Journal of Official Statistics
Vol. 2. No. 2, 1986, pp. 161–168

Implicit Longitudinal Sampling from Administrative Files: A Useful Technique

Alan B. Sunter[1]

Abstract: Official statistical agencies frequently use administrative files, in which the records have unique identifying numbers, as sampling frames for surveys and analyses which have longitudinal as well as cross-sectional dimensions. This paper demonstrates: that Poisson sampling may be used efficiently in cross-sectional sampling; that Poisson sampling provides a simple way of maximizing the number of sampled units common to successive cross-sectional samples in spite of changes in sample specification; and that the unique identifying numbers themselves, transformed by an appropriate hashing algorithm, may be used to generate the "random" numbers required to implement Poisson sampling.

Key words: Survey sampling; Poisson sampling; longitudinal surveys; random numbers; pseudorandom numbers.

1. Introduction

In countries with highly developed taxation and social security systems, official statistical agencies often have access to machine-readable administrative files that provide complete coverage of some population of interest. A major statistical application of such files is their use as sampling frames for more detailed enquiry through survey or through more extensive examination of the documents underlying the machine-readable records. Typically, the units (persons, corporations, etc.) in these files have unique identifiers, usually numeric but possibly an alphanumeric combination, such identifiers being essential to their various administrative functions.

The sample design may be addressed simply to the periodic compilation of cross-sectional population statistics. In this case it may be applied independently in each administrative cycle and without any particular regard to its efficiency for estimates of *changes* in the values of population parameters from one cycle to another. On the other hand, estimates of such changes are often of at least equal interest to those of current levels so that the sample design must seek some compromise between "cross-sectional efficiency" and "longitudinal efficiency." Furthermore, in addition to the requirement of longitudinal "descriptive statistics" (a term we use here to denote tabulations of population parameters), there is often an analytic requirement for samples of longitudinal records. The term "analytic" is used here to denote studies of individual behaviour or response usually in the context of some mathematical model purporting to describe such behaviour or response.

[1] President, A.B. Sunter Research Design & Analysis Inc., 63 Fifth Av., Ottawa, Canada, K1S 2M3.
The author is grateful to two anonymous referees for their careful reading and helpful suggestions.

However, *explicit* longitudinal samples, by which we mean procedures in which the samples selected in cycle t are tracked through cycles $t+1, t+2, \ldots$, tend to be both operationally difficult to maintain and to become increasingly inefficient for current cross-sectional purposes. Furthermore, we do not usually know *this* year what will be of analytic interest in, say, ten years so that we do know how to specify the longitudinal sample for analytic purposes. In such cases we cannot claim that any particular method of longitudinal sampling has optimal efficiency and, indeed, no such claim is made for the method described in this article. We claim merely that the method provides a simple compromise between the requirement for efficiency with respect to current estimates, on the one hand, and efficiency with respect to estimates of change over time, on the other.

The particular application referred to in this article illustrates some of these points. The administrative file contains records for about 15 000 000 persons submitting annual income tax statements. More information is available in the tax returns themselves than can or need be captured in machine-readable form for the primary administrative purposes of tax assessment. Furthermore, even for those data captured on a universe basis, the statistics from a relatively large sample (in this case, about 500 000 records) are considered adequate to the analytic and planning functions for which statistics are generated. A consideration in this case (and, usually, for other large administrative files) is that the main file is very active, while the consequences of errors in data processing tend to be serious, so that it is required that statistical applications be taken off-line with respect to the main application as early as possible in the processing cycle. The original function of the annual samples had been seen in terms of annual cross-sectional analysis only and the samples drawn each year were independent of those in previous years.

There is considerable change in the population from year to year, in terms both of births and deaths and of the characteristics of its members, so that maintaining appropriate sample representation for the domains of interest requires annual adjustment of the stratification and allocation parameters. Under these circumstances explicit longitudinal sampling, if given any consideration at all, had been dismissed as being too inefficient for the main purposes of current cross-sectional analysis. However, two emerging considerations led to a change in design:

(i) Some users of the statistics, at low levels of aggregation (e.g., small areas), are concerned primarily with change (e.g., of occupational distributions for small areas) and were dissatisfied with the longitudinal instability of the statistics being produced by the independent samples.

(ii) From time to time proposals for changes in tax regulations require a retrospective analysis to be made on what the impact on individual taxpayers of such changes would have been had they been introduced, say, five years ago. These analyses require longitudinal records for individuals in the class effected by the proposed changes. With independent sampling, however, longitudinal records for sets large enough for meaningful analysis were simply not available.

The technique described in this paper provides a simple way of using the unique identifier to embed an *implicit* longitudinal sample in a series of cross-sectional samples. It is not necessary to limit cross-sectional efficiency to satisfy the longitudinal requirement. The sample overlap from cycle to cycle is automatically maximized subject to the sampling probabilities imposed by the cross-sectional designs despite changes in stratification, changes in sampling rates, and movements of individual units between strata. Finally, the subsample common to any set of cycles is a probability sample of the population of units common to those cycles.

Since both the cross-sectional and longitudinal samples are stratified Poisson samples, we begin with a demonstration that Poisson sampling has efficiency, at least for large samples, essentially equivalent to that of simple random sampling. Strictly speaking, this demonstration applies only to the comparison of equal probability sampling methods but extends, of course, to stratified sampling provided the within-stratum samples remain large. It follows that Poisson sampling may be used in place of the more usual stratified simple random or systematic sampling to satisfy the requirements of cross-sectional efficiency. It then becomes very easy, as we shall see, to serve the interests of longitudinal efficiency by maximizing the overlap between successive samples.

2. Relative Efficiency of Poisson Sampling

Although our main concern in this article is with equal probability sampling (within strata) we can present this discussion, an abbreviation of a more extensive one in Sunter (1977), in rather more general terms.

Selection probabilities π_i are assigned to the units labelled $i=1,2,\ldots,N$. The units are then sampled independently with the assigned probabilities. In operational terms the usual procedure for this is as follows. As the record for each unit is read, the system determines its selection probability π_i, assigns a random number r_i in the range (0,1) from its stored random number table or pseudorandom number generator, and then selects the unit if $r_i < \pi_i$, rejects otherwise.

The realized sample size, n', is a random variable with expected value

$$n = E(n') = \sum_{i=1}^{N} \pi_i ,$$

and variance

$$V(n') = \sum^{N} \pi_i (1-\pi_i) . \qquad (1)$$

The unbiased estimate

$$\hat{Y} = \sum^{n'} y_i / \pi_i \qquad (2)$$

of the population total has variance

$$V(\hat{Y}) = \sum^{N} y_i^2 (1-\pi_i) / \pi_i . \qquad (3)$$

In the equal probability case $\pi_i = n/N$ this becomes

$$V(\hat{Y}) = (1-n/N)(N/n) \sum^{N} y_i^2 , \qquad (4)$$

which clearly does not compare favorably with the corresponding expression for SRS with the same (expected) sample size. It is natural then to consider the ratio form suggested by Brewer et al. (1972)

$$\hat{Y}_B = (n/n') \sum^{n'} (y_i/\pi_i) , \qquad (5)$$

in which we "adjust" for the discrepancy between expected and realized sample sizes. Note that we have assumed here, and will do so elsewhere, that $P(n'=0)$ is negligible. Keeping in mind that we are discussing large national surveys we can proceed in the usual way(see, for example, Sunter (1977)) to demonstrate that (5) is, for practical purposes, unbiased[2] and that it has variance given approximately by

$$V(\hat{Y}_B) = \sum^{N} (y_i - (Y/n)\pi_i)^2 (1-\pi_i)/\pi_i . \qquad (6)$$

[2] The bias is approximately

$$\mathrm{BIAS}(\hat{Y}_B) = (1/n) \sum^{N} (1-\pi_i) ((Y/n)\pi_i - y_i)$$

In the particular case of equal probability sampling this expression reduces to zero but, even in the general case, it is easily shown (see, for example, Cochran (1977), Section 6.8) to be negligible relative to the standard error, provided the c.v. of n' is less than, say, 0.1.

The nature of the reduction in variance is immediately clear from the comparison between (3) and (6). In the particular case of equal probability Poisson sampling, (5) reduces to the intuitively appealing estimate

$$\hat{Y}_B = N\bar{y}$$

and its variance (6) to

$$V(\hat{Y}_B) = (1-n/N)\, N/n \sum^{N}(y_i - \bar{Y})^2$$
$$= (1-n/N)\,(N(N-1)/n)S^2. \quad (7)$$

This result, rather surprising in that it is actually less than the variance under SRS for the same expected sample size (a consequence, presumably, of the approximation involved in its derivation), demonstrates the relative efficiency of Poisson sampling in large samples when the estimator (5) is used. It also tells us that we can *estimate* the variance, in equal probability sampling, *as if* the sample were SRS of size n.

3. Implicit Longitudinal Sampling

In what we have called the *usual procedure* for Poisson sampling, suppose that the random number associated with a unit on its first appearance in the administrative file (either at the inception of the sampling system or, subsequently, on its "birth" into the file) is then *permanently* associated with that unit, rather than a new number being assigned on each sampling occasion. This simple change immediately turns a system of independent samples into one that maximizes the number of sample units common to two or more occasions. This maximization is conditioned by changes in the sampling specification from one occasion to another. If there is no change in the sampling specifications with respect to stratification and sample allocation, the sample will remain (except for the addition of samples of births, the deletion of deaths, and changes induced by the migration of units from one stratum to another) unchanged. The term *implicit longitudinal sampling* is derived from this property of maximal overlap.

The proof that the overlap is maximal is trivial. It rests on the observation that if the sampling probabilities assigned by the cross-sectional specification to population unit i on occasions $t, t+1, t+2, \ldots$ are $\pi_{t,i}, \pi_{t+1,i}, \pi_{t+2,i}, \ldots$ then its joint probability of selection on all occasions is simply min $(\pi_{t,i}, \pi_{t+1,i}, \pi_{t+2,i}, \ldots)$. It follows that the longitudinal sample obtained in this way has maximal size given the selection probabilities assigned in each of the cycles.

For estimates applying to the population of units common to a number of occasions we may use the form (2), with a value assigned to π_i by the expression given in the preceding paragraph and with variances estimated by

$$v(\hat{Y}) = \sum^{n'}(1-\pi_i)y_i^2/\pi_1^2. \quad (8)$$

For estimates of the form (5) and variance estimates corresponding to (6), however, we would need (in order to determine n) to have tracked min $(\pi_{t,i}, \pi_{t+1,i}, \ldots)$ not only for the sample but for the whole population of units. (Estimates of the form (5) and (6) are, of course, available for current cross-sectional statistics.)

We emphasize, however, that the practical value of the scheme lies not so much in its capacity for unbiased estimation of the parameters of the "longitudinal population" but in the reduction of variances of estimates of change from one occasion to another. This is done without sacrificing current cross-sectional efficiency, at the same time assuring the availability of a set of longitudinal observations for analytic studies.

4. Poisson Sampling by Hash Number

In order to reduce the vulnerability of the statistical subsystem to changes in the administrative system (whose objectives, it should be remembered, will be primary), we now propose to modify the mode of implementation of Poisson sampling by using a random number generator, "seeded" by the unique unit identifier, to produce an integer in the closed range [000,999]. If unit i is assigned a selection probability $a_i/1\ 000$ (where a_i is an integer), either through stratification or as a function of one or more of its variate values, we then select if HASH(IDENT$_i$)$<a_i$, reject otherwise.

The functional notation HASH(IDENT) is appropriate since the "random" integer it represents is actually a deterministic function of the identification number[3] that generates it. We require that the random integers, or hash numbers, display the properties of a uniform distribution. Since the hash number *is* a function of the identification number, it can be reproduced whenever required.

The hash number routine itself must satisfy certain requirements. Loosely speaking, these are :

(i) that the population identification numbers and their hash numbers are uncorrelated. This requirement is intended to guarantee that selection probabilities are not affected by any information content of the identification number.

(ii) that the hash number derived from the population identification numbers should appear to be uniformly distributed, and this property (of uniformity) should continue to hold under restrictions on the values of the identification numbers.

Our hash number algorithm is based on the use of a random number generator of the multiplicative congruential type

$$x_{n+1} = kx_n \bmod(m)$$

seeded by the identification number. The parameters of the generator[4] are $k=7\text{exp}(6)= 117649$ and $m=2\text{exp}(31)-1=2147483647$. The hash number is constructed as follows :

$$x_0 = \text{ident}$$

$$x_1 = kx_0 \bmod(m)$$

$$x_2 = kx_1 \bmod(m)$$

$$\text{hash} = [1\ 000\,(\,(x_1 x_2)\bmod(m)\,)/m\,].$$

Verbally, we take the product of the first two numbers (mod m) in the sequence seeded by the identification number. We then transform this product to an integer in the required range in the usual way.

Tests of the hash number routine were performed as follows. Using randomly selected (by a different random number generator) starting points for sets of 500 9-digit numbers obtained by incrementing the starts by 1's, by 10's...., by 100 000's, we performed a standard χ^2 test for uniformity among the first two digits of the hash numbers corresponding to each set. For example, if we began with the 9-digit numbers 123456789 we calculated the hash numbers corresponding to the sequences :

[3] Alphanumeric identifiers will require the preliminary step of conversion to fully numeric unique identifiers to be included as part of the hash number function.

[4] This is not a paper about random number generators *per se* and no particular virtues are claimed for this generator other than that it satifies our requirements. Tests similar to those described in this paper should be made before using any other generator. For an extensive bibliography on random number generators and methods of testing them see Sowey (1972). A good discussion of the number theory underlying the choice of parameters in multiplicative congruential generators and of methods of testing will be found in Downham and Roberts (1967).

Incrementing by 1's :
123456790, 123456791,..., 123457289
Incrementing by 10's :
123456799, 123456809,..., 123461789
.....
Incrementing by 100 000's :
123556789, 123656789,..., 173356789

We then tested these sets of hash numbers for uniformity. Typical results are given in Table 1 below.

In addition to testing for uniformity we also tested for serial correlation using a procedure suggested by Good (1953). Using y_i to denote the hash number generated by the i-th member of the sequence we take each pair (y_i, y_{i+1}) as coordinates in a square of appropriate dimension. Dividing each dimension into 10 equal intervals and denoting frequencies in the cell (j,k) by f_{jk}, their expected values under the hypothesis of no serial correlation by N_{jk}, and the corresponding marginal values by $f_{j.}$ and $N_{j.}$ respectively, Good showed that the statistic

$$S^2 = \sum_{j,k=1}^{10} (1/f_{jk})(f_{jk}-N_{jk})^2 - (1/N_{j.}) \sum_{j=1}^{10} (f_{j.}-N_{j.})^2,$$

is asymptotically χ^2 on 90 degrees of freedom. Typical results for this test are also shown in Table 1. Finally, we calculated the correlation, within each set of 500, between the identification numbers and the hash numbers that they generated. In no case did this correlation exceed 0.003 in magnitude, a result from which we may conclude that the identification numbers and their corresponding hash numbers are uncorrelated.

In the ten trials were six cases (out of 60) in which the test statistic was below the 2.5% level for χ^2 on 99 degrees of freedom and three in which it was above the 97.5% level. None of these failures of uniformity occurred, however, for increments of 1 or 10 in the identifier and, for our purposes, we are satisfied with the performance of the hash number algorithm. The results for two of the ten trials are shown in Table 1.

Table 1. Uniformity and Serial Correlation in Hash Numbers for Arithmetic Sequences

Start	Increment	Uniformity	Serial Correlation
651860928	1	108.4	89.8
	10	89.2	111.5
	100	72.0*	63.9
	1 000	69.6*	78.9
	10 000	103.6	87.1
	100 000	106.4	102.7
490312768	1	98.8	73.6
	10	112.4	73.8
	100	134.8*	67.8
	1 000	84.4	75.5
	10 000	98.4	99.4
	100 000	94.4	109.4

* Significant at the 5 % level for a two-tailed test.

Our reason for interest in the lower end of the distribution of the test statistic is our concern that some numbers in the sequence of identifiers may be missing in some systematic way. This could occur, for example, if identifiers were made available in the administrative process in blocks for systematic assignement to individuals as they arrive in the system. Under these circumstances there would be a tendency for higher numbers in the blocks to be missing and a hash number algorithm that worked "too well" on a *sequence* of identifiers would probably fail to give a uniform distribution on the actual identifiers. A failure of this type would be suggested by values for the test statistic that are too low.

In the test for serial correlation we are concerned only with the high end of the test statistic distribution. In the ten sets giving 60 test statistics there was only one case in which the statistic exceeded the 95% value of χ^2 on 90 degrees of freedom. Hence we are satisfied that serial correlation in a sequence of identifiers disappears in the corresponding sequence of hash numbers.

A random number generator found to perform well by Downham and Roberts uses parameters k=8192 and m=67099547. This generator produced results comparable to, if not a little better than, those of Table 1. On the uniformity test, there were only three failures in 60 trials at the low end of the test statistic distribution and one at the high end. There were no failures on the test for serial correlation.

As an indication of the performance of our hash number algorithm for identifiers of different lengths, we performed the same tests as those used to produce Table 1 on identifiers of lengths five digits to 12 digits. The results for sequences of 500 identifiers with increments of one are shown in Table 2. The corresponding results using the Downham and Roberts generator were similar.

Table 2. Uniformity and Serial Correlation in Hash Numbers for Arithmetic Sequences of Identifiers of Different Lengths

Length (digits)	Start	Uniformity	Serial Correlation
5	65186	93.2	104.1
6	868861	102.8	124.3 **
7	7297624	82.8	69.0
8	79885296	87.6	95.9
9	073698040	108.0	85.2
10	4903127552	99.6	77.9
11	45451890688	84.8	87.1
12	107249557504	95.2	91.9

** Significant at the 5% level for a one-tailed test.

Despite the assurance given by the above results for "invented" sequences, a final test should be made on the actual identification numbers in any particular application. In our own application the identification numbers have 9 digits, in which the last digit is a "check digit," the first is assigned geographically, while the remaining digits are a mixture of meaningfully and sequentially assigned. Taking the ordered file in blocks of 500 identification numbers we applied the same tests as those described above. The results were completely consistent with those for the invented sequences.

5. Conclusion

The Poisson sampling techniques described in this article provide a simple compromise between the requirement for cross-sectional efficiency, with the implications of that requirement for changes in sampling specification from time to time, and the requirement for longitudinal stability in sample estimates.

A statistical agency, using an administrative file "owned" by another agency, may wish to reduce its vulnerability to changes in the file maintenance systems by using a hash number procedure that uses the permanent identifier associated with the file units to generate the "random" number used in Poisson sample selection. The hash number algorithm described in this article, using either of the two multiplicative congruential pseudorandom number generators discussed, appears to be satisfactory for identifiers of five to 12 digits.

6. References

Brewer, K.R.W., Early, L.J., and Joyce, S.F. (1972): Selecting Several Samples From a Single Population. Australian Journal of Statistics, 14(3), pp. 231–239.

Cochran, W.G. (1977): Sampling Techniques, third edition. John Wiley & Sons, New York.

Downham, D.Y. and Roberts, F.D. (1967): Multiplicative Congruential Pseudorandom Number Generators. Computer Journal, 10, pp. 74–77.

Good, I.J. (1953): The Serial Test for Sampling Numbers and Other Tests for Randomness. Proceedings of the Cambridge Philosophical Society, 49, p. 276.

Sowey, E.R. (1972): A Chronological and Classified Bibliography on Random Number Generation and Testing. International Statistical Review, 40(3), pp. 355–372.

Sunter, A.B. (1977): Sampling With Unequal Probabilities and Without Replacement. Statistisk tidskrift, 15(3), pp. 227–236.

Received August 1985
Revised March 1986