PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Systematic Sampling

F. Yates

Phil. Trans. R. Soc. Lond. A 1948 **241**, 345-377
doi: 10.1098/rsta.1948.0023

SYSTEMATIC SAMPLING

By F. YATES, Sc.D., *Rothamsted Experimental Station*

(*Communicated by H. G. Thornton, F.R.S.—Received* 17 *January* 1948)

CONTENTS

This paper gives an account of the results of an investigation into one-dimensional systematic sampling, i.e. the sampling of sequences of quantitative values by the use of sampling points equally spaced along the sequence.

New methods, using what are termed partial systematic samples, are evolved for estimating the systematic sampling error from short sections of sequences of completely enumerated numerical material. This gets over the difficulty, which previously existed, that the only estimates of the systematic sampling error of a numerical sequence, even when completely enumerated, were those provided by the actual deviations of the systematic samples of the whole sequence. Such deviations are few in number and by no means independent.

Simple end-corrections are proposed for eliminating the errors, due to trend, which are otherwise inherent in randomly located systematic samples.

It is demonstrated that it is impossible to make any fully reliable estimate of the sampling error from the systematic sampling results themselves, though if the continuous components of variation are not too marked, the sum of sets of terms taken alternately positive and negative, with suitable end adjustments, will provide a moderately satisfactory estimate, which will always be an overestimate provided there are no periodicities. This estimate is substantially better than the customary estimate based on successive differences.

Vol. 241. A. 834 (Price 8*s*. 6*d*.) 43 [Published 14 September 1948

In other cases supplementary sampling is required to furnish an estimate of error, and methods are described whereby estimates can be derived from supplementary samples at half-spacing, or at half and quarter spacing.

The performance of systematic sampling is investigated theoretically for certain mathematical functions, and also by the numerical analysis of certain numerical sequences. The mathematical functions investigated are (1) the two-valued function, $f(x) = 0$ or 1, corresponding to sampling for attributes, (2) the normal error function, which corresponds to sampling for density with material normally distributed about a point in a line, and (3) the one-term autoregressive function

$$y_{r+1} = by_r + a_{r+1}.$$

In the case of the two-valued function the relative performance of systematic and random samples is shown to depend on the lengths of the intervals of the function relative to the sampling interval. If these are small all forms of sampling are about of equal accuracy, but if they are large, systematic sampling is on the average twice as accurate as random sampling with one point per block, which is again twice as accurate as random sampling with two points per block. Similar results hold for the autoregressive function when $b \to 1$.

In the case of the normal function, numerical analysis shows that systematic sampling over the whole of the curve is remarkably accurate in determining the integral of the curve. Mathematical reasons why this should be so are put forward. The sampling of part of the curve by systematic sampling is also investigated, and is used to demonstrate the value of end-corrections. The effect on the sampling errors of departures of actual density distributions from the normal form due to random variations in the material are evaluated.

Numerical analyses are made of five numerical sequences: (1) 288 altitudes at 0·1 mile intervals along a grid line of a 1 in. O.S. map, (2) yields of 96 rows of potatoes, (3) 192 daily maximum screen temperature readings, (4) 192 soil temperature readings (9 a.m.) at 4 in., (5) 192 similar readings at 12 in. These analyses confirm the findings of the theoretical part of the investigation, and show that for these types of material the gain in precision with systematic sampling over stratified random sampling of the same intensity with one point per block is of the same order as the gain in precision with stratified random sampling with one point per block over stratified random sampling of the same intensity with two points per block, though the former tends to be larger in material of the more continuous type. The actual average ratios of the variances for the five sequences range from 1·26 to 2·99 in the first case, and 1·31 to 1·90 in the second.

The relation between the gain in precision and the gain in efficiency is evaluated. The latter is always smaller owing to decrease in accuracy per point for a given method of sampling with decrease in intensity. Consideration of the relation between sampling costs and the losses due to errors in the sampling results shows, however, that with a more precise method of sampling greater accuracy should be demanded in the results.

The danger of using systematic sampling in material about which nothing is known, or on material which may be subject to periodicities, is stressed, as is the importance in large-scale sampling investigations of making a preliminary investigation before instituting systematic sampling and of arranging for adequate control of error in the form of error estimates, with supplementary observations if necessary, in systematic sampling or stratified random sampling with one point per block. Control of this type should of course also be employed in stratified random sampling with two or more points per block, but in this case no special provisions are necessary, since valid estimates of error are always available from the sampling results themselves.

1. Introduction

Sequences of quantitative values are of very general occurrence, and the sampling problems to which they give rise are consequently of considerable importance. Such sequences may be temporal, as in the case of economic time series, meteorological data, quality determinations on material flowing through a delivery pipe, or articles coming off a production line, or they may be spatial, as in the case of the yields of successive plants in a row, diameters of a wire at different points along its length, or altitudes along a road.

The material itself may be continuous or discontinuous. Articles coming off a production line and plants in a row are examples of discontinuous material. If every individual is assessed for some quantitative characteristic we shall obtain a finite sequence of values which may be called the 'parent sequence'. In continuous material the quantitative characteristic has a value at every point (whether or not assessments at an indefinite number of points are practically possible), and is therefore mathematically definable as a function. Such a function may be called the 'parent function'. In many cases the methods of measurement generate a finite sequence from continuous material, often involving some form of integration of the basic material. Daily figures for rainfall, for example, are totals of the rainfall falling over the last 24 hr.

The distinction between sequences and functions, though not of great importance, must be preserved in any formal discussion of sampling theory, since the number of values in a finite sequence is itself finite, and if all are observed the sampling error (apart from errors of observation) will be zero. Errors of observation also require slightly different formal treatment in sequences and functions, though this will not enter into the present discussion.

In sampling a sequence of quantitative values, the sampling points may be located wholly at random, at random within subdivisions of the sequence, or at equal intervals along the sequence. Analogous procedures are available for material distributed spatially in two or three dimensions. These three types of sampling are known respectively as random, stratified random, and systematic. A one-dimensional systematic sample may either have a randomly located starting point, or a starting point distant half the sampling interval from the start of the sequence. The latter, which may be termed the centrally located systematic sample, is clearly in general only of advantage if the length of the sequence or function is a multiple of the sampling interval.

General considerations indicate that a one-dimensional systematic sample may be expected, except in certain special cases, to give a more accurate result than will be obtained from the same number of randomly located sampling points. This has also been demonstrated mathematically for certain types of sequence by Cochran (1946), and Quenouille, following Cochran's general line of approach, has recently extended Cochran's results to the two-dimensional case. Systematic sampling has also the considerable practical advantage of being simpler to execute. There are obvious advantages, for example, in being able to take samples at fixed time intervals in sampling of the quality control type, at fixed space intervals when sampling an agricultural crop or a forest area. The construction of a graph or map of the results is also considerably simplified and improved in accuracy when the sampling points are equally spaced.

There are, however, certain objections to systematic sampling. The most important is that very inaccurate results will be obtained if there are any periodicities in the parent sequence or function, and the sampling interval is a multiple of the basic period. Even damped periodic effects, subject to regeneration by random impulses, may considerably reduce the accuracy. Moreover, in periodic material the uniformity of the sampling results will give a false impression of precision. Systematic sampling must therefore not be used without thorough investigation on material which is liable to periodic effects.

There are two other disadvantages. First, there is no method, corresponding to that which is available for most types of random sampling, by which valid estimates of the sampling

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

errors can be obtained from the sampling results themselves; and secondly, in the case of randomly located systematic samples, any pronounced trend in the material will substantially reduce the accuracy of systematic relative to random sampling.

This last disadvantage, however, is not a serious one, since, as will be shown in the course of the present paper, the loss of accuracy can be avoided by the use of very simple end-corrections. Apart from this, the paper is mainly concerned with the investigation of the comparative accuracy of systematic and random sampling, and with the problem of the estimation of the errors of systematic sampling. There are two distinct aspects to the latter problem. The first is the determination of the 'expected' systematic sampling error in a sequence of which all the values are known. The second is the question of how far estimation of the error is possible from the sampling results themselves, and in so far as this is impossible, what is the best procedure (using supplementary observations) for obtaining a reliable estimate.

At first sight the former problem appears trivial, since it is always possible, when all the values of a sequence are known, to calculate the actual deviation of all systematic samples having a given sampling interval. The number of such samples, however, will be equal to the sampling interval and will therefore be small if the latter is small. Even if the interval is large the separate deviations, though more numerous, are by no means independent. Consequently the mean square deviation of the systematic samples, although an exact measure of the actual systematic sampling errors, will only provide a very rough estimate of the systematic sampling error which may be expected in similar sequences, and such questions as how the accuracy of systematic sampling compares with random sampling in such sequences, and how the accuracy changes with change of sampling interval, will be very sketchily answered.

2. Concepts and notation

(*a*) *General*

In the case of a sequence the number of terms in the parent sequence will be denoted by l. The terms will be numbered consecutively, beginning at 1, the value of the rth term being denoted by y_r. In the case of a function the value of the parent function at point x will be denoted by $f(x)$, with $0 \leqslant x \leqslant l$.

The variance per term within blocks containing d terms (or of length d) will be denoted by R_d. If l is an integral multiple of d the sequence or function can be divided into blocks without any residue, but in other cases there will be some residue. In any case, apart from end-conditions, the location of the block divisions must be regarded as arbitrary, and a somewhat more accurate estimate of R_d (regarded as a feature of the material) will be obtained if (in a sequence) all possible block divisions are taken.

The mean square of differences of terms d units apart (divided by 2 to bring it to a 'per term' basis) will be denoted by D_d. Again all possible differences of terms d units apart will provide the best estimate of D_d.

The value of R_d can be estimated indirectly from the values of D_s, $s = 1, \ldots, d-1$. This follows from the fact that the sum of the squares of the deviations of the d values $y_1, y_2, \ldots$ from their mean is equal to

$$\frac{1}{d} \sum_{q=1}^{d} \sum_{p=q+1}^{d} (y_p - y_q)^2.$$

This expression contains $d-1$ differences 1 unit apart, $d-2$ differences 2 units apart, etc. Consequently an estimate of R_d is given by

$$\frac{2}{d(d-1)}\{(d-1)\,D_1+(d-2)\,D_2+\ldots+D_{d-1}\}.$$

If all possible block divisions and all possible differences are used for estimating R_d and D_s, the two methods of estimation of R_d will give identical results except for end-disturbances.

The sampling variance of systematic samples at interval d (on a per term basis) will be denoted by S_d. Distinction must be made between the 'expected' systematic sampling variance, which can be evaluated in mathematical terms in certain types of function subject to known laws of variation, and estimates of this 'expected' variance which can be evaluated numerically by the use of partial systematic samples, in the manner explained below. To avoid unnecessary multiplication of symbols the same symbol S_d has been used for both these quantities, as no confusion is likely to arise once the distinction is recognized. There are also the further estimates of the 'expected' variance which can be derived from the errors of the actual systematic samples of a complete sequence. These, as already explained, are subject to large sampling errors of estimation and have been allotted a separate set of symbols, which are given in the next subsection.

There is one further variance which is of importance. This is the variance (on a per term basis) derived from the sum of terms at interval d taken alternately plus and minus, for which the symbol E_d will be used. In this case also distinction must be made between the theoretical value and the values derived from numerical data by the use of partial samples. (Symbolic definitions of S_d and E_d appropriate to numerical data are given in subsection $(2f)$.)

(*b*) *Randomly located systematic samples*

The sum of the terms of a systematic sample of every dth term of a sequence, starting at term r $(r\leqslant d)$, will be denoted by ${}_rH_d$, and the corresponding mean by ${}_rh_d$. The number of terms ${}_rn_d$ in ${}_rH_d$ will be the integral part of $(l+d-r)/d$. Thus

$${}_rH_d = y_r+y_{r+d}+y_{r+2d}+\ldots.$$

The d values of the sums and means of all possible samples at spacing d will be collectively denoted by H_d and h_d. If l is not an integral multiple of d it will be necessary to work with the means. Otherwise the choice of means or totals is a matter of convenience.

In considering the accuracy of systematic sampling we shall require some expression for the variance of the actual samples. The true errors of sampling will be given by the deviations of h_d from the mean of all the terms of the sequence. It appears best, however, to adopt the convention that

$$V[h_d] = \frac{1}{d-1}\sum_1^d ({}_rh_d-\bar{y})^2,$$

remembering that, with this convention, a factor $(d-1)/d$ must be introduced into the final estimate of error of a sequence to allow for the fact that we are sampling from a finite population. The corresponding variance of the sums H_d will be denoted by $V[H_d]$. If l is not a multiple of d, $V[h_d]$ will not be exactly equal to the variance of h_d, since the mean of h_d will not be exactly equal to the mean of all the terms of the sequence, but for long sequences the difference will be trivial.

It will often be more convenient to express these variances on a per term basis. These will be defined by

$$v[H_d] = v[h_d] = \overline{{}_r n_d}\, V[h_d].$$

If the terms are randomly distributed uncorrelated variates with variance σ^2 the deviations of h_d will provide an estimate of σ^2 based on $d-1$ degrees of freedom given by

$$s^2 = v[h_d].$$

Similar expressions hold for systematic samples of a function, d being taken to represent the sampling interval. In this case there are an infinite number of values in H_d, and the expression for the mean square deviation per term becomes

$$v[h_d] = \frac{1}{d^2}\int_0^d ({}_x h_d - \overline{f(x)})^2\, dx \,.\, \int_0^d {}_x n_d\, dx \,.$$

(c) *Centrally located systematic samples*

If l is an integral multiple of d and d is odd the centrally located sample will have each of its units at the centre of one of the l/d blocks of length d into which the sequence can be divided. The mean of this sample will be denoted by ${}_c h_d$. If d is even there is no centrally located systematic sample, but either of the two samples a half unit off centre can usually be taken as equivalent to a centrally located sample.

Since there is only one centrally located sample for a given sequence or function, no mean square deviation corresponding to $V[h_d]$ can be calculated for a single sequence or function. The sampling error can only be expressed in terms of the difference of the mean of the sample from the mean of the whole sequence or function.

(d) *End-corrections*

Instead of taking the ordinary mean ${}_r h_d$ for a randomly located systematic sample, allowance can be made for the fact that the first and last terms are situated at distances from the ends of the sequence or function which in general differ from $\frac{1}{2}d$. The simplest form of adjustment, and the only one which is ordinarily worth considering, is to assign a weight to the first term of $(\frac{1}{2}d+r-\frac{1}{2})/d$ in the case of a sequence and $(\frac{1}{2}d+r)/d$ in the case of a function, and a similar weight of $(\frac{1}{2}d+r'-\frac{1}{2})/d$ or $(\frac{1}{2}d+r')/d$ to the last term, where r' is the number or distance of the last term reckoned from the end of the sequence or function.

We will call such a mean the mean of a randomly located systematic sample with end-corrections and will denote it by ${}_r h'_d$. If l is not a multiple of d the divisor ${}_r n_d$ will be replaced by a non-integral divisor ${}_r n'_d$. The mean square deviation $v[h'_d]$ can be defined as before, retaining ${}_r n_d$. If the sequence is random, $v[h'_d]$ will not be an exact estimate of σ^2, but with a sequence of any length the difference will be of no consequence.

(e) *Supplementary observations at ends*

The further the starting or end-points of a randomly located systematic sample from the beginning or end of the sequence the more inaccurately will the terminal regions be determined. This situation can be improved by introducing a supplementary sampling point or points. With a single initial supplementary point the best location (for a function) appears to be at $\frac{1}{2}(r-\frac{1}{2}d)$ from the beginning, with a weight of $(r-\frac{1}{2}d)/d$ and unit weight for the first

ordinary term of the sample, no supplementary observations being taken if $r \leqslant \frac{1}{2}d$. A similar procedure can be followed at the end of the sample. Such samples will be termed systematic samples with supplementary observations, and their means will be denoted symbolically by ${}_rh''_d$.

Centrally located systematic samples and samples with end-corrections or supplementary observations will be biased in the sense that the mean of all the sample means will not be exactly equal to the population mean. Such bias will only be of importance if the end-terms are exceptional in some way (e.g. the edges of a field or forest). If circumstances are such that this bias must be eliminated, some special treatment of the ends of the sequence must be introduced. Thus the central part of the sequence, comprising a length which is an integral multiple of d, might be sampled by means of a systematic sample with random starting point, the ends being sampled by means of one or more randomly located points at each end, with appropriate weighting.

(*f*) *Partial systematic samples*

If a given sequence has been completely determined, $v[h_d]$, $v[h'_d]$ and $v[h''_d]$ will provide measures of the actual sampling errors in this particular sequence. As already pointed out, however, looked on as estimates of the probable sampling errors of similar material they will not be at all accurate, since in each measure the differences of only d quantities are involved, and these quantities will not be by any means independent if the systematic components of the variation are at all marked.

Consequently, in order to arrive at any satisfactory estimates of the probable systematic sampling errors to which a given type of material is subject, a long sequence must be split up into segments so as to obtain an adequate number of independent comparisons and thus utilize the material to best advantage.

In the choice of the length of these segments two conflicting factors must be considered. The shorter the segments the more numerous will be the available differences, but if the segments are too short the more long-term compensating effects of systematic sampling will be lost.

The shorter the segments, also, the greater will be the disturbance due to end conditions. This disturbance, however, can in the main be removed by making end-adjustments. We will define such an adjusted partial sample containing k or $k+1$ terms, and falling within the range $c+1$ to $c+(k+1)\,d$, as

$$ {}_rG_d(k,c) = \frac{r}{d}y_{c+r} + y_{c+r+d} + y_{c+r+2d} + \ldots + y_{c+r+(k-1)d} + \frac{d-r}{d}y_{c+r+kd}. $$

Usually it will not be necessary to specify k or the range, in which case ${}_rG_d(k,c)$ may be abbreviated to ${}_rG_d$.

In general, in the calculations carried out in this paper, the subdivision into segments will be made in such a manner that the $d-1$ terms involving fractional coefficients at the end of each segment will be used also in the next segment. Segments of $5d-1$ terms (or $4d$ terms excluding overlap at one end, $k=4$) have been taken as likely to represent a reasonable compromise between utilization of material and the need for elimination of the long-term compensating effects.

If there are b segments, the estimate of the sampling error of systematic samples at spacing d, on a per term basis, may be taken to be

$$S_d = v[G_d] = \frac{1}{\lambda b(d-1)} \sum_{\text{segments}} \sum_{1}^{d} ({}_rG_d - \overline{{}_rG_d})^2,$$

λ being chosen so as to give an unbiased estimate of σ^2 when the material is random. For this we must have, with $k = 4$, $\lambda = \frac{11}{3} + \frac{1}{3d^2}$.

In the calculation of E_d from terms at spacing d taken alternately plus and minus, a weight of $\frac{1}{2}$ will be assigned to each end-term, and $2k+1$ terms in all will be included. If there are b' such segments (which may be taken as overlapping in a range d), and all possible sets of terms in these segments are taken, we shall have ($k = 4$)

$$E_d = \frac{1}{7 \cdot 5 b' d} \sum_{\text{segments}} \sum_{1}^{d} (-\tfrac{1}{2}y_{c+r} + y_{c+r+d} - y_{c+r+2d} + \ldots + y_{c+r+7d} - \tfrac{1}{2}y_{c+r+8d})^2.$$

To save computation in the analysis of sequences for which all values are known and for which values of ${}_rG_{2d}$ have already been calculated, E_d may alternatively be estimated from the expression

$$E_d = \frac{1}{2\lambda' b' d} \sum_{\text{segments}} \sum_{1}^{d} ({}_rG_{2d} - {}_{r+d}G_{2d})^2,$$

where $\lambda' = \frac{11}{3} + \frac{1}{12d^2}$.

3. Sampling for attributes: case when $f(x) = 0$ or 1 only

If a line is divided into sections of which the alternate ones possess a certain attribute, and we know the points of division, the determination of the proportion of the line which possesses the attribute is a simple matter, involving merely summation. In certain types of material, however, the location of the points of division may be difficult, and sampling can then be resorted to, the presence or absence of the attribute being determined at the sampling points only.

If the proportion of the line possessing the attribute is small, or alternatively if the proportion is nearly unity, the relative efficiency of systematic and random sampling can be easily determined.

Consider a single section of length a possessing the attribute. This can be represented by the function

$$f(x) = 1 \quad (0 \leqslant x \leqslant a), \qquad f(x) = 0 \text{ elsewhere.}$$

In the case of systematic sampling at interval d randomly located with reference to the section, when $a < d$, ${}_rH_d$ will have the value 1 in a proportion a/d of the samples and 0 in $1 - a/d$ of the samples. The mean square deviation from the mean a/d is therefore

$$S'_d = V[H_d] = \frac{a}{d}\left(1 - \frac{a}{d}\right),$$

where the dash indicates that the total variance, and not the variance per point, is implied. When $kd < a < (k+1)\,d$ we have similarly

$$S'_d = V[H_d] = c(1-c),$$

where $c = (a/d) - k$, i.e. c is the fractional part of a/d.

In the case of random sampling with one sampling point per block of length d, if $a<d$ the section will lie wholly in one block in a proportion $1-a/d$ of block locations, the variance of the mean being in this case $a(1-a/d)/d$. In the remainder of the block locations a block boundary will cut the section. If a part x of a lies in one block the variance of the mean a/d will be

$$\frac{x}{d}\left(1-\frac{x}{d}\right)+\left(\frac{a-x}{d}\right)\left(1-\frac{a-x}{d}\right).$$

The mean value of this, $0\leqslant x\leqslant a$, is $a(1-2a/3d)/d$. Hence

$$R'_d=\frac{a}{d}\left(1-\frac{a}{d}\right)^2+\frac{a^2}{d^2}\left(1-\frac{2}{3}\frac{a}{d}\right)=\frac{a}{d}\left(1-\frac{a}{d}\right)+\frac{1}{3}\frac{a^3}{d^3}.$$

When $a>d$ a similar integration over all possible block locations gives

$$R'_d=\tfrac{1}{3}.$$

The procedure for blocks of $2d$ each with two sampling points is similar. In this case we find

$$R'_{2d}=\frac{1}{12}\frac{a}{d}\left[12-6\frac{a}{d}+\frac{a^2}{d^2}\right] \quad (a<2d)$$

$$=\tfrac{2}{3} \quad (a>2d).$$

Finally, we may note that if the sampling points are located entirely at random with a density of $1/d$ per unit length the number of points falling in the section will conform to a Poisson distribution with mean and variance equal to a/d, so that $R'_\infty=a/d$.

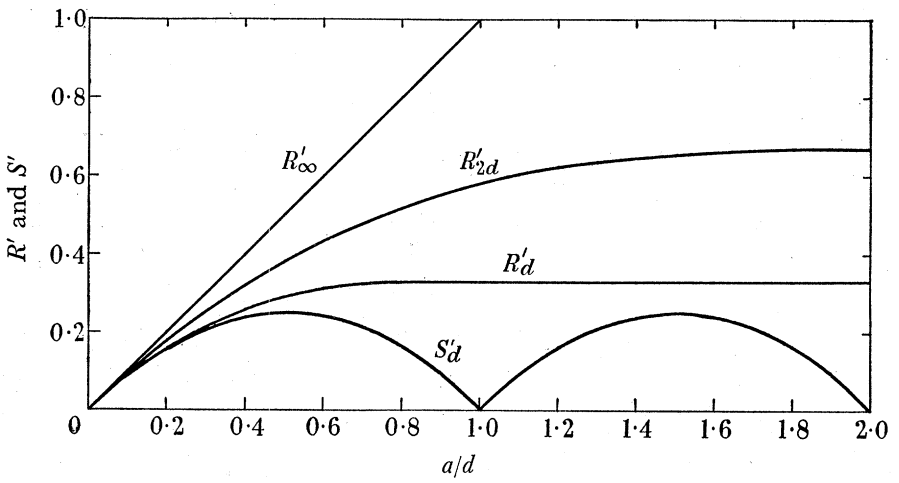

FIGURE 1. Sampling for attributes. Sampling variances of different methods of sampling a function $f(x)=1$, $0\leqslant x\leqslant a$, $f(x)=0$ elsewhere. S'_d, systematic sampling at interval d. R'_d, stratified random sampling with one point per block of length d. R'_{2d}, stratified random sampling with two points per block of length $2d$. R'_∞, random sampling with intensity $1/d$.

These results are exhibited graphically in figure 1. If the length of the section is small compared with d, all methods are of approximately the same accuracy, but as the section length increases the systematic sampling gains in relative efficiency. With a random distribution of section lengths systematic sampling has an average variance of $\frac{1}{6}$, which is one-half that of random sampling with one sampling point per block when the section length is

greater than d, and one-quarter that of random sampling with two sampling points for block when the section length is greater than $2d$.

If, therefore, the majority of sections are short relative to the sampling interval there will be little to choose between the various sampling methods, but if the majority of sections are greater than the sampling interval systematic sampling may be expected to be substantially more accurate than any form of random sampling.

The case in which the proportion of the line possessing the attribute is not small, or nearly unity, has not been investigated in detail, but the same general results may be expected to hold. The results established above for the limiting conditions will hold exactly when all section lengths are greater than $2d$ (or d if only systematic samples and randomly located samples of one per block are being considered) but are otherwise random.

It is of interest that under these conditions the relations between the errors of stratified random and systematic sampling, and between these errors and the mean square errors calculated from successive differences, and from terms taken alternatively positive and negative, are the same as those for an autoregressive function in the limiting case, when $b \to 1$. This function is discussed in § 5. In both cases

$$S_d = \tfrac{1}{2}R_d = \tfrac{1}{4}R_{2d} = \tfrac{1}{3}E_d = \tfrac{1}{6}D_d.$$

4. Sampling material of which the density distribution is in the form of a normal curve of error

(a) *Sampling the whole curve*

The behaviour of systematic samples from a parent function whose form is a normal curve of error brings out a number of points of general interest, and is also of interest in itself, since the problem of sampling material distributed in the form of a normal curve is one that occurs from time to time. The analogous two-dimensional problem was met with during the war, for example, when it was desired to sample aerial photographs to determine bombing patterns, the amount of damage in a city, etc.

We will first consider the sampling errors which occur with systematic samples spaced σ and 2σ, and with samples located at random in blocks of length σ and 2σ.

Table 1 gives values of the integral of the function estimated from samples spaced at 2σ. x represents the distance of the nearest sampling point from the centre of the distribution. It will be seen that even with these very widely spaced sampling points surprisingly high accuracy is attained, the error never being greater than 1·5 %.

Table 1. Systematic samples from a normal distribution function, with $d = 2\sigma$

x/σ	$2\ {}_xH_{2\sigma}$	x/σ	$2\ {}_xH_{2\sigma}$	x/σ	$2\ {}_xH_{2\sigma}$
0·0	1·014 3838	±0·4	1·004 4446	±0·8	0·988 3636
±0·1	1·013 6798	±0·5	1·000 0000	±0·9	0·986 3200
±0·2	1·011 6366	±0·6	0·995 5552	±1·0	0·985 6160
±0·3	1·008 4544	±0·7	0·991 5454		

If a single randomly located sampling point had been taken in each block of 2σ, the variance of the estimated integral of the function, which can be obtained by direct integration, would vary between 0·10762 when the block boundaries are at 0, $\pm 2\sigma$, $\pm 4\sigma$, etc.,

and 0·04863 when the block boundaries are at $\pm\sigma$, $\pm 3\sigma$, etc., with corresponding standard errors of 32·8 and 22·1 %.

The values of ${}_{x}H_{\sigma}$ for systematic samples spaced at interval σ can be obtained from table 1 by taking the means of samples at a half interval apart, i.e. of the values for 0·0 and 1·0, 0·1 and 0·9, etc. All these values will be found to lie between 0·999 9999 and 1·000 0001. The sampling error has therefore been virtually eliminated by the reduction of the interval from 2σ to σ.

The sampling error with randomly located sample points is still very large. If one point is taken in each block of σ the variance will vary between 0·01120 and 0·01118, giving a standard error of 10·6 %, while with two sampling points in each block of 2σ, which would be necessary in order to provide data for an empirical estimate of sampling error, the variance varies between 0·05381 and 0·02432, giving a standard error between 23·2 and 15·6 %.

(*b*) *Effect of random errors of distribution*

In the practical problem of sampling material with a density distribution of which the underlying form is a normal curve, there will be a further component of variation due to the divergence of the actual density distribution from the hypothetical infinite distribution. The nature of this divergence will depend on the material, but in the case in which the material is composed of separate units each of which is normally and independently distributed about a fixed point, the variance due to this cause can be easily assessed.

If there are N units in the material, and a fraction f of the whole range of the distribution is sampled, the variance of the number n falling in the sampled parts of the range will be made up of two parts, that due to the sampling variance of the total 'normal' probability P attributable to the sampled part of the range, and that due to the divergence of the actual density within the sampled part of the range from the 'normal' density. This latter will be given by a binomial distribution. We shall in fact have

$$V(n) = N^2V(P) + P(1-P)\,N.$$

The expectation of P is f, and its variance for the various methods of sampling is given in percentage terms by the results already obtained. If $N = 1000$ and $f = \frac{1}{10}$, for example, taking mean value, we have

$$P(1-P)\,N = 90{\cdot}0$$

$$N^2V(P) \begin{cases} \text{Systematic sample at spacing } \sigma & = 0{\cdot}0 \\ \text{Random sample, 1 from each block of } \sigma & = 111{\cdot}9 \\ \text{Random sample, 2 from each block of } 2\sigma & = 390{\cdot}6 \end{cases}$$

Thus in this case the use of a systematic sample, instead of a random sample with one point located in each block of σ, about halves the variance of the estimate of the total number of units in the distribution. A random sample of two points from each block of 2σ will have about five times the variance of the systematic sample. The advantage of a systematic sample will become progressively greater with increase in the number of units in the material.

The accuracy attained in the location of the centre point of the distribution can be calculated in a similar manner. With a single randomly located sampling point in each block of length σ (block divisions at 0, $\pm\sigma$, etc.) the variance of $\bar{x}$ due to variation in the location of the sampling points when N is infinite will be $0{\cdot}0173\sigma^2$. With $N = 1000$ and

$f = \frac{1}{10}$ the additional variance due to divergence of the actual density from the 'normal' density will be $\frac{1}{100}\sigma^2$. Thus in this case systematic sampling will be over twice as accurate as random sampling.

These results show that the use of an evenly spaced sampling grid for the determination of the location and number of bombs in bomb distributions, and in analogous problems, is fully justified. The gains in accuracy over any system of random sampling, though probably not so large as in the one-dimensional case, are nevertheless likely to be appreciable. In addition, the work of locating the sampling points is considerably simplified.

(c) *Relation to general theory of sampling continuous differentiable functions*

The very high accuracy attained by systematic samples of the normal distribution function, and the rapidity with which the errors are reduced with reduction in the sampling interval, are somewhat surprising, and it is worth considering how far this is a special feature of this distribution.

Let $f(x)$ be a function which is continuous and differentiable, and let the range x_0 to x_1 be sampled by r equally spaced points at interval d situated at points $\xi_1 = x_0 + \frac{1}{2}d$, $\xi_2 = x_0 + \frac{3}{2}d, \ldots$ with $x_1 - x_0 = rd$. Then by Taylor's theorem we have

$$\int_{x_0}^{x_1} f(x)\,dx = d\{f(\xi_1) + f(\xi_2) + f(\xi_3) + \ldots + f(\xi_r)\} + \frac{d^3}{4(3!)}\{f''(\xi_1) + f''(\xi_2) + f''(\xi_3) + \ldots + f''(\xi_r)\} + \frac{d^5}{16(5!)}\{f^{(\mathrm{iv})}(\xi_1) + f^{(\mathrm{iv})}(\xi_2) + f^{(\mathrm{iv})}(\xi_3) + \ldots + \tfrac{1}{2}f^{(\mathrm{iv})}(\xi_r)\} + \ldots.$$

The first term is evaluated by the sampling, while the subsequent terms represent the sampling error. As r is increased $f''(\xi_1) + f''(\xi_2) + \ldots + f''(\xi_r)$ tends to

$$\frac{1}{d}\int_{x_0}^{x_1} f''(x)\,dx \quad \text{or} \quad \frac{1}{d}\{f'(x_1) - f'(x_0)\},$$

and similarly for the other terms. Thus the sampling error tends to zero as

$$\frac{d^2}{4(3!)}\{f'(x_1) - f'(x_0)\} + \frac{d^4}{16(5!)}\{f'''(x_1) - f'''(x_0)\} + \ldots.$$

Since the normal curve of error is asymptotic to $x = 0$ at both ends, $f'(x_1) - f'(x_0)$, etc., will all be zero in any sample covering the whole of the distribution. We may thus anticipate that the sampling error will tend to zero very rapidly in the final stages.

The convergence in the case of the normal curve is, however, somewhat exceptional. Aitken (1939) has pointed out (p. 45) that the error of numerical integration at unit interval (with central ordinate) is much smaller for the function $f(x) = e^{-\frac{1}{2}x^2}$ than for the function $f(x) = 1/(1+x^2)$, the former being $0{\cdot}52 \times 10^{-8}$ proportionately, and the latter $0{\cdot}37 \times 10^{-2}$. D. G. Kendall (1942) has shown that, with suitable restrictions, it is necessary and sufficient for the truth of

$$\tfrac{1}{2}\lambda f(0) + \lambda \sum_{n=1}^{\infty} f(n\lambda) = \int_0^{\infty} f(x)\,dx$$

for all λ less than λ_1, that $f(x)$ has a Fourier cosine-transform $F_c(t)$ such that

$$F_c(t) = 0 \quad \text{for all } t > 2\pi/\lambda_1.$$

He also shows that the 'approximate' truth of the first of these equations is equivalent to the 'approximate' truth of the second. The Fourier cosine-transforms of the above functions are $F_c(t) = e^{-\frac{1}{2}t^2}$ and $F_c(t) = e^{-t}\sqrt{(\frac{1}{2}\pi)}$, and the relative accuracy of numerical integration in the two cases is thus connected with the relative smallness at infinity of their transforms.

(*d*) *Sampling a part of the curve*

If $f'(x_1) - f'(x_0)$ is not zero the sampling error may be expected to tend to zero with d^2. An example of this is given in table 2, which shows the effect of sampling the normal curve from $+\sigma$ to $+\infty$.

TABLE 2. SYSTEMATIC SAMPLES, CENTRALLY PLACED RELATIVE TO THE STARTING POINT, TAKEN FROM THE RANGE $+\sigma$ TO $+\infty$ OF THE NORMAL CURVE

sampling interval d/σ	location of first point x/σ	$_cH_d$ / true value	error / true value	predicted
1·0	1·5	0·932428	0·067572	—
0·5	1·25	0·983876	0·016124	0·016893
0·2	1·1	0·997452	0·002548	0·002580
0·1	1·05	0·999363	0·000637	0·000637

The values shown in the 'predicted' column are calculated by multiplying the previous error by the square of the ratio of the sampling intervals. Their agreement with the values in the error column indicates that the convergence is nearly proportional to d^2.

The errors in this case are very much greater than those obtained when the whole curve is sampled. This is, indeed, to be expected, since the curvature is continuously positive, so that the value at each sampling point must be less than the average of the range on which it is centred. Under such circumstances centrally located systematic samples will be biased, as will also randomly located systematic samples with end-corrections. The bias, however, will tend to zero as the square of the sampling interval, and is not likely to be of importance in the types of material to which systematic sampling is applied in practice.

Even in this case systematic sampling gives considerably smaller errors than random sampling. The ratio of the variance of a set of randomly located samples, one in each block of length σ, for example, to the square of the true value is 0·12938, corresponding to a proportionate standard error of 0·3596. The actual proportionate error of the corresponding centrally located systematic sample is only one-fifth of this.

Moreover, as shown above, the square of the sampling error of a centrally located systematic sample tends to zero as d^4. The variance of the ordinate of a point randomly located on a straight line with slope b and projected length d is $\frac{1}{12}b^2d^2$. As the length of the blocks is reduced the part of the function lying within any one block will approximate to a straight line. The number of sampling points is proportional to $1/d$. Consequently the variance of the corresponding stratified random sample tends to zero as d^3. For continuous functions, therefore, systematic sampling may be expected to become progressively more accurate relative to random sampling as the sampling interval is decreased.

The above theory is applicable only to centrally located systematic samples. If the function is asymptotic to zero in both directions all samples can be regarded as centrally located. In other cases a non-centrally located sample can be regarded as centrally located

for the part of the function beginning at the point $r-\frac{1}{2}d$ if this is positive, and $r+\frac{1}{2}d$ otherwise, with similar conditions at the end.

If no end-corrections are applied the additional errors due to these end-conditions can be large relative to the sampling errors of the rest of the distribution. With the end-corrections and supplementary observations described in §§ 2 (*d*) and 2 (*e*), however, they can be reduced considerably. Table 3 shows the effect of applying these corrections when sampling a normal function from 0 to $+\infty$.

TABLE 3. SYSTEMATIC SAMPLES OF ONE-HALF OF A NORMAL CURVE OF ERROR, WITH END-CORRECTIONS, AND OF A SIMILAR CURVE PRECEDED BY A SECTION FOR WHICH $f(x)$ IS CONSTANT

	half-normal curve			half-normal curve with constant section
r/σ	${}_rH_\sigma$	${}_rH'_\sigma$	${}_rH''_\sigma$	${}_rH'_\sigma$ and ${}_rH''_\sigma$
0·0	0·699 4711	0·500 0000		1·000 0000
0·1	0·662 1453	0·503 3643		1·002 5683
0·2	0·623 1469	0·505 8341	as ${}_rH'_\sigma$	1·003 4642
0·3	0·582 8448	0·506 5672		1·003 0563
0·4	0·541 6494	0·504 8224		1·001 7552
0·5	0·500 0000	0·500 0000		1·000 0000
0·6	0·458 3505	0·491 6730	0·498 1949	0·998 2447
0·7	0·417 1551	0·479 6059	0·496 5456	0·996 9436
0·8	0·376 8532	0·463 7607	0·495 1970	0·996 5359
0·9	0·337 8546	0·444 2887	0·494 2717	0·997 4316
1·0	0·300 5288	0·421 5142	0·493 8628	1·000 0000
	0·500 0000	0·486 0673	0·500 1729	1·000 0000

As is to be expected, the unadjusted set of samples with random starting point, H_σ, is very inaccurate, having an error mean square of 0·0140 (calculated from the tabulated values with half-weight to 0·0 and 1·0). This is somewhat greater than the variance 0·0056 of a random sample with one sampling point in each block of σ. Unadjusted randomly located systematic samples are necessarily inaccurate with any material which has a pronounced trend.

The use of end-corrections, without supplementary observations, considerably improves the accuracy, but the samples for which the starting point is widely separated from the beginning of the distribution are still rather unsatisfactory, owing to the rapid fall in value of $f(x)$ with increasing x. The error mean square of this set of samples, H'_σ, is 0·00081. The use of a supplementary sampling point at $x=\frac{1}{2}(r-\frac{1}{2}d)$ for samples for which $r>\frac{1}{2}d$ removes the major part of this error, as shown in the column H''_σ. The error mean square is now reduced to 0·000020.

If the half-normal curve is preceded by a section in which $f(x)$ is constant we obtain the values shown in the last column of table 3. (The length of the constant section, $-1{\cdot}253\ 3141\sigma$, has been chosen so as to give a total integral of unity.) In this case H'_σ and H''_σ are identical, since the adjustments are applied to the constant part of the function. The errors are appreciably smaller than those for H''_σ from the half-normal curve, indicating that end-corrections satisfactorily compensate for the variation that would otherwise occur in systematic samples when the function starts at one more or less stable value but subsequently falls to another more or less stable value. If the adjustments were not used the errors would be quite large, the value at $r=0{\cdot}7\sigma$ being 0·816 0974 and at $r=0{\cdot}8\sigma$ being 1·174 7378.

5. Sampling material varying quantitatively according to the autoregressive law

(a) *Properties of autoregressive sequences*

In an autoregressive sequence each term is in part determined by the preceding term or terms and in part by an independent random component. In the simplest case the autoregression may be expressed as a regression on the last preceding term only. The generating law may then be written

$$y_{r+1} = by_r + a_{r+1},$$

where b is the regression coefficient and a_{r+1} is a random variate, with mean zero and variance σ'^2 say. In order that the total variance σ^2 of successive terms of the sequence shall be constant we must have $1 > b > -1$ and

$$\sigma'^2 = \sigma^2(1-b^2).$$

An autoregressive sequence has the property that the sequence formed by terms s units apart will itself be autoregressive with a coefficient b^s. We may therefore define an autoregressive process which will generate a function which has the property that any sequence of equally spaced points on it, distance s apart, forms an autoregressive sequence with coefficient b^s. A function generated in this manner will be continuous but not differentiable. It therefore has certain properties which are essentially different from the continuous variate type of function.

The correlation between terms d apart in an autoregressive sequence is b^d. Consequently the correlogram is of the exponential form $\rho_d = e^{-\lambda d}$, where $b = e^{-\lambda}$.

Cochran (1946) has established that with such a correlogram the sampling variance of a systematic sample with spacing d of a function sufficiently long relative to the value of b for end-effects to be neglected is given on a per term basis by

$$S_d = \sigma^2\left\{1 - \frac{2}{\lambda d} + \frac{2}{e^{\lambda d}-1}\right\},$$

while that of a sample of the same density with one sampling point located at random in each block of length d is

$$R_d = \sigma^2\left\{1 - \frac{2}{\lambda d} + \frac{2}{\lambda^2 d^2} - \frac{2e^{-\lambda d}}{\lambda^2 d^2}\right\}.$$

The expression for the variance per term of a sample with two sampling points located at random in each block of length $2d$ can be obtained from the last expression by substituting $2d$ for d.

Cochran used these expressions to evaluate the relative efficiency of stratified random and systematic sampling. We reproduce his results (slightly modified) in table 4, with the addition of values for random sampling in blocks of $2d$.

Table 4. Relative efficiency of systematic and stratified random samples (one per block and two per block) from an autoregressive function

b^d	0·9	0·8	0·7	0·6	0·5	0·4	0·3	0·2	0·1
R_d/S_d	1·95	1·89	1·84	1·78	1·71	1·64	1·55	1·46	1·33
R_{2d}/S_d	3·80	3·59	3·38	3·17	2·95	2·71	2·47	2·19	1·86

If $d = 2^s$ the sum of squares between the d samples on which the systematic sampling variance is based can be split into a set of component parts which exhibit its structure. This

analysis is of interest in connexion with the general problem of estimating the error of systematic samples from observational material.

Following the ordinary procedure of partitioning the contrasts between 2^s quantities into a set of orthogonal contrasts of the type z_1-z_2, z_3-z_4, ..., $z_1+z_2-z_3-z_4$, ..., we have

D.F.	contrasts	range of r
2^{s-1}	$_rK_{\frac{1}{2}d} = {}_rH_d - {}_{(r+\frac{1}{2}d)}H_d$	1 to $\frac{1}{2}d$
2^{s-2}	$_rK_{\frac{1}{4}d} = {}_rH_d + {}_{(r+\frac{1}{2}d)}H_d - {}_{(r+\frac{1}{4}d)}H_d - {}_{(r+\frac{3}{4}d)}H_d$ $= {}_rH_{\frac{1}{2}d} - {}_{(r+\frac{1}{4}d)}H_{\frac{1}{2}d}$	1 to $\frac{1}{4}d$
2^{s-3}	$_rK_{\frac{1}{8}d} = {}_rH_d + {}_{(r+\frac{1}{4}d)}H_d + \ldots - {}_{(r+\frac{1}{8}d)}H_d - {}_{(r+\frac{3}{8}d)}H_d - \ldots$ $= {}_rH_{\frac{1}{4}d} - {}_{(r+\frac{1}{8}d)}H_{\frac{1}{4}d}$	1 to $\frac{1}{8}d$
...	.....	...
$\frac{1}{2^s-1}$	$_rK_1 = {}_1H_d + {}_3H_d + \ldots + {}_{d-1}H_d - {}_2H_d - {}_4H_d \ldots {}_dH_d$ $= {}_1H_2 - {}_2H_2$	

where $_rK_s$ represents the sum of the terms s apart, taken alternately positive and negative, and beginning with the rth term.

$(_rK_{\frac{1}{2}d})^2$, therefore, will contain $(2n-1)$ products of values $\frac{1}{2}d$ apart, each with coefficient -2, $(2n-2)$ products of values d apart, each with coefficient $+2$, etc.

The covariance of values kd apart is $\sigma^2 b^{kd}$. Consequently

$$V[K_{\frac{1}{2}d}] = \sigma^2\{2n - 2(2n-1)\,b^{\frac{1}{2}d} + 2(2n-2)\,b^d - \ldots - 2b^{(n-\frac{1}{2})d}\}$$
$$= \sigma^2\left\{2n\frac{1-b^{\frac{1}{2}d}}{1+b^{\frac{1}{2}d}} + 2\frac{b^{\frac{1}{2}d} - b^{(n+\frac{1}{2})d}}{(1+b^{\frac{1}{2}d})^2}\right\}.$$

In a long sequence, therefore, the variance (on a per term basis) tends to

$$E_{\frac{1}{2}d} = v[K_{\frac{1}{2}d}] = \frac{1-b^{\frac{1}{2}d}}{1+b^{\frac{1}{2}d}}\sigma^2.$$

Similarly
$$E_{\frac{1}{4}d} = v[K_{\frac{1}{4}d}] = \frac{1-b^{\frac{1}{4}d}}{1+b^{\frac{1}{4}d}}\sigma^2, \quad \text{etc.}$$

The variance of systematic samples at spacing d in a long sequence will therefore be given by

$$S_d = v[H_d] = \frac{\sigma^2}{2^s-1}\left[2^{s-1}\frac{1-b^{\frac{1}{2}d}}{1+b^{\frac{1}{2}d}} + 2^{s-2}\frac{1-b^{\frac{1}{4}d}}{1+b^{\frac{1}{4}d}} + \ldots + \frac{1-b^{d/2^s}}{1+b^{d/2^s}}\right].$$

Using the identity
$$\frac{1-x}{1+x} = 2\frac{1+x^2}{1-x^2} - \frac{1+x}{1-x},$$

we have
$$S_d = \frac{\sigma^2}{2^s-1}\left[2^s\frac{1+b^d}{1-b^d} - \frac{1+b^{d/2^s}}{1-b^{d/2^s}}\right]$$
$$= \frac{\sigma^2}{d-1}\left[d\frac{1+b^d}{1-b^d} - \frac{1+b}{1-b}\right].$$

In the case of a function, a subdivision in powers of 2 is possible for all values of d, since the function exists at all points. The process of subdivision can also be extended indefinitely,

the value of S_d being given by the limits of the above expression as s tends to infinity. In this case $2^s(1-b^{d/2^s}) \to -d\log_e b$ and therefore

$$S_d = \sigma^2\left[\frac{1+b^d}{1-b^d}+\frac{2}{\log_e b^d}\right].$$

This agrees with Cochran's expression.

Finally we have $$D_d/\sigma^2 = 1-b^d.$$

Table 5 gives the values of D_d/σ^2, E_d/σ^2 and S_d/σ^2 (for a function) for various values of b^d. The values of b^d, each of which is the square root of the preceding one, are chosen so as to exhibit the structure of S_d. Thus the value 0·90983 of S_d/σ^2 for $b^d = 2^{-32}$ is equal to

$$\tfrac{1}{2}.0{\cdot}99997+\tfrac{1}{4}.0{\cdot}99222+\tfrac{1}{8}.0{\cdot}88235+\ldots.$$

The values of S_d/σ^2 for the corresponding autoregressive sequences can also be quickly calculated from the table, using the first of the above formulae, provided d is a power of 2. Thus in a sequence with $b = 2^{-\frac{1}{4}}$

$$S_{16}/\sigma^2 = \frac{\tfrac{1}{2}.0{\cdot}6+\tfrac{1}{4}.0{\cdot}33333+\tfrac{1}{8}.0{\cdot}17157+\tfrac{1}{16}.0{\cdot}08643}{\tfrac{1}{2}+\tfrac{1}{4}+\tfrac{1}{8}+\tfrac{1}{16}} = 0{\cdot}43755.$$

This may be compared with the value of 0·41199 for the corresponding autoregressive function.

TABLE 5. VALUES OF D_d/σ^2, E_d/σ^2 AND S_d/σ^2 FOR AUTOREGRESSIVE FUNCTIONS

b^d	D_d/σ^2	E_d/σ^2	S_d/σ^2
2^{-32}	1·00000	1·00000	0·90983
2^{-16}	0·99998	0·99997	0·81969
2^{-8}	0·99609	0·99222	0·64717
2^{-4}	0·9376	0·88235	0·41199
2^{-2}	0·75	0·6	0·22397
2^{-1}	0·5	0·33333	0·11461
$2^{-\frac{1}{2}}$	0·29289	0·17157	0·05765
$2^{-\frac{1}{4}}$	0·15910	0·08643	0·02887
$2^{-\frac{1}{8}}$	0·08300	0·04330	0·01444
$2^{-\frac{1}{16}}$	0·04240	0·02166	0·00722
$2^{-\frac{1}{32}}$	0·02143	0·01083	0·00361

(*b*) *Estimation of the sampling error without supplementary observations*

We may now consider whether, in material of the autoregressive type, the sampling variance of a systematic sample can be calculated from the sample values, without any supplementary observations.

If the material is known to be truly autoregressive, the problem presents no difficulty. D_d and σ^2 may be estimated from the sample values (the latter, in a long sequence, being equal to the total variance of the individual values). Once these are determined b^d and S_d may be obtained from the above formulae. This procedure is tentatively suggested by Cochran in the paper referred to above.

Unfortunately, in practice it is unlikely that there can be any certainty that the autoregressive law is followed exactly. The effect of applying the above procedure to material which is not truly autoregressive must therefore be investigated.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Two simple types of departure from the autoregressive law may be considered: first, the case in which autoregressive variation is superimposed on a long-term variation, and secondly, the case in which the individual values of an autoregressive sequence are subject to a superimposed random variation.

In the first case, if the long-term variation is sufficiently gradual relative to the sampling interval, neither S_d nor D_d will be appreciably affected, but the overall variance will be increased. If the overall variance is σ_0^2 and the autoregressive variance is σ^2, with $\sigma_0^2 = \lambda\sigma^2$, denoting estimated values by the suffix e, we have

$$(b^d)_e = 1 - \frac{D_d}{\sigma_0^2} = \frac{\lambda - 1}{\lambda} + \frac{1}{\lambda} b^d.$$

If, for example, $\lambda = 2$ and $b^d = 0{\cdot}5$, $(b^d)_e = 0{\cdot}75$, and, for a function,

$$(S_d)_e = 0{\cdot}0481\sigma_0^2 = 0{\cdot}0962\sigma^2,$$

whereas the true value of S_d will be $0{\cdot}1146\sigma^2$.

The existence of long-term variation, therefore, though it markedly affects the correlation between neighbouring values, and therefore the estimate of b^d, does not seriously affect the estimate of the systematic sampling variance, since there is a compensating increase in the overall variance.

In the second case, S_d, D_d, and the overall variance, will all be increased by the random variance σ_1^2. In this case, if $\sigma_1^2 = \mu\sigma^2$,

$$(b^d)_e = \frac{b^d}{1 + \mu}.$$

If, for example, $\mu = \frac{1}{4}$ and $b^d = 0{\cdot}5$, $(b^d)_e = 0{\cdot}4$, and when d is large

$$(S_d)_e = 0{\cdot}1506(\sigma^2 + \sigma_1^2) = 0{\cdot}188\sigma^2, \quad S_d = 0{\cdot}1146\sigma^2 + \sigma_1^2 = 0{\cdot}365\sigma^2.$$

In this case, therefore, there is a very serious underestimation of the systematic sampling variance, even when the random component of variance is quite small relative to the autoregressive component. Nor is it likely that such deviations from the true autoregressive law will be detected by examination of the correlogram, particularly as the two types of disturbance produce effects which tend to mask each other in the correlogram.

An alternative approach would be to use the estimates of E_d and D_d, instead of σ^2 and D_d, in order to determine b_d and hence S_d. This eliminates disturbances due to long-term variation, but there is still substantial underestimation of S_d when there is a superimposed random component. S_d is also seriously affected by errors of estimation in E_d.

We therefore reach the conclusion, from the study of the effects of simple disturbances on the autoregressive type of variation, that no reliable estimate of the systematic sampling variance can be obtained from the values of a single systematic sample. The fact of the matter is that it is impossible to predict the behaviour of a sequence or function at intermediate values from the known values at the equally spaced sampling points, unless the mathematical law of variation is already known. The best estimate that is available from the values of a single systematic sample appears to be E_d. The general relation already established

$$S_d = \tfrac{1}{2}E_{\frac{1}{2}d} + \tfrac{1}{4}E_{\frac{1}{4}d} + \tfrac{1}{8}E_{\frac{1}{8}d} + \ldots$$

indicates that, provided E_d, $E_{\frac{1}{2}d}$, $E_{\frac{1}{4}d}$, ... are a decreasing sequence, E_d will always be an overestimate of S_d. The degree of overestimation in the case of an autoregressive function can be seen from table 5, the ratio E_d/S_d being only slightly greater than unity when b^d is small, but tending to the value of 3 as b^d approaches 1. If, however, there is a superimposed random variation, the ratio E_d/S_d will have a maximum at some intermediate value of b^d, and will tend to unity as b^d approaches 1, since the random component then becomes dominant. In the sampling of certain types of mathematical function, such as the normal curve, with no superimposed random variation, E_d will give an even greater degree of overestimation, but material of this type is not likely to occur very frequently.

It may be noted that, apart from errors of estimation, E_d will give a better estimate of the systematic sampling error than will D_d. In the autoregressive function, for example, the ratio D_d/E_d tends to 2 as b^d tends to 1. If the ratio D_d/E_d is large it may be taken as prima facie evidence that E_d is itself likely to be a substantial overestimate of S_d, but any quantitative assessment of the degree of overestimation based on the numerical value of D_d/E_d would be dangerous, as will be apparent from the numerical results reported in § 5 (*c*).

(*c*) *Use of supplementary partial systematic samples*

If an accurate estimate of S_d is required supplementary sampling must be undertaken. Two types will be considered, namely, the location of supplementary partial systematic samples at random relative to the main sample (using the same sampling interval) and the location of such samples systematically at the mid and quarter points relative to the main sample.

If the location is random then clearly an unbiased estimate of the systematic sampling variance will be obtained, apart from any errors introduced on account of the limited range of the partial samples. If the location is systematic the mid-point samples, in conjunction with the original sample, will provide an estimate of $E_{\frac{1}{2}d}$, and the quarter-point samples, if they are taken, will provide an estimate of $E_{\frac{1}{4}d}$, together with additional information on $E_{\frac{1}{2}d}$. In this case an unbiased estimate of S_d will not be available, but with quarter-point sampling the estimate (for a function)

$$S_d = \tfrac{1}{2}E_{\frac{1}{2}d} + \tfrac{1}{2}E_{\frac{1}{4}d}$$

is not likely to be seriously in error, the actual error being

$$\tfrac{1}{4}E_{\frac{1}{4}d} - \tfrac{1}{8}E_{\frac{1}{8}d} - \ldots,$$

i.e. less than $\frac{1}{4}E_{\frac{1}{4}d}$. In the case of a sequence the error is slightly less. When $d = 16$, for example, the error will be

$$\tfrac{3}{15}E_4 - \tfrac{2}{15}E_2 - \tfrac{1}{15}E_1,$$

i.e. less than $\frac{1}{5}E_4$. Even mid-point sampling will often give an estimate which is sufficiently accurate for practical purposes.

Values of the above expressions for an autoregressive function can be immediately obtained from table 5. Comparison of $E_{\frac{1}{2}d}/\sigma^2$ with S_d/σ^2 gives the error with mid-point sampling. Where $b^d = 0{\cdot}5$, for example, the true error is $0{\cdot}115\sigma^2$ against the value of $0{\cdot}172\sigma^2$ given by $E_{\frac{1}{2}d}$. Over a wide range, therefore, the error is very close to the limiting value of 50 % in excess. With quarter-point sampling and $b^d = 0{\cdot}5$ the estimate is

$$\tfrac{1}{2}(0{\cdot}172 + 0{\cdot}086)\,\sigma^2 = 0{\cdot}129\sigma^2,$$

i.e. close to the limiting value of $12\frac{1}{2}\%$ in excess. Any superimposed random component of variation will reduce the proportional error substantially. In a sequence with $\sigma_1^2 = 0{\cdot}25\sigma^2$ and $d = 16$ the corresponding excesses are 16 and 3 % respectively.

From the point of view of providing unbiased estimates of systematic sampling error, therefore, random location of the supplementary partial systematic samples is preferable. On the other hand, systematic location has certain advantages, in that the location of the sampling points is easier, the additional information is of greater value for such purposes as mapping, and the results can be incorporated as part of an additional systematic sample if subsequent increase in accuracy is required. Moreover, with systematic location the accuracy to be expected with less intensive sampling is somewhat more easily judged—with a supplementary sample at half-spacing only, for example, a good estimate of the sampling error at spacing $2d$ can be made. Somewhat better use can also be made of the supplementary observations in that overlapping partial samples can be constructed. One of the practical difficulties in the determination of the systematic sampling error by randomly located supplementary partial samples is that, even with samples of only four points, each set of four observations only contributes one degree of freedom to sampling error. (This point will be discussed in more detail in § 5 (*e*).)

(*d*) *Effect of variation in length of partial systematic samples of an autoregressive function*

In § 2 (*f*) it was stated that it was considered that a range $4d$ for the partial samples represented a reasonable compromise between economy of material and elimination of the long-term compensatory effects from sampling error. In the case of an autoregressive function it is possible to evaluate the errors that arise from this source.

Consider the estimation of $E_{\frac{1}{2}d}$ from contrasts between partial systematic samples of the type

$$ {}_rG_d(k,c) - {}_{r+\frac{1}{2}d}G_d(k,c).$$

The maximum and minimum errors will occur when $r = \frac{1}{2}d$ and $r = \frac{1}{4}d$ respectively, and it will therefore be sufficient to evaluate the expectations of the square of the above expression divided by the appropriate divisors, $2k-\frac{1}{2}$ and $2k-\frac{3}{4}$, for these two values of r. Denoting these expectations by A and A' respectively, and putting $b^{\frac{1}{2}d} = u$, we find

$$\frac{A}{\sigma^2} = \frac{1-u}{1+u} + \frac{2u(1-u)+u^{2k}(1-u)^2}{(4k-1)\,(1+u)^2},$$

$$\frac{A'}{\sigma^2} = \frac{1-u}{1+u} + \frac{2u(1-u)\,(3-u)-u^{2k-1}(1-u)^4}{2(8k-3)\,(1+u)^2}.$$

Similarly, if B denotes the expectation of the square of

$$ {}_{\frac{1}{4}d}G_d(k,c) - {}_{\frac{1}{2}d}G_d(k,c) + {}_{\frac{3}{4}d}G_d(k,c) - {}_dG_d(k,c)$$

(for which the error of estimation of $E_{\frac{1}{4}d}$ is maximum), divided by the appropriate divisor, $4k-\frac{5}{4}$, and $v = b^{\frac{1}{4}d}$, we find

$$\frac{B}{\sigma^2} = \frac{1-v}{1+v} + \frac{2v(1-v)\,(2-v+v^2)+v^{4k-2}(1-v)^2\,(1+v^2)^2}{2(16k-5)\,(1+v)^2}.$$

The last term in each case represents the bias. These biases are roughly inversely proportional to k. Their values in percentage terms when $k = 4$ are shown in table 6.

TABLE 6. PERCENTAGE BIASES IN VARIANCE ESTIMATES DERIVED FROM PARTIAL SYSTEMATIC SAMPLES OF AN AUTOREGRESSIVE FUNCTION

b^d	A	A'	B
2^{-16}	0·05	0·04	0·19
2^{-4}	2·67	1·90	0·99
2^{-1}	5·60	3·28	1·47
$2^{-\frac{1}{4}}$	6·51	3·44	1·66
limit	6·67	3·45	1·69

These biases are sufficiently small for practical purposes. In the case of certain mathematical functions, such as the normal curves of error, the biases in percentage terms will be much larger, but it is unlikely that empirical estimates of error by means of partial systematic samples will be required in such cases. Any superimposed random component of variation will reduce the biases. It therefore appears that the use of partial systematic samples of four terms for the estimation of sampling error will be satisfactory.

(*e*) *Use of overlapping partial samples*

When estimating such quantities as $E_{\frac{1}{2}d}$ from supplementary observations at half-spacing the partial systematic samples may either be overlapping or non-overlapping. If the partial samples are spaced at $4d$, so that there is no overlap except for the ends, there will be only one-quarter the number of contrasts that will be obtained with a spacing of d. These latter contrasts, however, will be by no means independent, and the question therefore arises as to how much additional information will be obtained by the use of overlapping samples. A similar situation arises when calculating D_d from successive differences of observations at interval d, where all possible differences or alternate differences may be taken.

A complete analysis of this problem in, for example, an autoregressive sequence would be complicated. The solution in the case of a random sequence, however, is simple, and will give an indication of the additional information in material subject to a large random component of error.

We will consider first the calculation of D_d. Denote the successive values by $y_1, y_2, \ldots$, with variance σ^2 and let $z_1 = \frac{1}{2}(y_1 - y_2)^2$, $z_2 = \frac{1}{2}(y_2 - y_3)^2$, ... Then $V(z) = 2\sigma^4$, and we find $\operatorname{cov}(z_1 z_2) = \frac{1}{2}\sigma^4$. Consequently in a sequence with $2n$ terms, when n is large,

$$V(\bar{z}) = \frac{1}{2n}\{V(z) + 2\operatorname{cov}(z_1 z_2)\} = \frac{3}{4n}V(z).$$

If alternate z's are taken, so that all the z's are independent,

$$V(\bar{z}) = \frac{1}{n}V(z).$$

There is thus a gain in information of one-third by taking all successive differences.

A similar procedure shows that with partial systematic samples and supplementary observations at half spacing, if $4n$ samples overlapping by 3 units are taken

$$V(\bar{z}) = \frac{0{\cdot}75}{n}V(z),$$

if $2n$ samples overlapping by 2 units are taken

$$V(\bar{z}) = \frac{0{\cdot}79}{n}V(z),$$

while if n samples only overlapping at the ends are taken

$$V(\bar{z}) = \frac{1{\cdot}001}{n} V(z).$$

The gain by doubling the number of contrasts by taking samples overlapping by 2 units is therefore somewhat over 25 %, but the additional gain by a further doubling is only of the order of 5 %.

6. Results of experimental sampling of numerical material

In the course of the present investigation trial samplings were carried out on various sequences of numerical material. A good deal of this work was exploratory, and the results obtained suggested a number of the procedures that have already been outlined. This numerical work will be briefly reported, as the results illustrate the various points at issue and provide practical demonstration of the validity of the conclusions that have been reached.

(*a*) *Altitudes*

This sequence consists of 296 approximate altitudes (of which the first 288 were retained for analysis) at intervals of 0·1 mile along the west-east grid line 1355 of sheet 96 (Hertford and Saffron Walden) of the 1 in. O.S. map of England and Wales (5th ed.). The land covered by the line is undulating with altitudes ranging from 200 to 450 ft.

The altitudes were obtained by plotting the points given by the 50 ft. contour intersections, and also the estimated locations and altitudes of the maxima and minima. A line was then drawn through these plotted points, and the altitudes at 0·1 mile intervals were read off.

The method of determining the altitudes has naturally resulted in a certain smoothing of the minor variations, and the decrease in variance with close sampling is therefore somewhat more extreme than would occur with actual altitudes. The results, however, may be taken as typical of the type of results that will be obtained when sampling material which is subject to the 'smooth-curve' type of variation.

In the tables which follow, all mean squares have been expressed on a 'per point' basis in units of $(1\,\text{ft.})^2$. No correction for 'finite population' has been introduced, as the 288 values can themselves be regarded as a sample of a continuous function.

In order to test the effects of end-corrections and supplementary observations the values of $v[h_d]$, $v[h'_d]$ and $v[h''_d]$ were obtained for all values of d from 2 to 20. In spite of the length of the sequence appreciable gains in accuracy were shown for most values of d, there being 14 positive and 5 negative values of the logarithms of the ratio of the variances of $v[h_d]/v[h'_d]$, and 12 positive and 6 negative values (+ one zero) of $v[h'_d]/v[h''_d]$. The mean gains (derived from the means of the logarithms) are shown in table 7. (The values for $d = 2, 3, 4$ have been omitted from this table.)

Table 7. Altitudes: gain in accuracy resulting from end-corrections and supplementary observations ($d = 5$ to 20)

	$v[h_d]/v[h'_d]$	$v[h'_d]/v[h''_d]$
$d = 6, 8, 9, 12, 16, 18$	1·40	1·19
remainder	1·07	1·06

As might be expected, the largest gains occur when l is a multiple of d. Under these circumstances there is an average gain of 40 % with end-corrections, and a further gain of 19 % with supplementary observations. Apart from this, and irregularities with small d, the gains do not show any marked association with sampling interval.

The estimates S_d of the systematic sampling errors given by partial systematic samples were also calculated for $d = 2$ to 6, 8 to 10, 12, 15, 16, 18, 20, using samples of length $k = 4$, with overlap only at the ends, and omitting the parts at the end of the sequence which did not make complete partial samples. (Complete coverage with overlap at the intermediate points would have been slightly more satisfactory.) The values of these estimates, together with the available values of R_d, R_{2d}, E_d and D_d, and the values of $v[h''_d]$ (black circles) are shown in figure 2, which has been plotted on a logarithmic scale in both directions.

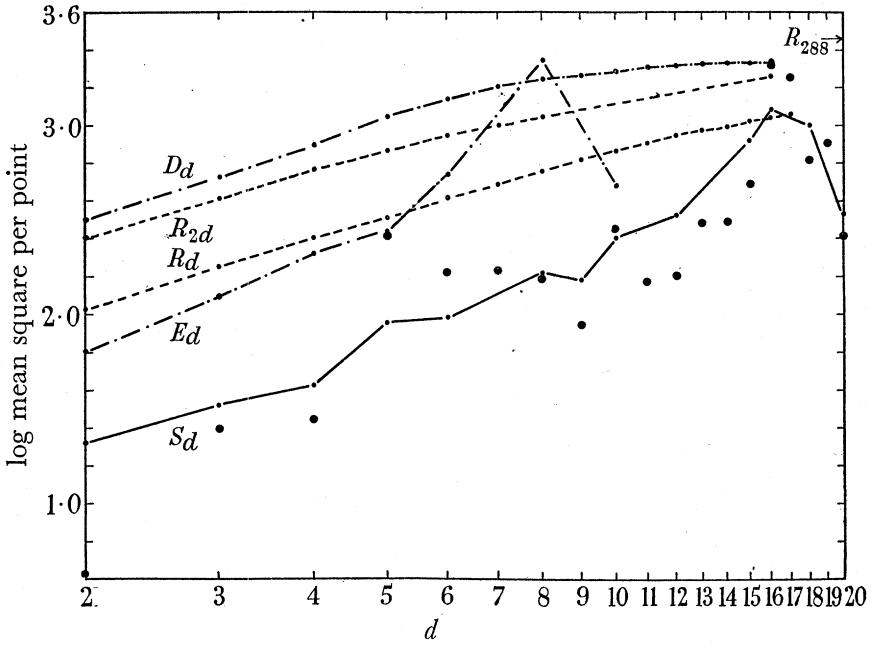

FIGURE 2. Altitudes: conspectus of variability (log mean squares per point). Actual systematic samples, $v[h''_d]$, are represented by black circles, partial systematic samples, S_d, by the full line, random samples in blocks of d and $2d$, R_d and R_{2d} by broken lines, differences, D_d, and sums taken alternately positive and negative, E_d, of points d apart by lines of alternate dots and dashes. The small black circles on all these lines indicate the computed values.

The regularity of R_d is due to the fact that the values, except for R_{32}, have been calculated from the values of D_s, and therefore approximately represent the values that would be obtained with all possible block boundaries. In blocks of 16, for example, the mean squares vary from 890 to 1310 according to the choice of block boundary, the mean being 1080, compared with the value of 1110 from D_s.

Various points are apparent from the figure. In the first place it will be seen that S_d conforms satisfactorily to the trend of values of $v[h''_d]$. The deviations are no greater than may be expected from chance causes, having regard to the small number of values of h''_d for small d, and the lack of independence of the values amongst themselves.

All forms of sampling increase rapidly in accuracy as the sampling interval is decreased, the precision per point being more than doubled, for example, for stratified random samples when the block size is decreased from 8 to 4 units.

Over the lower part of the range systematic sampling is about four times as accurate as stratified random sampling with one point per block. For higher values of d, however, the relative accuracy falls away, and there is a peak at $d = 16$ where systematic sampling, as judged by S_d, is no more accurate than stratified random sampling. This, and the even poorer performance of the actual systematic samples, is a reflexion of the fact that certain of the major features of the variation in altitude happen to coincide with the sampling interval, but this is scarcely likely to be more than a chance effect. It may be noted that even the least accurate of the values of h''_{16} has a discrepancy of less than twice the standard deviation of the corresponding stratified random sample of one per block, and of only 1·5 times the standard deviation of the stratified random sample with two per block. Undoubtedly, however, as examination of the altitude curve (not reproduced) indicates the advantage of systematic over random sampling is fading away as d is increased, as also is the value of stratification. A degree of instability is also entering into the estimates of the systematic sampling variance due to the shortness of the sequence relative to the sampling interval.

The figure shows that D_d is a very poor estimate of the systematic sampling error in material of this kind. In general, E_d provides a much better estimate, though it behaves poorly at $d = 8$, for the reason already discussed. In general, also, the discrepancy between D_d and E_d provides an indication that even E_d is likely to be a substantial overestimate.

The estimation of error from the use of supplementary samples at half and quarter spacing only was tested for $d = 8, 12, 16, 20$. The results are shown in table 8. (The value for $E_{\frac{1}{2}d}$ for $d = 8$ differs from that shown in figure 2, being derived from partial systematic samples for $d = 8$ only for the table, but $d = 8$ and 16 for the figure.)

It will be seen that supplementary samples at half and quarter spacing provide a satisfactory estimate of the error, whereas supplementary samples at only half spacing give a substantial overestimate. The foregoing theory indicates that this is what is to be expected in material of this kind.

TABLE 8. ALTITUDES: ESTIMATION OF ERROR FROM SUPPLEMENTARY SYSTEMATIC SAMPLES AT HALF AND QUARTER SPACING ONLY (MEAN SQUARES PER POINT)

d	all possible locations S_d	half spacing only $E_{\frac{1}{2}d}$	half and quarter spacing $\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})$
8	162	249	156
12	332	540	331
16	1232	2214	1190
20	347	476	368

(*b*) *Potato yields*

The data, due to Kirk, were taken from a paper by Kalamkar (1932), and consist of the yields of 96 rows of potatoes each 132 ft. long. Each row was harvested in six sections, but for the purpose of the present investigation only the yields of complete rows have been considered.

Systematic samples at spacing $d = 3, 4, 6, 8, 12, 16$ were taken, and end-corrections, and corrections for supplementary observations, were applied. The ratios of the resultant variances are shown in table 9.

The use of end-corrections has on the average reduced the sampling variance by about one-sixth, but supplementary observations have effected no further improvement.

Table 10 gives a conspectus of the values (on a 'per row' basis) of various variance estimates that were calculated for this sequence. No corrections for 'finite population' have been introduced. The variance within blocks, R_d, decreases considerably as block size is reduced, the value for $d = 3$ being less than one-tenth that for the whole sequence. Consequently, random sampling with one point per block is decidedly more efficient than random sampling at the same density with two points per block, the geometric mean of R_{2d}/R_d (for $d = 3, 4, 6, 8, 12, 16$) being 1·39. The systematic sampling variance, S_d, as estimated from partial systematic samples, is less than or equal to R_d for all these values of d, the geometric mean of the ratio R_d/S_d being 1·37. There is therefore an average further gain in precision of 37 % with systematic sampling over random sampling with one point per block. Without end-corrections the gain would be rather less than half this. $v[H'_d]$ shows reasonable agreement with S_d, the arithmetic mean of the ratio $v[H'_d]/S_d$ being 0·98.

TABLE 9. POTATOES: EFFECT OF END-CORRECTIONS AND SUPPLEMENTARY OBSERVATIONS

d	$v[H_d]/v[H'_d]$	$v[H''_d]/v[H'_d]$
3	1·09	0·83
4	1·87	0·88
6	0·90	1·03
8	2·59	0·79
12	0·86	1·14
16	0·73	1·17
geometric mean	1·20	0·96

TABLE 10. POTATOES: COMPARATIVE VALUES OF DIFFERENT ESTIMATES OF VARIANCE (LB.)^2

d	R_d	$v[H'_d]$	S_d	D_d	E_d	$E_{\frac{1}{2}d}$	$\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})$
3	32·0	10·1	24·2	51·1	45·9	—	—
4	40·0	32·3	40·0	57·5	35·6	29·5	45·2
6	37·5	54·1	29·8	83·2	48·0	39·0	—
8	50·4	18·7	36·2	112·3	80·1	36·8	31·4
12	67·2	48·7	44·6	166·6	—	48·4	53·2
16	88·5	59·0	47·0	228·5	—	59·4	41·0
24	75·0						
32	121·3						
48	267·2						
96	389·2						

The mean square difference, D_d, between successive values d apart greatly overestimates the systematic sampling error, there being some improvement, but still considerable overestimation, by using E_d. Finally, there is in this series no consistent difference between $E_{\frac{1}{2}d}$ and $\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})$, either set of values giving a tolerable estimate of S_d.

In conclusion, it should be noted that though no marked periodicities are shown by these data, cultivation often produces periodic variation in fertility in agricultural land, and for this reason systematic sampling of whole rows is not in general advisable for agricultural crops.

(c) *Air and soil temperatures*

Three sequences of daily temperature readings were taken from the Rothamsted records, each covering a period of 192 consecutive days. These are as follows:

(1) Air temperature (maximum in screen). 27 March 1945 to 4 October 1945.

(2) Soil temperature (under bare soil) at 4 in. 15 January 1945 to 25 July 1945.

(3) Soil temperature (under grass) at 12 in. 15 January 1945 to 25 July 1945.

These sequences were analyzed in a similar manner to the potato yield data.

The effects of end-corrections and supplementary observations are shown in table 11. No calculations for supplementary observations were made in the case of air temperatures, as it was apparent that no appreciable gain was likely to result.

TABLE 11. TEMPERATURES: EFFECT OF END-CORRECTIONS AND SUPPLEMENTARY OBSERVATIONS

	air	soil (4 in.)		soil (12 in.)	
d	$v[H_d]/v[H'_d]$	$v[H_d]/v[H'_d]$	$v[H'_d]/v[H''_d]$	$v[H_d]/v[H'_d]$	$v[H'_d]/v[H''_d]$
4	0·90	1·14	2·46	5·47	3·26
8	0·86	1·85	1·05	5·45	0·79
12	1·12	2·31	0·98	10·05	0·87
16	0·85	1·81	1·08	15·76	0·28
24	0·59	1·72	1·15	2·51	1·07
geometric mean	0·84	1·72	1·26	6·52	0·92

TABLE 12. TEMPERATURES: COMPARATIVE VALUES OF DIFFERENT ESTIMATES OF VARIANCE (DEG. F)^2

d	R_d	$v[H'_d]$	S_d	D_d	E_d	$E_{\frac{1}{2}d}$	$\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})$
			(1) air temperature				
4	17·3	15·0	14·3	37·0	26·4	12·8	15·0
8	25·7	15·0	21·0	59·4	34·9	26·4	17·2
12	33·4	22·1	21·0	62·2	84·5	19·4	23·4
16	43·2	32·1	28·1	56·7	93·0	34·9	30·4
24	46·5	37·5	52·5	35·5	29·1	84·5	48·8
32	46·6						
48	47·3						
192	57·6						
			(2) soil temperature (4 in.)				
4	7·7	5·5	4·5	13·9	9·7	6·4	3·6
8	10·3	11·0	7·4	23·7	16·5	9·7	7·8
12	14·8	6·3	7·6	29·2	23·8	7·1	8·2
16	14·4	8·9	13·1	27·5	34·4	16·5	15·0
24	20·6	27·2	16·2	25·3	15·6	23·8	15·3
32	24·0						
48	28·7						
192	113·6						
			(3) soil temperature (12 in.)				
4	1·43	0·58	0·42	3·76	1·62	0·62	0·32
8	2·49	1·43	1·20	8·10	3·27	1·62	1·10
12	5·44	0·62	1·33	11·08	7·94	1·56	1·56
16	4·73	0·39	2·37	11·76	13·38	3·27	2·53
24	6·93	8·04	4·80	12·93	7·84	7·94	4·70
32	9·10						
48	13·87						
192	86·03						

As is to be expected, end-corrections are of no value in the air temperature sequence, in which the average values at the two ends of the sequence are similar, but such corrections have given substantial gains in accuracy in the case of the 4 and 12 in. soil temperatures, where differences between winter and summer temperatures are involved. There is little further gain with supplementary observations; the apparent average gain of 26% for 4 in. soil temperatures is due to the high value for $d = 4$, which is almost certainly fortuitous.

The values of the various variance estimates are shown in table 12, which corresponds to table 10 for potatoes. R_d decreases markedly with block size, the greatest proportionate decrease being with 12 in. soil temperature, as is to be expected. (The values for air and soil temperatures are not strictly comparable owing to the different periods of the year covered by the air and soil temperatures.) Systematic sampling shows gains over random sampling with one point per block in almost all cases, the gain being greatest with the 12 in. soil temperatures, where the random fluctuations are least. $v[H'_d]$ shows reasonable agreement with S_d in all cases.

The various estimates of S_d behave in the expected manner. $\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})$ provides a satisfactory estimate in all cases, $E_{\frac{1}{2}d}$ is somewhat of an overestimate, while E_d and D_d are considerably in excess, the discrepancies being greatest, as is to be expected, where the random fluctuations are least.

7. Summary of numerical results

For purposes of comparison the results obtained in the analysis of the five sequences of the last section are summarized in table 13. The table shows average values of certain ratios for the various sequences. In general, geometric means have been taken over all values of d for which values of the ratio in question were available. Exceptions are as follows. In all cases R_{2d}/R_d is calculated for the same values of d as R_d/S_d. In the altitude data $v[H_d]/v[H'_d]$ and $v[H'_d]/v[H''_d]$ are taken from table 7, and $\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})/S_d$ and $E_{\frac{1}{2}d}/S_d$ from table 8; other ratios involving S_d cover the values of $d = 2$ to 6, 8 to 10, 12, 15, 16, 18, 20, values of R_d and D_d being obtained by interpolation where necessary.

The ratios of the table must not be regarded as estimates of parameters which have definite fixed values for the type of sequence in question. In part the actual values of the ratio will depend on the range and values of d included in their calculation, in part on the properties of the particular sequence chosen for analysis. Thus, for example, the low value of $v[H_d]/v[H'_d]$ for air temperature is due to the fact that the period covered was deliberately chosen to illustrate the case in which end-corrections would be of little or no value, whereas in the case of the soil temperatures the period was deliberately chosen so as to give maximum chance of improvement with end-corrections.

Nevertheless, in spite of these qualifications, a number of general conclusions emerge from the table, which may be summarized as follows.

(*a*) *End-corrections*, $v[H_d]/v[H'_d]$. The importance of end-corrections is apparent. In cases in which there is a marked trend, and a fairly continuous type of variation, failure to apply them may, when l is a multiple of d, result in systematic samples being less accurate than random samples with one point per block (4 in. soil temperature), or even than random samples with two points per block (12 in. soil temperature).

(*b*) *Supplementary observations*, $v[H'_d]/v[H''_d]$. The additional gain with supplementary observations is in general small, and on the evidence of these results it would appear that they are not worth making, except possibly in the case of sequences subject to a continuous type of variation which are also short relative to the sampling interval. (The values tabulated do not include any allowance for the additional observations required; if the cost of a supplementary observation is taken as equal to that of an ordinary observation the ratios must be reduced on the average by about 5 %.)

TABLE 13. SUMMARY OF NUMERICAL RESULTS

			temperatures			
	ratio	potatoes	air	4 in.	12 in.	altitudes
gain with end-corrections	$v[H_d]/v[H'_d]$	1·20 (*a*)	0·84 (*a*)	1·72 (*a*)	6·52 (*a*)	1·40 (*a*) 1·07 (*b*)
additional gain with supplementary observations	$v[H'_d]/v[H''_d]$	0·96 (*a*)	not calculated	1·26 (*a*)	0·92 (*a*)	1·19 (*a*) 1·06 (*b*)
validity of partial systematic samples for estimating sampling error	$v[H'_d]/S_d$	0·98	0·93	1·18	0·98	1·10
gain with random sampling of 1 per block over 2 per block	R_{2d}/R_d	1·39	1·31	1·43	1·75	1·90
gain with systematic sampling over random sampling of 1 per block	R_d/S_d	1·37	1·26	1·45	2·42	2·99
error from additional observations at half and quarter spacing	$\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})/S_d$	1·00	0·99	1·00	0·97	1·00
error from additional observations at half spacing	$E_{\frac{1}{2}d}/S_d$	1·06	1·16	1·26	1·40	1·58
error from systematic sample only	E_d/S_d	1·57	1·87	2·07	3·56	4·22
error from successive differences	D_d/S_d	2·79	1·97	2·62	5·84	7·78

(*a*) l a multiple of d; (*b*) l not a multiple of d.

(*c*) *Estimation of systematic sampling error from partial systematic samples*, $v[H'_d]/S_d$. In so far as the data are capable of providing an answer on this point, partial systematic samples appear to provide a completely satisfactory estimate of the systematic sampling errors. The weighted arithmetic mean of $v[H'_d]/S_d$ is $1{\cdot}05 \pm 0{\cdot}11$. It may be expected that the mean ratio will show a slight positive bias, owing to errors in S_d, but the results appear to exclude the possibility of any large overestimation of the sampling error because of incomplete elimination of the continuous components of variation from the partial samples.

(*d*) *Gains with systematic sampling*, R_d/S_d *and* R_{2d}/R_d. The gain with systematic sampling over random sampling with one point per block is of the same order of magnitude as the gain with random sampling of one point per block over random sampling with two points per block. As is to be expected, the gains are most marked with variation of the continuous type (12 in. soil temperatures and altitudes), and in this case the gain with systematic sampling is somewhat greater than that due to reduction in block size.

(*e*) *Estimation of systematic sampling error from additional observations at half and quarter spacing*, $\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})/S_d$. This is completely satisfactory in this material.

(*f*) *Estimation of systematic sampling error from additional observations at half spacing only*, $E_{\frac{1}{2}d}/S_d$. This is reasonably satisfactory for the material which has no large continuous component of variation (potatoes and air temperatures). The degree of overestimation becomes

more marked as the continuous component increases, though it is never very large (25 % for the standard error in the case of the altitudes).

(*g*) *Estimation of the systematic sampling error from the data of the sample only, E_d/S_d and D_d/S_d.* Differences taken alternately plus and minus give estimates E_d which are in excess (in terms of the standard error) by amounts ranging from 25 to over 100 %. E_d is, however, decidedly better than the estimate D_d derived from successive differences, the difference being more marked with the more continuous types of material.

8. Relation between the gain in precision with a given intensity of sampling and the gain in efficiency

At first sight it might appear that the gain in efficiency resulting from the use of random sampling with one point per block instead of two points per block, or the use of systematic sampling instead of random sampling with one point per block, is equal to the gain in precision with a given intensity of sampling. This, however, is not the case, because with a more accurate type of sampling a lower intensity will be required to attain a given accuracy, and consequently the block size or sampling interval will be increased. If, for example, random sampling with one point per block is twice as precise, for a given intensity of sampling, as random sampling with two points per block ($R_{2d}/R_d = 2$), halving the number of sampling points with change from random sampling from two points per block to one point per block would result in doubling the block size, in which case the variance per point would be the same as that for the original sampling. The intensity of sampling required to give the same accuracy with random sampling of one per block is in fact $1/\sqrt{2}$ that required for random sampling with two points per block.

In general terms, if the ratio $R_{2d}/R_d = \phi$ remains constant over the relevant range, so that

$$R_x = kx^b,$$

where $b = \log\phi/\log 2$, the ratio i_2/i_1 of the intensities of sampling required to give the same accuracy in the final results with random sampling of two points per block and one point per block respectively is given by

$$\log\frac{i_2}{i_1} = \frac{\log\phi\,\log 2}{\log 2\phi}.$$

The situation with systematic sampling is somewhat more complicated, since, in addition to the ratio $R_d/S_d = \psi$ of the precision of systematic sampling to that of random sampling of one point per block with a given intensity of sampling, the change in R_d with change in block size is involved. The ratio i_1/i_s of the intensities required to give the same accuracy in the final results with random sampling of one point per block and with systematic sampling is given by

$$\log\frac{i_1}{i_s} = \frac{\log\psi\,\log 2}{\log 2\phi}.$$

Values of these ratios for the five numerical sequences investigated are shown in table 14, the values of ϕ and ψ being taken from table 13.

These results are obtained on the assumption that the accuracy required in the results is laid down at the start, and that the only problem concerning the statistician planning the sampling is to attain the required accuracy at minimum cost. This, however, although

common practice, is an over-simplification of the whole problem. If less costly methods of sampling are available, it will pay to increase the accuracy of the results, thereby reducing the magnitude and therefore the cost of errors.

TABLE 14. RELATIVE INTENSITIES OF SAMPLING REQUIRED TO GIVE THE SAME ACCURACY WITH RANDOM SAMPLING OF TWO POINTS PER BLOCK (i_2), RANDOM SAMPLING WITH ONE POINT PER BLOCK (i_1), AND SYSTEMATIC SAMPLING (i_s)

sequence	ϕ	ψ	i_2/i_1	i_1/i_s	i_2/i_s
potatoes	1·39	1·37	1·25	1·24	1·55
air temperature	1·31	1·26	1·22	1·18	1·43
4 in. soil temperature	1·43	1·45	1·27	1·28	1·62
12 in. soil temperature	1·75	2·42	1·36	1·63	2·22
altitudes	1·90	2·99	1·40	1·77	2·46

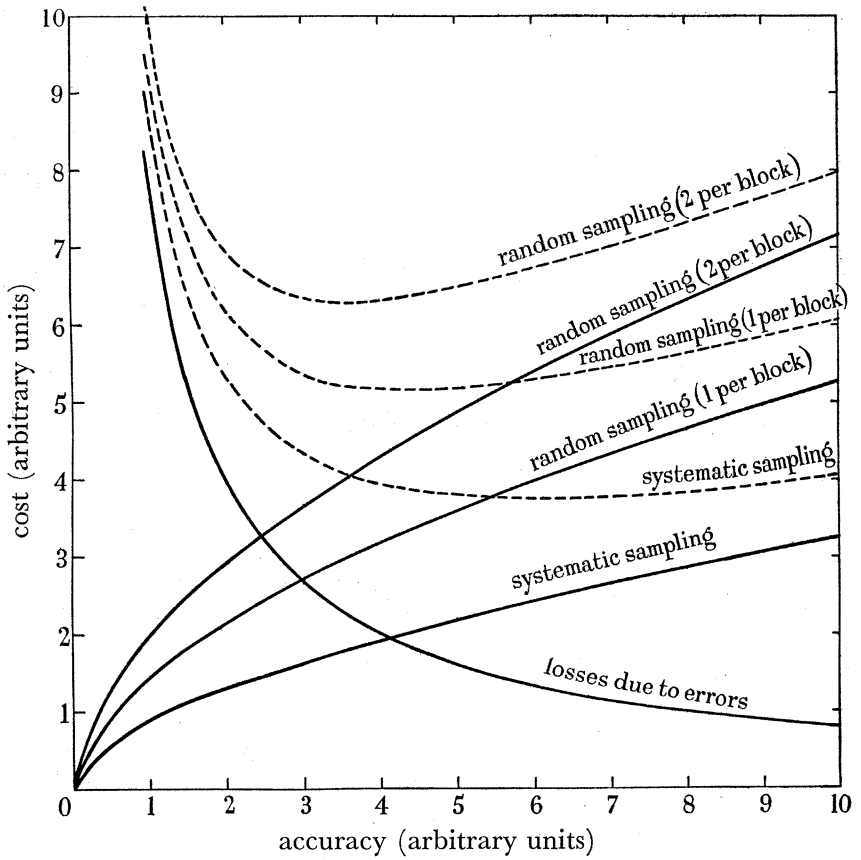

FIGURE 3. Diagram illustrating factors influencing the choice of the optimum intensity of sampling. The full lines through the origin show the cost of the sampling required to give varying degrees of accuracy in the case of the 12 in. soil temperatures ($\phi = 1{\cdot}75$, $\psi = 2{\cdot}42$), for systematic sampling and for random sampling with one and two points per block. The other full line gives a typical loss function of the errors in the results, assuming that costs of errors are proportional to their squares. The broken lines show the total costs of sampling and resultant errors.

The exact relationships for maximum efficiency will depend not only on the laws governing the various types of sampling error, and the cost of different types of sampling, but also on the form of the loss function due to errors in the results. As a simple example we have taken the case in which the additional cost per sampling point is the same whatever the type and intensity of sampling, and in which the loss due to (or cost of) an error is proportional to the square of that error.

The various cost functions were calculated on these assumptions for the values $\phi = 1{\cdot}75$ and $\psi = 2{\cdot}42$, which are those obtained in the 12 in. soil temperature sequence, and are shown in full lines in figure 3, together with a typical loss function for errors. (Apart from the values of ϕ and ψ the relationships are the same whatever the magnitude of the numerical constants.) The broken lines show the sums of the costs of the various types of sampling and the loss due to errors. Their minima indicate the accuracies which are required for maximum efficiency for the various types of sampling. The accuracy at the minimum for random sampling of two per block has a value of 3·57 units, that for random sampling of one per block has a value of 4·37 units, and that for systematic sampling has a value of 5·97 units. The relative intensities of sampling, which are given by the ordinates of the full lines at these points, are 4·04, 3·32 and 2·42 units, i.e. 1:0·82:0·60. The minimum total costs (sampling + loss due to errors) are 6·28, 5·15 and 3·76 units. The losses due to errors are decreased in the same proportion as the total costs, the actual costs of these losses being 2·24, 1·83 and 1·33 units respectively.

9. Conclusions

The present investigation shows that over a wide range of conditions one-dimensional systematic sampling of sequences and functions is more accurate than stratified random sampling, the average gain with systematic sampling over random sampling with one unit per block being of the same order as the gain with random sampling with one unit per block over random sampling with two units per block. These conclusions, however, only apply to systematic samples with random starting point if the effects of any trend in the sequence or function is removed by the use of end-corrections of the type described in § 2 (*d*). Since these corrections are easily applied, and require no supplementary observations, they should always be used in material subject to any appreciable trend. They are of particular importance when the total length of the sequence is a multiple of the sampling interval.

The use of supplementary observations at the ends of the type described in § 2 (*e*) is shown to be worth while only in exceptional circumstances such as arise in sampling material in which a long-term continuous component of variation is dominant (e.g. the half-normal curve described in § 4 (*d*)). In such material the use of centrally located systematic samples (which avoid the necessity of any form of end-adjustment) is usually more practical.

The investigation also confirms that no method of obtaining a really valid estimate of error from the sampling results themselves can be hoped for. The laws of variation to which different types of material are subject are too ill-defined, and too varied, for it to be possible to base any reliable estimate of error on assumptions of some definite law of variation. One of the outstanding advantages of random sampling is that the validity of the estimation of error is largely unaffected by the actual laws of variation followed by the material.

For this reason random sampling can be applied with complete confidence to material about which nothing is previously known. The degree of accuracy actually attained can be judged from the results. It may, of course, be found that the amount of sampling has been inadequate to give the accuracy required, in which case further sampling will have to be undertaken, but no false conclusions will be reached. If systematic sampling is used in such cases unsuspected periodicities in the material may lead to grossly misleading results.

Usually, however, a good deal is known about material on which sampling is to be undertaken. Even if it is not, it is often profitable, when extensive sampling investigations are being planned, to make a fairly thorough study of the type and degree of variability to which the material is subject by detailed investigation of typical parts of it. In such cases the relative advantages of systematic and random sampling can be properly investigated, and the dangers inherent in the uncritical use of systematic sampling can be avoided. The methods based on partial systematic samples described in the present paper make possible the estimation of the expected systematic sampling error with reasonable economy of material, either by the calculation of S_d directly (where all the values of a sequence are known) or by its estimation from the E series.

If it is decided to use systematic sampling for an investigation, it is important that some controls over the actual sampling errors should be available, whether or not a preliminary investigation has been undertaken. The nature of the control will depend on the material: in material which has only a moderate continuous component of variation E_d may be considered adequate control, in which case no supplementary observations will be required. As has been demonstrated, E_d, i.e. the mean square based on the sum of a sequence of terms of the sample taken alternately positive and negative, with suitable end-adjustments, is a decidedly better estimate of the systematic sampling error than is D_d, the mean square based on successive differences. Where there is any considerable continuous component the basis of control will have to be $E_{\frac{1}{2}d}$ or even $\frac{1}{2}(E_{\frac{1}{2}d}+E_{\frac{1}{4}d})$, in which case arrangements must be made to take supplementary observations where necessary. This, of course, will somewhat lower the efficiency of systematic sampling.

In planning the error control a certain degree of overestimation of the sampling error can often be tolerated. Naturally, also, by no means the whole of the material need be covered. In large-scale investigations it is only rarely that estimates of the sampling error are made from the whole of the material, even when the method of sampling is such that fully valid estimates are possible. Nevertheless some degree of control—analogous to quality control of an industrial product—is always advisable.

When, for one reason or another, random sampling is considered preferable to systematic sampling, there remains the question of whether such sampling should be two units per block or one unit per block. The only advantage of random sampling of two per block is that an estimate of error is automatically available for every part of the material if required, but this is not a very cogent reason if such estimates are not in fact computed. If, as is usual in extensive sampling, some form of error control is all that is required, it may well be better to use random sampling of one per block, supplementary observations at randomly located points being taken in a sufficient proportion of the blocks to give adequate error control. An alternative procedure, in material in which the block divisions are arbitrary, i.e. virtually located at random, is to estimate the error mean squares of differences of terms $s = 1, 2, 3, \ldots, d-1$, units apart from differences between terms in neighbouring blocks, building up the mean square within blocks of d from the formula given in § 2 (*a*). If d is at all large, however, the mean squares for small s, which have greatest weight in the final estimate, may not be sufficiently accurately determined without some supplementary observations. These observations may be located so as to give information only on the differences for small s. Extrapolation from higher values of s may also occasionally be used, though this is inadvisable if

the material is such that there may be considerable negative correlation between neighbouring terms of the sequence, such as arises, for example, from competition of plants in a row.

It has already been pointed out that the use of systematic sampling is dangerous on material with unsuspected periodicities. When the periodicities are known, however, it can be effectively used, by proper choice of sampling interval, either as a means of eliminating the effect of the periodicities from the sampling error, or for the purpose of providing comparative results all affected to the same degree by the periodicities. Many meteorological observations, such as daily readings of temperature taken at a fixed time of the day, provide examples of the latter type of sampling.

Finally, the findings of § 8 may be emphasized, which show the importance of taking into account the relative cost of the various sampling methods, and their relation to the losses due to errors in the results, when determining the degree of accuracy that is required. The true gains with systematic sampling over random sampling of one point per block, and of the latter over random sampling of two points per block are not as great as the gains in precision with sampling of a given intensity, owing to the loss of accuracy with decrease in intensity, which is more than proportionate to the decrease in number of sampling units. They are, however, appreciably greater than the gains which result if the intensity is adjusted so that the same overall accuracy is achieved. It always pays, when a more accurate method of sampling is available, to increase the overall accuracy over that which would be optimum for the less accurate method.

Finally, I should like to acknowledge the great assistance in the extensive numerical calculations rendered by various workers in my Department, particularly Mrs I. Mathison, Mrs R. O. Cashen, Miss P. M. Clarke and Mr H. D. Patterson. I also wish to thank Mr F. J. Anscombe and Mr M. H. Quenouille for their continuing interest in the progress of the investigation.

References

Aitken, A. C. 1939 *Statistical mathematics*. Edinburgh: Oliver and Boyd.
Cochran, W. G. 1946 *Ann. Math. Statist.* **17**, 164–177.
Kalamkar, R. J. 1932 *J. Agric. Sci.* **22**, 373–383.
Kendall, D. G. 1942 *Quart. J. Math.* **13**, 172–184.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES