SOME EXPERIMENTAL TESTS OF THE RESPONDENT-GENERATED INTERVALS APPROACH TO SAMPLE SURVEYS[1]

S. James Press, University of California at Riverside and Judith M. Tanur, State University of New York at Stony Brook
Judith M. Tanur, Sociology Department, State University of New York at Stony Brook, Stony Brook, NY 11794-4356

Key Words: Bayes Estimation, Brackets, Cognitive Science, Hierarchical Model, Range Information, Recall, Response Bias, Survey.

1. INTRODUCTION

Consider a factual question that requires a respondent (R) to recall a numerical amount, perhaps the amount or frequency of a behavior. Assume that R is sincerely trying to recall the quantity accurately, not attempting to deceive, and actually has information about the answer. We refer to the quantitative item being requested as *usage*. R must render a subjective assessment of his/her usage. It is common in surveys to have Rs answer within a preassigned interval rather than specifically, especially for sensitive questions. The current procedure is different, however, in that the interval is not preset but generated by R (Respondent Generated Interval, RGI), and used not to protect R's sensibilities, but to provide a tool for estimation.

The RGI protocol for questionnaire design has its origins in Bayesian assessment procedures wherein an entire prior distribution for an individual is assessed by connecting a collection of points on the individual's subjective probability distribution via a sequence of elicitation questions (see, e.g., Schlaifer, 1959, Ch. 6; Hogarth, 1980, App. B and C, and Press, 1989, Ch.IV).

R's degree of belief (subjective probability) about the correctness of a factual recalled quantity is characterized by an entire probability distribution for R, not just a single point. For example, R may have a normal subjective probability distribution for the number of doctor visits in the last year, say N(4,1), so that s/he believes that it is most likely that he/she visited 4 times last year (modal value), with a standard deviation of 1. But usually we ask for just one point on this distribution. Perhaps we can improve the measurement of R's recall by measuring several points on his/her recall distribution. It would be ideal to ask R many additional questions about his/her recall of the usage quantity to assess many points on his/her recall distribution. But respondent burden, the cost of added interviewer time, etc., argue against many additional questions. We therefore describe a procedure that involves adding just two bounds-questions, and we examine the possible benefits accruing from such an addition.

Other survey procedures request that Rs provide bounds information under certain circumstances. Usually, they ask Rs to select their response from among several (analyst-generated) pre-assigned intervals (sometimes called brackets). Kennickell (1997), however, described the 1995 Survey of Consumer Finances (SCF), carried out by NORC, as including opportunities for the Rs who answered either "don't know" or "refuse" to select from eight pre-assigned ranges or to provide their own upper and lower bounds ("volunteered ranges"). Another related technique that has been proposed is called "unfolding brackets" (Heeringa, Hill and Howard, 1995). Here Rs are asked a sequence of yes/no bracketing questions that successively narrow the range in which the R's true value might lie.

2. POINT ESTIMATION OF THE POPULATION MEAN

We present several possible point estimators of the population mean for a question using the RGI protocol. Which estimator is best, in which circumstances, is an empirical issue addressed below. Let X_i denote the reported usage quantity for R_i in the survey, and let a_i and b_i denote the lower and upper bounds given by R_i, $i = 1,...,n$. Traditional point estimators for the population mean are the sample mean, $\overline{X}$, and the sample median, X_{med}. While these estimators are simple, quickly calculated, and have many well-known useful properties, they don't take advantage of the additional bounds information provided by RGI that might help reduce bias.

[1] The authors are especially grateful to Dr. Kent Marquis, a co-author on the reports of the Census experiment, for helpful discussions and suggestions. They are also grateful for the technical assistance rendered by several research assistants at the University of California at Riverside, at the State University of New York at Stony Brook, and at the US Bureau of the Census.

2.1 Weighted Average Estimation

We can generate a family of point estimates of the population mean by using a weighted average of the assessed average bounding endpoints, $\bar{a}$ *and* $\bar{b}$, where

$$\bar{a} \equiv \frac{1}{n}\sum_i a_i, \quad \bar{b} \equiv \frac{1}{n}\sum_i b_i \ .$$

Such a weighted average point estimator of the population mean is given by:

$$\bar{m} = \omega\bar{a} + (1-\omega)\bar{b}, \quad 0 \le \omega \le 1.$$

But how should we select ω? If ω and $(1-\omega)$ are each taken to be ½, we have an equal weighting, and $\bar{m}$ becomes a *midpoint estimator* $= (\bar{a} + \bar{b})/2$.

To account for respondent error in assessing the bounds, we could calculate the standard deviations of the bounds and use weights that are proportions of total variances, a common practice (see, e.g., Kish, 1965, p. 432). We could also express the weights in terms of the *precisions* of the average bounds, that is, the reciprocals of the variances of the average bounds. An unequally weighted estimator would be sensible in situations in which we expect there might be substantial discrepancies between the uncertainties in assessing a_i versus those in assessing b_i.

2.2 Bayesian Point Estimation

A vague information model might adopt the assumption that the usage quantities are uniformly distributed within the bounds. The analysis for this ideal model is not addressed here because its result cannot be implemented numerically at this time. Instead we consider a modified model where we suppose it is reasonable to assume that the sampling model for X_i is normal, so that

$$(X_i \mid \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2),$$

and that the X_i are mutually independent. We define the n-vector $\theta = (\theta_i)$, and the (nxn) diagonal matrix

$$D_{\sigma^2} \equiv \text{diag}\,(\sigma_1^2, ..., \sigma_n^2).$$

Then, in more compact form, we have, for $\mathbf{X}=(x_1,\ldots,x_n)$,

$$\mathbf{X} \mid \theta, \boldsymbol{D}_{\sigma^2} \sim N(\theta, \boldsymbol{D}_{\sigma^2}).$$

We refer to the distribution of X_i as the recall distribution for survey respondent i. Note that the mean usage quantity, θ_i, is generally different for each R, as is the variation, or uncertainty in reported usage, σ_i^2. We assume for the moment that the recall variances, the σ_i^2, are all *known* (they are estimated by $(b_i-a_i)/6$, as most of the mass is included within 3 standard deviations of the mean in a normal distribution, and the sample size is assumed large), but the θ_i are *unknown*. Adopt the exchangeable prior distributions, for i = 1,...,n,

$$(\theta_i \mid \theta_0, \tau^2) \sim N(\theta_0, \tau^2),$$

with $h = \tau^{-2}$. We assume independence, a priori, of θ_0 and h. Adopt the higher order prior distributions:

$$p(\theta_0) \propto \text{ constant},$$

$$\text{and } p(h) \propto h^{\alpha-1}\exp(-\beta h), \ \alpha > 0, \ \beta > 0.$$

Here α *and* β are hyperparameters that must be assessed. Combining densities and using Bayes' theorem, yields the joint posterior density

$$p(\theta_0 \mid \mathbf{x}) = \frac{\iint p(\mathbf{x}\mid\theta)\,p(\theta\mid\theta_0,h)\,p(h)\,p(\theta_0)\,dh\,d\theta}{\iiint p(\mathbf{x}\mid\theta)\,p(\theta\mid\theta_0,h)\,p(h)\,p(\theta_0)\,dh\,d\theta\,d\theta_0}$$

Note that the denominator integral is just a numerical constant (depending upon the data and the pre-assigned hyperparameters). The Bayes estimator of the population mean is taken to be the posterior mean:

$$\bar{\theta}_0^{\,*} = E[\theta_0 \mid \mathbf{x}] = \int \theta_0\, p(\theta_0 \mid \mathbf{x})\, d\theta_0.$$

All integrals over the components of θ are taken over the same ranges, namely $(-\infty, +\infty)$; integrals over h are taken over $(0, \infty)$. The numerical evaluation of this Bayes estimator is effected by means of the Gibbs sampler. (See Evans and Swartz, 1995, for related discussions about evaluating such ratios of integrals.) The one-dimensional conditional densities of all of the variables are readily obtained by conditioning in the joint density. Because we know that the joint density exists (we know it explicitly up to proportionality constants), this is all we need to be able to apply the Gibbs sampler. To do so, we used the WinBUGS program, Version 1.2

(see Spiegelhalter et al., 1999). Conveniently, it is not necessary to work out the joint or conditional distributions in any particular format, but only necessary to input the three stages of distributions using the proper WinBugs syntax.

A variety of Bayesian estimators may be generated by using various types of assessments for the hyperparameters, α and β. By using values for these parameters obtained from a previous survey, we can generate a true Bayesian estimate; if we assess these hyperparameters from sample values obtained in the current survey, we generate an empirical Bayes estimate. Bayesian estimators are likely to have smaller associated credibility intervals than those found by confidence interval estimation, and they are likely to be more accurate than traditional estimators because they account directly for prior information about the population mean. These empirical questions are addressed below.

2.3 Point Estimation and Multiple Imputation

It might be of interest to impute missing usage information from respondents who do not provide usage information, but provide bounds information. In such cases, we could impute the missing usage data and use the bounds information to constrain the multiple imputation (Kennickell, personal communication). We have used the midpoint estimator to impute values for item nonresponse (without use of multiple imputation). The results, discussed below, seem very promising.

3. INTERVAL ESTIMATION

We evaluate several competing interval estimators in the RGI context: traditional confidence intervals, Bayesian credibility intervals, and the ARGI, or Average Respondent-Generated Intervals, $[\overline{a}, \overline{b}]$. Because the extremes of respondent belief are reflected in the intervals provided by RGI respondents, the ARGI will typically cover the true population values (minimizing response bias), while the other two interval estimators are less likely to cover the true population values. We have found these expectations to hold in the experimental studies described below.

4. THE EXPERIMENTS

Several empirical studies were designed to examine the functioning of RGI under a broad range of conditions. Questions that we have attempted to address in our experiments are:

a) How would RGI work with sensitive questions, such as "income"?
b) Does the ordering of the basic usage and bounds questions matter?
c) Does an option to choose between usage and bounds questions affect response rate?
d) Can wording of the bounds questions be improved to aid respondents' understanding?
e) Can we ask only for bounds information (RGI) without asking for basic usage and still improve accuracy and response rate?

4.1 Student Surveys

At each of our campuses we carried out a paper-and-pencil survey, asking students about quantitative aspects of their campus life that could be verified by appropriate campus offices.

The usage question was always asked before the bounds question. The form of the bounds question was "Please fill in the blanks – There is almost no chance that the number of credits I earned by the beginning of this quarter was less than _____, and almost no chance that it was more than _____." Sample sizes for many questions were reduced because (1) we used half the sample to test another form of the bounds question, since discarded and (2) not all students gave permission to access their records for verification. For four questions about fees, however, all respondents were asked the RGI format and no individual verification was needed as these fees are uniform across students. Thus we have much bigger sample sizes for these questions.

4.2 The Census Experiment

This experiment was designed to vary the order of asking the bounds and usage questions, test if the RGI proceedure can be used in a telephone interview, test its usefulness for sensitive questions, and broaden our universe of Rs past college students.

This experiment used extensive cognitive pretesting for the form of the interval question. In a split-panel experiment 75 percent of the Rs were asked the two bounds questions first, followed by the usage question and 25% received the reverse ordering.

From a frame of households that filed joint tax returns having wage and salary income for the previous five consecutive years, a sample of about 2000 households was drawn. From this sample the Census Bureau obtained a quota of 500 CATI interviews. Rs answered questions about their income from salary/wages and from interest/dividends for the past two years, and about

the change in both types of income over the previous five years. Since the frame information also included data from administrative records about household income, we could verify the responses.

4.3 The HMO experiment

A new experiment is being fielded in order to test if Rs are willing to answer the bounds question without being offered the usage question, and to explore which option they choose if permitted a choice between the bounds and usage questions.

Mail questionnaires will go to 3000 female members of an HMO asking questions for which the answers can be verified from the HMO files. There will be five groups of Rs: a control group asked the usage quantity only, another control group asked the questions in the form currently used by the HMO (respondents classify themselves into one of several predetermined interval options), a group that will receive only the bounds questions, and two groups that will be offered a choice of answering the bounds or the interval question (bounds offered first to one group, second to the other).

5. RESULTS

5.1 Accuracy of point estimates

We calculated the signed percentage of the true values that the errors (deviations from truth) constitute. These percents vary widely, from less than 1 percent to close to 500 percent. We found that several quantities seem to be estimated very badly regardless of estimation method used, notably both the change variables in the Census experiment, especially when the usage question is asked before the bounds question. We speculate that respondents find it difficult to report these changes, as they must not only recall two quantities but also carry out a calculation, a process fraught with opportunities for error. Also estimated particularly badly are the number of traffic tickets on both campuses and the number of library fines at SUSB. These are the most sensitive questions on the questionnaires for the campus experiments, both asking for reports of negative behaviors.

In 19 of the 30 questions tabulated (18 in the campus experiments and 12 in the Census experiment), one of the estimators arising from the RGI procedure had the smallest absolute percentage error. There was also a remarkable similarity of performance of the estimators (with the possible exception of the Bayesian estimator that uses the sample median for the prior mean). In 22 out of 30 cases the estimators are unanimous in either under-estimating or over-estimating truth.

5.2 Accuracy of interval estimates

Because our experiments were all designed to offer validation data for the group being studied, we can see whether the intervals being calculated cover the true values. Note that for this argument we are making a rather unusual use of intervals. Rather than asking whether an interval calculated for a sample covers the true population value, in this case we are thinking about our samples rather as if they were populations and asking whether the calculated intervals cover the average true values for the group of people questioned.

For the interval estimates, we found that the ARGI covered the average true value in 23 of the 30 questions, while the traditional 95% confidence interval covered the average true value for only 17 of the 30 questions.

Next we consider the order variation in the Census Experiment, asking if the length of the ARGI is different if the usage question is asked before or after the bounds questions. We had hypothesized that the length would be shorter when the respondent has a chance to anchor the bounds on the usage question. For the questions on salary and wages (and their five-year change) the ARGI is always smaller when the questions are ordered with the basic usage question first, as hypothesized. These questions are ones for which the information is probably best known to the respondents. The trend is exactly reversed, however, for the questions about interest and dividends (and their five-year change) where the information is probably not as well known to respondents. (Interest and dividends do not appear on regular paychecks, and often a respondent's only information about them may come from a year-end summary used to prepare income taxes.) For these questions, the shorter ARGI is found when respondents are asked the bounds questions first. We must refine our hypothesis and speculate that respondents give shorter intervals in the usage-quantity-first condition when they can utilize their usage response as an anchor, if they are confident of their usage response. When respondents are not sure about their recall of the usage quantity, however, that anchoring effect is either not available or not useful.

5.3 Reduction of item nonresponse

To investigate whether the RGI procedure reduces item nonresponse we use data from the paper-and-pencil campus experiments. Those Rs who gave an interval but did not give a usage quantity constitute an appreciable percentage of those who did not give a usage quantity

and thus were potential nonresponders to each item. Indeed, those percentages are never less than 4% and twice are over 40%. We can interpret these results as estimated conditional probabilities of giving an interval among those who did not give a usage quantity. We can use the midpoint of the RGI as a point estimator and the ARGI as an interval estimator, for those respondents who offered interval but no usage quantity responses, and inquire into the accuracy of these estimates for the fee data (where sample sizes are large and verification data unnecessary because of the uniformity of the fees across respondents). We find that the average midpoints overestimate usage for 3 or the 4 cases, but the ARGI cover the true value in all cases.

Thus in the Campus Experiments in a substantial proportion of cases, Rs who do not supply an estimate of usage quantities do supply intervals which are reasonably accurate, thus reducing the amount of item nonresponse appreciably. In the Census Experiment, although many Rs did not supply usage quantities, in only a few such cases did they supply bounds information. Why these differences? There may be an effect of the sensitivity of the questions, sensitive questions about income in the Census Experiment, less sensitive questions in the Campus Experiments. There may also be a mode effect. In the paper-and-pencil Campus Experiments it was easy to fill in part of a question; it is less easy to answer part of a question posed by an interviewer over the telephone. The type of respondent, type of interviewer, and survey sponsor may matter. The Campus Experiments involved undergraduate student Rs, students distributing questionnaires, and an "academic" survey. The Census Experiment interviewed Rs from established households, who were presented with questions from professional interviewers representing the US Census Bureau. Overall, there was greater respondent cooperation in this government survey by telephone than we found in our earlier campus-based experiments.

6. CONCLUSIONS

We have found that the RGI technique yielded promising results in improving the accuracy of point and interval estimation and in reducing item nonresponse in cases of low cooperation.

BIBLIOGRAPHY

Evans, Michael and Swartz, Tim (1995). Methods for Approximating Integrals in Statistics With Special Emphasis on Bayesian Integration Problems, *Statistical Science*, Vol. 10, No. 3, 254-272.

Heeringa, Steven G.; Hill, Daniel H.and Howell, David A. (1995). Unfolding Brackets for Reducing Item Non-Response in Economic Surveys, *Health and Retirement Study Working Paper Series*, Paper No. 94-029, Institute for Social Research, University of Michigan, June.

Hogarth, R. (1980). *Judgment and Choice: The Psychology of Decision*, New York: John Wiley and Sons, Inc.

Kennickell, Arthur B. (1997). Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances, Jan., 1997, Board of Governors of the Federal Reserve System, Washington, D.C. 20551.

Kish, Leslie (1965) *Survey Sampling*. New York: John Wiley and Sons.

Marquis, Kent H., and Press, S. James (1999). Cognitive Design and Bayesian Modeling of a Census Survey of Income Recall, *Proceedings of the Federal Committee on Statistical Methodology Conference*, Washington, DC, Nov. 16, 1999, pp.51-64 (see http://bts.gov/fcsm).

Press, S. James (1989). *Bayesian Statistics: Principles, Models and Applications*, New York: John Wiley and Sons, Inc.

----------(1999). Respondent-Generated Intervals for Recall in Sample Surveys, manuscript, Department of Statistics, University of California, Riverside, CA 92521-0138 , Jan., 1999. http://cnas.ucr.edu/~stat/press.htm

Press, S. James and Marquis, Kent H. (2002a) Bayesian Estimation in a U.S. Government Survey of Income Using Respondent-Generated Intervals. *Proceedings of the Sixth World Meeting of the International Society for Bayesian Analysis*, May, 2000, Crete, Greece; Eurostat, in press.

---------- and ----------(2002b) Bayesian Estimation in a U. S. Census Bureau Survey of Income Recall Using Respondent-Generated Intervals. *Journal of Research in Official Statistics*, Eurostat, in press.

Press, S.James and Tanur, Judith M. (2000a). Respondent-Generated Interval Estimation to Reduce Item Item Nonresponse, *Applied Statistical Science V*, Nova Science Publishers, See also, http://cnas.ucr.edu/~stat/press.htm

---------- and ---------- (2000b) Experimenting with Respondent-Generated Intervals in Sample Surveys, with discussion. Pages 1-18 in Monroe G. Sirken (ed.) *Survey Research at the Intersection of Statistics and Cognitive Psychology*, Working Paper Series #28, National Center for Health Statistics, U.S. Department of Health and

Human Services, Center for Disease Control and Prevention.

Schlaifer, H. (1959). *Probability and Statistics for Business Decisions*, New York: McGraw-Hill Book Co., Inc.

Spiegelhalter, David; Thomas, Andrew; and Nicky Best (May, 1999). "WinBUGS Version 1.2 User Manual", MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, UK. http://www.mrc-bsu.cam.ac.uk/bugs.