ESTIMATION OF AREA IN CROPS ALTERNATIVE TO COCA IN THE CHAPARE, BOLIVIA: A CONSULTANT'S REPORT

Charles Proctor
1110 Blenheim Drive, Raleigh, NC 27612

Key words: Area sampling, Dual frame, Nonresponse, Weighting, Acculturation

Introduction

The ultimate and concrete objective is to do an enumerative sample survey to furnish the estimate of numbers of hectares in alternative crops. We are (as of August, 2000) still planning the survey so we do not know even if it will be possible, much less how well it will be done. In the talk the most we can hope to do is explain how we plan to do the survey. Our viewpoint is the specialized and narrow one of the consultant. (We adopt a bee's eye view in accord with the *New Yorker* cartoon that shows the worker, queen, drone and consultant bees.)

The report will, thankfully, be incomplete as the details of even this brief tour can take more time to relate than it took to live through. We will cover the following points:

- Gadgets
- Mental background in language, culture and area sampling skills
- Pre-trip briefing in library and with local experts to acquire point sampling skills
- Scope of work and first day ploy
- Actualization of sample selections
- Review of previous years' sample surveys
- Nonresponse Possibilities
- Dual Frame Design
- Conclusions

Gadgets

You can't see too well the special equipment the consultant bee carries. Actually most of it is in his head, and we'll describe that shortly. He does have some gadgets. A laptop and a printer are standard; stapler, staple extractor, calculator, erasers, Exacto knife, scotch tape, rulers, clips, pens, pencils, and refills are in a ditty bag. I also took a CD player and diskettes, but left my 220-to-110 transformer behind and that was a mistake. I took camera and tape recorder but misused them - - better luck next time. The GPS IIPlus (Garmin™ geopositioning equipment) was very helpful in benchmarking the map the office furnished, which only had kilometer coordinates, but I did not use it in sampling. A page of random numbers is always useful. Even though computer generated ones can serve in many instances, it seems simpler to verify or audit how the numbers were used when starting from the printed page rather than having to trace some machine code.

Mental Background

Spanish language skill was essential and experiences in Latin America helped. Both MSU and NCSU have active international programs that I participated in. I enjoy tracing my statistics training back to P. C. Tang, whom I never met, but who had just passed through Costa Rica in 1951 and set up an elegant sampling scheme for their Ag. Census. It had five 1/100th replicated subsamples. We quickly got the overall estimates (by Census Office clerks punching the cards and using IBM accounting machines) and the five estimates. Then we ground out the standard errors on a (what?) Monroe calculator. Remembering this experience, I always try to furnish replicate subsamples. At a tender age I also spent a few months in Ann Arbor as a clerk in Leslie Kish's Sampling Section of the Institute for Social Research at the U. of M. I delineated area segments from Master Sample of Agriculture materials for their survey of Consumer Finances. There I learned that the concrete product of a sampler was writing out the selected addresses. Thus when I arrived in North Carolina in 1960 I could well appreciate Al Finkner's contributions to area sampling (his and John Monroe's *Handbook of Area Sampling*, is long out of print but making a copy in the library is well worth while) and have continued to admire the work done at RTI by Bob Mason and Jim Chromy. I have sent in some papers on area sampling to ASTM and SRMS of ASA and have published one in Estadistica (1991:27-51) that represents what I take to be the ideal area sampling survey design. Bill Wigton gave a paper at the ASA in Baltimore (1999) on his recent survey in Malawi that reinforces the current relevance of the method.

Here is the image of the ideal area sample survey in terms of 1) Frame, 2) Arithmetic, 3) Equal probability, Multi-start, Systematic, Random selection, 4) Replicated subsamples, 5) Simple raising factors and 6) Report in on time. Each of these topics deserves its own seminar but a few words on each will have to suffice.

- A frame is what? Ans.: An indexed list of addresses for taking measurements plus the materials (maps, directories or other) referred to in the addresses. This definition will someday become standard but now it is implicit in sampling practice. When Dr. Deming saw it in an ASTM proposed standard guide he wrote me, "There is no secret about frame . . ." and he enclosed copies of his papers. You should be familiar with his *Sample Design in Business Research* (Wiley, 1960). While we are naming textbooks you should also know Frank Yates' *Sampling Methods for Censuses and Surveys*, which came from his work at the UN and is appropriate for work in underdeveloped settings.
- Area sampling arithmetic is the calculation applied to size information to find out how many area segments to make of each count unit. It is so deceptively simple it can only be termed arithmetic. However, it must be done in just the one way. The frame materials must define count units which partition the study area and which are larger than the sampling units and which have unambiguous boundaries and to which sizes can be assigned. This is a tall order and acquiring frame materials will occupy the consultant's first hours, days, weeks or months. Once all sizes are in one column, cumulate them, note the total, divide by segment size to get a rough number of sampling units and round this to get a convenient value of N. Now multiply the cumulated sizes by N/total-of-sizes and round to get cumulated numbers of sampling units in the next column. These are sacred numbers as they determine (by differencing) how many SU's in each CU. (At this point at least two possibilities for forming the SU's arise. One is to "cruise" the CU and list housing units or fields or whatever, and then create systematic samples as SU's. The other is to cluster adjacent housing units into compact clusters, number then and re-randomly select one for interviewing.)
- Selection can be done from the column of cumulated SU's by imagining one list of N units and using multi-start systematic selection to get selection numbers of SU's to form replicated subsamples.
- Each enumerator team is assigned to just one subsample so subsample differences will also reflect measurement errors from teams.
- The basic raising factor is N/n but factors that correct for nonresponse and for subsampling in the field are also included.
- In order to get out the report on time one must begin entering data as it arrives. I favor doing a field edit for completeness by the supervisory enumerator, and then an office edit whereby entries on the questionnaire are transferred to a tabulation sheet and then data is entered into the computer from the tabulation sheet. Just multiply by raising factors and add to get estimated totals and then take ratios, etc. for the final estimates. The Jackknife device can usually be applied to advantage with the replicated subsamples if complex estimates are used.

Pre-trip Briefing

As usual I went to the library. We found the book titled Bolivia by Olen Leonard and a report on agricultural development in the Chapare. Consuelo Arellano (former statistics student) loaned me a report written by Larry Szott (program leader in the Chapare) on agro forestry in the Chapare. I also looked into Steve Monteith's NCSU PhD thesis on soils of the Chapare. I found INE's (Bolivia's institute of Statistics) web page and got numbers of households for 1992 (in Villa Tunari=13101, Tiraque=6402 and Carrasco=7451). In e-mails from Larry Szott we learned the study area was 165 km by 45 km (=742,500 ha) with 100,000 Ha of crops and roughly 25000 families. They also Faxed copies of maps but the quality was dubious

I spent several days looking into spatial sampling. When Dr. Nelson and I gave a course in El Salvador we mentioned dot sampling from aerial photos (which I did routinely with NCSU students in sampling classes) but did not recommend point sampling. In March of 1998 I attended Ag 2000, a conference of ag statisticians where we saw that the Europeans were using point sampling. Ray Garibay of USDA reported on a trial in Nicaragua of point sampling using a one-operator segment with PPS to size of operation. Larry Szott also was planning to use satellite imagery, so we went over to visit with Hugh Devine, GIS expert in forestry at NCSU, to ask about such data. There we were encouraged to carry a geopositioning system (GIS). I even started to develop an image of the ideal point sampling method using hexagonal plots but it was not tried in the field.

Scope of work

As a consultant you seem to spend an inordinate amount of time defining your role before signing a contract. I have saved some 49 pages of copied e-mails. (Some say computers will tend to destroy historical records but in this case there is an excess of records.) Finally we settled on three tasks for the scope of work. (1) Review current survey methods and bring up to date the sampling scheme, (2) Assess data already gathered, and (3) Design a survey of zero-coca farmers. Then I signed away all rights to anything tangible I might produce, to N. C. State who seems so far to approve of my explaining these operational steps, and took off for the Chapare.

Whenever I am asked to design a sample survey I find it useful to request some kind of materials from the client that will keep them busy so you have time to nose around. In this case I had determined to ask that the study area be given a clean and clear outer boundary. However, Larry Szott and Eduardo Valerde (the computer expert in Bolivia) assured me that the list of sindicatos (farming communities) defines the study area and the map of them would necessarily have a somewhat jagged outline. When I asked for a diskette with the list it was supplied in minutes. So my ploy for delay was foiled and I plunged into selection of a sample using sindicatos as count units and 1992 population counts as sizes.

Sample Selections

Since my return to Raleigh I have copied the map and cut the copy into pages along kilometer divisions. It is now easier to work with. In Cochabamba I could only look at it, and wonder about the GIS (geographic information system software) that made it. I regret not learning more about their GIS but perhaps another time. The list on the diskette was in Excel and I easily put it, by columns, onto the clipboard and then into SAS. There was an identifying number (called CODOS) for each sindicato and a number of families as of 1992 (NUMERO) to serve as size measure. In reflex mode according to the above image of ideal area sampling practice, I applied the area sampling arithmetic. Some 377 of the 1239 sindicatos had zero or missing sizes and so we supplied 3 for all of them. (Homework exercise: Can you tell what effect this will have on bias and variance relative to a case of having the actual numbers? Ans: No bias but usually some slight increase in variance. Why?)

Since I made a programming mistake on this initial dry run it served to test the method and I will report on the current selections. The total of size measures was 44240 and with 10 families per SU (an optimum determined by some Smith's b values, for which see below in the section on dual frame design) this would imply N around 4424. Since we were planning to use around 80 or 90 SU's as sample size we rounded to N=4452 to allow 4 subsamples each of 21. Notice that the zone size is 4452/21=212. That is, the entire list of 4452 SU's is viewed as broken into paper zones each of 212 SU's. From each zone come 4 selections.

Our table of random numbers showed: 4)(7056 25632 59620 51225 26618. We needed start numbers in the range 1 to 212. Thus we used 3 digit numbers mod 500 and wasted any over 212. We also rejected any start within 5 of an already selected start. Please verify that the starts are: 205, 125, 132, 96, plus 26 and 118 for a fifth and sixth subsample. (For example, 705 mod (500) =

205. The selections for, for example, subsample #1 are 205, 417, 699 and so on to 4445.

In order to uncover SU #205 on the ground, we look at the list under CUMSU, the cumulative sampling units, to see where 205 appears. There is a 204 at CODOS=40705 and then a 215 at CODOS=10510. We thus know that SU #'s 205, 206, . . ., 215 are in CODOS=10510. The enumeration team must go there and define 11 segments, then number them on the sketch map and then randomly select one for interviewing. Before considering this delineating and re-randomizing operation let's look at a rather more complex case presented by SU number 4445. CODOS=30405 was 4444, and CODOS 40606, 40901 and 40903 all have 4445 before CODOS=40904 changes to 4446. The repetitions of 4445 tell us to combine the three sindicatos into a single unit which (since 4444 changes just to 4445) is just one SU.

Segmenting is easy when few segments have to be made and all end up having around 10 chacos. If more than 7 or 8 segments are required than it may be best to divide the count unit (on a sketch map of the sindicato) into parts, assign segments to the parts and randomly select a segment which will point to just one part and then segment in detail just that one part. If the selected segment has more than, say, 25 chacos then subdivide it into segments with closer to 10 chacos each and randomly select one for interviewing. Be sure to keep a record of this subsampling so as to modify the raising factors.

We visited the study area in the Chapare for only a day and a half, but it was enough time to verify that delineating segments and re-randomizing would be possible - - actually would be fairly simple. All sindicatos are laid out along roads with each chaco having 100 m, or 150 m or 200 m frontage with a distance of 1km into the forest. Thus segments can be clearly defined as a collection of more or less adjacent chacos. Re-randomizing must be done by a certified random device and from our experience in Africa we are recommending coin flips.

Review of Previous Year's Surveys

The design used previously was described as "two-stage' and "self-weighting'. In fact sindicatos were drawn with probabilities proportional to NUMERO (number of families from the 1992 Census) and 10 interviews were conducted at each selected sindicato. For example, with a total of 40000 for NUMERO and a sample size of n=120 sindicatos (or of 1200 chacos) the basic raising factor would be calculated as RF0=33.33333 . . ..if only 7 interviews were obtained then RF1=RF0(10/7)=47.619 would be used. If the current number of chacos were to differ from the 1992 number then the basic factor should be changed. If the 1992 number was 80 but there are 70 today then RF0 should be 33.3333(70/80)=29.167 rather than 33.3333. If only 7 were interviewed and not 10 the RF1=41.6667, no? We reported on these matters in a note called, "Reponderacion" (Reweighting).

As best we can determine from what we attempted to achieve by way of non-threatening conversations with Eduardo Valerde, who did the tabulations, the actual method for selecting the 10 chacos to be interviewed was done by judgement. Not all were of the same type and they were likely not all near neighbors. This kind of sampling is sometimes systematized as quota sampling. In the bulk sampling field analogous instructions may simply read: "Don't take all increments (scoops) from the same place."

Thus the design as it was implemented suffered from two defects. There was no correction for changes in numbers of families from 1992 to the present year and there was non-frame and non-random selection of chacos. Do you notice that this design could have been worked correctly in accord with the first of the two options for cluster formation that we described above? That, is, a current list of all occupied chacos could be constructed and every 4th or 5th or whatever could be made into the selected segment. When households have mailing addresses this kind of listing can be objective and clear, but when chacos are not easily individually identified, as actually appears to be the case in the Chapare, then the compact segment seems an easier type of unit to objectify. In fact the SU's in the new sample selection can be used for years to come if they are defined on two or four corners by the GPS.

Nonresponse Possibilities

Although we are critical of interviewing only at chacos whose farmers volunteer to answer, what will we do if there is a refusal? Both Eduardo and Larry said some sindicatos would not cooperate and we noticed from our trip to the Chapare that some chacos seemed deserted. Let's consider the following three cases separately. (1) If the field work is handled by INE as planned, then acceptance of that agency as a means of getting government financial help will allow entry to all the sindicatos and this was suggested by people from INE. (2) We will need to pretest the method, but we may be able to deal with the apparently deserted chacos by asking neighbors for upper bounds. What is the most pina they could have? This would be done only for the five alternative crops. (3) Refusals can be handled by the usual device of boosting the raising factors of the other chacos in the sampling unit.

The solutions in the first two cases are only pious hopes that will need to be pretested. We were able to furnish some analysis for the third case in a note called "El Asunto de Falta de Respuestas" (The Issue of Nonresponses). Data from earlier surveys showed that about a fourth of all land is in crops and a fourth of crop land is in the alternative crops. Thus we created a population with a responding stratum at 27% and a nonresponding stratum at 22% for the variable of interest. Let P_R (=.27) and P_{NR} (=.22) represent the two proportions and let π be the proportion of nonrespondents. The actual estimate will have expected value P_R but it should be $\pi P_{NR}+(1-\pi)P_R$. Sampling variance is about $P_R\,(1-P_R)/[(1-\pi)n]$, is it not?

Now when an estimate is unbiased, sampling variance is a good index of merit but when it is biased then mean square error (MSE) is the index of merit (check into the Cochran textbook (Sampling Techniques, Wiley, 1977) for similarities in confidence interval properties between SE and $\sqrt{}$MSE). Table 1 shows the two contributions to MSE for 12 cases of small, medium and large sample sizes with a range of four amounts of nonresponse.

For n=1000 the bias contribution is still less than variance, even with 20% nonresponse. We were assured that nonresponse would not exceed 15% and so we recommended that the survey go ahead. We might just mention parenthetically that the method of proof for these results on bias and variance is by manageable numerical example rather that by a statement of a theorem or by a simulation or by other means and this is rather a common state of affairs when working on one-of-a-kind surveys in such places.

Dual Frame Design

This design feature of also using a list frame with the area sample has become well established in U. S. Ag surveys. In Cochabamba we found out about lists of names of farmers generated by a U. S. sponsored program of payments for signing up for zero coca and additional lists of members of farmers organizations. In the U. S. the listed farmers have larger acreages and the same might be expected here but on a smaller amount of difference.

When we guessed at these differences and then calculated anticipated sampling variances (area alone versus dual frame) the savings were a modest 15%. The details are in another note, "Muestra de Dos Marcos" (Dual Frame Sampling). In the course of writing this note we received into our laptop the data of the 1998 survey and were able to judge optimum area segment size from some Smith's b values. For comparing the designs we needed to optimize sample sizes between the two frames. We can only hope that our conclusions agree with H. O. Hartley's original analysis.

The recommended design calls for n=160 from the list, which we determined to select in 16 clusters of 10 names per cluster. The list itself is grouped into 103 organizations and we ran the area sampling arithmetic to create integer numbers of clusters in each association. We then wrote SAS code to form the clusters of around 10 names each and printed them out. All is explained in the note. The SAS program shows the code but, as mentioned earlier, is a little difficult to verify.

Conclusions

We have mentioned the three notes and the written addresses as "deliverables" in the lingo of these programs. We also spewed forth a 4-

page report that referred to 12 such notes. Another 4 notes were included as well. Remember that our tasks included designing a survey of members of the farmers associations. Although we can be content with the consulting episode, I am far from satisfied with prospects for getting useful estimates from the actual survey.

In general, collecting data does not mix with regulating. In this case data on crops is mixed up with regulations against coca. We wish only to estimate hectares in alternative crops and do not want to know about hectares in coca. How can we communicate this to farmers? We will need to do extensive pretesting of the measurement operations as well as of the segment selection operations. Getting a solid sample is quite possible but getting the data seems illusive. I fear that the measurement problems might interfere with getting a solid sample. Enumerators tend to value getting data a good deal above objective placement of the sampling unit boundaries. Not all enumerators are susceptible to such temptations, but there seem always to be a few who are willing to stretch the rules. Here is where professionalism of the statistician's role is most helpful and we will be curious to see what will be INE's position in these issues. I am sorry I did not have a chance to meet with statisticians from INE but maybe we will get together eventually.

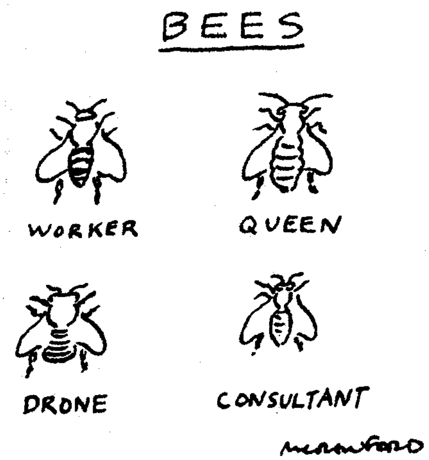

Table 1. Contributions to Mean Squared Error from Squared Bias (Sesgo) and from Variance (Variancia) for Differing Proportions of Nonresponse (10% to 40%) and Sample Sizes (200, 1000 and 2000) as Reported in a Note.

Contibuciones de sesgo y de variancia por diferentes tamanos de muestra y diferentes cantidades de falta de respuesta				
% de	n=200		n=1000	n=2000
Norespuesta	Sesgo^2	Variancia	Variancia	Variancia
10%	.000025	.00110	.00022	.00011
20%	.000100	.00123	.00025	.00012
30%	.000225	.00141	.00028	.00014
40%	.000400	.00164	.00033	.00016