55589

C2

Sampling Methods for Repeated Establishment Surveys

K.P. Srinath and R. Carpenter

1 INTRODUCTION

Establishment surveys, especially those of farms and businesses, are usually repeated periodically. The period could be a week, a month, a quarter, a year, or some other time interval of interest. For several reasons, chief among them being increased accuracy (see Sigman and Monsour, Chapter) and reduction of response burden, it may be desirable that the sample change from one survey occasion to the next, though in practice, the same sample is often used for several survey occasions. Our focus in this chapter will be on sampling from a list frame in a particular stratum of a repeated survey. Every unit in the stratum will have approximately an equal probability of selection. Although the composition of the stratum itself may change over time due to "births" (the introduction of new establishments into the stratum) and "deaths" (the removal of establishments from the stratum), the definition of the stratum and its desired sampling fraction will stay roughly constant.

Sigman and Monsour (Chapter) briefly described one method of sampling from a list frame for a repeated survey -- the rotation group method. They did not, however, address how births

---

The authors thank Philip Kott, whose constructive comments led to a number of improvements in the manuscript. The views expressed by the authors do not necessarily reflect the policies of Statistics Canada

should be handled with this method. Ohlsson (Chapter) discusses the use of Permanent Random Number sequential sampling. That methodology, which is identical to collocated sampling (Brewer et al, 1972) for a fixed population, features a clever technique for handling births in a repeated survey.

Section 2 of this chapter provides a more in depth discussion of the rotation group method than Sigman and Monsour (Chapter). It also reviews a method of handling births with collocated sampling suggested by Tambay (1988). Finally, it introduces a new method of sample selection that combines rotation group and collocated sampling techniques. Section 3 contains some simulations, and Section 4 some concluding remarks.

2 SELECTION AND ROTATION PROCEDURES

In this Section, we consider three methods of sample selection within a particular stratum in a repeated survey with births and deaths. The first, rotation group sampling, is used in retail and wholesale trade surveys in the U.S. and Canada. The second, repeated collocated sampling, has been studied internally at Statistics Canada. The third, which we introduce here, combines elements of the first two methods.

2.1 Rotation Group Sampling

In rotation group sampling the population units in each stratum are randomly divided into a number of mutually exclusive rotation

groups of equal or nearly equal size. A without replacement simple random sample of rotation groups is then selected for the sample.

To allocate the stratum units to rotation groups for the first sampling occasion, the following method is often used. Suppose there are to be P rotation groups numbered from 1 to P. A random permutation of this ordering is used for allocation and is called the <u>assign ordering</u>. The first unit in the population is assigned to the first rotation group in the assign ordering, the second population unit to the second rotation group in the assign ordering, ..., the Pth population unit to the Pth rotation group in the assign ordering; the (P+1)th unit is again assigned to the first rotation group in the assign ordering, and so on.

The original numbers of the rotation groups before permutation form what is called the <u>rotation ordering</u>. This ordering is used for purposes of selection and rotation of units. For example, rotation groups numbered 1 to p in the rotation ordering may be included in the sample on the first occasion; on the second occasion, rotation group 1 rotates out of the sample while the (p+1)th rotation group rotates into the sample. For example, let let the population size N in a stratum be 16, the sample size n be equal to 8. Therefore the sampling fraction f is equal to 0.5. Suppose we want the units to stay in the sample for four occasions, then the total number of rotation groups in the population is 1/f x 4 which is equal to 8. That is 16 units in the populaiton are randomly divided into 8 rotation groups

numbered 1 to 8, each rotation group containing 2 units. A random permutation of the numbers 1 to 8 is used for assigning units to the rotation groups. If rotation group 6 in the rotation ordering is the first rotation group in the assign ordering, then the first unit in the population is assigned to the 6th rotation group and so on. Rotation groups 1 to 4 before permutation are slected in the sample for the first occasion. On the second occasion rotation group 1 rotates out and rotation group 5 in the rotation ordering rotates in.

Suppose there are N population units in the stratum on the first sampling occasion. If N/P is not an integer, then the P rotation groups will not all be of equal size. Thus, the number of units in a sample of p rotation groups will be random variable with expected value $n = (p/P)N$. The probability of selection for any unit in the population would be p/P. If f is the desired sampling fraction for the stratum, then p and P should be chosen so that f is approximately p/P.

Many surveys that employ rotation group sampling do not use the <u>unconditional</u> unit selection probability, p/P, in estimation. Instead, they rely on <u>conditional</u> selection probabilities given realized (actual) sample sizes. It is not difficult to see that given a realized sample size of m in the stratum, each unit has a m/N probability of being in the sample. Conditional probabilities, like m/N, can easily be calculated for any of the sampling designs discussed in this chapter.

The following method is used to handle births in some of

Statistics Canada's business surveys. Births are assigned to rotation groups using the assign ordering. For example, suppose on the first sampling occasion that the last (Nth) unit in the stratum was assigned to the qth rotation group in the assign ordering. The first new birth for the second sampling occasion would then be assigned to the (q+1)th group in the assign ordering. On each sampling occasion, the rotation group to which the last birth was assigned is noted, and subsequent births are assigned to rotation groups starting with the next rotation group in the assign ordering.

Some business surveys at Statistics Canada use a one-level rotation scheme (see Sigman and Monsour, Chapter) in which one rotation group is rotated out of the sample and another is rotated into the sample on each new sampling occasion. Considerable effort has been extended so that small units that have been in the sample stay out of the sample for designated periods of time. This can lead to certain complications like rotation groups without any members. The interested reader is referred to Hidiroglou and Srinath (1993) for an in depth discussion. For our purposes, it suffices to recognize that it is a relatively simple matter with rotation group sampling to assure that no rotation group, and thus no sampling unit, is sampled too often due to random chance.

2.2 Repeated Collocated Sampling

Ohlsson (Chapter) discusses collocated sampling in an

unequal selection probability context. We describe below a procedure first given in Tambay (1988) which uses equal selection probability collocated sampling on the first sampling occasion. The handling of births, however, differs from the treatment in Ohlsson (Chapter) We will call this procedure repeated collocated sampling.

Let the units within a stratum be arranged in a random order. A sample selection number, SSN(i) is assigned to the ith ranked unit in the population as follows:

$$SSN(i) = (i - \epsilon)/N, \tag{1}$$

where ϵ is a number randomly drawn from the uniform distribution on the unit interval (0, 1). All sample units whose sample selection numbers lie within the interval (0, f), where f is the desired sampling fraction, are included in the sample.

Rotation of the sample is done by simply shifting the sampling interval (0, f). For example, the sample on the second sampling occasion might be every unit with sample selection number in the interval (s, s+f). The value of s can be chosen to either eliminate sample overlap (if possible) or to create a degree of overlap that would be near optimal for estimation purposes.

On each sampling occasion, new births are assigned a sample selection number independent of the numbers assigned to older units. This means that on each sampling occasion, the Q new births since the last sampling occasion are assigned sample selection numbers:

$$SSN(i) = (i - \epsilon)/Q,$$

where ϵ may be the same value as in equation (1) or a new draw from the unit interval.

Equally spacing the new births on the unit interval will alter the equal spacing of the stratum units as a whole and may lead to variations in the sample size over time. Ohlsson's Permanent (PRN) Random Number sequential SRSWOR (Chapter) uses a similar approach that keeps the sample size constant over time. The disadvantage of that approach is the PRN's of births are not equally spaced on the interval (0, 1). This may lead to births being over or underrepresented in the sample due to random chance.

It can be shown that rotation group sampling method is conditionally equivalent to simple random sampling without replacement (SRSWOR). The same can be said about repeated collocated sampling because births are selected independently each time. Estimating the variance of an estimator based on a repeated collocated sample may not be a trivial matter. This is a minor point. If anything, the SRSWOR variance estimator will be slightly conservative (because repeated collocated sampling does a better job at proportionally representing the births in the population than SRSWOR).

2.3 Modified Collocated Sampling

In this subsection, we propose a new sampling approach that is a mixture of rotation group sampling and collocated sampling.

Instead of rigidly spacing the sample selection numbers on the unit interval, the selection numbers are assigned as follows. First the interval (0, 1) is divided into N intervals (if N is very large, then it is possible to put a limit on the number of intervals created). A random permutation of the intervals is obtained. The ordering obtained by the random permutation of the numbers 1, 2,..., N can be thought of as an assign ordering and used to assign random numbers in the corresponding intervals. The first unit in the population is assigned a number randomly selected from a uniform distribution on the first interval in the assign ordering, the second unit is assigned a random number in the second interval in the assign ordering, and so on.

The following example may help clarify this procedure. Let the stratum population size, N be 20 and the desired sample size, n, 5. The unit interval, (0, 1), is divided in to twenty intervals of length 0.05 each. Let the random permutation of the numbers 1 to 20 be 9, 15, 1, 4, 20, 19, 14, 2, 10, 8, 17, 7, 11, 18, 3, 13, 12, 16, 6, and 5.

Now the first unit in the population is assigned a random number between 0.40 and 0.45; that is, the first population unit is assigned to the ninth interval. The second unit in the population gets a random number between 0.70 and 0.75 (the 15th interval), the third unit gets a random number between 0.90 and 0.95, and so on. Each unit gets a unique random number in the interval to which that unit has been assigned as shown in Table 1. All units having random numbers in the interval 0 to 0.25

(the desired sampling fraction, 5/20) are selected in the sample. This means that the units assigned to first five intervals in the original ordering namely, units 3, 4, 8, 15 and 20 are in the sample. Sample rotation is achieved in exactly the same manner as described in the previous subsection.

Births are assigned random numbers in the intervals given by the assign ordering. In the example above, the first birth on the second occasion is assigned to the 9th interval and receives a random number between 0.40 and 0.45, the next birth is assigned to the 15th interval, etc. Each birth is sequentially assigned to the intervals in the assign ordering. This assignment is independent of the sampling process, and any birth in the sampling interval for that occasion is included in the sample.

A disadvantage of this method may be a slight variability of the sampling fraction from one sampling occasion to the next because the sample selection numbers are not equally spaced. The principal advantage of modified collocated sampling over rotation group sampling is realized when more than a single repeated survey is based on the same list frame. Interval selection methods like repeated and modified collocated sampling simplify the coordination of sample selection across surveys (see Ohlsson, Chapter)

2.4 Removal of Deaths

A simple unbiased procedure to remove population deaths (establishments that cease operation) from the sample and the

nonsampled part of the population is to identify them as such through an external source that is independent of the sample survey. Once they are identified as dead units by this external source, the units are dropped from the sampling frame. This method is applicable to all the selection and rotation procedures described above. The removal of deaths may cause an imbalance in the number of units in a rotation group in rotation group sampling or in the sampled intervals in the method described in Subsections 2.2 and 2.3.

Deaths identified from the sample survey should be treated in a different manner. Such dead units remain in the frame. When they are sampled, their survey values are all zero. For a more detailed discussion on the removal of deaths see Hidiroglou and Srinath (1993).

2.5 Changes in Classification

Changes in the classification variables like industry, geography, or size of businesses over time are common in repeated business surveys because of the dynamic nature of the business population. These changes are observed more often in the sampled units than in the nonsampled portion of the population. Consequently changes in the frame can be implemented, if they are identified by a source that is independent of the sample survey. As we have suggested already, a simple way to implement such changes is to treat the units coming into a stratum as births and the units going out of a stratum as deaths.

This may not cause any problems in rotation group sampling except, perhaps, in creating an imbalance in the number of units in each rotation group. In the procedure described in sections 2.2 and 2.3, this method of treating classification changes will result in the assignment of new random numbers for the old units. If is desired to retain the old numbers for sample coordination purposes (see Ohlsson, Chapter), then the units can be shifted to new strata with their old numbers.

2.6 Sequential Sampling

Given the sample selection numbers of sections 2.2 or 2.3, one has the ability to come closer to the desired sampling fraction by replacing interval sampling with sequential sampling (Ohlsson, Chapter). That is to say, rather than sampling all units with sample selection numbers within an interval, one could sample the first $n \approx fN$ units to the right (or left) of a designated number on the unit interval.

Not only does sequential sampling provide closer to ideal sample sizes, it removes the need for sample selection numbers to be nearly equally spaced on the unit interval (this is the distinction between the sampling section numbers in this chapter and the permanent random numbers in Ohlsson (Chapter) . The problem with sequential sampling, and the reason it is not used by Statistics Canada, is that it is not conducive to the establishment of strict time-out-of-sample rules for previously sampled units. With interval sampling, we can avoid the

selection of certain intervals and thus the units within them. We have no such mechanism with sequential sampling.

3 SIMULATION RESULTS

A short scale simulation was carried out to study the long term behaviour of the rotation group sampling method, the proposed modified method of collocated sampling, and Bernoulli sampling (see Ohlsson, Chapter). Repeated collocated sampling was not included in the study as it was thought that the results may be similar to the modified collocated sampling. The realized sampling fraction, defined as the number of units in the sample divided by the number of units in the population, was compared with the desired sampling fraction to determine the variability of the sampling fractions over time. Also, the realized sampling fraction for births was compared with the expected fraction to determine whether new births are properly represented in each period.

Three population sizes of 50, 500, and 1000 were used. Two sampling rates, 0.10 and 0.20 were considered for a population of 50 while a sampling rate of 0.05 was used for populations of size 500 and 1000. After the initial sample selection, it was assumed that the number of births and deaths on each occasion follow a Poisson distribution with means of 20 and 15 respectively for populations of size 50 and 500 and means of 200 and 185 for population of size 1000. The deaths are removed from the populations only if they are identified as such by an

external source. Though deaths occur each month, the probability of identifying a death through an external source was assumed to be 0.70.

The overall sampling fractions and the sampling fractions just for births were studied assuming sample rotation over 50 occasions. The results of the simulation with respect to maintaining a constant sampling fraction are shown in figures 1a to 4b. Figures 1a, 2a, 3a, and 4a show the overall sampling fraction while figures 1b, 2b, 3b, and 4b show the realized sampling fractions just for births. The means assumed for the occurrence of births and deaths for each population are shown in the figures as E(birth) and E(death).

It is easily seen that Bernoulli sampling has the largest variation in sampling fractions between occasions for all the three population sizes. Rotation group sampling and modified collocated sampling provide sampling fractions closer to the desired sampling fraction. For a population of size 50 with a sampling fraction of .10 rotation group method produced a slightly lower sampling fraction than desired for later occasions while the modified collocated sampling provides sampling both above and below the desired sampling fraction. This may be due to the fact that all the units in a rotation group rotate out or rotate in whereas only a fraction of the units in an interval rotate in or rotate out depending on the shift With a sampling fraction of .20 rotation group sampling provided samples which result in sampling fractions closer to the desired sampling

fraction. . Rotation group method and modified collocated sampling give sampling fractions closer to the expected for populations of size 500 and 1000.

It is seen from the graphs that the sampling fractions for births vary widely for Bernoulli sampling. For larger populations both the rotation group method and the modified collocated sampling do well in representing births in the sample.

4 CONCLUDING REMARKS

If it is desired to maintain a roughly constant sampling fraction within a sampling stratum and to avoid undue variation in sample size unrelated to population size, then Bernoulli sampling without any controls over the allocation of births is not a desirable procedure for sampling though it is a strictly an unbiased procedure. The changes in the sample size may result in varying levels of reliability of the estimates in addition to difficulties in managing the sample for data collection. Moreover, the realized sampling fraction for births may not be the same as for the older units. As a result, there is a possibility of under or over representing births in the sample on any given occasion.

Rotation group and modified collocated sampling methods provide sampling fractions closer to the desired fractions especially when the population sizes are large. Both the procedures do not attempt to exactly proportionately represent births in the sample

because of the spacing of births by allocation panels or sub-intervals.One minor disadvantage of the rotation group method is the difficulty of changing the sampling fraction from one sampling occasion to the next as it would require the creation of additional rotation groups or dropping of some rotation groups. The number of rotations groups is directly linked to the sampling fraction and the number of occasions a unit is expected to be in the sample. Repeated collocated sampling samples births independently on each occasion and therefore provides better representation for births. But the spacing of the units on the first occasion which is directly dependent on the number of units for that occasion and births on subsequent occasions may lead to not realizing the desired sampling fraction if sampling fractions are changed during later occasions.

Modified collocated sampling, introduced here, has many of the desirable attributes of rotation group sampling and repeated collocated sampling including a mechanism for assuring that units with small selection probabilities are not sampled too often by pure random chance. Unlike rotation groups sampling, however, this sampling scheme would have no trouble accommodating a change in the desired sampling fraction from one sampling occasion to the next. Moreover, it is conducive to the coordination of several survey samples sharing the same sampling frame.

REFERENCES

Brewer, K.R.W., Early, L.J. and Joyce, S.F., "Selecting Several Samples from a Single Population," Australian Journal of Statistics, Vol. 14, 1972, pp. 231-239.

Finkner, A.L. and Nisselson, H., "Some Statistical Problems Associated with Continuing Cross-Sectional Surveys," in Survey Sampling and Measurements, N.K. Namboodri (ed), Academic Press, New York, 1978.

Hidiroglou, M.A., Choudhry, G.H. and Lavallée, P., "A Sampling and Estimation Methodology for Sub-Annual Business Surveys," Survey Methodology, Vol. 17, 1991, pp. 195-210.

Hidiroglou, M.A. and Srinath, K.P., "Problems Associated With Designing Sub-Annual Business Surveys," Journal of Business and Economics Statistics, 1993, forthcoming.

Konschnik, C.A., Monsour, N.J. and Detlefsen, R.E., "Constructing and Maintaining Frames and Samples for Business Surveys," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1985, pp. 113-122.

Ohlsson, E., "Co-ordination of Several Samples Using Permanent Random Numbers," Proceedings of the International Conference on Establishment Surveys, American Statistical Association, pp---

Sigman, R and Monsour, J., "On Selecting Samples From Frames Based on Lists of Establishments," Proceedings of the International Conference on Establishment Surveys, American Statistical Association, 1993 pp.----

Srinath, K.P., "Methodological Problems in Designing Continuous Business Surveys: Some Canadian Experiences," Journal of Official Statistics, Vol. 3, 1987, pp. 283-288.

Tambay, J.L. (1988), Collocated Sampling, Statistics Canada, Unpublished Manuscript.

Table 1: Assignment of random numbers

Sample Unit	Random Permutation	Corresponding Interval	Sample Unit	Random Permuation	Corresponding Interval
1	9	0.40-0.45	11	17	0.80-0.85
2	15	0.70-0.75	12	7	0.30-0.35
3	1	0.00-0.05	13	11	0.50-0.55
4	4	0.15-0.20	14	18	0.85-0.90
5	20	0.95-1.00	15	3	0.10-0.15
6	19	0..90-0.95	16	13	0.60-0.65
7	14	0.65-0.70	17	12	0.55-0.60
8	2	0.05-0.10	18	16	0.75-0.80
9	10	0.45-0.50	19	6	0.25-0.30
10	8	0.35-0.40	20	5	0.20-0.25

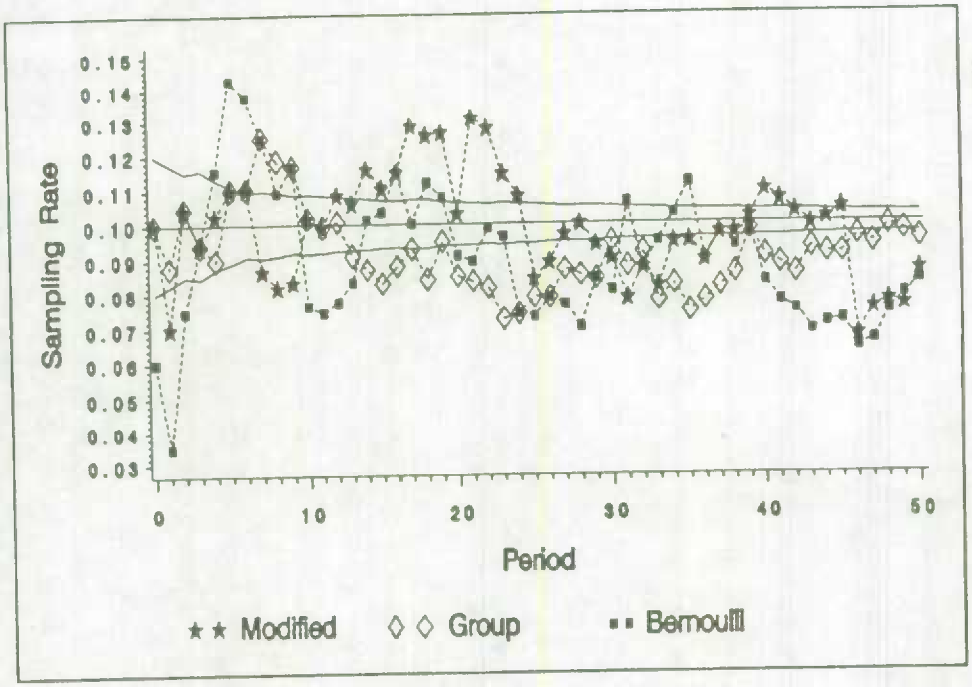

Figure 1a Initial population = 50, sampling fraction = 0.10, E(birth) = 20, E(death) = 15, Prob {death removed}=0.70

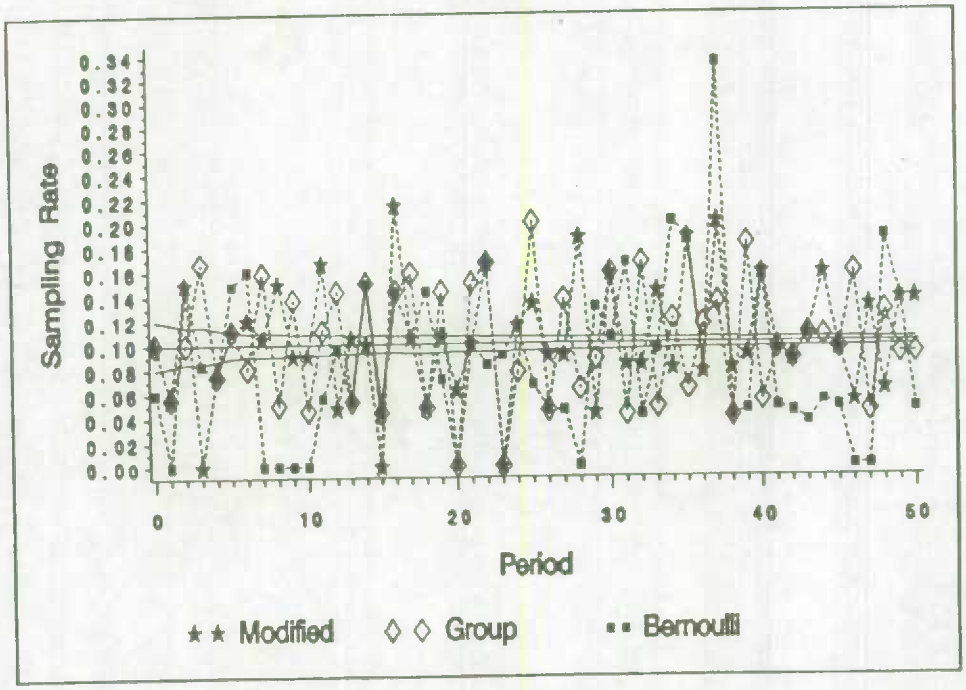

Figure 1b NEW BIRTHS ONLY: initial population = 50, sampling fraction = 0.10, E(birth)= 20, E(death)= 15, Prob {death removed}= 0.70

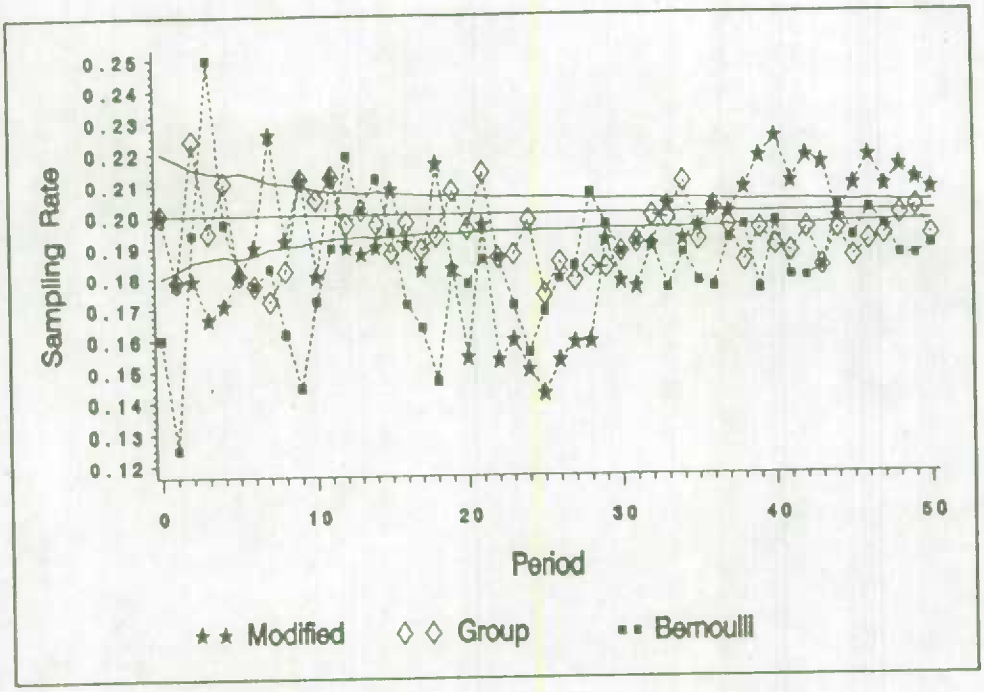

Figure 2a Initial population = 50, sampling fraction = 0.20, E(birth) = 20, E(death) = 15, Prob (death removed)=0.70

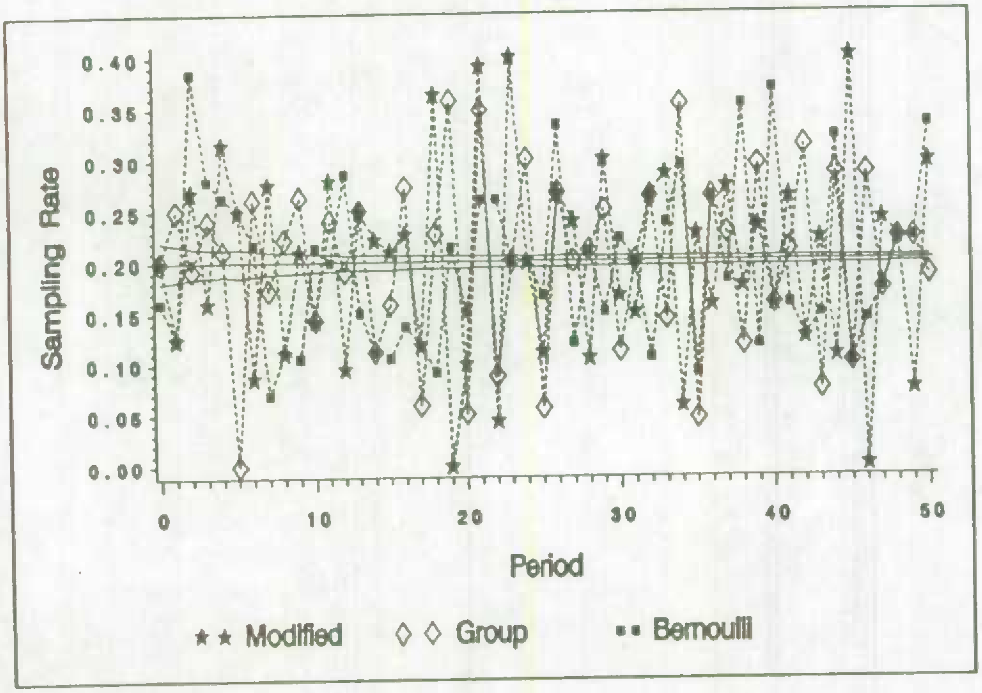

Figure 2b NEW BIRTHS ONLY: initial population = 50, sampling fraction = 0.20, E(birth) = 20, E(death) = 15, Prob (death removed)=0.70

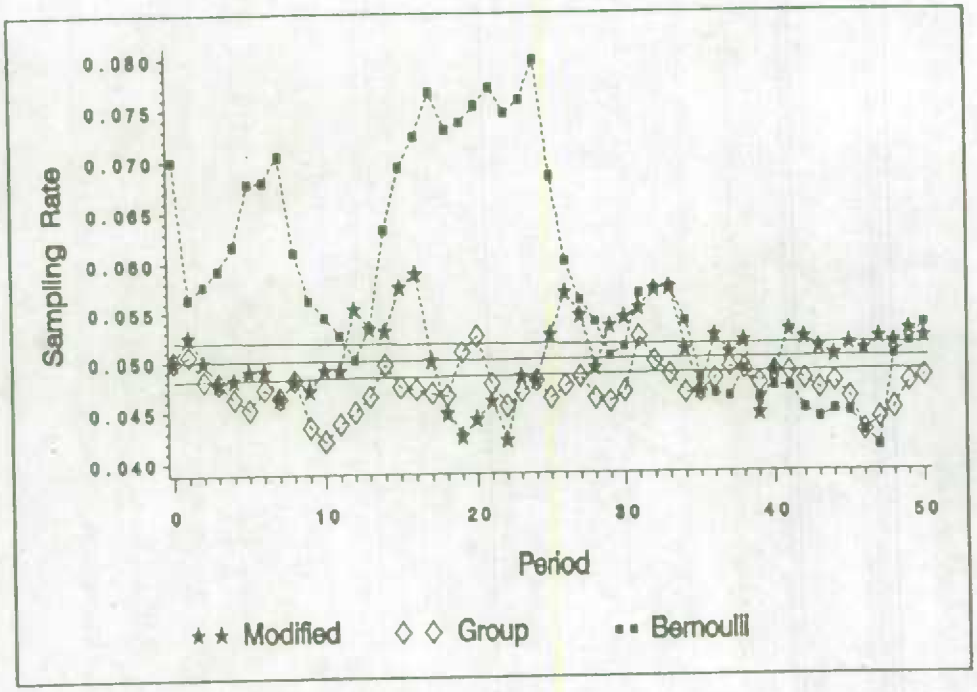

Figure 3a Initial population = 500, sampling fraction = 0.05, E(birth) = 20, E(death) = 15, Prob {death removed}=0.70

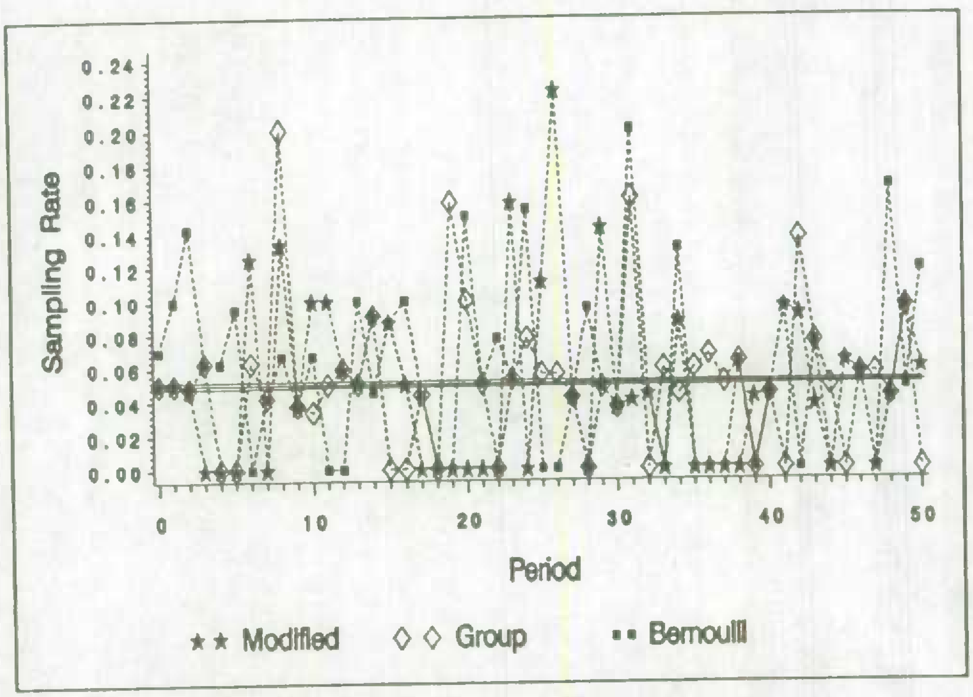

Figure 3b NEW BIRTHS ONLY: initial population = 500, sampling fraction = 0.05, E(birth) = 20, E(death) = 15, Prob {death removed}=0.70

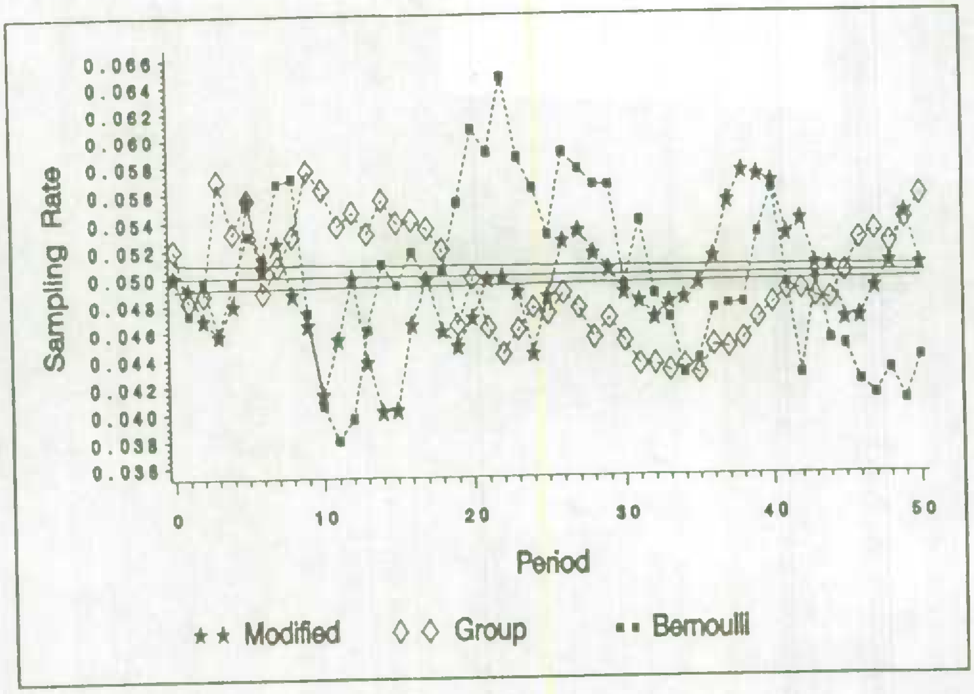

Figure 4a Initial population = 1000, sampling fraction = 0.05, E(birth)= 200, E(death)= 185, Prob {death removed}= 0.70

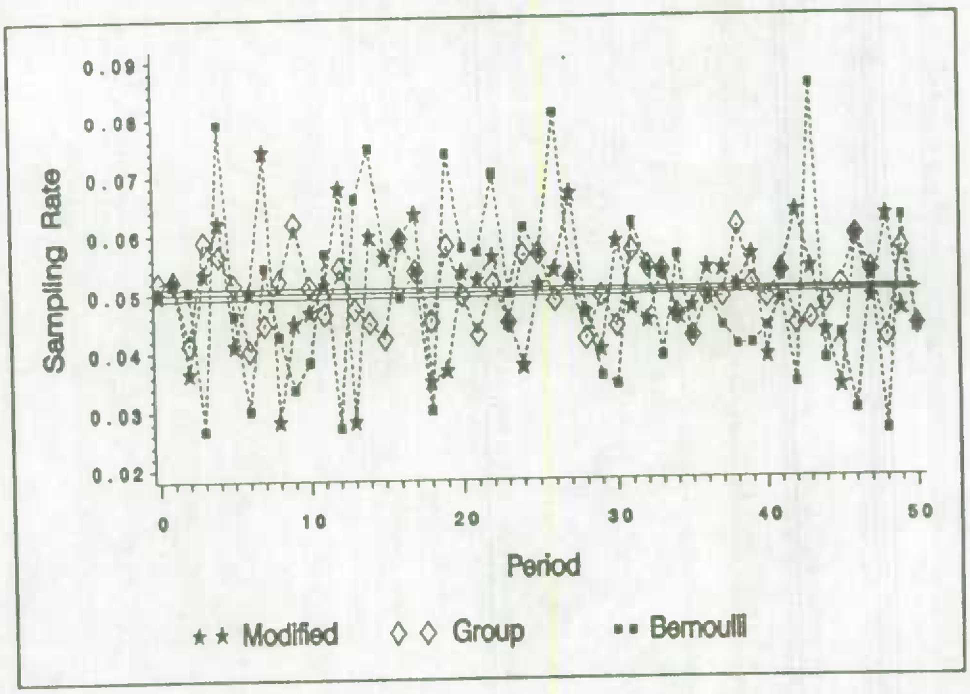

Figure 4b NEW BIRTHS ONLY: initial population = 1000, sampling fraction = 0.05, E(birth)= 200, E(death)= 185, Prob {death removed}= 0.70