n}\left(\frac{1}{n'} - \frac{1}{N}\right)\bar{\sigma}_w\bar{\sigma}_B \\ &\cong \frac{1}{n}\left[(\bar{\sigma}_w)^2 + \frac{1}{n'}\bar{\sigma}_w\bar{\sigma}_B\right], \end{aligned} \qquad \text{plan E}$$

the latter form useful if N is large relative to n'.

Sequential classification of units into strata. We now describe two plans in which the sample-sizes, n_i, are reached sequentially, with considerable saving under appropriate conditions.

Plan F: Determine the desired sample-sizes, n_i, as in plan B. Draw by random numbers one unit at a time from the frame, and classify it into its proper stratum. Continue until the quotas, n_i, are all filled. Form X as in plan B; $\operatorname{var}\bar{x}$ will be the same as for plan B.

Plan G: This is the same as plan F except that the sample sizes, n_i, are fixed as in plan C. Form X as in plan C; $\operatorname{var}\bar{x}$ will be the same as for plan C.

P_i not known in advance. When the proportions, P_i, in the frame are unknown, estimates thereof must come from a sample, usually a preliminary sample of size $N' > n$, where n is the size of the final sample.

Plan H: Decide with the help of eq. (55) the size, n, for the final sample. Compute the optimum size, N', of the preliminary sample by the formula

$$(53) \qquad \frac{n}{N'} = \frac{\sigma_w}{\sigma_b}\sqrt{\frac{c_1}{c_2}},$$

where c_1 is the average cost of classifying a sampling unit in the preliminary sample, and c_2 is the average cost of the final investigation of one sampling unit.

The procedure is to draw as in plan A a preliminary sample of size N' and to classify it into strata. Treat the preliminary sample as a frame of size N'. Then thin all strata of the preliminary sample proportionately to reach the final total size, n.

Carry out the investigation of every sampling unit in the final sample. An unbiased estimator of a is

$$\bar{x} = \frac{x}{n}, \tag{54}$$

where x is the total x-population in the sample. Then

$$\text{var}\,\bar{x} \cong \frac{\sigma_w^2}{n} + \frac{\sigma_b^2}{N'} = \frac{\sigma^2}{N'} + \left(\frac{1}{n} - \frac{1}{N'}\right)\sigma_w^2, \tag{55}$$

plan H,

is an excellent approximation if N be large relative to N'.

Plan I: Decide with the help of eq. (59) the size, n, for the final sample. Compute the optimum size, N', of the preliminary sample, using the equation (Neyman 1938)

$$\frac{n}{N'} = \frac{\bar{\sigma}_w}{\sigma_b}\sqrt{\frac{c_1}{c_2}}. \tag{56}$$

Draw as in plan A a preliminary sample of size N'. Classify it as in plan H. Thin the strata differentially to satisfy the Neyman ratios

$$\frac{n_1}{N_1'} : \frac{n_2}{N_2'} : \cdots = \sigma_1 : \sigma_2 : \cdots \tag{57}$$

and to reach the desired final total sample-size, n. Carry out the investigation of every sampling unit in the final sample. An unbiased estimator of a is

$$\bar{x} = \frac{1}{N'}\left(\frac{N_1'}{n_1}x_1 + \frac{N_2'}{n_2}x_2\right), \tag{58}$$

for which

$$\text{var}\,\bar{x} \cong \frac{(\bar{\sigma}_w)^2}{n} + \frac{\sigma_b^2}{N'}, \tag{59}$$

plan I,

is an excellent approximation if N be large relative to N' and to n.

One may use plan F or plan G in combination with plan H or plan I to reap the benefit of many strata without actually classifying the entire preliminary sample, N' (Koller 1960).

Gains of stratified sampling. Gains of stratified sampling can be evaluated by comparing variances. Denote by A, B, and C the variances of the estimators of a calculated by the plans A, B, and C. Then

$$\frac{A-B}{A} = \frac{\sigma^2 - \sigma_w^2}{\sigma^2} = \frac{\sigma_b^2}{\sigma^2} = \sum_{i<j} P_i P_j \left(\frac{a_j - a_i}{\sigma}\right)^2, \tag{60}$$

$$\frac{B-C}{B} \cong \frac{\sigma_w^2 - (\bar{\sigma}_w)^2}{\sigma_w^2} = \sum_{i<j} P_i P_j \left(\frac{\sigma_j - \sigma_i}{\sigma_w}\right)^2. \tag{61}$$

For example, if $P_1 = .6$, $P_2 = .4$, and $\sigma_w^2 = .8\sigma^2$, $(A-B)/A$ would be $(1-.8)/1 = .2$, meaning that 100 interviews selected according to plan B would give rise to the same variance as 125 selected according to plan A.

The gains of plans F and G over plan A are the same as the gains of plans B and C over plan A. The average gains in repeated trials of plans D and E are less. If σ_R^2 and $\bar{\sigma}_w\bar{\sigma}_R$ are large, plans D and E will usually not be good choices. For large samples, however, in circumstances where σ_R^2 and $\bar{\sigma}_w\bar{\sigma}_R$ are not large, the gains of plans D and E may be almost equal to the gains of plans B and C, at considerably less cost.

Eqs. (60) and (61) show that the gain to be expected from the proposed formation of a new stratum, i, will not be impressive unless its proportion, P_i, be appreciable, or unless its σ_i or its a_i be widely divergent from the average.

Stratification to estimate over-all ratio. The case to be used for illustrating stratified sampling to estimate an over-all ratio consists of three strata: stratum 1 for large units (for example, high incomes or large farms), stratum 2 for medium-sized units, and stratum 3 for small units. Here stratum 1 is to be covered 100 per cent; obvious modifications take care of the case in which stratum 1 is not sampled completely.

First take as an estimator of ϕ

$$f = \frac{X}{Y} = \frac{A_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{B_1 + N_2\bar{y}_2 + N_3\bar{y}_3}, \tag{62}$$

in the notation of tables 1 and 2, with B_i as the value of the y-characteristic in stratum i of the frame. Optimum allocation to strata 2 and 3 is very nearly reached if both

$$n_2 = \frac{B_2 s_2}{B_2 s_2 + B_3 s_3} \tag{63a}$$

and

$$n_3 = \frac{B_3 s_3}{B_2 s_2 + B_3 s_3}, \tag{63b}$$

wherein s_2 and s_3 are the standard deviations of the ratio of x to y in strata 2 and 3.

If, as is often the case, s_2 and s_3 do not differ much, or if little is known about them in advance, one can still make an important gain in efficiency by setting $n_2 : n_3 = B_2 : B_3$ or $n_2 : n_3 = A_2 : A_3$.

Another estimator of the ratio ϕ is

$$f = P_1 f_1 + P_2 f_2 + P_3 f_3, \tag{64}$$

wherein $P_i = B_i/B$ and $f_i = \bar{x}_i/\bar{y}_i$. This estimator is sometimes preferred when f_i varies greatly from stratum to stratum, and when there can be no trouble with small denominators. The allocation of sample for this estimator is, for practical purposes, the same as in eq. (63) (Cochran [1953] 1963, p. 175; Hansen et al. 1953, vol. 1, p. 209).

Sequential adjustment of size of sample. It is sometimes possible, when decision on the size of sample is difficult, or when time is short, to break the sample in advance into two portions, 1 and 2, each being a valid sample of the whole frame. Portion 1 is definitely to be carried through to completion, but portion 2 will be used only if required. This may be called a two-stage sequential method. It is practicable where the investigation is to be carried out by a small number of experts that will stay on the job as long as necessary but not where a field force must be engaged in advance for a definite period.

Modifications for differing costs. If investigating a sampling unit in a particular stratum is three or more times as costly as the average investigation, it may be wise to decrease the sample in the costly stratum and to build up the sample in other strata (Deming 1960, p. 303).

Considerations for planning. In order to plan a stratified sample, certain assumptions are necessary. Fair approximations to the relevant ratios, such as $\sigma_{ir} : \sigma$, $\bar{\sigma}_{ic} : \sigma$, $\bar{\sigma}_{ic} : \sigma_b$, $\sqrt{c_1 : c_2}$, will provide excellent allocation. On the other hand, bad approximations to these ratios, or failure to use theory at all, can lead to serious losses.

The required good approximations to these ratios may come from prior experience, or from probing the knowledge of experts in the subject matter. For example, the distribution of intelligence quotients in the stratum between 90 and 110, if rectangular, would provide $\sigma^2 = (110 - 90)^2/12$, or 33, whence $\sigma = 5.7$. Other shapes have other variances, but shape is fortunately not critical (Deming 1950, p. 262; 1960, p. 260). A stratum with very high values should be set off for special treatment and possibly sampled 100 per cent.

Stratification for analytic studies. As mentioned earlier, the aim in an analytic study is to detect differences between classes or to measure the effects of different treatments.

The general formula for the variance of the difference between two means, $\bar{x}_A$ and $\bar{x}_B$, derived from independent samples of sizes n_A and n_B drawn by random numbers singly and without stratification from, for example, two groups of patients, A and B, is

$$(65) \qquad \text{var}\,(\bar{x}_A - \bar{x}_B) = \sigma_A^2/n_A + \sigma_B^2/n_B,$$

wherein σ_A^2 and σ_B^2 are the respective variances between the patients within the two groups.

For such analytic studies the optimum allocation of skill and effort is found by setting

$$(66) \qquad \frac{n_A}{n_B} = \frac{\sigma_A}{\sigma_B}\sqrt{\frac{c_B}{c_A}},$$

wherein c_A and c_B are the costs per case. Note that the sizes of the groups do not enter into this formula and that it is different from the optimum allocation in enumerative problems.

In many analytic studies σ_A and σ_B will be about equal, and so will the costs c_A and c_B. In such circumstances, the best allocation is

$$(67) \qquad n_A = n_B.$$

Regression estimators

We have already seen reduction in variance resulting from use of prior or supplementary knowledge concerning the frame. Use of prior knowledge of N to form the estimator $X = N\bar{x}$ is an instance. Prior knowledge of B to form the ratio estimator is another instance. This section, on *regression estimators*, describes other ways to use prior or supplementary knowledge concerning the frame. Regression estimators include the simple estimator, $\bar{x}$, and the ratio estimator, fb, as special cases, but they also include many other estimators, some of them highly useful. Like the ratio estimator of a total, these additional estimators are applicable only if independent and fairly reliable information is available about the y-population in the frame. Any estimator that takes advantage of supplementary information may have considerable advantage over the simple estimator, $\bar{x}$, if the correlation, ρ, between x_i and y_i is high, but this condition is not in itself sufficient.

Specific forms of regression estimators. Assume simple random selection and write the regression estimator in the form

$$(68) \qquad \bar{x}_i = \bar{x} + m_i(b - \bar{y}),$$

wherein b is known independently from some source such as a census. The subscript i on $\bar{x}_i$ here differentiates the several specific forms of regression estimators.

Regression estimators are closely allied with the analysis of covariance [*see* LINEAR HYPOTHESES, *article on* ANALYSIS OF VARIANCE]. The four cases to be considered here are taken largely from Hansen, Hurwitz, and Madow (1953).

Simple estimator. If m_i is taken as zero, the regression estimator obtained is $\bar{x}_1 = \bar{x}$, seen earlier. This procedure makes no use of supplemental information. Under the assumption that N is large relative to n, the variance of this estimator is

$$(69a) \qquad \text{var}\,\bar{x}_1 = \sigma_{\bar{x}}^2;$$

likewise,

$$(69b) \qquad \text{var}\,\bar{y}_1 = \sigma_{\bar{y}}^2.$$

Difference estimator. The estimator $\bar{x}_2$, often called the difference estimator, is practicable if

Table 4 — Rel-variances when estimator of b is subject to sampling error

Estimator	Case I: sample of size n is drawn as a subsample of n'	Case II: samples of size n and n' are independent
$\bar{x}_1$	$C_{\bar{x}}^2$	Same as in Case I
$\bar{x}_2$	$C_{\bar{x}}^2[1 - \rho^2(1 - e^2)(1 - n/n')]$	$C_{\bar{x}}^2[1 - \rho^2(1 - e^2) + \rho^2(1 + e)^2 n/n']$
$\bar{x}_3$	$C_{\bar{x}}^2[1 - \rho^2(1 - n/n')]$	Same as in Case I
$\bar{x}_4$	$C_{\bar{x}}^2 - (2\rho C_{\bar{x}} C_{\bar{y}} - C_{\bar{y}}^2)(1 - n/n')$	$C_{\bar{x}}^2 - (2\rho C_{\bar{x}} C_{\bar{y}} - C_{\bar{y}}^2)(1 - n/n') + (2C_{\bar{y}} n/n')(C_{\bar{y}} - \rho C_{\bar{x}})$

prior knowledge (such as prior surveys of a related type) provides a rough approximation to the regression coefficient $\beta = \rho\sigma_{\bar{x}}/\sigma_{\bar{y}}$. This estimator is

$$\bar{x}_2 = \bar{x} + m_2(b - \bar{y}), \tag{70}$$

where m_2 is any approximate slope not derived from the sample under consideration. The variance of $\bar{x}_2$ is

$$\begin{aligned} \operatorname{var} \bar{x}_2 &= \sigma_{\bar{x}}^2(1 - \rho^2) + \sigma_{\bar{y}}^2(m_2 - \beta)^2 \\ &= \sigma_{\bar{x}}^2(1 - \rho^2 + \rho^2 e^2), \end{aligned} \tag{71}$$

where $\beta = \rho\sigma_{\bar{x}}/\sigma_{\bar{y}}$ and $e = (m_2 - \beta)/\beta$. (Note that $\rho e = (\sigma_{\bar{y}}/\sigma_{\bar{x}})(m_2 - \beta)$ even if $\rho = 0$.)

Least squares regression estimator. If m_i is chosen as

$$m_3 = \frac{\sigma_{\bar{x}\bar{y}}}{\sigma_{\bar{y}}^2} = \rho\,\frac{\sigma_{\bar{x}}}{\sigma_{\bar{y}}}, \tag{72}$$

then the equation

$$\bar{x}_3 = \bar{x} + m_3(b - \bar{y}) \tag{73}$$

gives the so-called least squares regression estimator. The variance of this estimator is

$$\operatorname{var} \bar{x}_3 = \sigma_{\bar{x}}^2(1 - \rho^2) + R. \tag{74}$$

Here R is a remainder in the Taylor series involving $1/n^2$ and higher powers; this remainder will be negligible if n is large.

Ratio estimator. If m_i is chosen as

$$m_4 = \bar{x}/\bar{y} = f, \tag{75}$$

the ratio estimator is

$$\bar{x}_4 = fb. \tag{76}$$

The variance of $\bar{x}_4$ is

$$\operatorname{var} \bar{x}_4 = \sigma_{\bar{x}}^2(1 + C_{\bar{y}}^2/C_{\bar{x}}^2 - 2\rho C_{\bar{y}}/C_{\bar{x}}) + R', \tag{77}$$

R' being another remainder. For large n and for $C_{\bar{x}} \cong C_{\bar{y}}$,

$$\operatorname{var} \bar{x}_4 \cong 2\sigma_{\bar{x}}^2(1 - \rho). \tag{78}$$

It follows that for large n and for $m_2 \cong \beta$ and $C_{\bar{x}} \cong C_{\bar{y}}$,

$$\frac{\operatorname{var} \bar{x}_4}{\operatorname{var} \bar{x}_2} \cong \frac{2}{1 + \rho}. \tag{79}$$

Comparison of regression estimators. If the correlation, ρ, between x_i and y_i is moderate or high, but the line of regression of x on y misses the origin by a wide margin, then the estimator $\bar{x}_3$ will show substantial advantages over $\bar{x}_4$ and $\bar{x}_1$. If the y-variate shows relatively wide spread (that is, if $C_{\bar{y}}$ is much greater than $C_{\bar{x}}$), the ratio estimator $\bar{x}_4$ may be far less precise than the simple estimator $\bar{x}_1 = \bar{x}$, even when ρ is high, especially if the line of regression misses the origin by a wide margin. On the other hand, if the line of regression passes through the origin ($\rho C_{\bar{x}} = C_{\bar{y}}$), or nearly through it, $\bar{x}_3$ and $\bar{x}_4$ will have about the same variance, but $\bar{x}_4$ may be much easier to compute.

Estimator of b subject to sampling error. It often happens that the y-population per sampling unit is not known with the reliability of a census but comes instead from another and bigger sample. This circumstance introduces additional terms into the variances. Let n be the size of the present sample and n' the size of the sample that provides the estimate of b. We suppose that the variance of this estimate of b is $n\sigma_{\bar{y}}^2/n'$. The resulting variances of the regression estimators are shown in Table 4.

W. EDWARDS DEMING

BIBLIOGRAPHY

BOWLEY, ARTHUR L. (1901) 1937 *Elements of Statistics.* 6th ed. New York: Scribner; London: King.

CHEVRY, GABRIEL 1949 Control of a General Census by Means of an Area Sampling Method. *Journal of the American Statistical Association* 44:373–379.

COCHRAN, WILLIAM G. 1946 Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations. *Annals of Mathematical Statistics* 17:164–177.

COCHRAN, WILLIAM G. (1953) 1963 *Sampling Techniques.* 2d ed. New York: Wiley.

DALENIUS, TORE 1962 Recent Advances in Sample Survey Theory and Methods. *Annals of Mathematical Statistics* 33:325–349.

DEMING, W. EDWARDS (1943) 1964 *Statistical Adjustment of Data.* New York: Dover.

DEMING, W. EDWARDS 1950 *Some Theory of Sampling.* New York: Wiley.

DEMING, W. EDWARDS 1960 *Sample Design in Business Research.* New York: Wiley.

DEMING, W. EDWARDS; and KEYFITZ, NATHAN 1965 Theory of Surveys to Estimate Total Population. Volume 3, pages 141–144 in World Population Conference, Belgrade, August 30–September 10, 1965, *Proceedings.* New York: United Nations.

FISHER, R. A. (1922) 1950 On the Mathematical Foundations of Theoretical Statistics. Pages 10.307a–10.368 in R. A. Fisher, *Contributions to Mathematical Statistics*. New York: Wiley. → First published in Volume 222 of the *Philosophical Transactions*, Series A, of the Royal Society of London.

FISHER, R. A. (1956) 1959 *Statistical Methods and Scientific Inference*. 2d ed., rev. New York: Hafner.

GAUSS, CARL FRIEDRICH 1823 *Theoria combinationis observationum erroribus minimis obnoxiae*. Göttingen (Germany): Dieterich. → A French translation was published in Gauss' *Méthode des moindres carrés* (1855). An English translation of the French was prepared as *Gauss's Work (1803–1826) on the Theory of Least Squares*, by Hale F. Trotter; Statistical Techniques Research Group, Technical Report, No. 5, Princeton Univ., 1957.

HANSEN, MORRIS H.; and HURWITZ, WILLIAM N. 1943 On the Theory of Sampling From Finite Populations. *Annals of Mathematical Statistics* 14:333–362.

HANSEN, MORRIS H.; and HURWITZ, WILLIAM N. 1946 The Problem of Non-response in Sample-surveys. *Journal of the American Statistical Association* 41: 517–529.

HANSEN, MORRIS H.; HURWITZ, WILLIAM N.; and MADOW, WILLIAM G. 1953 *Sample Survey Methods and Theory*. 2 vols. New York: Wiley.

KISH, LESLIE 1965 *Survey Sampling*. New York: Wiley. → A list of errata is available from the author.

KOLLER, SIEGFRIED 1960 Aussenhandelsstatistik: Untersuchungen zur Anwendung des Stichprobenverfahrens. Pages 361–370 in Germany (Federal Republic), Statistisches Bundesamt, *Stichproben in der amtlichen Statistik*. Stuttgart (Germany): Kohlhammer.

LEVEN, MAURICE 1932 *The Income of Physicians: An Economic and Statistical Analysis*. Univ. of Chicago Press.

MAHALANOBIS, P. C. 1944 On Large-scale Sample Surveys. Royal Society of London, *Philosophical Transactions* Series B 231:329–451.

MAHALANOBIS, P. C. 1946 Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society* Series A 109: 326–378. → Contains eight pages of discussion.

MOSER, C. A. 1949 The Use of Sampling in Great Britain. *Journal of the American Statistical Association* 44:231–259.

NEYMAN, JERZY 1934 On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society* Series A 97:558–606.

NEYMAN, JERZY 1938 Contribution to the Theory of Sampling Human Populations. *Journal of the American Statistical Association* 33:101–116. → See especially equation 49 on page 110.

QUENOUILLE, M. H. 1959 *Rapid Statistical Calculations*. London: Griffin; New York: Hafner. → Pages 5–7 show estimators of the standard deviation by use of the range.

SATTERTHWAITE, F. E. 1946 An Approximate Distribution of Estimates of Variance Components. *Biometrics* 2:110–114.

SHEWHART, WALTER A. 1939 *Statistical Method From the Viewpoint of Quality Control*. Washington: U.S. Department of Agriculture, Graduate School.

STEPHAN, FREDERICK F. 1936 Practical Problems of Sampling Procedures. *American Sociological Review* 1:569–580.

STEPHAN, FREDERICK F. 1948 History of the Uses of Modern Sampling Procedures. *Journal of the American Statistical Association* 43:12–39.

STEPHAN, FREDERICK F.; and MCCARTHY, PHILIP J. 1958 *Sampling Opinions: An Analysis of Survey Procedure*. New York: Wiley.

STUART, ALAN 1962 *Basic Ideas of Scientific Sampling*. London: Griffin; New York: Hafner.

SYMPOSIUM ON CONTRIBUTIONS OF GENETICS TO EPIDEMIOLOGIC STUDIES OF CHRONIC DISEASES, ANN ARBOR, MICHIGAN, *1963* 1965 *Genetics and the Epidemiology of Chronic Diseases*. U.S. Public Health Service, Publication No. 1163. Washington: Government Printing Office. → See especially "Selection techniques for rare traits," by Leslie Kish, pages 165–176.

YATES, FRANK (1949) 1960 *Sampling Methods for Censuses and Surveys*. 3d ed., rev. & enl. New York: Hafner. → Earlier editions were also published by Griffin.