A BAYESIAN APPROACH TO SECOND GUESSING "UNDECIDED" RESPONDENTS

S. James Press
University of Chicago

Chung-fang Yang
Leo Burnett Company

We are interested in the problem of estimating category probabilities in a sample survey when people are asked to answer a question with categorical answers and, for some reason, some of the respondents either answer "I don't know" or refuse to give any answer at all. Of course, for some factual questions like "What is the capital of Zambia?", a lot of people would answer "I don't know" because they really don't know. However, in some other situations, for instance, when you ask somebody his opinion about a very socially controversial question, such as, "Are you opposed to school busing?", we think that people who answer "I don't know" to this question would be very different from those people who answer "I don't know" to the previous question. We call the latter type of "I don't know's," "Undecideds." We have reasons to believe that these "Undecideds" are not really neutral.

Traditional ways of handling these "Undecided" respondents are the following three:

1) Simply to throw them out of the sample,
2) to allocate them to the unambiguous categories according to the proportions of respondents who originally, unambiguously were assigned to each of the categories,
3) to allocate them equally to each of the categories.

We think none of these methods is satisfactory theoretically. However, we do not elaborate here.

In this paper, we propose a new way of handling the problem. We assume the following model:

1) There is one question, at a time, that we are mainly interested in. We refer to it as the "main" question. We assume the main question has category responses.
2) There are some other questions which are either being asked to the respondents at the same time when the main question is asked, or which are being asked to the same group of respondents at different times. I will refer to these questions as subsidiary questions. We assume that all respondents answer the subsidiary questions unambiguously, although only some of the respondents answer the main question unambiguously (the others are the "Undecideds").
3) We assume the subsidiary questions are related to the main question, either theoretically or empirically so that they can be used to predict the respondent's answers to the main question from the way they answered these subsidiary questions.

The method of estimating the true category probabilities on the main question is a Bayes approach. It uses several types of information.

1) Subjective prior information for the category probabilities.
2) Sample frequencies for those respondents who did answer the main question unambiguously.
3) Response pattern on the subsidiary questions from the respondents who answered the main question unambiguously.

The "Undecided" respondents will be "second guessed" on the main question i.e. effectively, they will be classified into one of the unambiguous response categories on the basis of their answers to the related subsidiary questions.

Estimators of the true category probabilities on the main question can be calculated in terms of both the "decided" and second guessed "undecided" respondents. The result is the following.

Suppose n_i subjects responded unambiguously in category i of the main question, $i = 1, \ldots, M$. If m subjects are "undecided" on the question, a Bayes point estimator of the probability, q_i, that a randomly selected subject will fall into category i is given by the mean of the posterior distribution:

$$E(q_i | n_1, \ldots, n_{M-1}; z^{(1)}, \ldots, z^{(m)}) \equiv \hat{q}_i = \frac{(n_i+\alpha_i) + \sum_{j=1}^{m} P\{\Pi_i | z^{(j)}\}}{m + \sum_{j=1}^{M} (n_j+\alpha_j)}, \quad i = 1, \ldots, M, \tag{1}$$

where α_i's are parameters of the prior density for the q_i's ($\alpha_i = 1$, if we take a vague prior), $P\{\Pi_i | z^{(j)}\}$ is the marginal predictive probability for classifying the j^{th} "undecided" respondent into category Π_i of the main question, given his response on the subsidiary questions, $z^{(j)}$.

This marginal predictive probability is shown in the paper to be expressible in the form:

$$P\{\Pi_i | z^{(j)}\} = \frac{(n_i+\alpha_i)\, h(z^{(j)} | \Pi_i)}{\sum_{k=1}^{M} (n_k+\alpha_k) h(z^{(j)} | \Pi_k)} \tag{2}$$

where $h(z^{(j)} | \Pi_i)$ denotes the marginal predictive density of the response to the subsidiary questions for the j^{th} "undecided" respondent, given he belongs to category Π_i on the main question.

Therefore, $\hat{q}_i$ can be expressed as:

$$\hat{q}_i = \left[\frac{n_i + \alpha_i}{m + \sum_{j=1}^{M} (n_j + \alpha_j)} \right] \times \sum_{j=1}^{m} \left[\frac{h(z^{(j)} | \Pi_i)}{\sum_{t=1}^{M} (n_t+\alpha_t) h(z^{(j)} | \Pi_t)} \right]. \tag{3}$$

Variances for the category probability estimators

are developed in the paper, as are Bayesian credibility or confidence intervals.

The only problem remaining is to evaluate the predictive density $h(z^{(j)}|\Pi_i)$. Three cases are discussed in the paper. We discuss the cases when the $z^{(j)}$'s (i.e. the responses to the subsidiary questions) are all continuous, all discrete (categorical), and the mixed case (with some subsidiary questions having continuous responses and some having discrete responses). For illustrative purposes, we now consider explicitly the case when $z^{(j)}$ is discrete, and give the predictive density for that case.

Suppose there are q subsidiary questions with discrete type responses. Let S denote the number of possible joint responses to this set of questions. For instance, suppose there are two questions with two possible answers for each question. Then there are $S = 2 \times 2 = 4$ possible joint responses to these two questions. Let u_k, an S dimensional unit vector with a "1" in the k^{th} place and all other places are "0", denote the k^{th} joint response pattern to the subsidiary questions. Then the predictive density is given by

$$h(z^{(j)} = u_k|\Pi_i) = \frac{x_{k|i} + \delta_{k|i}}{n_i + \sum_{t=1}^{S} \delta_{t|i}}, \qquad (4)$$

where $k = 1, \ldots, S$, $i = 1, \ldots, M$, $\Delta_i = \sum_{t=1}^{S} \delta_{t|i}$, $n_i = \sum_{k=1}^{S} x_{k|i}$, $x_{k|i}$ denotes the number of respondents who answered unambiguously in category i of the main question and who were in cell k of the subsidiary questions; $\delta_{k|i}$ denotes the parameters of the natural conjugate Dirichlet prior distribution for the cell probabilities of the responses to the subsidiary set of questions; n_i is the number of respondents who anwered unambiguously in category i of the main question.

Example

Suppose that out of 100 respondents to a sensitive question, 20 people respond in each of the three possible unambiguous categories and 40 are "undecided." Moreover, suppose that there is just one subsidiary question (with four response categories) which is used, and the "decided" group on the main question respond according to the table below.

		Subsidiary Question				
		(1)	(2)	(3)	(4)	Totals
Main Question	(1)	17	1	1	1	20
	(2)	5	5	5	5	20
	(3)	1	1	1	17	20

Thus, there are 17 subjects who responded in category "1" of the subsidiary question, given they responded in category "1" on the main question, etc.

For the "undecided" group on the main question, suppose 25 respond in subsidiary category "1", and 5 respond in each of the remaining categories. Results of the analysis are given below assuming vague priors for both the q_i's and for p_i.

i	Ignoring "Undecideds"	$\hat{q}_i$	σ_{q_i}	90% credibility Interval
1	.33	.40	.04	(.34, .48)
2	.33	.34	.04	(.28, .40)
3	.33	.26	.04	(.18, .32)

Thus, if the "undecided" group had been ignored and if the q_i's were estimated on the basis of sample frequencies, column (2) would have resulted. Use of (3) yielded column (3) while column (4) gives the standard deviations. The last column was obtained by using the beta approximation described in the paper with 5% probability in each tail of the approximating distribution.

Interested readers may refer to the complete paper for further details. It is scheduled to appear in Journal of the American Statistical Association, March, 1974.