ARE THE BIASES AND INEFFICIENCIES OF THE EPI METHOD OF SAMPLING RELATIVELY UNIMPORTANT WITH REASONABLE BETWEEN VILLAGE VARIANCES?

David J. Fitch, Rafael Flores, and Jorge Matute, INCAP
David J. Fitch, Apartado Postal 1188, Guatemala, Central America

Because we face here in Central America the problem of how to sample in developing country environments and want to consider candidate methods, we have undertaken a series of studies to try to better understand the advantages and disadvantages, the potential for bias and inefficiency, of one such method, the one growing out of the successful WHO program to eradicate smallpox, and currently widely used in their Expanded Programme on Immunization (EPI). In this EPI method, simply put, interviewers are instructed to go to the center of a selected village, spin a coke bottle, count houses in that direction, and to select one of these at random as the first house in which to interview. The second house is the one closest to the first, etc., until 7 houses, or 7 people of the population sought, are found. Initially we were only interested in convincing ourselves that, in fact, biases were possible, and we tried to think about different conditions under which EPI might yield biased results. There seemed to be three or four such possibilities.

To investigate whether or not our intuitions might be correct we undertook a series of simulations. We produced a layout of a simple Guatemalan village with a north-south path and an intersecting east-west path. The village had 80 dwelling units (DU's), 20 to the north and 10 each to the south, east, and west, with 30 lying off the main paths. We imagined that the purpose of the survey was to estimate average expenditures for food in the DU population and we generated, for each of the 3000 villages of the layout described, random normal deviates with a village mean and standard deviation of 200 and of 50 Quetzales, the Guatemalan currency. Then we introduced some local homogeneity, i.e. DU's closer together tended to be more alike than DU's further apart. Finally we restandardized so that each village had a mean of 200 and a standard deviation of exactly 50. We were able to illustrate though such simulations four potential possibilities for bias with the EPI method (Fitch, Matute, Flores, 1992).

A problem of equal or greater importance is the increase in variance with the EPI method, as compared with simple random sampling or a systematic procedure, due to local homogeneity. Any of the methods that select from only one part of each selected cluster is subject to this problem. Just how realistic was the local homogeneity that we introduced, we can not say. We hope some day to be able to undertake studies on real villages of Central America.

Having convinced ourselves that biases were possible, we began to try to think about the implications in more realistic situations. For example one would not expect every village to have, e.g., the more well-to-do DU's on the more populated paths, a condition that leads to bias. Sampled villages would have conditions that would give biases for different reasons, and sometimes positive and sometimes negative. If the sampled villages gave a mix of biases, this would tend to increase the variances of the estimates and that is what we found (Fitch, Flores, Matute, 1993), although the increase under the conditions of our simulations was not great.

Now as we have said our main goal has been to illustrate the potential for bias of EPI and to better understand those situations under which such would occur. For this reason we, up to now, have restandardized after introducing changes to illustrate each bias possibility, so that each village has a mean of exactly 200 and standard deviation of exactly 50. We have done this because we wanted to control for confounding factors. But real world villages don't have zero between variances. It seems, in fact, reasonable to think that in most surveys the main part of the variance in the estimates comes from between cluster variance and not from within cluster variance. We wondered if the increased variance within clusters of the EPI method might be relatively unimportant where there was realistic between cluster variance. We wondered if such between cluster variance would, for all practical purposes, wipe out the, perhaps comparative slight, inefficiencies of the EPI method. To obtain information on this question, we have undertaken the present study.

CHARACTERISTICS OF THE PRESENT STUDY

Computer programs: The previous studies were programmed in BASS, a SAS-like language after finding that SAS/IML language, although easy to use seemed to be feasible only with problems smaller than what we were contemplating. And then BASS made for much more transportable programs. In the end the basic BASS programming, enough to run our programs, and all our programs could fit on one diskette, which could have run on a small 286 computer. But there were two disadvantages. One, the complete set of programs took about 10 hours on our 386/16, and with power

irregularities, both here and elsewhere, it was, would be, hard to get a run to finish. The second problem had to do with limitations on the size and complexities of what we could do with BASS.

Although we were able to illustrate four bias problems with villages of size 80, it seemed like such could be more clearly done with larger ones. Wanting to increase the size to 160, to speed our runs, and in a sense to simplify the programming - as we could hold everything needed in the memory at the same time - we switched for the present study to FORTRAN, using the Lahey F77-EM/32 5.01. Our computers have 8MB of memory and this FORTRAN uses a DOS extender which allows the full memory to be address directly, which makes it very fast, even on our 386/16. Runs are of course much faster on the 486/66. Our programs often hold two 3000 by 160 matrices plus other large matrices and this uses not much more than half the available memory. This FORTRAN can produce programs with the DOS extender which can be run, without the need to pay royalties, on a 386 or 486, and are available.

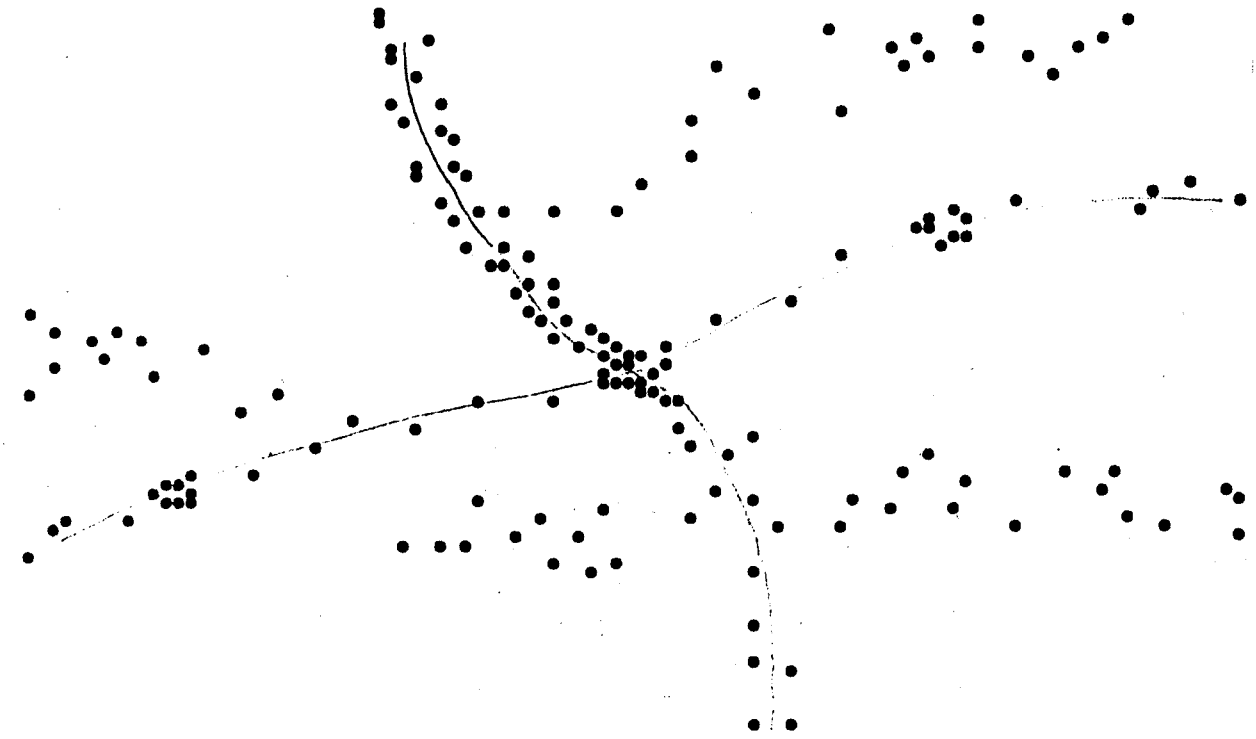

Figure 1. Village layout used for all villages of all data sets.

Generation of the "DU monthly food expenditures" data sets to illustrate the potential for biases with EPI: The starting data set of 3000 simulated villages used in the present study comes from selecting 300 from each of the 10 data sets, each of 3000 villages, constructed to illustrate the five bias problems, of which we are now aware, with the EPI method - five sets where each problem resulted in a positive bias, and five a negative bias. In describing the construction of these 10 data sets we will expand on the methods, introduced above, by which the sets were generated in order to illustrate those possibilities for bias that we have discovered to data - four previously and a fifth to be noted there. As a starting data set from which the 10 sets were created, we adjusted the initially created set - the set with the random normal deviates and local homogeneity as described above - so that the EPI method gave zero bias. We did this through adding and subtracting constants to the on, and the off main path DU's. The desired adjustment could be made in this way because the off path DU's, simulating EPI, were undersampled. Then we modified this basic set by adding and subtracting constants so as to illustrate each problem. A final restandarization, before computing sample means and variances, gave each village a mean of 200 and a standard deviation of 50. Figure 1 shows the layout of each of the 3,000 villages. The following 10 modifications of the base data set were made to produce 10 sets - 3,000 villages with 160 DU's in each.

1. Center DU's low. Those DU's near the point where the two paths cross will be oversampled by the EPI

method for two confounded reasons. A DU say to the east and lying near the center can be selected in the 2nd through 7th selection by an initial spin of the bottle pointing to the north or south or west, as well as the east. This becomes less likely and then impossible as the initial selection moves out from the center. This means that DU's in such a center location will be oversampled. If people in such DU's tend to spend less money for food, as compared with others in the village, then EPI sampling will show a negative bias. We will note the other reason in 5. below. So to illustrate the problem, center DU's were lowered and those non-center DU's on the main paths were raised so that, after restandardization to 200 and 50, the difference between the means for these two on-path groups of DU's was about 50, i.e. one standard deviation. Constants were sought which would both achieve the 50 Quetzal difference and also would keep, except in cases 2 and 7, the off-main-paths DU's, after restandardization, at about where they had originally been, about 194, or to rise to an average of 200, the mean for the whole village.

2. Off path DU's low. As we have noted, the EPI method, as we understand it, would not select as a starting DU one that was off-path, and would be less likely, in 2-7 selections, to select an isolated DU. Hence off-path DU's would tend to have a lower probability of selection. Constants were sought which would give a 50 Quetzal difference between on and off-path DU's. We missed it by a bit, the difference being 50.7.

3. Periphery DU's low. DU's on the periphery can be reached only by going in an outward direction, whereas others have a chance to be reached going either in or out. The lower probability of selection of such DU's means that if they tend to be different, EPI sampling will be biased.

4. Populated routes low. We located 40 DU's to the north but only 20 each to the south, east, and west. This means that a house to the north will have only half the probability, as compared with the other three directions, of being initially selected, as well as roughly only half the overall probability of being selected.

5. Clusters low. All DU's that lie in one direction will have an equal chance, in the EPI method, of being the DU initially selected. But where DU's are clustered, once selection is from within a cluster, all remaining selections will tend to be from within such a cluster and hence such clustered DU's have a higher probability of being selected.

6. Center DU's high.

7. Off path DU's high.

8. Periphery DU's high. This was the only exception to creating a 50 Quetzal difference. Without having planned it that way, in the starting data set, the periphery DU's were already more than 50 Quetzales more than the remaining on-path DU's. In this case we increased the difference to 100.

9. Populated routes high.

10. Clusters high.

Realistic between village variances: We obtained, courtesy of MACRO International, the data from the 1987 Guatemalan Encuesta National de Salud Materno Infantil (National Maternal Child Health Survey) and undertook analyses of the data from the 123 rural clusters, comparing the average within cluster standard deviation to the standard deviation of the cluster means. In our simulation studies to data we have fixed the mean of each village at 200, i.e. there has been no between village variance. The standard deviations within each village has been fixed at 50 and we planned to keep them at that, but now we wanted to introduce between village variances which would be within a range which would be, relative to the within variance, similar to what holds in real data. We used the variable "Ideal family size". One rural cluster had only a single case. The average standard deviation within the remaining 122 clusters was 2.31, and the standard deviation between the 122 means was 1.12, or about half. On the basis of this finding we generated random normal deviates with a mean of 0. and a standard deviation of one half of the 50, the standard deviation within each village, i.e., 25. Generating 3000 such deviates, one was added to each of the 160 DU's of a village, thus achieving the desired variance, i.e. 25^2, between village means. The process was repeated using standard deviation values around 25, i.e., 10, 20, 30, 40, and 50.

RESULTS AND IMPLICATIONS

As indicated, the 3000 villages for this study were 1-300 from set 1 above, 301-600 from set 2, etc., through 2701-3000 from set 10. As we showed in our earlier study (1993), mixing these data sets together - some yielding with EPI a positive bias and some a negative bias - increases the variance, but not a whole lot. In these simulations the increase in average variance was from 40.41, before mixing, to 43.99 after mixing the 10 part sets, each where EPI gives the bias as described.

Although the purpose of the present study was primarily to investigate the effects of between village variance, not EPI bias and variance problems per se, we give results in Table 1 which show these problems.

Table 1.
Means and variances of the means for the base data, and with 10 modifications to illustrate potential bias problems with the EPI method of sampling. Each value is based on the mean of 100 samples, each of 30 villages.

	EPI		Systematic		SRS	
	Mean	$v(\bar{y})$	Mean	$v(\bar{y})$	Mean	$v(\bar{y})$
Base data	200.00	38.32	200.30	10.59	199.60	11.47
Center low	194.60	45.49	200.30	10.78	199.63	11.50
Off road low	214.64	31.67	200.39	8.87	199.68	11.32
Periphery low	207.17	31.68	200.00	10.16	199.43	11.38
Populated route low	206.60	46.21	200.32	9.68	199.79	11.38
Clusters low	188.92	39.55	200.26	10.73	199.60	11.50
Center high	207.07	32.91	200.17	10.01	199.52	11.44
Off road high	179.48	28.91	200.10	8.03	199.62	11.72
Periphery high	193.20	45.15	200.39	10.76	199.74	11.45
Populated route high	190.46	48.11	200.21	9.12	199.47	11.44
Clusters high	204.44	53.20	200.29	10.43	199.63	11.48

Each mean and variance is an average from the 100 iterations. Note that for EPI the mean absolute difference between the obtained means and the true mean of 200 is 9.34. This, if it held in an actual sample would be very serious bias indeed. We are later going to show from our simulations an estimated variance of the EPI estimated mean of about 63. Should the bias actually be this large, i.e., 9.34, it would increase the mean square error to $63 + 9.34^2$, or about 150. In practice we would expect the biases to partly cancel each other out so as to be considerably less serious.

Table 2 shows results of introducing between village variance. The increase in EPI variance of the mean from 43.99, with no between village variance, to 63.36 with what our analyses of real data suggests is realistic between village variance, i.e. 25^2, is less than we guessed, although we might well have been able, in a theoretical analysis, to anticipate this relatively small increase. As is well known, all of the estimate of variance in with-replacement sampling, and simulations can be considered with replacement, comes from between the means obtained in the sampling within the clusters. One might think that where all clusters have the same mean, such as the 200 which we used in all our previous work, that introducing variance between clusters would greatly increase the variance estimates. Not so. The answer to our question " Are the biases and inefficiencies of the EPI method of sampling relatively unimportant with reasonable between village variances? " is No!

Finally let us look at the implications, in the case of a between village variance of 25^2, of a sampling method, EPI, that yields a variance of 63.36 as compared with a variance of 28.12 from a systematic method of sampling, assuming as we usually do that it is ok to use simple random sampling equations with systematically collected data, which is of course theoretical not exactly correct. The variance estimates in Table 1 are averages of 100 estimates and hence reasonably good estimates. Each of the 100 is an estimate of the variance of a mean, estimated from 30 means, each based on a sample of size 7 from villages all of the same size.

Table 2.
Means and variances of the means with between village variances of 0, 10^2, 20^2, 25^2, 30^2, 40^2, and 50^2. Each value is based on the mean of 100 samples, each of 30 villages.

	EPI		Systematic		SRS	
Between village variance	Mean	v($\bar{y}$)	Mean	v($\bar{y}$)	Mean	v($\bar{y}$)
0	197.86	43.99	200.01	9.54	200.06	11.34
10^2	197.81	47.93	199.95	12.98	200.01	14.95
20^2	197.78	57.65	199.92	23.35	199.98	24.73
25^2	197.70	63.36	199.85	28.12	199.90	31.03
30^2	197.41	72.83	199.56	40.27	199.61	42.10
40^2	197.70	97.45	199.84	65.21	199.90	66.24
50^2	197.98	127.52	200.12	94.24	200.18	96.02

This "all of the same size", is important. It means that the mean of each village, selected *srs* from some population of villages, is an unbiased estimate of this population mean, i.e. we have 30 unbiased estimates. Using $\bar{y}$ for the mean of the 30 village means and $v(y)$ for the variance of the distribution of the 30 means, we have $v(y)=\Sigma(y_i-\bar{y})^2/(30-1)$. Then the estimated variance of the estimated mean, $v(\bar{y}) = v(y)/30$. Working with the variance 63.36 with EPI, $v(\bar{y}) = 63.36 = v(y)/30$. So $v(y)=1900.8$. What would the size of the sample of villages have to be, with such EPI sampling under the conditions specified and with this $v(y)$, in order to obtain a $v(\bar{y})$ of 28.12? It would have to be about 68 as 1900.8/67.6=28.12.

To the extent our assumptions and methods reflect the real world of villages in the developing world, and they may not to a considerable degree, but if they do, we could obtain equally accurate estimates at perhaps half the cost - sampling 30 as opposed to 68 villages, and this is without thinking about the increase in *mse* with the likely EPI bias. And sampling only 7 from each village is probably not optimum. We would guess that 10-15 would be more efficient. But such questions will have to await further research. The most worrisome question we have about our simulations is the degree to which the local homogeneity that we have built into our data are realistic. Perhaps we will program for a range of such homogeneity and see how the variances are affected. We would like to be able to undertake research where we would use the EPI method, and compare the findings with a method of sampling that we have been developing (Matute & Fitch, 1991) which uses a hand held computer to draw a systematic sample in a sample of villages in Central America. With recent census data such as in Guatemala and working with INE, the government agency that conducted the census, we would only need to sample and collect identifying data.

REFERENCES

Fitch, D.J., Flores, R., & Matute, J. (1993, August). Excessive variance in EPI sampling due to biases illustrated by simulation. Paper presented at the Joint Statistical Meetings, San Francisco, California.

Fitch, D.J., Matute, J., & Flores (1982, August). Problems with EPI sampling illustrated by sampling simulated Guatemalan villages. Paper presented at the Joint Statistical Meetings, Boston, Massachusetts.

Matute, J., & Fitch, D.J. (1991, September). A hand held computer method of sampling households within selected PSU's. Paper presented at the 48th Session of the International Statistical Institute, Cairo, Egypt.