SOME THEORY AND APPLICATIONS OF LATTICE SAMPLING

David E. Abbey, Loma Linda University

A type of two-way stratified sampling, named "LATTICE SAMPLING" by Yates (1960), appears to be a promising sampling design when small area or intra-universe estimates as well as estimates for the entire universe are desired. Another useful application of lattice sampling is sampling in time where it is often desirable to have balance because of periodicity.

The use of lattice sampling is analogous to the use of Latin squares in experimental design. In experimental design Latin squares are used to eliminate row and column effects from treatment comparisons. In lattice sampling we use the Latin square type restriction to eliminate the between row and column sources of variance from the sampling variance of the overall estimate and to enable estimates of row and column means to be made.

In simple lattice sampling, the population to be sampled is classified into a two-way arrangement of P row strata and Q column strata useful for a sampling frame. Each member of the population then falls into one of the PQ cells of the two-way table thus formed. If a sample was drawn from each of the cells, we would have ordinary or one-way stratified sampling, but lattice sampling permits samples from only some of the cells subject to the constraint that a fixed number of cells be drawn from each row and column of the table. This enables us to draw a small number of population elements and yet have a large number of strata, and because of the constraints on the way the cells are drawn, estimates can be made for each stratum. Thus estimates can be made for each sub-universe if the stratification is made so that each sub-universe corresponds to a stratum, that is a row or column of the table. We will assume that the sub-universes for which estimates are to be made correspond to rows of the table, and that the columns are formed by stratification on another variable, usually one that would increase the precision of the overall estimate.

A hypothetical example for a lattice sampling scheme which could be used for small area estimates is illustrated in figure 1. The universe for this example is the eight western states of Washington, Oregon, Montana, Idaho, California, Nevada, Utah, and Arizona. Eight socioeconomic strata are formed, and counties in each state are grouped into the cells of the two-way table formed by cross-classifying socioeconomic strata with states.

The first stage of sampling in this lattice sampling design is to sample cells from the table which are clusters of counties. In the illustration a sampling fraction of one-fourth was chosen so that two cells per row, and two per column were selected. There are different algorhythms available for selecting the lattice sample (16). All equal probability methods, however, are subject to the constraint that there be an equal number of cells selected from each row, and an equal number selected from each column, in this example, two per row, and two per column. This example could obviously be extended to include all fifty states, and the design of the lattice sampling frame does not have to be square.

Since a few sample elements are taken from each sub-universe it is possible with lattice sampling to take into account "local effects" that is effects which are inherent only within the sub-universe. For example, in estimating health characteristics for the nation (the universe), as well as for individual states (sub-universes), a state's own health care program would be a local effect. Most current schemes for small area estimation use estimators with standard sampling techniques which do not control the number of elements obtained in each sub-universe. Thus, each sub-universe may or may not contain part of the sample, and one cannot take into account local effects for the small area estimates. Thus, in the above example using most schemes which have been developed, one cannot compare the effectiveness of different state health programs as these programs would be "local effects". (5, 6, 8, 9, 10, 13, 14, 15, 17, 21, 22)

In contrast to the small area estimation schemes above which are used with non-controlled sampling methods, Brooks (1978), found in a Monte-Carlo study that using a Goodman-Kish controlled selection procedure in the CURRENT POPULATION SURVEY conducted monthly by the Bureau of the Census resulted in gains in the relative precision of the between primary sampling unit components of variance when compared to non-controlled independent selection for many of the variables estimated. Gains were more substantial for state estimates than for regional or national estimates and seem to depend on the proportion of controlled selection used in the area. This would suggest that sample designs using controlled selection may be a promising way to increase the accuracy of small area estimates.

Another useful application of lattice sampling is sampling in time where it is often desirable to have balance because of periodicity. Often when sampling over time, there may be double periodicity, and the lattice sampling will allow balance in the sample design on both types of periodicity. For example, in sampling across weeks we will have periodicity in the day of the week with Mondays tending to be similar, and periodicity within the day, with a given hour of a day tending to be similar to a given hour of any other day. A specific application of lattice sampling to this type of periodicity is discussed below.

Lattice sampling is also useful in sampling in time and space, where a space stratification can be the rows of a table and a time stratification, the columns. Yates (1960), proposed lattice sampling for sampling road traffic in this type of application.

1. HISTORY OF LATTICE SAMPLING. In 1942, Frankel and Stalk used a three-way form of lattice sampling to estimate national unemployment rates from a sample of a small number of counties. Tepping, Herwitz, and Demming (1943), discussed a number of multiple stratification techniques employing the Latin square type restriction which they refer to as "deep stratification". They compared the biases and variances obtained with deep stratification schemes with those of unrestricted random and one-way stratified sampling on block population data for Wilmington. They found that their deep stratification designs usually gave substan-

tial gains over one-way stratified sampling.

Patterson (1963), derived the sampling variances of the overall sample mean for different methods of lattice sampling with equal numbers of units in each cell. From the analytical expressions he derived, it can be seen that for estimating the overall mean, lattice sampling will give gains in precision over one-way stratified sampling if the F statistic for row effects from a two-way analysis of variance is greater than 1.

Delenius (1963), suggested that lattice sampling should prove useful for mapping surveys which require estimates for each of the individual non-sampled units on the basis of information gained from the sample unit.

Vos (1964), gives grand mean estimators and derives their variances for two dimensional lattice-like sampling schemes.

Unlike the before mentioned schemes which select cells with equal probability, Bryant, Hartley, and Jessen, (1960), considered a two-way stratification scheme where the cells in the two-way classification are selected with probability proportional to the product of the row and column stratification sizes. Both biased and unbiased estimators are given for the overall mean, and row and column means, and a method is given for obtaining essentially unbiased estimates of their variances. Their methods are particularly effective if the population cell frequencies are proportional to both marginal frequencies. However, if this is not the case, large biases and losses of efficiency can result.

Jessen (1970), presented some methods for sampling from a multi-dimensional universe where cells are selected with probability proportional to a measure of size. His scheme satisfied the constraints that the marginal totals of the sample are proportional to the marginal totals of the universe, and size of cells need not be proportional to the product of marginal sizes.

Abbey (1972), used alternative linear models with lattice sampling to obtain alternative sub-universe estimates, some of which attempted to make use of the structure of the population to "borrow" information from sample units outside of the sub-universe. Results of this work are summarized below, and then an application of the lattice sampling method to sampling in time is considered and the variance of the estimator for the overall mean derived.

2. RESULTS. In many sample surveys, in addition to having estimates for the entire universe it is desirable to have intra-universe estimates where the sample sizes within many of the sub-universes may be too small to give adequate accuracy. It may be possible to increase the accuracy of an intra-universe estimate by making use of the structure of the population to "borrow" information from sample units outside of the sub-universe. As a means of doing this, three alternative linear models or classifications of the universe - zero-way, one-way, and two-way for use with lattice sampling were considered.

Estimators for intra-universe means were developed under each of the three models using a number of different estimation methods. For the case when the cells in the two-way stratification are of equal size and are selected with equal probability according to a two-stage lattice sampling scheme, the sampling variances and biases of the estimators were derived. The schemes used were the "simple" (which used simple sample averages to estimate the parameters), the least squares, the parametric (which used the parameter estimates for every cell), and the missing cell scheme (which used the observed sample means in sampled cells and parameter estimates in missing cells). Those schemes which led to estimators having the lowest mean square error averaged over the sub-universes were termed best. By comparing analytical derivations of the average mean square errors supplemented with empirical tests, carried out on several small synthetic populations, it was found that these mean square errors depend largely on the structure of the populations dealt with.

It was found that:

1. The estimators using the two-way model are never worse and are usually better than those using the one-way model.
2. The two-way model is better than the zero-way model unless sub-universe differences are very small or the within cell variance is large compared to the between cell variance.
3. The simple method is better or at least as good as the least squares method except when the two-way model fits the data very closely.
4. The missing cell methods were better than the parametric methods when used with the two-way model, the same for the one-way model, and usually worse for the zero-way model.
5. The best combination of methods to use is usually the two-way model with the simple missing cell techniques.

For the case of equal cell sizes all of the different intra-universe estimators investigated gave the same estimate of the grand mean, namely the simple average of the cell means.

The lattice methods of estimation used with lattice sampling were compared with two synthetic estimators used with one-way stratified sampling by columns, since synthentic estimators are generally used with a one-way stratified sampling design where the sub-universe are not coexistent with strata. For the case of equal cell sizes, both synthetic estimators, one developed by the National Center for Health Statistics (1968), and another by Waksberg (1970), were identical with one of the lattice sampling estimators. The only difference between methods then, was the sampling technique used. It was found that both methods of estimation have the same bias. The lattice method would never have a larger variance than the synthetic estimators and would usually have a substantially smaller variance.

3. APPLICATION TO SAMPLING INCOMING TELEPHONE CALLS. The lattice sampling method was used for a sample design to sample incoming telephone calls for Tel-Med, a telephone dial access health information library. The universe was all incoming calls for a ten week period of time. First a stratification was made by week with each of the ten weeks being a stratum. Within each week a two-way stratified sample was drawn according to two criteria of stratification: 1) day of week, Monday through Friday; and 2) time of day, the hours of operation for Tel-Med on each day being divided into 5 strata: 9 a.m. to 11 a.m.; 11 a.m. to 1 p.m.; 1 p.m. to 3 p.m.; 3 p.m. to 5 p.m.; and 5 p.m. to 8 p.m. Each day two of the time periods were selected to sample incoming calls for telephone

interviewing.

Two sampling periods were chosen from each of the five time strata and each of the five days of the week according to a lattice sampling scheme so that a balanced representation of the sampling periods and days of the week would occur. The sample was selected by first randomizing rows and columns of a five-by-five Latin square. For each of the ten weeks two of the letters of this randomized Latin square were selected. There are ten ways in which two letters can be selected from five. They are:

12, 21, 12, 21, 12, 21, 12, 21, 12, 21

ab, ac, ad, ae, bc, bd, be, cd, ce, de .

The numbers 1 and 2 were assigned to these letters alternating the "1" and the "2" being first and the allocation of letters was put in random order and assigned to each of the ten weeks consecutively. This method of allocation assures a further balance on the sample design in that each given time period will occur on the same numer of Mondays, Tuesdays, Wednesdays, Thursdays, and Fridays for the ten week period, as will the number "1" and the number "2" for that time period. The sample design for each week is shown in figure 2. The number "1" in the cell of the table, indicates the first letter of the allocation pair and the number "2" indicates the second letter of the allocation pair. These numbers indicate replications and are used to estimate the sampling variance. Only cells with numbers are selected for interviewing. Having the number "1" occur in the same cell in the table the same number of times as the number "2" avoids confounding the sampling variance with any possible interaction between day and time.

For the 5 p.m. to 8 p.m. time period, two interviewers were to conduct interviews simultaneously. For the remaining time periods one interviewer would conduct interviews. This refinement of the sampling scheme was done to assure a more constant sampling fraction of the callers since the density of callers during the evening hours was approximately double that of the other hours.

Interviewing was to be conducted continuously throughout the sample time period with the interviewer selecting calls in a sequential fashion, taking the next available call as soon as an interview is completed.

The telephone operator records each call that comes in on a daily log for all time periods, whether sampled or not, during the ten week period. The tape number requested, the sex of the caller as estimated by the tone of the voice and whether the caller is a child or adult as estimated by tone of voice is recorded. The number of calls coming in for each time period, whether sampled or not, is essential data as it comes into the estimation formulas. The other data is optional but useful to provide a check on the representativeness of the sample.

4. SAMPLE ESTIMATES AND VARIANCES. Each week in which interviewing is done is a stratum. We first give the estimates and estimates of variance for a given week. The overall estimates for the ten week period are weighted averages of the stratum estimates.

Stratum Definitions. The following definitions apply to the particular week for which estimates are being made. Let Y be a variable for which an estimate is desired. Let y_{ijk} be the value of Y for the k^{th} interview during the j^{th} time period on the i^{th} day of the week. Let m_{ij} be the number of calls interviewed in time period j in day of week i. Let M_{ij} be the total number of calls which come in during time period j on day of week i.

Let us say that in general we have a PxP sampling frame from which we are taking a lattice sample of p cells per row. For the present case P = 5, p = 2.

The sampling frame can be regarded as a table with P rows, P columns and P^2 cells, each cell representing a time period for a given day. Thus for the present sample design we have five rows - days, and five columns - time period classifications, with 25 cells and we are sampling two "cells" or time periods each day with a total of ten cells being sampled from the entire week. Denote the sampling average over the interviews for the j^{th} time period on the i^{th} day by

$$\bar{y}_{ij} = \frac{1}{m_{ij}} \sum_{k=1}^{m_{ij}} y_{ijk} \quad .$$

Define the indicator variable

$$a_{ij} = \begin{cases} 1 \text{ if the } j^{th} \text{ time period on the } i^{th} \text{ day is sampled.} \\ 0 \text{ if it is not sampled.} \end{cases}$$

Let

$$w_{ij} = a_{ij} M_{ij}$$

and

$$f_1 = \frac{P}{P}$$

denote the primary sampling fraction of cells in the lattice sample design, and

$$f_{2ij} = \frac{m_{ij}}{M_{ij}} \quad ,$$

the within cell sampling fraction in the ij^{th} cell.

Formula For Estimating The Mean For One Week. An unbiased estimate of the mean for the entire stratum or week is

$$(1) \quad \hat{\bar{Y}} = \frac{1}{f_1 M_{..}} \sum_{i=1}^{P} \sum_{j=1}^{P} w_{ij} \bar{y}_{ij} \quad .$$

Formula For Estimating The Variance of $\hat{\bar{Y}}$. Let $\bar{Y}^{(1)}$ denote the estimate corresponding to that obtained in equation (1) above by only using cells in which an "1" occurs. Similarly, let $\bar{Y}^{(2)}$ denote the estimate obtained using only cells with "2". Then by applying Patterson and Yates rule (see Yates 1960, page 227), we have an unbiased estimate of the variance

$$(2) \quad v(\hat{\bar{Y}}) = (\hat{\bar{Y}}^{(1)} - \hat{\bar{Y}}^{(2)})^2 + \sum_{i=1}^{P} \sum_{j=1}^{P} \frac{w_{ij}^2}{f_1 M^2} (1-f_{2ij}) \sum_{K=1}^{m_{ij}} \frac{(y_{ijK} - \bar{y}_{ij})^2}{m_{ij}(m_{ij}-1)}$$

Formula For Estimating Overall Mean And Variance For Entire Period. The overall estimate of the mean for the entire ten week period is the weighted average of the strata estimates.

$$(3) \quad \hat{\bar{Y}}... = \sum_{\ell=1}^{10} M_{\ell..} \hat{\bar{Y}}_{\ell..} / M...$$

where M_{ℓ} denotes the total number of calls for the ℓ^{th} week and $\bar{Y}_{\ell..}$ denotes the estimate of the mean for the ℓ^{th} week given by (1), and

$$M_{...} = \Sigma_{\ell=1}^{10} M_{\ell..}$$

For the estimate of the variance for the entire ten week period, we have

$$(4)\quad v(\hat{\bar{Y}}...) = \Sigma_{\ell=1}^{10} M_{\ell..}^2 v(\hat{\bar{Y}}_{\ell..}) / M^2...$$

where $v(\hat{\bar{Y}}_{\ell..})$ denotes the estimate of variance for the ℓ^{th} week given by (2).

Note that many callers call Tel-Med more than once to listen to different tapes. As a result, some may be interviewed more than once. For those variables on the questionnaire which evaluate the tape just listened to, estimates will be made taking the element of the target population to be the telephone *call*. All interviews, even of those persons who are interviewed more than once, will be included in the analysis. For the remaining variables on the questionnaire which assess characteristics of the caller and the overall program, only the first interview will be counted. The interviewer asks each interviewee if this is the first interview. Here the element of the target population is the "*caller*".

The above estimation formulas apply strictly only to the first case where the element is the *calls* since M_{ij} is the number of calls during a given time period. The number of different *callers* which call during a given time period is not known and it is not feasible for the switchboard operator to attempt to determine this. However, the above formulas can still be used since $\frac{m_{ij}}{M_{ij}}$, the ratio of calls sampled to total *calls* during a time period, is an estimate of the corresponding ratio of *callers* sampled to total *callers* calling during a time period. So for both cases the weighting factors can be taken as $\frac{m_{ij}}{M_{ij}}$ the ratio of calls sampled to total calls. The y_{ijk} will differ. For the first case repeated interviews will be included; for the latter case, they will be excluded.

5. FUTURE RESEARCH. In some applications of intra-universe estimation, it may be desirable to obtain estimates for subuniverses which correspond to cells of a two-way stratification rather than rows or columns. Unbiased estimates for non-sampled cells cannot be obtained, but if interactions were not too large biases may be small. It would be interesting to investigate the biases and variances of different types of estimators of means of missing cells used with lattice sampling. A useful criterion of comparison would possibly be the mean square error averaged over all missing cells. Sometimes an investigator may be as interested in obtaining row estimates as in column estimates. A useful criterion in this situation would be the mean square error averaged over both rows and columns, the average margin mean square error. This can be obtained by first obtaining the average column mean square error by duality from the results given by Abbey (1972) and then averaging the average row and column mean square errors together. Another possibility for the use of the estimators would be to use different estimators for different rows to take advantage of differing row population structures.

The variances and biases of the different estimators for the case of unequal cell sizes with cells selected with unequal probabilities needs to be derived, so that comparisons between the different estimators can be made for this case. The performance of the estimators needs to be compared for some actual data. This could not be easily done for the equal sized cell case as populations with equally sized cells rarely exist.

The best estimators for the unequal sized cell case should be modified to take advantage of multiple regression techniques for small area estimation, and compared with other intra-universe estimation techniques.

Since lattice sampling appears to be a promising sampling scheme for the problem of small area estimation - one that is of continual concern in nationwide sampling, it would seem that some study would be warranted as to how National Sample Designs might be modified to incorporate lattice sampling, so as to improve on small area estimation, while at the same time being able to take into account the local effects of the small areas such as states or counties for which estimates are also needed.

REFERENCES

(1) Abbey, David E. Some estimators of subuniverse means for use with lattice sampling. Unpublished Ph.D. thesis, University of California, Los Angeles, 1972.

(2) Brooks, Camilla. (1978) The effect of controlled selection on the between PSU variance. Proceedings of the Social Statistics Section of the American Statistical Assoc. 336-339.

(3) Bryant, E.C., Hartley, H.O., and Jessen, R.J. (1960), Design and estimation in two-way stratification. Journal of the American Statistical Association, 55:105-124.

(4) Dalenius, T. (1963), Contributions to Statistics: Lattice Sampling by means of Lahiri's sampling scheme. Pergamon Press. Oxford Press.

(5) Crosetti, A.H. and Schmitt, R.C. (1956), A method of estimating the intercensal population of counties. Journal of the American Statistical Association, 37:77-80

(6) Ericksen, E.P. (1971), A method for combining sample survey data and symptomatic indicators to obtain estimates for local units. Unpublished Ph.D. thesis, University of Michigan.

(7) Frankel, L.R., and Stock, J.S. (1942), On the sample survey of unemployment. Journal of the American Statistical Assoc., 37:77-80

(8) Gonzales, M.E. (1972), Analysis of annual and monthly synthetic unemployment estimates for SMSA's. U.S. Bureau of the Census. Unpublished report. 28 pp. July 17, 1972.

(9) Gonzales, M.E., and Hoza, C. (1978), Small-area estimation with application to unemployment and housing estimates. Journal of the American Statistical Assoc., 73:7-15.

(10) Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), Sample Survey Methods and Theory. Vol. I and II. Wiley, New York.

(11) Jessen, R.J.(1970), Probability sampling with marginal constraints. Journal of the American Statistical Association, 65:776-796.

(12) Jessen, R.J. (1971), Studies in intra-universe estimation. Unpublished report to the Bureau of the Census. 45 pp. Revised January 1972.

(13) Levy, P.S. (1971), The use of mortality data in evaluating synthetic estimates. Proceedings of the Social Statistics Section of the American Statistical Assoc. 328-331.

(14) Madow, L.H. (1956), U.S. TV household by region, state and county. March 1956. An Advertising Research Foundation Report.

(15) National Center for Health Statistics. (1968), Synthetic state estimates of disability. PHS Publication 1759. U.S. Government Printing Office, Washington, D.C.

(16) Patterson, H.D. (1954), The errors of lattice sampling. Royal Statistical Society Journal, Series B (Methodological), 16:140-149.

(17) Rosenberg, H. (1968), Improving current population estimates through stratification. Land Economics, 44:331-338.

(18) Snow, E.C. (1911), The application of the method of multiple correlation to the estimation of post-censal populations. Royal Statistical Society Journal, 74:576-594.

(19) Tepping, B.J., Herwitz, W.N. and Deming, W.E. (1943), On the efficiency of deep stratification in block sampling. Journal of the American Statistical Assoc., 38:93-100.

(20) Vos, J.W.E. (1964), Sampling in space and times. Review of the International Statistical Institute, 32:226-241.

(21) Waksberg, J. Analysis of synthetic estimates. Unpublished memo to T.B. Jabine, December 10, 1970.

(22) Woodruff, R.S. (1966), Use of a regression technique to produce monthly national estimates of retail trade. Journal of the American Statistical Assoc., 61:496-504.

(23) Yates, F. (1960), Sampling Methods for Censuses and Surveys. (3rd edition.) Charles Griffin and Company, Limited, London. pp. 356-364.

FIGURE 1

LATTICE SAMPLING FRAME FOR 8 WESTERN STATES

	SOCIOECONOMIC STATUS							
STATES	1	2	3	4	5	6	7	8
WASHINGTON		X			X			
OREGON				X			X	
MONTANA	X				X			
IDAHO			X					X
CALIFORNIA		X				X		
NEVADA				X			X	
UTAH	X					X		
ARIZONA			X					X

FIGURE 2

LATTICE SAMPLE DESIGNS FOR TEL-MED

Randomized Latin Square.

Day of Week	Time of Day 9-11	11-1	1-3	3-5	5-8
M	E	C	A	B	D
T	C	A	D	E	B
W	A	D	B	C	E
Th	D	B	E	A	C
F	B	E	C	D	A

Sample designs for each week.

WEEK 1

Day of Week	Time of Day 9-11	11-1	1-3	3-5	5-8	DE
M			1		2	
T		1	2			
W	1	2				
Th	2			1		
F				2	1	

WEEK 2

Day of Week	Time of Day 9-11	11-1	1-3	3-5	5-8	DE
M	1				2	
T			2	1		
W		2			1	
Th	2		1			
F		1		2		

FIGURE 2 (continued)

WEEK 3 — AB

Day of Week	9–11	11–1	1–3	3–5	5–8
M			1	2	
T		1			2
W	1		2		
Th		2		1	
F	2				1

WEEK 4 — AE

Day of Week	9–11	11–1	1–3	3–5	5–8
M	1		2		
T		2		1	
W	2				1
Th			1	2	
F		1			2

WEEK 5 — BC

Day of Week	9–11	11–1	1–3	3–5	5–8
M		2		1	
T	2				1
W			1	2	
Th		1			2
F	1		2		

WEEK 6 — BD

Day of Week	9–11	11–1	1–3	3–5	5–8
M				2	1
T			1		2
W		1	2		
Th	1	2			
F	2			1	

WEEK 7 — CD

Day of Week	9–11	11–1	1–3	3–5	5–8
M		2			1
T	2		1		
W		1		2	
Th	1				2
F			2	1	

WEEK 8 — CE

Day of Week	9–11	11–1	1–3	3–5	5–8
M	2	1			
T	1			2	
W				1	2
Th			2		1
F		2	1		

WEEK 9 — BE

Day of Week	9–11	11–1	1–3	3–5	5–8
M	2			1	
T				2	1
W			1		2
Th		1	2		
F	1	2			

WEEK 10 — AC

Day of Week	9–11	11–1	1–3	3–5	5–8
M		1	2		
T	1	2			
W	2			1	
Th				2	1
F			1		2