INCAP Publication CI/020

Draft (24 July 1977)of a paper being developed for presentation at the 51 Session of the International Statistical Institute, Istanbul, 18-26 August 1997. The views expressed here are those of the author and should not be construed as views of the Instituto de Nutricion de Centro América y Panamá.

SAMPLING IN DEVELOPING COUNTRIES: CAN BIASES BE CORRECTED, INEFFICIENCIES OVERCOME?

David J. Fitch, INCAP, Apartado Postal 1188, Guatemala, C.A., dfitch@incap.org.gt

Sampling in the developing world is often biased and inefficient. I will present methods for controlling two biases in surveys of vaccination coverage, and suggest one way to improve efficiency of survey work in the developing world. Finally I will note some things I've seen in developing world sampling, both problems and ways in which sampling in poor countries is, or can be less biased and more efficient than survey work in developed countries.

Removing Two Biases from EPI Selection When Estimating Vaccination Coverage

Likely the method of selecting dwelling units (DU's) most widely used in all the world - and probably many of you have never heard of it - is the EPI method. It developed out of efforts of the World Health Organization to monitor immunization coverage in, mostly villages. Some 5 years ago their Expanded Program on Immunization office in Geneva that keeps track of such things had registered some 5,000 EPI surveys and they probably receive reports on only a fraction. The procedure involves selecting 30 villages with probability proportional to size. Within a selected PSU, DU's are selected as follows. The interviewer goes to the center of the village, throws a pencil into the air, walks off in that direction counting houses to the edge of the villages. Then with a random number procedure she selects one at random as the starting house. The second DU is the one closest to the first, the third the one closest to the second, etc. She continues in this fashion until 7 children, say under 5 have been located.

So as is clear this is not a probability method. It is subject to biases. For example there may be twice as many houses to the north than to each of the other directions so these houses have only half the probability of being selected. There would seem to be no practical way of controlling this bias. But there is another bias, and likely a very serious one, that could be easily controlled. It is a consequence of the EPI practice, in coverage surveys, of continuing to select DU's until a fixed number of children, 7, have been found. Unless estimates are correctly made from such data, which is not the case with standard EPI procedures, results will be biased. A little example. Let's say that our population is made up of an equal number of two kinds of villages, all of the same size and with the same number of mothers in each. In the first kind mothers have half as many children as in the second but they are 100% vaccinated. Children in the second type of village are only 40% vaccinated. EPI will say the population rate is 70% - 40+100=140/2 - while it is actually only 60% - 100+2*40=180/3=60.

Statistically the correction for this bias is rather simple and straight forward. It involves the collection and use of counts of number of DU's selected and the number of children in these selected DU's. The interviewer could make these counts in each PSU and record them on forms. Then in the home office they could be entered into a computer along with the obtained vaccination data and analyzed. But it would be a lot easier, and there would be many advantages in using a little hand held computer both for data collection and data analysis. We have used the Casio FX880P, programmed in BASIC. It sells for about $150.

In order to keep the notation simple let's assume the only survey question is, "Has the child been vaccinated against measles?" Measles can be a big killer in our part of the world. Our estimate of the proportion of children vaccinated will be a ratio - a weighted sum over the sum of the weights. The weight for a child is the inverse of the probability by which the child was selected and will be the product of the probability by which the village is selected times the probability of selecting the child, given the selection of the village. The key to eliminating the biases of.concern here is to collect and use information which will allow these probabilities to be computed. Villages are selected with the probability $30M_i/M$, where M_i is the size figure, in terms of number of DU's, for the ith village and M the population sum of the M_i's. There is typically no problem here. But we need to introduce new procedures into EPI in order to have correct child selection probabilities. If we continue to select DU's in a particular village until k_i children, say 7, have been found in m_i selected DU's, the estimated number of children in the village, assumming M_i is correct is $k_i(M_i/m_i)$. Having found k_i of an estimated $k_i(M_i/m_i)$ children the estimate for the probability of a child being selected is $k_i/k_i(M_i/m_i)=m_i/M_i$. Now it will be likely that M_i is an out of date figure and that a better figure can be obtained in the village, either by making a count or by asking village leaders how many houses there are in the village. If the size is M_i^* then the correct probability estimate is m_i/M_i^*. Finally there may be some children for which we are unable to get vaccination information. We can adjust for such nonresponse by multipying our probability by k_i/k_i^*, where k_i^* is the total number of children in the m_i DU's selected and k_i the number of children for which vaccination information is obtained. So recapitulating, the probability for each child of the population being selected <u>and</u> vaccination information obtained is

$$\frac{30M_i}{M}\frac{m_i}{M_i^*}\frac{k_i}{k_i^*}\text{ , so the weight for each child of the ith village is }\frac{M}{30M_i}\frac{M_i^*}{m_i}\frac{k_i^*}{k_i}$$

Our ratio estimate of the proportion of children vaccinated against measles in our population is the following weighted sum over the sum of the weights, expressed here also in a nonstandard $\hat{Y}_i$ and $\hat{M}_i$ notation.

$$\bar{y}_r = \frac{\sum_{i=1}^{30}\sum_{j=1}^{k_i} \frac{M}{30M_i}\,\frac{M_i^*}{m_i}\,\frac{k_i^*}{k_i}\,y_{ij}}{\sum_{i=1}^{30}\sum_{j=1}^{k_i} \frac{M}{30M_i}\,\frac{M_i^*}{m_i}\,\frac{k_i^*}{k_i}} = \frac{\sum_{i=1}^{30} \hat{Y}_i}{\sum_{i=1}^{30} \hat{M}_i} \tag{2}$$

In usual notation $\hat{Y}_i$ is the estimated total for the *i*th PSU but here it is the estimated total for that part of the population "represented" by the children sampled from the *i*th PSU. Similarly $\hat{M}_i$ hat is the estimated number of children of the population "represented" by the children sampled from the *i*th PSU. Note that *M*/30 appears in both the numerator and denominator so in programming this, one would likely not include this constant.

A linear approximation for each $\hat{Y}_i/\hat{M}_i$ is

$$\hat{Y}_i\hat{M}_i^{-1} \approx \bar{Y}\bar{M}^{-1} + \bar{M}^{-1}(\hat{Y}_i-\bar{Y}) - \bar{Y}\bar{M}^{-2}(\hat{M}_i-\bar{M}) \ . \tag{3}$$

The estimated variance of the estimated mean $v(\bar{y}_r)$ can be relatively easily programmed in BASIC for the hand held computer and is as follows. Details of this development and the programming are give in Fitch (1997).

$$v(\bar{y}_r) = \frac{s_r^2}{n} = \frac{\hat{\bar{M}}^{-2}\left[s_{\hat{Y}}^2 + \bar{y}_r^2\, s_{\hat{M}}^2 - 2\bar{y}_r\, cov_{\hat{Y},\hat{M}} \right]}{n} \tag{4}$$

The use of the hand held computer makes practical these bias corrections. It would make practical continuing to work in the village until it was time to return, not just until 7 children had been located. It also might facilitate a transition in EPI selection to a probability design. For example if EPI selected every 4th DU, using their procedures, better sampling would result. Even better would be a system whereby the computer randomly selected one DU out of every four hence removing unconscious tendencies of the interviewer to influence selection. From such a procedure it would seem to be only a small step to a full probability design where the interviewer systematically walked through the entire village or say a randomly selected half, and let the computer make all the selection decisions based on probabilities given it. One DU would be randomly selected from each succesive interval, the interval being the inverse of the selection probability.

Improving Efficiency with More Optimum "Takes"

In sampling people or households one typically first samples clusters of DU's and then, within each selected cluster, a sample of DU's and interviews people of the households that live in the selected DU's. A common inefficiency is to sample too many DU's, i.e., to use "takes" which are too large. An organization carrying out a survey under contract, although they may know better - but they may not - will attempt to impress their customer by the size of their "sample". They talk of sample sizes, *n*'s, as number of people interviewed. In an important sense sample size is not the number of people, but rather the number of clusters. Perhaps that's why in the equations we use in such sampling, *n* is the number of clusters sampled, not the number of people.

Let's see if we can better see the problem by looking at a particularly bad sample design recently proposed. The little survey was a part of an effort to control dengue, a disease spread by mosquitos. The decision had been made to sample DU's in 5 towns of our Pacific coastal area and the design proposed for use in each town was to sample 6 clusters, and within each cluster to sample 60 DU's for a "sample size" of 6*60=360 in each town. We suggested this "take" size would be inefficient. In Esquintla, the first town to be surveyed and the largest, we used 29 clusters with "takes" of 10 DU's, obtaining data from approximately 9. As we had maps.for each town showing each DU we might well have used a simple random sample design, and this is what we did in the other 4 towns. But because we wanted to get started right away, and happily for this paper, we used the two stage design in Esquintla and thus have some statistics from which to estimate here the efficiency of different "take" sizes. We will use as our variable the total number of breeding sites within each DU. The variable, about the only one we have, is not ideal in that its distribution is very far from normal but let's forget this and just work with the variances obtained as if the varianble was normally distributed.

First let's review some statistics and make some simplifying assumption. We will be working with two variances, the within cluster variance and the between cluster variance. Now the variances within different clusters will not all be the same but we will assume that they are, and we estimated this variance by computing the variance of our variable within each of the 29 clusters. We will use the mean of the 29 within cluster variances as s_w^2, the within variance. The variance of the cluster means contains two components, the between cluster variance, s_b^2, which is what we are presently seeking, and s_w^2/m, which is the within cluster variance divided by the cluster "take", for which we are using the symbol *m*, about 9. The sum of these two variance, s_w^2/m and s_b^2 is s^2, the variance of the cluster means. Obtaining a within variance s_w^2 of .3160, with $s_w^2/9$=.0351 and a variance of the sample means of .0677, leads to a between cluster variance s_b^2 of ,0677-.0351 = .0326. Where all clusters are of the same size, sampled with replacement, and the "take" from each of the *n* sampled cluster is *m* then we can write the estimator of our sample mean variance in the following forms where y_i is the mean of the *i*th cluster.

$$v(\bar{y}) = \frac{s^2}{n} = \sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n(n-1)} = \frac{\frac{s_w^2}{m}+s_b^2}{n} = \frac{\frac{.3160}{9}+.0326}{29} = .0023. \tag{5}$$

Now let's use this equation to examine the consequences of different "take" sizes, assuming data are collected from the proposed 360 DU's. First let's look at a "take" of size 1, a simple random sample.

$$v(\bar{y}) = \frac{s^2}{n} = \frac{\frac{s_w^2}{m}+s_b^2}{n} = \frac{\frac{.3160}{1}+.0326}{360} = .00097 \qquad SE = .0311 \tag{6}$$

And now with a "take" size of 10.

$$v(\bar{y}) = \frac{s^2}{n} = \frac{\frac{s_w^2}{m}+s_b^2}{n} = \frac{\frac{.3160}{10}+.0326}{36} = .00178 \qquad SE = .0422 \tag{7}$$

Let us say we want a confidence interval that can be expected to hold the population mean 95% of the time. From the table of the normal curve we see this interval goes from 1.96 SE's below the sample mean to 1.96 SE's above the sample mean, so assuming the variable is normally distributed, which it was not but let's assume it was, for simple random sampling this interval is 2*1.96*.0311=.122. With a "take" of 10 this interval is 2*1.96*.0422=.165. The ratio squared of these two intervals $(.165/.122)^2$=1.83 is known as the design effect. The design effect can be interpreted as the increase in number of ultimate units needed with "takes" of 10 over the number with simply random sampling, in order to yield equal precision. Now let's see what kind of SE we get with a "take" of 60.

$$v(\bar{y}) = \frac{s^2}{n} = \frac{\frac{s_w^2}{m}+s_b^2}{n} = \frac{\frac{.3160}{60}+.0326}{6} = .0063 \qquad SE = .0794 \tag{8}$$

So if we computed the confidence interval as before it would be 2*1.96*.0794=.311 for a design effect of $(.311/.122)^2$=6.51. But with an *n* of only 6, i.e., 5 degrees of freedom this interval and design effect are not correct. We must use the table for the *t* distribution as Kalton (1995) has pointed out.. Here the SE beyond which .025 percent of the cases lie is 2.57, so the confidence interval is 2*2.57*.0794=.408, which means a design effect of $(.408/.122)^2$=11.18.

It would seem important to give careful thought to these questions of "take" sizes especially for more efficient work in the developing world where they may be as high as 40 or 50, and in the case just described the plan had been for 60. We saw just how very inefficient that would have been. In the developed world from the vantage point of Washington, DC the "take" trend over the last 20 years has been like from 10 to 8 to then 6 in the 1987 National Medical Expenditure Survey. There is some research and opinion that 4-5 would be optimum. Cochran (1977) does not consider optimum “take” for household surveys. Des Raj (1968) does, as have certainly others of whom I am not aware. His derivation for *m*, the “take” optimum in two stage sampling is based on the more complicated, without-replacement-at-the-first-stage variance estimator. Now it is true that in practice we sample without replacement, but generally we seem to now use the more simple, with replacement equations. We need not ask our developing world sampler - and the donor or bank that is funding the work - to learn and use more complicated equations than are used in sophisticated Washington agencies. But I’d like to think that those providing the money and advice for efforts with the goal of saving kids’ lives would appreciate a simplified development of optimum “take”, so let’s go through such.

Any such equation will involve within and between cluster variances, plus costs. One can get quite involved with these costs but as goal here is to arrive at an understanding of things without getting bogged down in details we will say our unit cost is the cost of one DU interview. We will use *c* such units as the cluster cost, i.e., the average of the costs per cluster of selecting, mapping, and getting to the cluster. Let's say we have 1000 units to spend on a survey. With *c* being the cost of a cluster, $nm+cn = 1000$ so that $n = 1000/(m+c)$. Substituting this value of *n* into the second expression for $v(\bar{y})$ above, taking the derivative with respect to *m*, setting it equal to 0, and solving for *m* we optain an expression for the optimum "take" size *m*, developed as follows

$$v(\bar{y}) = \frac{\frac{s_w^2}{m}+s_b^2}{\frac{1000}{m+c}} = \left(\frac{s_w^2}{m}+s_b^2\right)\left(\frac{1000}{m+c}\right)^{-1} = \left(s_w^2 m^{-1}+s_b^2\right)\left(\frac{m+c}{1000}\right) . \qquad (9)$$

$$\text{Letting } u = \left(s_w^2 m^{-1}+s_b^2\right), \text{ and } v = \left(\frac{m+c}{1000}\right), \quad \frac{d(uv)}{dm} = u\frac{dv}{dm} + v\frac{du}{dm} . \qquad (10)$$

$$\text{Now } u\frac{dv}{dm} = \left(s_w^2 m^{-1}+s_b^2\right)\frac{d\left(\frac{m+c}{1000}\right)}{dm} = \frac{\left(s_w^2 m^{-1}+s_b^2\right)}{1000} , \qquad (11)$$

$$\text{and } v\frac{du}{dm} = \left(\frac{c+m}{1000}\right)\frac{d\left(s_w^2 m^{-1}+s_b^2\right)}{dm} = \left(\frac{c+m}{1000}\right) -s_w^2 m^{-2} = \frac{-cs_w^2 m^{-2} - s_w^2 m^{-1}}{1000} , \qquad (12)$$

$$\text{so}\quad \frac{d(uv)}{dm} = \frac{s_w^2 m^{-1} + s_b^2 - c s_w^2 m^{-2} - s_w^2 m^{-1}}{1000} = \frac{s_b^2 - c s_w^2 m^{-2}}{1000} \ . \qquad (13)$$

Setting this derivative equal to 0 in order to solve for optimum m we have

$$0 = s_b^2 - c s_w^2 m^{-2}\ , \quad m^{-2} = \frac{s_b^2}{c s_w^2} = m^2 = \frac{c s_w^2}{s_b^2}\ , \qquad m = \sqrt{c s_w^2 / s_b^2} \qquad (14)$$

Although this simplifies the real situation, and to use it one needs cost and variance information, it is likely that "takes" in the range of 5-10 would usually be optimum where often in surveys in the developing world we see "takes" of like 40-50, likely very inefficient. Let's say the s_W^2 / s_B^2 ratio is 8, which is typical of what we've found in our limited investigations. And let's say that the cost associated with each cluster, getting there and doing whatever needs to be done for sampling, but not including the locating of sampled units and data collection, is equal to 6 times the costs of locating and interviewing one unit. With a "take" of m per each of n clusters the total cost is thus $n(6+m)$, or with n=100 and m=40, 4600 for a $v(\bar{y}) = (8/40+1)/100 = .012$. With "takes" of 7, near optimum, the n needed for such a variance is about 179 for a cost of 179(6+7)=2327, or about half.

Some Examples of Developing World Sampling and Hopes for Improvements

So the answer to my question is "Yes". Some biases can be corrected, some inefficiencies overcome. But if this is to happen I think that we of the IASS will need to find ways of taking a more active role with the donor organizations. In my opinion they are typically weak in sampling expertise. Let me give some examples.

1. The 1987 DHS sampling manual endorses "takes" of 40 households in rural areas. Now the 1995 DHS work in Guatemala was the best that I have seen in my 8 years but they were apparently still using 40. My guess is that equally accurate estimates could be obtained at half the cost with "takes" of say 7.. With a 43 million dollar contract there are possibilities for saving a lot of money, a lot of kids' lives. It would be a relatively easy effort to obtain cost and variance information from which to calculate more optimum "takes".

2. A major donor organization contracted with a firm that set up an office in Guatemala. Their project was to teach departmental health offices how to sample. The organization has no statisticians. They taught the EPI method, but with a modification. In the name of rapid assessment health officers were instructed to skip the effort of finding a random starting house. Selection, they taught, was to start in the middle of every village.

3. A survey was carried out in Guatemala funded by an international organization to estimate the

vitamin A status of children. There were a number of problems. The same 2-3 year old maps that were used for the DHS survey without updating were used. [DHS usually updates maps (Lê, 1997) but in a few cases has not.] When a mother in a sampled house had say gone to the market and would not be back until later, the team would look around for children playing nearby and would go there. For selecting the one child of the family from which to draw blood, the procedure seemed to pick the one least afraid. When it started to rain in the afternoon work was ended in the village.

4. I found myself advising on a 7 million dollar, child health survey to be financed by one of the international banks but there were two problems. The people who wrote the specifications had very limited knowledge of sampling, assuming, e.g., that simple random sampling equations were correct for the 4 stage design specified. And the group bidding didn't really have a sampling statistician. She said that sampling one child within one family within one DU would give a self weighting sample. But how could I could encourage them to submit a probability design as it seemed reasonably clear that reviewers would not understand the need for the costs associated with such a design. It's an unfinished story as the project has been suspended.

Let me end on an optimistic note. We in a developing country such as Guatemala have opportunities for less biased, more efficient sampling than is possible in a large developed country such as the US. On bias, the refusal rates, at least outside of Guatemala City are very low. I referred to the survey of mosquito breeding sites on our south coast. We knocked on some 1,000 doors and less than 1% refused our request to look around inside their houses for larva. What would be the refusal rate in New York City to surveyors wanting to look for cockroaches in peoples' kitchens.

With regard to efficiency, as all survey work is out of the capitol, the cost of PSU offices which limits the number of strata in the US to typically about 50 is not a concern here. For census purposes Guatemala is divided into about 13,000 sectores of 100-200 DU's. These can be used as PSU. My computer will hold a triangular matrix of similarity scores between some 5,500 such PSU's which could be based on census data, and develop strata therefrom. This means, selecting 10 DU's per PSU with a total national sample of about 7,000 DU's, some 300 strata could be created in three runs. If there were no particular reason to use the same PSU's for a new survey, stratification could be tailored to increase the efficiency of the particular survey.

References

Cochran, W. G. (1977). Sampling Techniques, 3rd ed., New York, Wiley.

Des Raj (1968). Sampling Theory, New York, McGraw-Hill.

Fitch, D.J. (1977) Some sampling theory for work in developing countries. Manuscript submitted for publication.

Kalton, G. (1995) Variance estimation with few degrees of freedom. Proceedings, ACTES International Association of Survey Statisticians, ISI. Beijing