VARIATIONS IN RESPONSE ERRORS INDUCED BY CHANGING INSTRUCTIONS TO ENUMERATORS

Charles H. Proctor, North Carolina State University

0. Abstract

This article reports on some research into response errors that was done on the Agricultural Enumerative Surveys in North Carolina. The data on response errors were collected by comparing farm information obtained by follow-up enumerators with that turned in by initial enumerators. These initial enumerators varied their interviewing procedures in ways that were suspected to produce varying amounts of discrepancy. Three kinds of discrepancies were distinguished and a probability model of the data collection operation is introduced. The model has three parameters: π_s-stability, π_v-vigilance, and α-scatter, to represent the three kinds of discrepancy.

We calculate relative mean square error as a function of π_s, π_v and α and then suggest cost functions to use in optimizing the levels of the three. This reasoning goes beyond our data in that the cost parameters have not yet been estimated. We then return to the data of the experiment-in-a-survey and indicate those variations in interviewing techniques which seemed to result in more favorable levels of the response error parameters. These, in short, are the official rather than the friendly approach and the field-also rather than the house-only location instruction.

1. Introduction

As large-scale surveys become more widely used, the response error portion of the variance is getting more attention in survey design. This is quite understandable since large sample sizes have reduced the uncertainty arising from sample selection, while the use of supplementary information, as in ratio estimates, has further cut down the sampling error. Thus, although response error or measurement error may have been only of minor importance in early stages of survey work or with small samples, now it has grown in relative prominence. This is true even in factual surveys such as agricultural enumerative surveys of crop acreages and production and of livestock numbers.

The present study, an experimental survey, grew out of the interest of the Standards and Research Division, Statistical Reporting Service, USDA and the North Carolina State Statistician's office who wanted to learn about how the data were gathered during the June and December Enumerative Surveys and how this affected its quality. Two of us at the Institute of Statistics, starting in 1961, talked with supervisors, accompanied enumerators and tried interviewing using the forms and instructions of these surveys. We gradually came to recognize an "optimum interviewing style". This may be defined as a course of action by the enumerator which would somehow squeeze the most information from the situation presented at each sample segment.1/

This style we found embodied in supervisory enumerators and other "successful" enumerators who were fully acquainted with the schedules and the importance of their work, and who were lucky enough to adopt just the right way of handling problems in conversation with farmers.2/ Such a subjective conception is, of course, too hazy for precise experimental control but we decided to use the results of supervisory enumerators as a standard to compare to those of other enumerators working under a variety of instructions. Thus in the discrepancies between the results of two enumerations of the same tract we planned to measure "response error", and here we met quite a problem.

At each tract two numbers were recorded.3/ One, to be written x, was obtained by the non-supervisory, initial enumerator under particular instructions and the other, to be written x', was found by the supervisory, follow-up enumerator under optimum interviewing style. The problem was to characterize these differences in some convenient, and hopefully, suggestive way so as to reflect the main components of response errors. The following "solution" may look technical, and it is, but it is also very close to the data - a grossly empirical approach.

1/The sampling unit is a land area for these surveys. This area is used either in the "closed segment" approach or the "open segment" way. Under the open segment definition all land operated by only those persons living in the sampled area is covered in the questionnaire, while under the closed segment approach all acreages and livestock on only the sampled area are covered. A tract is defined as that portion of a closed segment which is operated by one person.

2/Our findings at this stage were reported in Progress Report 31, of the Institute of Statistics and USDA. This report covered the work done from August 1961 to February 1962. These semi-annual mimeographed reports, which we will refer to only by number from now on, are available upon request to the Institute of Statistics, North Carolina State University, Raleigh, N. C.

3/In point of fact, two collections of "pencil patterns" were recorded and we in the office translated these to the numbers x and x'. We did not use in full the editing instructions for these surveys and omitted "don't know" responses from many of the comparisons. Progress Report No. 33 described our early attempts to tabulate data from initial and follow-up interviews.

2. Response Error Model

To be specific let's consider livestock numbers, e.g. "cattle and calves of all ages". In some cases the x and x' values were identical-a stable response. Where they were not equal, two types seemed worth distinguishing: those having $x = 0$ with $x' \neq 0$ (the reverse case $x \neq 0$ and $x' = 0$ was rare but present, although we will ignore it for now) and those having $x' > x$ and both non-zero (again the reverse case of $x > x'$ will be ignored for now since it was infrequently observed). When both x and x' were non-zero the variation in the difference was greater the greater the level of the x and x' values.

These three types of discrepancies can be represented by a random process with three parameters, π_s, π_v and α.4/ The following fanciful story shows the process. Suppose the farmer has some cows. When the enumerator starts to talk with the operator about a tract he takes out a short roofing nail, flips it and if it lands point up, he writes "no cows" without bothering to ask the farmer. If the nail lands on its side (which it does with "high" probability, π_v) he then gives the farm operator a thumbtack. The farm operator flips the thumbtack and if it lands point up (which it does with probability π_s) he tells the enumerator the "correct" number but if it lands on its side he selects a number r say in the range 0 to α (where α is something like .8) and multiplies the "correct" number by $1 - r$ and tells the enumerator this "shrunken" result.

An initial enumerator who operates with $\pi_v = 1.00$ and either $\pi_s = 1.00$ or $\alpha = 0$ (or both) will produce results as free from response error as a follow-up enumerator. Deviations of these parameters from their extreme values shows deterioration of interviewing. Using rather naive techniques of estimation and a sample of 108 tracts we estimated $\pi_v = .95$, $\pi_s = .60$ and $\alpha = .8$ for "cattle and calves of all ages on (this tract)".

3. Relative Mean Square Error

Having such a model of response errors one can calculate the influence of changes in the parameter values on, say, the mean square error of estimates and also design and conduct experiments to see how easily or painfully the parameter values can be improved. We began work in both of these directions and will describe the findings shortly. Another more pressing problem would seem to be to find alternative, possibly simpler and more realistic models. Although we feel that this is an important job we left it undone.

4/Progress Report 34 describes a more elaborate model of which the present one is a special case.

To calculate the mean square error of the mean of n x-values from farms with non-zero x' value we partition the square as follows:

$$(1)\quad E(\bar{x}-E(\bar{x}'))^2 = E[\bar{x}-\bar{x}') + (\bar{x}'-\mu')]^2$$
$$= E(\bar{x}-\bar{x}')^2 + 2E(\bar{x}-\bar{x}')(\bar{x}'-\mu') + E(\bar{x}'-\mu')^2 .$$

Here we are using $E(\bar{x}') = \mu'$. The last term is the sampling variance of the optimum interviewing value mean, which we may call σ^2/n. The quantity $\bar{x}-\bar{x}'$ can be written as $\frac{1}{n}\sum_{j=1}^{n}(x_j-x_j') = \frac{1}{n}\sum_{j=1}^{n} Z_j x_j'$ where the Z's are random quantities, sort of percent bias quantities, whose distributions are given by the response error model and its parameters. In evaluating $(\frac{1}{n}\sum_{j=1}^{n} Z_j x_j')^2$ we will require to calculate $E(Z_j)$ and $E(Z_j^2)$ when x_j' is non-zero.

Notice that $Z_j = -1$ if there is lack of vigilance (because $x_j = 0$ while $x_j' \neq 0$), $Z_j = 0$ if there is stability and vigilance and $Z_j = -r$ if there is vigilance but not stability. We took the distribution of r to be "parabolic" rather than uniform and in this case $E(r) = .40\alpha$ with $E(r^2) = .32\alpha^2$. That is, the frequency function for r was chosen as $f(r) = 1.5(r-2)^{1/2}\alpha^{-3/2}$ for r in the range 0 to α. We used the maximum observed value of r to estimate α.

$$(2)\quad E(Z_j) = 1(1-\pi_v) + 0(\pi_v\pi_s) + E(-r)\pi_v(1-\pi_s)$$
$$= [1 - .40\alpha(1-\pi_s)]\pi_v - 1 = B \text{ say.}$$

$$(3)\quad E(Z_j^2) = (1-\pi_v) + .32\alpha^2\pi_v(1-\pi_s) = C \text{ say.}$$

$$(4)\quad E(\bar{x}-\bar{x}')^2 = (\tfrac{1}{n})^2 E\left[\sum_{j=1}^{n} Z_j^2\, x_j'^2 + \sum_{i\neq j}\sum Z_i Z_j x_i'\, x_j'\right]$$
$$= (\tfrac{1}{n})^2[Cn(\sigma^2+\mu'^2) + B^2n(n-1)\mu'^2]$$
$$= \tfrac{1}{n}[C(\sigma^2+\mu'^2) + (n-1)B^2\mu'^2] .$$

The middle equals sign in (4) depends on the Z_j's and x_j''s being independent and doing the sampling with replacement - neither of which are true strictly, but actual cases will not be too far from this. By a similar computation the middle term of equation 1) becomes $2B\sigma^2/n$ so that the mean square error becomes $(C+2B)\sigma^2/n + C\mu'^2/n + B^2(n-1)\mu'^2/n + \sigma^2/n$ .5a/

5a/For the general case when only a proportion, P say, of the farms have non-zero x' values this mean square error is $P\left\{(C+2B+1)\frac{\sigma_+^2}{n} + (C-PB^2+2BQ+Q)\frac{\mu_+^2}{n} + PB^2\mu_+^2\right\}$ where σ_+^2 and μ_+ are the variance and mean of the non-zero x' values.

The relative mean square error appears when we divide this by μ'^2. Thus rel-MSE = $(C+2B)V^2/n + C/n + (n-1)B^2/n + V^2/n$, where V^2 is σ^2/μ'^2, the population rel-variance of the optimum interview results.

4. Cost and Worth Functions

The quantity B^2 so dominates the result, when n is even moderate in size, that in practice we would be concerned to reduce it first.5b/ The further B^2 is reduced the more valuable are the results, while the closer π_v and π_s are to 1.00 and α to zero the more costly, presumably, is the interviewing. If we could provide these cost functions then maximizing the value of the survey's worth would give us some notion of what level of vigilance, stability and scatter we should aspire to.

In our experience vigilance, π_v, was .95 and to reduce the .05 of non-vigilance remaining to .025, say, would seem to me to require doubling the cost. Getting $\pi_v = 0$ is almost free and to obtain $\pi_v = 1.00$ would be priceless. The function $\$.05/(1-\pi_v)$ which becomes very large as π_v approaches 1 and doubles from \$1 to \$2 as π_v goes from .950 to .975 may serve to represent this. A value more appropriate than \$.05, call it ν in general, may be found after experimentation, such as we will describe shortly, has been carried out. Similar cost functions can be designed to include π_s and α.

The worth of the estimate would increase upon decrease in B^2 but not without bound for values of B near zero. At the other extreme it could become negative, I suppose, if the results were dangerously misleading. A worth-per-interview function with some claim to applicability is $w_o - w|B|$ where w_o and w are appropriate values. If an interview in a survey which provides an unbiased estimate is valued at \$10 and one in a survey with a 5% bias is deemed worthless then w_o may be near \$10 and w near \$200.

5. Optimizing π_v

To illustrate the application of this representation consider determining the economic optimum level of vigilance when α and π_s are fixed, at α^* and π_s^* say. We wish to maximize, by judicious choice of π_v, worth per interview minus cost per interview for the survey or $(w_o-w|B|) - \nu/(1-\pi_v)$ or $(w_o-w+\gamma\pi_v-\nu(1-\pi_v)^{-1}$, where $\gamma = w[1-.4\alpha^*(1-\pi_s^*)]$. The answer by differentiation is to set $(1-\pi_v) = \sqrt{\nu/\gamma}$. This argues for increasing vigilance if it costs less (i.e., ν decreases) or if reduction in bias becomes deemed more worthwhile (i.e., w increases) or if bias from other sources is introduced (i.e., $\alpha^*(1-\pi_s^*)$ increases).

Plugging in the values suggested, namely $\nu = \$.05$, $w = \$200$, $\alpha^* = .8$ and $\pi_s^* = .60$, we find that π_v should be .95. This may only verify that we have chosen the parameter values to be reasonable in the light of actual practice. The net worth per interview in this case is -\$5.3, however and this does not sound too reasonable. It appears that bias is very large in this example. If vigilance drops to .90 then the net worth per interview becomes -\$33.5 while if it is increased to $\pi_v = .99$ then the net worth per interview is -\$21.3. Thus we see that the optimum is "sharp" to this extent.

6. Styles of Interviewing

Our speculations about cost functions are, we believe, of some interest and may argue for further work on the numerical values of the constants. The investigation which we did in fact carry out, was an experiment-within-a-survey designed to find out what effect changing the style of interviewing would have on the response error parameters.6/ We drew a sample of 144 tracts in a way so as to parallel the sample drawn for the December 1962 Enumerative Survey in North Carolina. Our sample was not used to make estimates of livestock and crops, although the data were collected in exactly the same way.

There were six different initial enumerators, two in each of three sub-regions of the state. An observer accompanied each initial enumerator; he timed sub-tasks of the interview and "coded utterances."7/ The observers' data were collected to allow us to see whether the various styles of interviewing differed greatly in time requirements. Except as we will mention below, they did not and thus we can concentrate directly on the variation in response errors by differences in style of interviewing.

Each initial enumerator was trained to interview in 8 different ways. These formed a 2^3 factorial arrangement. The three factors were called location instruction, respondent instruction and approach instruction. The two levels of location were fields-also and house-only. Under the fields-also instruction the enumerator asked the respondent if he might go to the fields as the data were collected on crops and livestock, while when he interviewed under the house-only instruction he carried on the conversation in the yard or house.

5b/If bias were removed by some kind of adjustment, say a subsample was reinterviewed and a correction made then one would focus on the quantity C.

6/Progress Report 34 carries the more detailed description of this work.

7/During the June 1962 Enumerative Survey we carried out an "Observational Study of Interviewing Technique..." which is discussed in Progress Report 32.

Variations in the respondent instruction were called *best* and *first*. Under *first*, the enumerator began asking the schedule questions as soon as he found himself talking to some one connected to the tract, while under *best* he was to determine, by each section of the questionnaire if it looked possible, who would be best informed and best able to answer and talk only with them.

The approach instruction had two alternatives, *friendly* and *official*. Under the *friendly* approach the enumerator looked at the respondent's eyes while asking the questions and down at his papers while the respondent answered. This was reversed under the *official* approach. Also under the *friendly* approach the enumerator moved closer and side-by-side versus more distance and head-on for the *official*. Extra topics of conversation were discouraged under the *official* but pursued more naturally under the *friendly* approach.

7. Effects on Behavior during the Interview

The effectiveness of these instructions in producing changes in the enumerators behavior is open to some question. For example, the interviewer may have attempted to ask questions of the housewife or young son when he was under the *first* instruction but in a majority of cases he would have been referred to the operator himself and ended up talking with the same person that he would have under the *best* instruction. However, we can tell by the increased number of don't know responses under the *first* respondent instruction, that there was some effect on the interview.

Although for six of the eight treatments, 25 to 29 minutes was the average time per interview there was a decided difference between the *official* and *friendly* approaches within the *fields-also* and *best* instructions. There the *official* approach "cost" only 22 minutes while the *friendly* approach cost 37 minutes per interview.8/ This shows some influence of the instruction and also it is important effect in that the *fields-also* with *best* is an attractive combination.

8. Effects on Response Errors

We examined the response errors on four kinds of items: field crops and livestock by the closed and open segment basis. We also distinguished, where applicable,9/ the three kinds of response errors. Thus we looked at many cross-classifications and, when dealing with scatter, at many analyses of variance. We found relatively few showing significance, and our chances of making a type I error by reporting them all is high. But their significance level is not the only reason that we think these relationships should be considered as suggestive. Thus we'll report on all of them and try to guess which ones may hold up in future survey experience.

The respondent instruction was practically never shown to be a source of change in level of response error. There was a suggestion that *first* responses were more stable than *best* responses on hog items for the entire farm (open segment). The data showed that of 17 unstable responses 13 were made under the *best* instruction, while the 34 stable responses were split evenly. This result is so contrary to what would be expected that it cannot be taken too seriously. There, however, is a suggestion that the respondent who is selected under the *first* instruction may turn out to be the "best" informant on hog items; that is, the farm operator may not know the hog enterprise as well as he does the rest of the farm.

The location instruction, *fields-also*, reduced response error in a number of places relative to the *house-only* instruction. *Fields-also* seems to cause higher stability, larger π_s, for acreages of "other land" in the tract. This category includes house lot, "woodland not pastured, swamp, pond, idle land...". Of 47 *fields-also* interviews, 37 were stable while of 45 *house-only* interviews only 24 were stable on the other land acreages. It is reasonable to expect that looking at the land itself would enable the interviewer to better classify the areas in accord with the schedule instructions and particularly when the category is as broad as this one. In practice the reported acreage under "other land" suffers heavy office editing so perhaps improvement here is not worth attempting to get.

The number of tracts with non-zero hog numbers, 28, was too small to allow us to estimate all three response-error parameters so we used a combined index of discrepancy, namely the difference in the two reports divided by the largest number reported. The *fields-also*

8/Table 4 of Progress Report 34 contains these data and also shows the breakdown of time by subtasks within the interview. This shows that the extra time was largely taken up in "introductory conversation" rather than in "other topics". Other topics did show a slight increase under the *friendly* approach, however.

9/The final formulation of the response-error model came while we were doing this analysis so that much of our work would have to be redone if we were to be faithful to the current model. We were using what were called D-scores as measures of response errors. We distinguished "stable", "mid-scores" and "high" D-scores and this corresponds to the stability, scatter and vigilance breakdown of the current representation. See Progress Reports 33 and 34.

instruction had a smaller mean value for this index than did house-only. The fields-also instruction seems to bring out more vigilance and higher stability on the principal crop item. This question asks for the principal crop and its acreage for the entire farm, an open segment definition. In both of these cases the fields-also effect may be present and desirable but we do not find any compelling reason for it. There may be a generalized effect from asking to see the fields that puts the respondent in a more serious frame of mind.

In general on the crop items the official approach showed up as more beneficial than the friendly. Data were collected on 584 fields and 251 of these were done under the official approach. Of these 251, there were 13 of them with a different crop reported initially than at the follow-up enumeration. Of the 333 fields done under the friendly approach 33 were reported as a different crop. This kind of error is of the vigilance type. Under the official approach 27% of the reported yields per acre (on 112 fields) were stable while this was only 16% under the friendly approach (used on 134 fields). There was also some evidence, based on the index of discrepancy mentioned above, that the official approach is superior to the friendly on hog items but inferior on cattle items. Thus while the picture is mixed for livestock items the official approach seems to work on the crop items.

In the course of searching for effects on response errors we uncovered some apparent interactions among the instructions. There was more stability in acreages by tenure status when best was paired with fields-also and first with house-only than when best went with house-only and first with fields-also. The latter two combinations would be expected to be more awkward to handle and to cause a loss of rapport, so that when the questions on land rented out and so on arose they could have more easily been misunderstood.