WEIGHTING OF SURVEY DATA: IS THE VIEW THAT SUCH SHOULD NOT BE ATTEMPTED IN DEVELOPING COUNTRIES REASONABLE?

David J. Fitch, Instituto de Nutrición de Centro América y Panamá (INCAP) and Universidad del Valle de Guatemala (UVG), Roberto Molina and Tomás Barrientos, UVG.
David J. Fitch, INCAP, Apartado Postal 1188, Guatemala, Guatemala 01011. dfitch@incap.org.gt

Key Words: Efficiency, Bias, EPI, DHS, ENIGFAM

Introduction

Chris Scott was a key member of the World Fertility Survey team (Verma, Scott, & O'Muircheartaigh,1980). Later he did well regarded survey work, primarily in Africa. In a sampling manual (Scott, 1987) produced to guide work in the Demographic and Health Survey (DHS) program of the United States Agency for International Development (USAID), he gives his four principles for survey work in developing countries. One states in part, "*Samples should be self-weighting unless there is good reason to depart from the principle in specific cases. In countries where statistical offices are new or lack resources and/or personnel, the use of weights may present problems. The need to compute weights and carry them as part of the database, the need to assess when and how they should be applied, and to correctly report their use, can be an appreciable burden on staff*". This has been the typical wisdom. One reason for not using weights when Scott formulated his principles was that at that time there were much more limited computer facilities than now. An additional reason might be that the donor agencies funding developing country work, and the contractors they select, either don't understand weighting and the frequent benefits of such, or find it easier not to do weighting. We would like to make a case for weighting.

With the correct use of weights, estimation can be improved. We of the statistical community, have a big challenge in how best to work with groups like the Expanded Programme of Immunization (EPI) people of the World Health Organization (WHO), and groups sponsoring survey work within the USAID, in order to help them understand the benefits and possibilities of improving estimation and saving money through the use of modern survey methods. At the same time, we need to be sensitive to the difficulties of people working in developing countries, realizing that what we might see as good survey practices may not be as easy to undertake as we might think. The benefits of better sampling practices in developing country work are enormous. It is in such countries that an awful lot of children are dying, that shouldn't be dying, that wouldn't be dying if there were better information available to help the relevant ministries and concerned individuals spend their limited money better.

The EPI Method of the WHO

In estimating children's vaccination rates using the EPI method, the interviewer goes to the center of a sampled -with probability proportional to its size- village, throws a pencil into the air, and goes off in the direction to which it points, counting houses. (WHO,1991). Using some random number procedure - say getting a serial number from a note- one of the counted house is selected as the starting house. The second house is the one closest to the first, the third the closest to the second, etc. The interviewer continues until 7 children are found and then quits the village. Thus EPI does not use probability sampling. We have considered this problem elsewhere (Fitch, Matute & Flores, 1992). Here we are concerned with a bias problem that could be controlled through the use of weights. As EPI is practiced, children in villages where women have more children, are undersampled. Take an extreme example of two villages, each of size 140 dwelling units (DUs) but there being twice as many children per DU in the first compared with the second. And let us suppose that only 40% of the children in the first village are vaccinated, as compared with 100% in the second village. Let us imagine that the interviewer goes to 14 DUs to find her 7 children in the first, but 28 in the second village. To weight the two villages equally, as does EPI, will give biased estimates. If our population of villages is made up of half of the one kind and half of the other, EPI will estimate the population's vaccination rate as (40%+100%)/2=70%, where the true expected estimate is (40%*2+100%)/3=60%. Well, the example is extreme but it illustrates the problem. The solution is to use weights. One would count the number of DUs visited. The weight would be proportional to the inverse of the number DUs visited to find the 7 children. In making EPI estimates using such weights, variance estimation would be more complicated, but as nowadays EPI estimates are made using computer programs such as COSAS (1992), this would be no problem. The programs would only need some revision. If one were to modify EPI procedures using a Pocket PC in order to do probability sampling, the computer used for sampling could be programmed to do all the estimations. We should note that going until one has found 7 children rather than going to some fixed number of DUs, means that our weighting scheme is not exact but we will not further consider this small problem.

The 1995 Guatemalan DHS

Our second example of where weighting would improve estimation is from the 1995 Guatemalan DHS, one of a large number of surveys carried out by USAID in developing countries. We find five design problems in this survey, which we list at the end of this section, and describe the weighting we used and the improved estimation resulting from this weighting. But first some background. Guatemala in the early 1990's was planning to conduct a census. With hopes that this might be carried out in 1993, an operation was conducted in 1992 by the cartography group of the National Institute of Statistics (Instituto Nacional de Estadística, INE) to form census enumeration districts, called sectores in Guatemala, and to map each sector showing each DU, or in Spanish vivienda. There was a delay in carrying out the census so that the mapping continued into 1993. In the spring of 1994 the census was conducted. Sectores for the DHS were selected, as Primary Sampling Units (PSUs), with probability proportional to their size figures found at the time the maps were drawn, some 2-3 years out of date. The DHS plan had been originally to select 501 PSUs and within each PSU to select 25 DUs. However plans changed. About 34 DUs were to be selected from now 407 PSUs. Census DU size figures became available well before the DHS data collection started, and showed considerable size changes. In general, the rural, traditional PSUs -with less well educated women and having more babies- decreased in size and the urban PSUs increased in size. Table 1 gives a summary of these statistics.

Table 1

Number of strata: 196	
Number of PSUs: 407	
Aimed-for number of DUs per PSU: Mean 34.3, Range 9 - 71 and Standard deviation 5.89	
Number of DUs in which interviewing took place per PSU: Mean 22.4, Range 1 - 49 and Standard deviation 6.31	
Number of women interviewed per PSU: Mean 30.5, Range 1 - 67 and Standard deviation 9.66	
Mean number of women per DU:	1.40
Undersampling in rural (urban) PSUs:	0.37 (0.28)
Undersampling in PSUs with women with less (more) education:	0.38 (0.30)
Undersampling in PSUs without (with) electricity:	0.38 (0.32)
Undersampling in PSUs with women having more (fewer) children:	0.38 (0.31)
Undersampling in PSUs with more (fewer) children dying:	0.37 (0.33)
Undersampling in PSUs with women less (more) knowledgible about contraceptives:	0.38 (0.32)
Undersampling in PSUs where women less (more) likely to hear message on radio:	0.38 (0.31)
Undersampling in PSUs where ideal number of children more (less):	0.38 (0.32)

With these size figures available, in order to have selfweighting samples, the original number of DUs -about 34 per PSU- was changed to a new aimed-for numbers of DUs, $m'_i = 34\frac{M'_i}{M_i}$ where M_i is the PSU size at the time of the map making and M'_i the size found in the 1994 census. The new sizes retained a mean of about 34 per PSU, but the numbers of DUs varied considerably with a range of 9 to 71 and a standard deviation of 5.89. Selected PSUs were not remapped. This meant that in marking the selected DUs on maps of the PSUs that had decreased in size, women in such PSUs would be undersampled. Perhaps we should spell this out. Consider an extreme example where a PSU had gone from 200 DUs at the time the map was drawn to 100 at the time of the census. The DHS needed now to select not 34 but the women in 17 DUs, in order to keep a design where weighting would not be required. On the map showing 200 DUs, 17 were marked for visiting but as half of the 200 were no longer there or were unoccupied, interviewers were likely to be able to find women in only 8 or 9, and not 17 DUs . Of course, there would be other reasons that might account for undersampling in rural and more traditional PSUs. Such would tend to be isolated and hence both harder to get to and more time consuming to get from, leaving less time for the interviewers to find the women in all of the selected DUs. Poorer families would be more likely to need to take employment harvesting sugar cane on the South Coast and hence be temporarily absent. There would be other possible reasons for the undersampling. In any event, it would be reasonable to attempt to partly correct for bias due to this undersampling, through adjusting weights using the aimed-for DU "take" size m'_i, divided by m^*_i, the number of DUs in which interviewing actually took place. This is what we did and the results are shown in Table 2.

The key results are, for the 7 variables considered, our estimates of the sample size in terms of number of PSUs that would be needed to obtain the same *mse*s with unbiased estimates, as compared with the biased estimates from the DHS design. We need, before going into the explanation of the computations understood that our goal is not to arrive at better estimates than the ones found in the DHS published report. Our goal is to show weighting possibilities which might well contribute to more accurate estimation in developing country work. Second, what we call unbiased estimates are not really unbiased estimates. They are only the estimates obtained using the weights which we used. There would be other sources of bias not controlled, and/or not controllable. For example the design used, as we understand it, did not allow for women living in newly constructed housing to be included in the sample. Third, our biased estimates, made without the weight adjustments, will not match exactly the published estimates since we imputed for missing data.

Table 2

Variable	$\bar{y}$	$\bar{y}'$	b	$v(\bar{y})$	$v(\bar{y}')$	$v(\bar{y}-\bar{y}')$	$b^2 - v(\bar{y}-\bar{y}')$	mse	n'
Nc	3.708	3.782	-0.0738	0.00168	0.00140	0.00020	0.00525	0.00693	122
Edu	0.997	0.953	0.0433	0.00054	0.00049	0.00075	0.00180	0.00234	103
Luz	0.641	0.605	0.0362	0.00036	0.00034	0.00004	0.00128	0.00164	105
Cb	2.857	2.939	-0.0082	0.00181	0.00172	0.00030	0.00647	0.00828	125
Cd	0.310	0.323	-0.0124	0.00010	0.00010	0.00002	0.00014	0.00023	215
Kc	2.353	2.287	0.0655	0.00073	0.00065	0.00008	0.00421	0.00492	83
Radio	0.472	0.456	0.0163	0.00005	0.00005	0.00001	0.00026	0.00031	101

Estimation of n', the number of the 407 PSUs needed to give equal *mse* with the bias controlled by weights.
nc- ideal number of children; edu- education; luz- electricity; cb- children born; cd- children died; kc- knowledge of contraceptive; radio- heard on radio.

The PSU sample sizes, the n' s, given in Table 2, were obtained as follows. Believing that the best estimates we could make -ones which would take into account the complexities of the design and the data collection- would come from taking systematic samples from the 407 PSUs, and computing variance estimates using the Taylor series linearization method; our program selects at random a PSU within the first $407/n$ PSUs, where n is the number of PSUs to select, and then, using this interval, n-1 more PSUs are selected. The selected PSUs are divided into urban and rural PSUs. The first two urban PSUs form the first stratum, the second two the second, etc. If there is an odd number of PSUs in either group then the last stratum of the group has three PSUs. Stratum numbering continues through the rural PSUs. This formation of strata mimics that used in the survey. Variance estimation for the ratio means -weighted sum divided by the sum of the weights were computed- was checked against variances obtained using PC CARP. For each n, 50-130 and 160-220, 1,000 iterations were made. There was still considerable variation in means based upon 1,000 estimates but the n' s shown in Table 2 should be reasonably good.

Now let us describe the *mse* computations involved in Table 2. The biased estimates, the $\bar{y}$ s, are with the use of the DHS weights, unadjusted. The $\bar{y}'$ s are with our weight adjustment, $\frac{m'_i}{m^*_i}$. The bs are their differences. The $v(\bar{y})$ s and the $v(\bar{y}')$s are the variance estimates from PC CARP. Now, b^2 would not be an unbiased estimate of the bias squared term of the MSE as each estimate has sampling error. We have from Phil Kott (personal communication, and we believe the appendix contains a proof) that a unbiased estimate of the bias squared term, which would enter into the MSE estimate, is $b^2 - v(\bar{y}-\bar{y}')$. So the estimated variance of the estimated difference is also given in the table. PC CARP can be used to get the variance of the difference between two ratios, but has no provision for different weights for the two ratios. We handled this by computing and specifying variables that gave what was wanted. The next column gives this bias squared term. The *mse* is the sum of this and $v(\bar{y})$. The numbers in the final column are the estimated numbers of PSUs, needed to give unbiased estimates of equal accuracy to the biased estimates found without adjusting the weights. We found this number, on average for the 7 variables with which we worked, to be about 122. This suggests that using weights would give equally accurate MSE estimates, sampling less than one third the number of PSUs.

So, our analyses suggest that efficiency could have been considerably increased with the use of weights. But we might note that the survey had a series of design problems and weights helped only with the differential undersampling. The problems are: (1) The aimed-for number of DUs , 34 or about 47 women, is too large for efficiency. The optimum number is likely to be in the 5-15 range. USAID would do well to initiate research here. (2) PSUs should have been selected with probability proportional to the size figures available from the 1994 census -one year old- and not with the three year old figures obtained at the time the maps were drawn. Using these old figures, and adjusting with the census data, led to an aimed-for DU "take" range of 9-71. Such is quite inefficient. (3) Selected PSUs should have been remapped as called for in the Scott manual (Scott, 1987), and as is standard practice. Remapping would have given non-zero probability to women in DUs constructed since the 1992 mapping as well. The failure to include these women means further but uncorrectable bias. (4) With the failure to remap they should have at least marked 34 DUs on the old maps of PSUs that had shrunk rather than say 17, where a PSU was half the size it was in 1992 -as was pointed out by Gary Shapiro, discussant of the session. And finally, (5) failing in these design considerations, they should have at least done weighting similar to that reported here.

Let us for the moment engage in speculations with the idea of seeing possibilities with the use of better survey designs and better analysis procedures, including the use of weights, in surveys carried out in developing countries. Let us imagine that our results using weights with 7 variables from one DHS survey might hold generally for all DHS variables for all DHS work -hardly likely but let us so imagine. And remember we have not attempted to estimate the consequences from inefficiently large, aimed-for "takes" -some 47 women in 34 DUs, on the average, in the Guatemalan survey. If data collection costs for all of the DHS surveys that have been conducted, could have been reduced to one third of what they actually were -as suggested by our findings in the one survey and looking at a very limited number of variables- and if data collection costs are half of the total costs of the program that has spent some 120 million US dollars to date, then we could argue that we of the statistical community could have saved the US Government 40 million dollars. Not likely! But a few million dollars. Quite likely!

We have a suggestion for people doing surveys in a country like Guatemala. While selecting from maps that are three years old is certainly extreme, it's likely that selection from maps or listings that are more than a few months out of date is not uncommon in developing country work. In quality work done in developed countries, personnel are sent to the selected clusters shortly before interviewing is to take place. They list all the DUs that are in the cluster at that time, and it is from such lists that DUs are selected and interviewing assignments are made. Developing country work needs some practical way of doing something similar, i.e., listing and selecting just prior to interviewing. Our suggestion is to have the sampling and interviewing done by two teams. Team 1 goes into a selected village a day before Team 2 arrives to do the bulk of the interviewing. After getting permission from the village officials, they undertake sampling using a Pocket PC. The supervisor-sampler, with her assistant, goes systematically through the village "pointing" the computer at each DU and touching the enter key. If the computer selects it then inquiries are made as to whether the DU is currently occupied. If so, the computer assigns a questionnaire number. This number will be written on a new questionnaire along with the name of the head of the household, and information on when the people living there are likely to be at home on the following day. Team 1 will have camping equipment and will stay in the village overnight, interviewing in the evening women who will not be at home during the following day. This sampling team will go with the best maps available, but will not limit their sampling to the parts of the village that existed at the time the maps were drawn. The Pocket PC will be programmed for a variety of possibilities and will keep track of all the selection probabilities. For example, if a village is quite spread out, the sampler could locate a road or stream that divides the village. She could obtain an estimate from people of the village as to what percent of the DUs were in say the larger part, and enter this percent into the computer. The computer would select one part with probability proportional to size. Then only this part would be sampled, although there would be the possibility of sampling the other part, perhaps at a reduced rate, if it were later judged that the interviewers would have time to work in both parts. If sampling at the initial rate was found to be selecting too many or too few DUs for the time available, the sampler could change the selection probabilities. The PC would keep track of all of these probabilities and when the interviewing has been completed and Team 1 meets up with Team 2 in the next village, any non-response will be reported by Team 2 and the computer will compute weights which will be entered onto the questionnaires.

In our limited experience we have found that people in developing countries, especially village people, seldom refuse to be interviewed. A system such as we have here suggested might well increase the response rate -percent of selected DUs in which interviewing takes place- from the 65 percent of the 1995 Guatemalan DHS to the high 90s. The weighting we did is, we believe, good and reasonable, but far better would be procedures which reduced the non-response, as here defined.

We want to end this part on the 1995 DHS in Guatemala, by reminding the reader that we have been concerned only with the possibilities of improving estimation with the use of weights to correct for non-response. Whereas we believe the type of weighting we did would have been helpful, we have not attempted to evaluate the work generally. We want to say that, in our opinion, the quality of the data collection as it was observed in two villages of Momostenango was excellent. With regard to the design, good stratification procedures were used. And the contractor does an excellent job of making the data available.

The 1999 ENIGFAM

Our last example of where weighting could have improved estimation is the Guatemalan 1999 Family Income and Expenditure Survey (Encuesta de Ingresos y Gastos Familiares: ENIGFAM). Let's start with some background. The interviews of the survey were carried out from March 1998 to June of 1999. There were selected approximately 430 PSUs with probability proportional to their sizes -sizes at the time of the 1994 national census- and with replacement. From each selected PSU there were selected, with simple random sampling and without replacement, 15 households (HH), or in Spanish hogares, for the rural PSUs and 18 HHs for the urban ones. With this sample of approximately 7,350 HHs, estimates were computed. In preparation for the survey, during 1997 the selected PSUs were visited by the personnel of the cartography department of the INE, to update the PSUs' maps and their size figures.

At the publication of the results and the technical notes of the statisticians in charge (ENIGFAM, 1999), we notice that the surveyed PSUs were selected using size figures that were at least 3 years old. We understand that these same size figures were used, without adjusting them with the updated size figures, to weight the answers of the interviewed HHs. Instead, the estimates for totals in each stratum were adjusted using an estimate of the stratum's new size. Let us look at the problem with such weighting. Let's say that a particular stratum has two PSUs -two villages each with 100 HHs at the time of the 1994 census. The survey would have selected in 1998, 15 HHs in each PSU for interviewing. But let's say that in 1998 one PSU has only 50 HHs while the other has 150 HHs, and so economic conditions in the two villages must have been quite different with regard to the ENIGFAM variables. Were we not to take into account these size changes through weighting, our estimates would be biased. The 15 HHs of the 150 HHs village should have weights three times the size of the 50 HHs village.

Initially we sat as our objective the computation of a set of estimates adjusting the original weights as we did for the DHS data. We believe that our proposed adjustment of the weights would produces less biased and more efficient estimators, which would mean smaller samples and less expensive surveys. Unfortunately, our solicitude of the survey data was rejected by INE. This solicitude was actually a process of several verbal and written communications to explain the data we needed and the purpose of our research. We asked INE for the surveyed PSUs size figures at the time of the census and at the cartographic update, the number of family members of the sampled HHs, and the values of three variables like family income and expenses, and number of members laboring, but with no identifying information. We got a letter from the Head of INE together with a juridical dictum of INE's juridical office, saying that our solicitude attempts against the confidentiality of statistical information and so violates Guatemala's national Constitution. We understand INE's concern for confidentiality, but surely there is a way they could have given the data to us without problems of confidentiality. Surely too, it is beneficial to examine what was done and from this examination, to propose how could survey work in Guatemala be improved. With no data, we can just comment on the adjustment proposed.

We don't see a straight forward solution to the problem of sampling with PSUs size figures that are at least 3 years out of date. But we believe that adjusting the weights with the updated PSUs size figures would at least produce less biased estimates. Considering any stratum and denoting M_i and M'_i the number of HHs in the i-th sampled PSU at the time of the census and the cartographic update respectively, M the total number of HHs in the stratum at the time of the census, n the number of sampled PSUs in the stratum and m the number of HHs sampled in each sampled PSU. In the survey, we understand a total for the stratum was estimated as $y = \frac{T'}{T}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{M}{mn}y_{ij}$, where $T' = \sum_{i=1}^{n} M'_i$ and $T = \sum_{i=1}^{n} M_i$. In particular, the total number of HHs in the stratum at the time of the survey was estimated as $M' = \frac{T'}{T}M$. The quotient $\frac{T'}{T}$ actually adjusts the original weights keeping them equal for all HHs in the stratum. But weights should be equal only if all PSUs in the stratum had changed in size at the same rate during the period from the census to the survey, which of course did not happened. We note that each HH sampled from the i-th sampled PSU of the stratum was actually selected with probability $\frac{m}{M'_i}$, and so any total for the stratum should be better estimated as $y' = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{MM'_i}{mnM_i}y_{ij}$, and M' should be estimated as $\sum_{i=1}^{n}\frac{M}{nM_i}M'_i$. We should note that y' is also a biased estimate since $\frac{mnM_i}{MM'_i}$ is not quite the probability of selecting a HH of the i-th PSU, because of the HHs that existed at the time of the census but not at the time of the survey, and the HHs that didn't exist at the time of the census but they did at the time of the survey. The squared difference of this two estimates is

$$(y-y')^2 = \frac{M^2}{n^2}\left[\sum_{i=1}^{n}\left(\frac{T'}{T}-\frac{M'_i}{M_i}\right)^2 \bar{y}_i^2 + 2\sum_{i=1}^{n}\sum_{k>i}^{n}\left(\frac{T'}{T}-\frac{M'_i}{M_i}\right)\left(\frac{T'}{T}-\frac{M'_k}{M_k}\right)\bar{y}_i\bar{y}_k\right]$$

where $\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$.

Appendix

In general, if we have a biased estimator U of a population parameter θ, $Bias[U] = E[U] - \theta \neq 0$ and $MSE[U] = V[U] + Bias[U]^2$. If we also have a second estimator W, an unbiased estimator of θ, we have that $B = U - W$ is an unbiased estimator of $Bias[U]$. Now, to estimate $MSE[U]$ we actually need to estimate $Bias[U]^2$. Our first guess B^2, is a biased estimator of $Bias[U]^2$, in fact $E[B^2] = Bias[U]^2 + V[U-W]$, as it is shown bellow.

$$\begin{aligned}
E[B^2] &= E\left[(U-W)^2\right] \\
&= E\left[(U - E[U] + E[U] - \theta + \theta - E[W] + E[W] - W)^2\right] \\
&= E\left[(U-E(U))^2\right] + E\left[(E[U]-\theta)^2\right] + E\left[(E[W]-\theta)^2\right] + E\left[(W-E[W])^2\right] \\
&+ 2E[(U-E(U))(E(U)-\theta)] - 2E[(U-E(U))(E[W]-\theta)] - 2E[(U-E(U))(W-E[W])] \\
&- 2E[(E[U]-\theta)(E[W]-\theta)] - 2E[(E[U]-\theta)(W-E[W])] + 2E[(E[W]-\theta)(W-E[W])] \\
&= V[U] + Bias[U]^2 + V[W] - 2CV[U,W] \\
&= Bias[U]^2 + V[U-W]
\end{aligned}$$

Thus If we have an unbiased estimator of $V[U-W]$, say S^2_{U-W}, then $B^2 - S^2_{U-W}$ is an unbiased estimator of $Bias[U]^2$. If we also have an unbiased estimator of $V[U]$, say S^2_U, then $S^2_U + B^2 - S^2_{U-W}$ is an unbiased estimator of $MSE[U]$.

References

COSAS (1992). EPICENTRE, 8, rue St Sabin, 75011 Paris

Fitch, D.J., Matute, J., & Flores, R. (1992, August). Problems with EPI Sampling Illustrated by Sampling Simulated Guatemalan Villages. Paper presented at the Joint Statistical Meetings, Boston, Massachusetts.

Instituto Nacional de Estadística (1999). Encuesta de Ingresos y Gastos Familiares. INE, Guatemala.

Scott, C.(1987) Sampling Manual. Institute for Resource Development, Columbia, Maryland

Verma, V., Scott, C., & O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. J. R. Statist. Soc. A, 143, Part 4, 431-473.

World Health Organization (1991). Training for Mid Level Managers: The EPI Coverage Survey. WHO Expanded Programme on Immunization, Geneva, Switzerland. (Document WHO/EOI/MLM/91.10).