Respondent-Generated Intervals In An HMO Survey: Preliminary Results

by[1]
Diane Miller and S. James Press
University of California, Riverside

Key Words: survey, response error, respondent-generated intervals, bounds, ranges, Bayesian.

ABSTRACT AND SUMMARY

We describe a factorial experiment in the use of the RGI protocol for asking questions involving recall of answers to factual questions in sample surveys. The experiment is designed is ask the same questions of different groups of survey respondents in different ways to compare the results regarding non-sampling error and response rate. The experiment is embedded within a sample survey to investigate quality of health care.

1. Introduction

We are concerned with a new method for asking factual recall questions in sample surveys. The method is called Respondent-Generated Intervals (RGI). The paper presents the design and preliminary results of an experiment embedded within a survey in the context of exploring quality of health care at a health maintenance organization (HMO). This research is part of a Ph.D. dissertation in statistics at the University of California, Riverside. It involves a collaborative effort with an HMO to collect appropriate data for which the true answers to relevant questions about health care are known and provided through record checks. We are concerned with using the RGI protocol for asking questions to improve accuracy of estimation of population means by reducing non-sampling error, and with increasing response rates. Our objectives in this study are:

1) to be able to compare answers given by HMO members with their true values to establish empirically what the best estimators are for this type of factual data;

2) to help the HMO to make more informed management decisions by providing a better way to elicit information, by increasing estimation accuracy at an aggregate level, and thereby increasing the quality of the information that the survey provides about quality of health care.

2. Background

Press (2000) proposed a method for asking recall questions in sample surveys called RGI, for Respondent-Generated Intervals. For recall-type questions such as "How often have you visited the doctor in the last six months?" that require a quantitative answer, he proposed that the respondents not only answer the question (give a "usage quantity"), but also give values that bracket the answer. That is, the respondent would be asked to provide a number reflecting how many doctors' visits that s/he believed was almost certainly the smallest it could be, and also a number that s/he believed was almost certainly the largest it could be. In some situations, the intervals may be easier for the respondent to construct than the standard simple point quantitative response. The cognitive processes involved in generating the intervals may aid in recall and may seem less threatening to answer than are point estimates in the case of sensitive questions. If such an assumption is correct, it is reasonable to expect a higher response rate for RGI-style sensitive questions.

While earlier research is promising in addressing both the accuracy and response rate issues, several questions still need to be answered within this new method.

Q1 -- Is asking for the point estimate as well as the bounds necessary? That is, is it reasonable to ask for just the bounds?

[1] We are grateful to Drs. Diana Petitti and Valerie Crooks, and to Vicki Chiu for their help and support in providing data required for this study, and to Dr. Judith M. Tanur for helping to design the study.

Q2 -- Would the response rate when asking for the bounds alone be higher than the rate when asking for a point estimate alone?

Q3 -- Does the response rate depend on the sensitivity of the question posed? That is, are respondents more likely to respond to a sensitive question with RGI?

Q4 -- Does the sequencing of the questions matter? Is the response rate the same if the order of question styles is reversed?

Q5 -- Which of the two methods is perceived as easier to answer by the respondent? Does the perceived ease depend on the type of question asked?

Q6 -- When given a choice between these options for question style, will respondents choose to answer the interval more readily than the point estimate? Does type of question matter in this choice?

We have carried out a survey in which we have embedded an experiment designed to answer these questions. We plan to ask questions involving recall about health care issues for which we can determine the "true" values from record checks. Knowing the "true" value is crucial to establishing how accurate each of the estimated population parameters will be in comparing differing question styles. We plan to address these various questions by giving different versions of our questionnaire to five different groups. A complete explanation of the groups proposed and the differing question styles can be found in Section 3.

In traditional survey methodology, the point estimate for some usage quantity is usually the sample mean, and the interval is the 95% confidence interval based upon this average. Press (2000) proposed two new point estimators and two new interval estimators based upon the RGI. The first of these point estimators is the average of the midpoints of the bounds given by each individual. This is called the midpoint estimator. The second point estimator uses Bayesian hierarchical modeling, starting with the average midpoint as a starting point and adjusting it based on certain model assumptions (see Press, 2000). The first type of new interval estimator is known as the ARGI (Average Respondent-Generated Interval). This estimator averages the lower bounds across all survey respondents and uses this value as the lower bound of the interval. Similarly, the average across respondents of the upper bounds serves as the upper limit of the interval. The second of the interval estimators enlarges the interval formed by the averages of the bounds by adding one standard deviation of the respondents' upper bounds to the upper limit of the interval and subtracting one standard deviation of the respondents' lower bounds from the lower limit. This is referred to as the ARGI (1σ). We will compare the statistics developed using the RGI with the results obtained through standard survey methods.

Press (2000) developed and empirically tested these estimators on data obtained from a survey of students at the University of California at Riverside (UCR). For those students who gave permission, school records were used to compare the answers they provided with the true values obtained from university records. Accuracy of this format of questionnaire can be compared to traditional survey methods.

In the UCR survey one of the questions was: "At the beginning of this quarter, how many credits had you earned?" The 129 students were also asked to provide an upper and lower bound to this quantity. We found that the sample mean is the most biased (least accurate) of all the point estimates generated (i.e., it is the furthest away from the true value) and the midpoint estimator was the best. We also saw that the confidence interval based on the sample mean does not actually include the true value for the number of credits earned while the other interval estimators enclosed the true population mean.

Judith Tanur carried out a similar study involving a survey of undergraduates at the State University of New York at Stony Brook (SUNY/SB). This second survey was designed to parallel many of the same questions used in the UCR survey. Again, true values were obtained so that the accuracy of the new format can be assessed. The two campus surveys included a total of 18 questions. The first striking result was that the sample mean was the least accurate of the estimates generated for 10 out of the 18 questions. The interval results were similar; the 95% confidence intervals covered the true values for the usage quantities only 9 times; the ARGI interval covered the true values

15 times; and the ARGI (1σ) interval covered the true values 17 out of 18 times. Therefore, the method using RGIs seems an improvement over the current method of relying on sample means. The two campus surveys are summarized in Press and Tanur, 2000a. Press and Marquis, 2002a and 2002b, who used the RGI protocol in a survey of income carried out by the US Census Bureau, conducted additional research.

3. Design of the HMO Survey

3.1 Treatment Groups

We extended the concepts of the earlier research to embed an experimental design within a survey as discussed by Fienberg and Tanur (1988). We used this technique to answer the above questions by randomly splitting respondents into seven different treatment groups:

Group I consists of people answering just a point estimate. This serves as the first control group. We generate our sample means and 95% confidence intervals in the standard manner.

Group II answers with bounds only, and not the point estimate. We can find the "midpoint estimator" by finding the midpoint between the average lower bound and the average upper bound. We can explore taking weighted averages of the bounds.

Group III asks for both bounds and point estimates with the point option presented first. This provides the most information about the respondent's knowledge and certainty of the question asked.

Group IV asks for both types of information also, but with the bounds requested first. A comparison with the Group III will allow us to examine sequencing effects.

Group V asks the respondents to choose whether to provide an interval or a point estimate, offering the point estimate option first. Thus, we can ascertain if certain types of questions lend themselves to being answered by a point estimate or by an interval.

Group VI reverses the procedure followed in Group V, giving the respondents a choice between question forms but offering the interval estimate option first.

Group VII has the respondents answer in one of several intervals pre-assigned by the questionnaire designer, instead of intervals generated by the respondent. This is the second of the control groups (Group I was the first). For example, "How long have you belonged to this HMO? (_____5 years; _____5-6 years; _____6-7 years; etc.).

3.2 Preliminary Classroom Experiment

We decided that it would be prudent to first try out our design in a preliminary way in a classroom environment. We presented these seven groups to several classes of statistics students at the University of California, Riverside to pilot study the procedure and look at preliminary results. We asked the students two substantive questions: (1) how many points have you earned so far this quarter on all midterms taken in this class, and (2) how many points have you earned on all quizzes and homework assignments combined? The first question was deemed to be fairly easy as it should be 1 or 2 quantities to recall accurately and should be salient to the students. The second question was deemed to be very complex as it is the sum of many small quantities and somewhat less salient. We have small responses rates within each of the seven groups because no class was over 150 students and we have seven questionnaire groups.

From the second question on quizzes, we asked the respondents in groups requiring both question constructs which of the types of data they were required to provide did they believe was more accurate, which was less personal (threatening), and which was easier? From the overall sample of 175 we have the following data.

	Point (usage)	*Interval (bounds)*
More Accurate?	5 1	116
Less Personal?	95	84
Easier?	59	101

The results above show that the respondents believed that the bounds were more accurate, the single point usage value was less personal, and the bounds were easier to provide.

For groups requiring a choice between question constructs we asked the students which option they chose. We then asked whether their choice was selected because they felt it would be more accurate, less personal, or easier. This table is from the 140 students in those two groups. In general, the students felt that the answers provided were easier and more accurate.

	YES	*NO*
More Accurate?	111	12
Less Personal?	16	75
Easier?	96	16

3.3 HMO Experiment

We next extended the classroom experiment to the world of examining quality of health care provided by an HMO. We decided upon a mail questionnaire. We would have preferred to use the 7 groups described above but because of limited resources, only a total of 3000 questionnaires could be mailed. We decided, based upon expected response rates, to limit our groups to five; we dropped the two groups which required both point and bounds constructs (Groups III and IV).

We worked interactively in designing the questionnaire to focus on areas of concern to the HMO. It was decided to ask quantitative questions whose answers could be checked against patient records. We used the following questions:

- Date of last blood test to measure cholesterol levels
- Level of most recent total cholesterol
- Date of most recent child born in the HMO hospital
- Birth weight of this child
- Date of most recent pap smear
- Date of most recent mammogram
- Date of most recent influenza vaccination
- Length of continuous membership within the HMO.

We see from the above list that some questions are gender-specific, and some are age-specific. (It was decided that because males have a lower response rate, we could lower our costs by using only females; moreover, younger females have a lower response rate than older females, so an age delineation also seemed appropriate.) The age is classified into 2 groups: 25 to 50 years old and 50 years old and older. Therefore, within each of the five groups described above, we have 2 subgroups: Younger Females and Older Females for a total of 10 different questionnaires.

One of our primary interests was the investigation of the estimation in the context of sensitive questions. Such questions could include: weight; number of cigarettes smoked per week; number of sexual partners; etc. We were hampered by our need to compare what we found with the true value. While weight is probably measured at every doctor visit, the HMO computer system is not updated due to constant fluctuations in weight measurements. Thus, with the possible exception of total serum cholesterol level (it is desirable to have a reasonably low level), no other sensitive question with resultant truth could be asked.

The specific protocol followed was that a mail survey should be conducted with a postcard follow-up. We wished to examine the HMO members who have had at least five years of continuous membership. This allows time for the procedures that can be verified to have happened within the members' enrollment. The HMO has the ability to subdivide its membership into Younger Females and Older Females. From each of these subdivisions (younger and older females), we took a random sample of 300 drawn for each of the five groups of the experiment. Thus, we sent 300 questionnaires for each of the 10 groups for a total of 3000 survey questionnaires.

4. Response Rate Results

We have just begun to analyze the data obtained from the HMO survey and the truth data obtained from the HMO. We want to answer the questions that drove the design of the survey as well as to explore other hypotheses that arose in regard to collecting RGI protocol data. We have begun analysis with examination of response rates, which is all we are able to report on in this work.

Preliminary results on this data set reveal the following response numbers per group, per stratum.

	Young	Older	Totals
Usage Only	81	155	236
Bounds Only	68	103	171
Choice of *Usage Only* or *Bounds Only* (so ordered)	76	122	115
Choice of *Bounds Only* or *Usage Only* (so ordered)	66	115	181
Pre-Assigned Intervals	103	161	264
Totals	394	656	1050

We see that there is a substantial difference between the response rates of the two age strata. Younger women have a 26.3% response rate (394/1500) and older women have a 43.7% (656/1500) response rate.

We look at the results from the two-way ANOVA and the factor mean plots in Figure 1. One factor is "age strata" and the other factor is "questionnaire version" or "experimental group"; the dependent variable is response rate.

Analysis of Variance for Response Rate

```
Source      DF        SS        MS        F        P
Age          1   0.07627   0.07627    64.36    0.001
Version      4   0.03388   0.00847     7.15    0.041
Error        4   0.00474   0.00118
Total        9   0.11489
```

```
                  Individual 95% CI
Age      Mean  ---------+---------+---------+---------+--
old      0.437                    (-----*------)
young    0.263  (------*-----)
               ---------+---------+---------+---------+--
                    0.280     0.350     0.420     0.490
```

```
                  Individual 95% CI
Version  Mean  ---------+---------+---------+---------+--
1        0.393            (--------*---------)
2        0.285  (---------*--------)
5        0.330       (---------*---------)
6        0.302    (---------*---------)
7        0.440                      (---------*---------)
               ---------+---------+---------+---------+--
                    0.280     0.350     0.420     0.490
```

Figure 1: ANOVA RESULTS FOR RESPONSE RATES

We see that age results are as expected. The group whose questionnaires provide pre-assigned intervals has the largest response rate, with 44% on average, and it is larger than that of Group II or Group VI. We see that sequencing between choice styles generates no real differences in response rates, and our "bounds-only" version, Group II, isn't different from any questionnaire format but from the pre-assigned interval version (VII).

We also found some low response frequencies in specific categories. This is because not all women have had all the procedures or events asked about. One question asked of all women that has nearly universal response is when they had their last pap smear. We used that question to prompt for more information on the choices they made and why. When we examined the choice for the constructs, we see the following results for the Groups V and VI in both strata.

Response Rates

	Point	Bounds
Young, V (choice-usage first)	52	19
Young, VI (choice-bounds first)	38	22
Older, V (choice-usage first)	82	17
Older, VI (choice-usage first)	73	25

We see that the majority of patients chose to give the usage point estimate instead of the bounds. We can view the above as a factorial design with the factors of interest being whether the respondent chose point or bounds, the sequence these items are presented, and the age stratification. Analysis has shown that the response rates do not rely on the sequence presented and the age strata.

Similar to the classroom experiment, we want to know about perceived accuracy and ease. We examined the reasons why the choice was made and found the following results.

	Young, P→B	Young, B→P	Older, P→B	Older, B→P
More Accurate				
Yes	60	49	72	84
No	3	0	5	0
More Private				
Yes	3	1	4	5
No	21	18	21	29
Easier				
Yes	30	21	35	35
No	9	7	12	13

We see that the reasons respondents made the choices they made are due to the belief that it is more accurate and easier; not due to whether the number(s) are less threatening or more private.

5. Conclusions

While it is still too early to draw any conclusions about implications from this RGI survey regarding accuracy, the survey does suggest that most respondents prefer to answer questions for which pre-assigned intervals are provided, and least like to answer questions for which they must provide bounds.

References

Blair, E. A., and Burton, S. (1987). "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions". *Journal of Consumer Research.* **14**: 280-288.

Bradburn, N. M., Rips, L. J., and Shevell, S. K. (1987). "Answering Autobiographical Questions: The Impact of Memory and Inference On Surveys". *Science.* **236**: 157-161.

Burton, S., and Blair, E. (1991). "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys". *Public Opinion Quarterly.* **55**: 50-79.

Fienberg, S. E., and Tanur, J. M. (1988). "From the Inside Out and the Outside In: Combining Experimental and Sample Structures". *Canadian Journal of Statistics.* **161**: 135-151.

Kennickell, A. B. (1997). "Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances". Manuscript available on the Internet.

Menon, G. (1994). "Judgments of Behavioral Frequencies: Memory Search And Retrieval Strategies". In Norbert Schwarz and Seymour Sudman, (Eds.) *Autobiographical Memory and the Validity of Retrospective Reports.* (pp. 161-172). New York: Springer-Verlag.

Press, S. James (2000). "Respondent-Generated Intervals for Recall in Sample Surveys"; manuscript submitted

Press, S. James, and Marquis, Kent H. (2002a) "Bayesian Estimation in a U.S. Government Survey of Income Recall Using Respondent-Generated Intervals", *Proceedings of Conference of International Society of Bayesian Analysis*, Crete, Eurostat.

Press, S. James, and Marquis, Kent H. (2002b) "Bayesian Estimation in a U.S. Census Bureau Survey of Income Recall Using Respondent-Generated Intervals", *Research in Official Statistics*, Eurostat.

Press, S. J., and Tanur, J. M. (2000). "Experimenting with Respondent-Generated Intervals in Sample Surveys", with discussions; in Survey Research at the Intersection of Statistics and Cognitive Psychology. Monroe G/ Sirken, Editor, National Center for Health Statistics, Center for Disease Control and Prevention. Working Paper Series #28, pp. 1-18.

Press, S. J., and Tanur, J. M. (2001). "Respondent-Generated Interval Estimation to Reduce Item-Nonresponse"; in Applied Statistical Science V, Nova Science Publishers, edited by M. Ahsanullah, J. Kennyon, and Sarkar, S. K.