5

Some Statistical Principles for Efficient Design of Surveys and Experiments

W. EDWARDS DEMING, PH.D.

Reprinted from Kallmann's GENETICS IN PSYCHIATRY (New York, 1962)

5

Some Statistical Principles for Efficient Design of Surveys and Experiments

W. EDWARDS DEMING, PH.D.

SKILLS AND RESOURCES for research are always limited. Hence, an ever-present problem in the planning stages of any investigation is how to use the available skills and resources to the best advantage. To the statistician, this is the problem of efficient design, namely, to maximize the amount of information per unit cost, or per man-day, within any restrictions imposed. Statistical theory is the statistician's tool for design.

AIM AND SCOPE OF THE REPORT

It may be pointed out that *sample design* is much more than a procedure for the selection of cases for investigation. Sample design includes, in the planning stages, organizational assistance to the experts in the subject matter (human genetics, biochemistry, agricultural science, etc.). The aim is to formulate the problems of the survey in statistical terms so that the information to be obtained from the survey will be maximally useful.

Sample design includes the choice of sampling units as well as the search for procedures and formulas for calculating estimates of whatever characteristics constitute the aim of the survey. It requires, moreover, plans of selection and estimation that will permit, without undue labor and expense, calculation of standard errors and tests of significance. It includes tests to measure the biases of nonresponse, tests to detect and measure the effects of human errors in carrying out the prescribed procedures, including the recording, coding, punching, and tabulating.

Finally, statistical theory and experience are helpful in the interpretation of results. At the completion of the study, the investigator wishes to draw whatever conclusions seem warranted. Statistically, it will have to be determined, therefore, whether certain hypotheses are apparently confirmed, or left in doubt, or require re-examination.

The prescription of aims for a study is never statistical. No amount of statistical knowledge will indicate the need of an investigation in medical research, nor in any other kind of research. The perception of a problem and the initial prescription of aims for the survey come from substantive knowledge in such fields as medicine in general or one of its specialities. The statistician's part in a study is to maximize its utility and to protect it from unwarranted and uneconomical expenditures—in short, to obtain the most possible information for the skill and resources that are available.

This report gives two illustrations of the use of statistical theory directed toward efficient design. The illustrations will come from work in several departmental research projects. However, the theory and principles have wide application to other studies in medical research, as well as to demography, sociology, current government statistical series, censuses, and to other problems that the statistician encounters. The wide applicability of statistical theory comes from its abstract nature, as the symbols in any theory are unmindful of the uses that man may make of them.

The first illustration will deal with the design of an *enumerative* study[1,ch.7] where the aim is to count the number of people that have certain characteristics. The second illustration will deal with the design of an *analytic* study[1] where the aim is to detect causes of differences, and to measure their effects. Both illustrations will bear on the problem of allocation of skill and of other resources.

ILLUSTRATION OF ALLOCATION OF RESOURCES IN AN ENUMERATIVE STUDY

As stated before, in an enumerative problem, the aim is to count. Let us take for illustration the study of the fertility of schizophrenics which we are now working on. One enumerative aim of the study, for example, is to estimate the number of schizophrenics by sex and age in the hospital population admitted (or readmitted) during 1934-36. Another enumerative aim is to estimate the number of children born to schizophrenics before onset, before first admission, before second admission; likewise for admissions during 1954-56.

The frame in the study is the hospital admissions recorded in mental hospitals in the State of New York during the years 1934-36, and during the years 1954-56. The sampling unit is the hospital admission. The admissions in any hospital have serial numbers; hence, one may draw from any hospital, by random numbers, any desired number of admissions.

Any hospital admission may be schizophrenic, or may not be. If we were to study a 100 per cent sample of all hospital admissions in the years 1934-36 (or 1954-56), and discard those that are not schizophrenic, the remainder would be all the schizophrenics that were admitted (or readmitted) to hospitals over that period. Hence, to draw a sample of one in eight schizophrenics of any age or sex, admitted during the years 1934-36, it is only necessary to draw a sample of hospital admissions for that period; then to study each admission by means of hospital records to decide whether the case was schizophrenic or otherwise. If the case is schizophrenic, retain it for investigation; if not schizophrenic, reject it as a blank, as it does not fall within the scope of this investigation.

As the definition of schizophrenia is a medical and not a statistical problem, the examination of each case requires the proficiency of a psychiatrist. The rules for examination would be the same whether we were to use sampling or a complete coverage. It may be mentioned in passing that, for medical reasons, first admissions 60 years old or over at the time of admission were not eligible for the study, nor cases 14 years or under.

Formula for the Sampling Variance of the Mean in an Enumerative Study

Assume that there are $\bar{N}$ schizophrenics in each hospital, and that we draw with random numbers, without stratification, a sample of $\bar{n}$ schizophrenics from each of m hospitals, the m hospitals themselves being drawn at random from the M mental hospitals in the State of New York. On page 38, a simple modification of the formula takes care of variation in size. Let $\bar{x}$ be the average of some characteristic per sampling unit. Then the formula for the variance of $\bar{x}$ will be

$$\text{Var } \bar{x} = \frac{M-m}{M-1}\frac{\sigma_b^2}{m} + \frac{\bar{N}-\bar{n}}{\bar{N}-1}\frac{\sigma_w^2}{m\bar{n}}$$

The symbol σ_w^2 is the average variance between the $\bar{N}$ sampling units within a hospital, and σ_b^2 is the variance between the means of the sampling units in the M hospitals. For example, to illustrate numerical values of σ_w and σ_b, we may work with the number of children ever born to schizophrenic females ever married and in the age group 20-39. The number of children will vary from 0 to possibly 7 or 8 per patient in any one hospital. The standard deviation σ_w between patients[2,p.260] would then be about $.24 \times 8$, call it 2 for the sake of simplicity. A few women might have more than eight children, but their effect on the variance may be neglected, as their number is small. The mean number of children per schizophrenic female in the various hospitals in New York State might range normally from perhaps 1.5 to 2.5; σ_b would then be about 1/6. These numerical values will be used later on.

The multiplier $(\bar{N} - \bar{n})/(\bar{N} - 1)$ in the above equation reduces the second term on the right to 0 if the sample includes all the $\bar{N}$ patients in each of the m hospitals, for then $\bar{n} = \bar{N}$, and $\bar{N} - \bar{n}$ will be 0. The multiplier $(M - m)/(M - 1)$ reduces the first term to 0 if all M hospitals are put into the sample, for then $m = M$, and $M - m$ will be 0.

A Simple Cost-Function

Whatever be the numerical values of σ_b and σ_w, we may derive with little effort a very important principle of sampling for enumerative purposes. We start with an oversimplified assumption concerning costs, namely, that the cost of drawing one more schizophrenic into the sample will be c dollars, regardless of which one of the M hospitals the case comes from. This assumption is about right for some kinds of surveys, although we can improve on it for the present study (see next section). We proceed with it here, nevertheless, because it leads unmistakably and with ease to a very important principle of sample design, namely, to disperse the sample, even in circumstances where costs are not so simple to represent.

The problem is to find the optimum values of $\bar{n}$ and m. The total sample will be

$$n = m\,\bar{n}$$

On the simple assumption just described, the total cost of the study will be

$$C = c\,n = c\,m\,\bar{n}$$

regardless of which hospital any of the n cases may come from.

The question is how to vary m and $\bar{n}$ so as to minimize Var $\bar{x}$ while keeping the cost C constant. The variance of $\bar{x}$, for a fixed cost C, will now be

$$\text{Var } \bar{x} = \frac{\sigma_b^2}{m} + \frac{c\sigma_w^2}{C}$$

because $m\bar{n} = C/c$. We observe that the second term is constant; it neither increases nor decreases as we vary the number m of hospitals in the study. This is so because if m increases, $\bar{n}$ must decrease to keep the cost C constant. The first term, however, decreases as m increases. The only way to decrease Var $\bar{x}$ is obviously to add more hospitals to the study, i.e., to make m big and $\bar{n}$ small. In fact, the best possible sample, under the cost-function assumed, would be to take $\bar{n} = 1$. In other words, we should take only one schizophrenic from each hospital, and go into enough hospitals (i.e., make m big enough) to reduce Var $\bar{x}$ to the level desired.

Stated simply, what our theory tells us is to disperse the sample, rather than concentrate it into a few hospitals. A refinement in the cost-function, as we shall soon see, leads to the same principle.

Refinement in the Cost-Function

Our assumption about costs may be refined by taking account of the fact that, in the study of schizophrenics, it costs less to admit a new case into the study if it is drawn from a hospital where we are already working, rather than from a hospital not yet in the study. It should be borne in mind that, in this study, there will be the cost of adding a hospital. To arrive at a workable rule, we assume that it costs c_1 dollars, on the average, to bring into the sample one more hospital, and that this cost c_1 will be the same whether we draw one schizophrenic or 10 or 100. The cost c_1 is the average cost of paying a visit to the director of a hospital to describe the purpose of the study, to meet the supervisor who will later on assist our fieldworkers when they come to draw the sample of admissions; and to study the hospital records; and to discover, in advance, any special problems with respect to the records or use thereof.

We may further suppose that it costs c_2 dollars, on the average, to study one more admission in a hospital, once the hospital is in the sample. The cost c_2 will be chiefly the average cost of studying the hospital records to decide whether a case is schizophrenic or not, plus the average cost of transcribing the information required for an admission, and of carrying out any fieldwork that may be necessary to complete the records on a patient. The cost c_2 also includes the costs of tabulation. The total cost of the study will now be

$$C = mc_1 + m\bar{n}c_2$$

Compared with the previous cost-function, the special feature of this one is that it divides the cost into two parts: one part for the m hospitals in the study, and another part for the $m\bar{n}$ cases within these hospitals.

The question arises again how to minimize Var $\bar{x}$ by varying m and $\bar{n}$ while keeping the cost C constant. For an answer to this question, one may show by the ordinary methods of the calculus that the optimum value of $\bar{n}$ is

$$\bar{n} = \frac{\sigma_w}{\sigma_b}\sqrt{\frac{c_1}{c_2}}$$

This formula was first given simultaneously by W. A. Shewhart[3,p.389] and by L. H. C. Tippett[4,p.177] in 1931. As before, the optimum number m of hospitals to take into the study is whatever number reduces Var $\bar{x}$ to the level desired, or which consumes the allowable budget C. Thus, in summary, the procedure of allocation would be the following:

1. Calculate $\bar{n}$, based on plausible values of $\sigma_w : \sigma_b$ and of $c_1 : c_2$.
2. Substitute this number $\bar{n}$ into the cost-function $C = mc_1 + m\bar{n}c_2$ to find what number m of hospitals will consume the allowable budget C, or substitute $\bar{n}$ into the formula for Var $\bar{x}$ to find what number m will yield the precision desired.

Application to an Enumerative Survey

For a numerical example of the equation under discussion, we may return to the proposed values $\sigma_w = 2$ and $\sigma_b = 1/6$ as plausible standard deviations for the number of children ever born to married female schizophrenics. Suppose further that $c_1 = \$225$ and that $c_2 = \$25$. Then

$$\bar{n} = \frac{2}{1/6}\sqrt{\frac{225}{25}} = 36$$

The meaning of this finding is that, for any age group for which we require separate estimates, we should take about 36 women from the average hospital, and enough hospitals to build up the total sample required.

With this procedure, our refined assumption leads us to almost the same conclusions that we derived from the very simple cost-function assumed in previous paragraphs, but not to the extreme recommendation that $\bar{n}$ should be 1.

Some characteristic other than the number of children ever born would possibly have a different ratio for $\sigma_w : \sigma_b$, and might lead to some $\bar{n}$ slightly different. What one does in practice is to estimate in advance what $\sigma_w : \sigma_b$ might be for several important characteristics: then to observe by a number of such calculations, what appears to be a reasonable number for $\bar{n}$. The important point is the principle that comes forth; namely, that we should take a fairly small sample of cases from each hospital, and disperse the sample to enough hospitals.

This is a very important principle. Where it is used, it saves thousands of dollars every day in research and in industrial and agricultural production. It tells us that if the variability within a hospital is large, compared with the difference between hospitals, then we should take a fairly large sample from each hospital. On the other hand, as is more usual, hospitals are considerably different in character, for various reasons. They are under different directors. Some hospitals are located in metropolitan districts, while others are located in rural areas with distinct cultural and religious backgrounds. These differences between hospitals enlarge the denominator σ_b, reduce $\bar{n}$, and enlarge m, the optimum number of hospitals in the sample.

Let us consider how costs affect the allocation in an enumerative design. Our equation for $\bar{n}$ contains the factor $\sqrt{(c_1 : c_2)}$, which tells us that if it costs a great deal to go into a hospital to open it up for investigation (that is, if c_1 is large), then we should increase $\bar{n}$ and reduce the number m of hospitals in the sample to hold the cost C to the allowable budget. Converse remarks hold for the quantity c_2, the cost of investigating a single case. If the cost c_2 is fairly large, then we care not so much where the additional case is located, which means that we reduce $\bar{n}$ and go into more hospitals.

In short, the foregoing theory tells us what to do and, more specifically, what statistical characteristics of the sampling units we need to know approximately in the planning stages in order to design an efficient sample. It tells us that we only need to know the ratios $\sigma_w : \sigma_b$ and $c_1 : c_2$.

Parenthetically, it may be mentioned that a very helpful principle applies in the statistician's efforts to achieve efficiency. It so happens, as further theory shows, that rough values for the ratios $\sigma_w : \sigma_b$ and $c_1 : c_2$ are as good as exact values for allocation of skills and resources. The explanation comes from the fact that the graph of Var $\bar{x}$ against these ratios is very flat in the neighborhood of the optimum value of $\bar{n}$. This statement is especially true with respect to the ratio $c_1 : c_2$, because $c_1 : c_2$ occurs under the root-sign.

Thus, where our calculations lead to $\bar{n} = 36$, any size of sample between 28 and 50 would yield almost as much precision for the same cost for this characteristic as $\bar{n} = 36$. This wide range of acceptable size of sample usually takes care of the fact that different characteristics lead to slightly different optimum values of $\bar{n}$. Choice of some number over such a range is a lot different from taking $\bar{n} = 100$ or 150, as one might be tempted to do without the aid of statistical theory.

Two remarks may be interjected at this time. First, calculations like those that we have just made almost always lead to much smaller samples ($\bar{n}$) from each hospital than people are in the habit of taking, when they fail to use theory. Thus, suppose that without the aid of statistical theory, one were to draw a sample of 125 schizophrenics in an age group, instead of some number in the range 28 to 50. The loss of information would then be perhaps as much as 50 per cent. In other words, half the skills and resources expended on the study would be wasted.

Second, good research requires one to abide by the recommendations of theory. To do otherwise is to throw away some portion of the skills and physical resources available for the study. If the results of calculations lead to values of $\bar{n}$ and m that seem strikingly different from the values that one might have prescribed without the aid of statistical theory, then he must choose between two alternatives: (a) look carefully at his arithmetic and at the ratios $\sigma_w : \sigma_b$ and $c_1 : c_2$ that he used in his calculations; and possibly repeat the calculations with new ratios; or (b) resign his conscience, if possible, to accept a lower degree of precision in his estimates and comparisons than would be possible with values of $\bar{n}$ and m closer to optimum. There is no third possibility.

When the hospitals are not all the same size, the optimum allocation is embodied in the equation

$$\frac{n_i}{N_i} = \frac{\sigma_{wi}}{\sigma_b}\sqrt{\frac{c_{1i}}{c_{2i}}}$$

Here the subscript i refers to Hospital i. Usually the standard deviations σ_{wi} and the costs do not vary appreciably from one hospital to another, in which case, as in this study, the equation simplifies to

$$\frac{n_i}{N_i} = \text{constant}$$

indicating proportionate allocation of the sample. Thus, if we take a sample of 1 in 10 admissions in one hospital, we should take a sample of 1 in 10 admissions in another hospital.

ILLUSTRATION OF ALLOCATION OF RESOURCES IN AN ANALYTIC STUDY

The aim in an analytic study is to discover causes or sources of variation, or to learn whether some treatment or environmental condition produces an effect, and if so, how much.

Formula for the Sampling Variance in an Analytic Study

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

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

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

The optimum allocation of resources in analytic studies is different from the optimum allocation in enumerative problems. For analytic studies, the optimum allocation of skill and effort is found by setting

$$\frac{n_A}{n_B} = \frac{\sigma_A}{\sigma_B}\sqrt{\frac{c_B}{c_A}}$$

wherein c_A and c_B are the costs per case. One may note that the sizes of the groups A and B do not enter into this formula.[1,p.240]

It so happens that in most studies σ_A and σ_B will not be greatly different, nor the costs c_A and c_B. Under such circumstances, one may write the useful approximation

$$n_A = n_B$$

This is a very simple equation. It tells us that if we wish to find the effect of two different treatments, we simply take equal numbers of patients from each group, regardless of how many patients are in each group.

Of course, if we already have on hand observations on 100 patients from one source and on 1000 patients from another source, we do not discard data to make the two groups equal in size. We take what data we have, as it is too late to plan samples that would have been economical.

The enumerative and analytic problems are competitive; that is, the allocations are different. In an enumerative study, the size of sample is proportionate to the size of the hospital. For an analytic study, the size of sample would be the same for one group as for another, even though one group is ten times as big as the other. In carrying out a study that is simultaneously enumerative and analytic, we are confronted with the fact that efficient sample design for enumerative purposes may not be the best allocation for analytic purposes. Therefore, some compromise is usually necessary in a study that serves both aims. It may happen, of course, that the two groups under comparison are, by good fortune, about the same size.

Application to an Analytic Problem

A further aim of the study of schizophrenics is analytic, to discover differences between different types of orientation, or to discover differences between communities, or differences between the two periods, 1934-36 and 1954-56, or between the upper and lower parts of New York State. One might ask, for example, whether the number of readmissions per patient has increased between the two periods, or whether the proportion of married schizophrenics has changed, or the average age of admission (including readmission), or the number of children ever born. One might ask whether these characteristics differ between groups of hospitals, or between the upper and metropolitan parts of New York State. Again, we may bear in mind that the statement of aims in any problem is not statistical. Here, they belong to medical genetics.

The decision in the present study was to give preference to the enumerative aims of the study in each period. The sample was accordingly distributed among the hospitals in proportion to their sizes. Fortunately, because the number of admissions in the upper and lower portions in the state are not greatly different, being in the ratio of about one-third in the upper part to two-thirds in the lower part, this allocation caused no serious loss for analytic use.

ALLOCATION OF THE SAMPLE IN THE STUDY OF SCHIZOPHRENICS

Thus, on the basis of the foregoing theory, we arrived at the following steps for the selection of the sample of schizophrenics:

1. The size of sample will be 8000 admissions in 1934-36, and the same number in 1954-56.
2. Distribute the sample proportionately among all the hospitals in the study.
3. Decide the number of subsamples (10 in this case).[2]
4. Compute the zoning interval.
5. Draw 10 random numbers in each zone for each hospital. Each random number is a hospital consecutive number.

Use of Subsamples

The random numbers in the table draw the sample, not as one big sample, but as ten independent smaller samples, called subsamples, or — according to Mahalanobis[2,pp.186-187] who introduced their use in 1936 — an interpenetrating network of samples. There is a great advantage in the use of independ-

ent subsamples, as they facilitate estimation of the standard errors, and detection of human errors. Many illustrations of the use of subsamples appear in the author's book on sample design.[2] I include for illustration a portion of one of the sampling tables.

SAMPLING TABLE FOR THE METROPOLITAN INSTITUTIONS 1954-1956

Zone	1	2	3	4	5	6	7	8	9	10
0001 - 0080	0032	0015	0050	0055	0058	0041	0021	0079	0011	0054
0081 - 0160	0139	0087	0130	0096	0159	0129	0148	0106	0086	0095
0161 - 0240	0211	0228	0192	0197	0189	0217	0202	0193	0214	0163
etc.										

The sample of hospital admissions was naturally bigger than the desired number of schizophrenic patients that we hope to investigate. Actually, it was by intention about four times as big. In the first place, some trial studies showed that only about half the admissions are schizophrenic, and in the second place, the prediction was that it would be possible to follow up on only about half the cases selected but not now in the hospital.

In relation to the task of thinning the sample for the fieldwork—both for patients not now in the hospital and for relatives — it may be mentioned in conclusion that there is a further contribution which statistical theory can make toward conservation of resources. Many cases in this study, possibly 70 per cent of them, have left the hospital before they ended their reproductive ages. They will require further investigation, beyond information obtainable from hospital records, to ascertain whether any more children were born to these patients. Such information will be fairly easy to obtain when records of the Department of Mental Hygiene indicate that a patient was readmitted to a state hospital after 1936 or after 1956, respectively.

However, investigation of some cases can proceed only by attempts to find the patient, or to find some relative or other possible informants. Letters addressed to informants or relatives named in the hospital records, or to the patient himself, will undoubtedly yield some valid addresses. On the other hand, some letters will come back marked "Unknown," or "Deceased." Social workers may in some cases be able to find a patient. As a last resort, actual scouting to the last-known address of a relative, or of an informant, or of the patient himself, may be necessary. This is expensive, and requires time and skill.

At the point where the cost of further information suddenly rises to a relatively high figure, statistical theory indicates that one may wisely reduce the work-load by retaining for investigation only a randomly selected portion of such patients. The proportion to retain is determinable by consideration of costs. The formula for the optimum thinning ratio is $1 : \sqrt{r}$, where r is the ratio of the two costs. For example, if the average cost of further information rises to nine times the cost of acquiring the same information from records in the hospital, or by mail, then the optimum thinning-ratio for patients that require scouting would be 1:3.

The actual thinning would be done by random numbers. Thus, for a ratio

of 1:3, random numbers would retain one patient from every three consecutive patients. Examples of balanced thinning are given in the book previously mentioned.[2,p.337]

One would of course not do any thinning if it were necessary to retain all cases in order to achieve a prescribed goal of precision. What the statistical plan of thinning does is to yield the maximum amount of information (greatest precision) for a total allowable expenditure of skills and of resources for the entire investigation.

REFERENCES

1. Deming, W. E.: Some Theory of Sampling. New York, Wiley, 1950.
2. ——: Sample Design in Business Research. New York, Wiley, 1960.
3. Shewhart, W. A.: Economic Control of Quality of Manufactured Product. New York, Van Nostrand, 1931.
4. Tippett, L. H. C.: Methods of Statistics. London, Williams & Norgate, 1931.