Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 589–593
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Item Count Technique in estimating the proportion of people with a sensitive feature

Arijit Chaudhuri[a,*,1], Tasos C. Christofides[b]

[a]*Applied Statistics Unit, Indian Statistical Institute, 203 BT Road, Kolkata-700108, India*
[b]*Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus*

Received 5 July 2004; received in revised form 21 April 2005; accepted 5 January 2006
Available online 21 February 2006

Abstract

In assessing the prevalence of a sensitive attribute like habitual heroin consumption in a community of people, indirect questioning is a necessity to extract truth on ensuring protection of privacy. The current literature seems to need supplementary specification of a relevant practical and theoretical justification for one possibility by what is called an Item Count Technique. This method can be easily incorporated in large scale sample surveys where the medium of collecting information is a structured questionnaire. This feature will make this technique attractive to social survey researchers. In this article we present an amendment to the currently available technique rendering it well-equipped with a provision to protect privacy and also a sound theoretical foundation.

MSC: 62D05

Keywords: Indirect questioning; Protecting privacy; Randomized response; Sophisticated estimation; Unequal probability sampling

1. Introduction

In a survey to estimate the proportion of people bearing a stigmatizing characteristic like habitual gambling, marijuana consumption, experience of induced abortion, tax evasion, rash driving, etc., truth is suspected to be a casualty in generating direct responses (DR). A classical alternative, namely the randomized response (RR) technique introduced by Warner (1965) and developing rapidly even now is often criticized not only because of its exacting demands on the skills of the respondents in handling the required devices, but mainly because this technique asks respondents to provide information that seems useless or even tricky. Simply because a respondent does not understand the mathematical machinery behind the technique, then the entire procedure seems suspect and the interviewee may think that there is in fact a way for the interviewer to find his/her exact status regarding the sensitive characteristic by processing the response he/she provides. Three alternatives, namely the Item Counting, the Nominative Technique and the Three Card Method are proposed in the literature. Details can be found in Droitcour et al. (1991, 2002), and Miller (1985), respectively. All three of the alternatives to RR are so designed that participants need not be aware that a special estimation technique

* Corresponding author. Tel.: +91 33 2575 2816; fax: +91 33 2577 3104.
E-mail addresses: achau@isical.ac.in (A. Chaudhuri), tasos@ucy.ac.cy (T.C. Christofides).
[1] His research is partiallysupported by CSIR Grant no. 21(0539)/02/EMR-II. This work was done during his visit to the University of Cyprus.

0378-3758/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.01.004

is being used. In addition, the respondents provide answers that make sense to them and at the same time they are fully aware of what they are revealing about themselves.

The current status of the Item Count Technique, in our view requires amendments vis-a-vis its practicability and theoretical sophistication. Section 2 elaborates on these. Concluding remarks constitute Section 3.

2. Description of applicability and theory of the Item Count Technique

As per the available literature to date, the Item Count Technique consists of taking two independent samples, asking each person in one of them to report to the investigator the number out of a given list of say, G items that are applicable to him/her. All these items are supposed to be innocuous. Each person in the second sample from the same population is requested to report the number of items applicable to him/her out of the same list to which one "stigmatizing" item is added. The sample mean number calculated from the second sample minus that from the first sample is taken as an estimate of the proportion of people bearing the stigmatizing $(G+1)$st item in the community. The role of the G innocuous items is to induce a better respondent's cooperation. They should be so chosen and worded in adequate numbers as to ensure enough variability in the numbers in which they apply to the people in the community.

It is clear that this technique can be incorporated into self-administered questionnaires for large scale sample surveys. This feature makes the technique appealing to social survey researchers. However, in our view we are yet to be satisfied about how to maintain secrecy for the respondents to whom all the $(G+1)$ items may apply or also, if none applies, especially for those in the second sample. One way to minimize the chance of having respondents indicating agreement with all $G+1$ items, and thus revealing that they posses the sensitive characteristic, is to include in the list of items, at least one item whose prevalence is extremely low or alternatively multiple low prevalence items. Of course, even in this case, there is no guarantee that a respondent will not be found in a situation where he/she has to report agreement with all items.

A second minor disadvantage of the technique is that no precaution against a possible "negative" value for the estimate has emerged as yet.

We propose the following courses of action. From a given community of N people, adopting a suitable common design p choose two independent samples s_1 and s_2, say, of the same average sample size, say, $v=\sum_s v(s)p(s)$ where $v(s)$ is the number of distinct units in a sample s.

Every unit in s_1 is then presented with a list of $(G+1)$ items of which the first G are innocuous and the last or $(G+1)$st item stands for "either one of a stigmatizing i.e. "tainted" type, say, "T" or a fresh item, say "F" which is non-stigmatizing" or both of them. The person then is to give out the number of items that are true for him/her. Obviously, this number must be one of $0, 1, \ldots, G, G+1$. It is clear that for such a person either T or F or both T and F or neither of T and F may apply. Similarly, to every person in the second sample s_2 is given a list of $(G+1)$ items, of which the first G items are exactly the same as those above, but the $(G+1)$st item or the last item stands for "either the non-applicability of T or the non-applicability of F", including the non-applicability of both T and F. Thus, from a person in the second sample, the number given out must be one of the numbers $0, 1, \ldots, G, G+1$ as well. Clearly, the respondent must understand that "either the complement of T, say, T^c or the complement of F, say, F^c" or both T^c and F^c together may be applicable to him or her, or neither T^c nor F^c as well.

Let for a typical member i of the community chosen in s_1, the number given out be denoted as y_i and for a typical member j of the community chosen in s_2, the number given out be denoted as x_j.

Let π_i $(\pi_i>0)$ denote the inclusion probability of a unit i of the community in a sample chosen according to a design p.

Let further, the fresh item F be so chosen that assuming the size N of the community as at least moderately large, the proportion θ_F of the people in the community possessing F may be supposed to be a known number. For example, F may be taken to denote that the year of birth of a person in s_1 is an odd number so that θ_F may be taken to be $\frac{1}{2}$. Of course, F^c then denotes an even number as the year of birth of a person.

Now applying the Horvitz–Thompson (1952) method of unbiased estimation of a total and writing

$$Nt_1 = t(s_1) = \sum_{i\in s_1} \frac{y_i}{\pi_i} \quad \text{and} \quad Nt_2 = t(s_2) = \sum_{j\in s_2} \frac{x_j}{\pi_j}$$

we may state the following result.

Theorem 1. *$\hat{\theta} = t_1 - t_2 + 1 - \theta_F$ is an unbiased estimator for the unknown proportion θ of people bearing the tainted characteristic T in the community.*

Proof. Writing E_p as the expectation operator with respect to the design p we get

$$\begin{aligned}
E_p(\hat{\theta}) &= E_p(t_1) - E_p(t_2) + 1 - \theta_F \\
&= \{\text{proportion of people in the community bearing } (T \cup F) \\
&\quad \text{in combination with } 1, 2, \ldots, G \text{ or none of the } G \text{ innocuous items}\} \\
&\quad - \{\text{proportion of people in the community bearing } (T^c \cup F^c) \\
&\quad \text{in combination with } 1, 2, \ldots, G \text{ or none of the } G \text{ innocuous items}\} + 1 - \theta_F \\
&= \{\text{proportion bearing } T \text{ with or without any of the } G \text{ items}\} \\
&\quad + \{\text{proportion bearing } F \text{ with or without any of the } G \text{ items}\} \\
&\quad - \{\text{proportion bearing } T \text{ and also } F \text{ with or without any of the } G \text{ items}\} \\
&\quad - 1 + \{\text{proportion bearing } T \text{ and also } F \text{ with or without any of the } G \text{ items}\} \\
&\quad + 1 - \theta_F \\
&= (\theta + \theta_F - 1) + 1 - \theta_F \\
&= \theta,
\end{aligned}$$

on noting the DeMorgan's law that $T^c \cup F^c = (T \cap F)^c$; the proof is complete. □

Observe that by construction, $\hat{\theta}$ is unlikely to turn out negative.

Recalling from Chaudhuri and Pal (2002) that for the total $Y = \sum_k y_k$ the Horvitz–Thompson's estimator $t = t(s) = \sum_{k\in s}(y_k/\pi_k)$ based on a sample s chosen with a design p from a population of N units has variance

$$V_p(t) = \sum_{k=1}^{N} \sum_{l=k+1}^{N} (\pi_k \pi_l - \pi_{kl}) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 + \sum_{k=1}^{N} \frac{y_k^2}{\pi_k} \beta_k, \tag{1}$$

for

$$\pi_{kl} = \sum_s I_{sij}\, p(s) \quad \text{and} \quad \beta_k = 1 + \frac{1}{\pi_k} \sum_{l=1, l\neq k}^{N} \pi_{kl} - v,$$

where $I_{sij} = 1$ if both i and j belong to s and zero otherwise and V_p denotes the variance operator with respect to the design p.

An unbiased estimator of $V_p(t)$ is

$$v_p(t) = \sum_{k\in s} \sum_{l\in s, l>k} \left(\frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}} \right) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 + \sum_{k\in s} \frac{y_k^2}{\pi_k} \frac{\beta_k}{\pi_k}, \tag{2}$$

assuming throughout that $\pi_{kl} > 0, \forall k \neq l$.

So we may write the following:

Theorem 2.

(a) $$V_p(\hat{\theta}) = V_p(t_1) + V_p(t_2)$$
$$= \frac{1}{N^2}\left[\sum_{k=1}^{N}\sum_{l=k+1}^{N}(\pi_k\pi_l - \pi_{kl})\left\{\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 + \left(\frac{x_k}{\pi_k} - \frac{x_l}{\pi_l}\right)^2\right\} + \sum_{k=1}^{N}\frac{\beta_k}{\pi_k}(y_k^2 + x_k^2)\right]$$

(b) $$v_p(\hat{\theta}) = \frac{1}{N^2}\left[\sum_{k\in s_1}\sum_{l\in s_1, l>k}\left(\frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}\right)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 + \sum_{k\in s_2}\sum_{l\in s_2, l>k}\left(\frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}\right)\left(\frac{x_k}{\pi_k} - \frac{x_l}{\pi_l}\right)^2\right.$$
$$\left. + \sum_{k\in s_1}\frac{y_k^2}{\pi_k^2}\beta_k + \sum_{k\in s_2}\frac{x_k^2}{\pi_k^2}\beta_k\right]$$

Proof. Straightforward from (1) and (2). □

3. Concluding remarks and comments

With the above formulations the Item Count Technique is hereby claimed to be a viable indirect questioning method adequate for estimation with a respondent's privacy well protected. It can be incorporated into large scale sample surveys and this feature should make it quite appealing to social survey researchers.

It is clear from Section 2, that the mathematical formulation of the technique does not involve the value G, i.e., the number of statements about innocuous items. Therefore, the onus is upon the designer of the survey to select this value and most importantly the statements themselves. Common sense suggests that this value should not be either very small or very large, so that the cooperation of the respondents is not jeopardized and the statements should be chosen in such a way that all of the numbers $0, 1, \ldots, G, G+1$ would be possible as potential answers. Further, the statement involving the stigmatizing attribute should not necessarily appear as the last of the $G+1$ statements, but rather it is preferable to appear somewhere in the middle. In addition, if this technique is a part of a larger sample survey covering issues other than the study of the tainted characteristic T as well, then those statements should not appear anywhere else, for example as single statements in a questionnaire where the respondent indicates precisely his/her answers.

For the Item Count Technique, to increase the sense that the list of items serves a meaningful purpose and therefore increase the level of cooperation of the participants, the items should seem to blend together and give the impression that the number reported to the interviewer is a meaningful piece of information. Having this in mind, the G innocuous items should not be totally unrelated to the stigmatizing $(G+1)$st item. In addition, some of the innocuous statements could be phrased in a way similar to the statement regarding the stigmatizing characteristic, as it is done in the sample questionnaires below.

The way the item in the list regarding the stigmatizing characteristic for both samples is phrased may create suspicions or confusion similar to those in randomized response technique. However, these disadvantages may be eliminated with the appropriate layout of the lists. In addition, clear instructions should be given emphasizing the fact that what is important is the total score reported and not the answers to individual statements. Below are given two sample lists which can be used to estimate the proportion of marijuana users in a certain community. The items in the lists are not totally unrelated to each other. On the contrary, they may be considered as items related to, say, childhood asthma, or health in general. Thus, one may regard the lists as part of an extended questionnaire on health issues. A questionnaire should be given to a participant with the clear instruction that it should not be returned to the interviewer. So the instructions could be the following:

> For each one of the following statements give a score of 1 in the right column if the statement applies to you and a score of 0 if not. If a statement consists of two substatements, such as in Statement 2 or Statement 4, a score

of 1 should be given if at least one of the substatements applies and a score of 0 if none of them does. Count the number of 1's put in the right column. This is the total score. Report the total score and nothing else. Do not return the questionnaire. It is given to you for your convenience.

Questionnaire 1

Number	Statement	Score
	I have never been hospitalized	
2	Substatement 2a: I do have hay fever	
	Substatement 2b: I do have eczema	
3	I have taken antibiotics during the last two years	
4	Substatement 4a: I make use of marijuana	
	Substatement 4b: During my childhood I had asthma	
5	At least one of my parents was a smoker during my childhood	
6	I consider smoking no less harmful than the use of marijuana	
	Total Score	

Questionnaire 2

Number	Statement	Score
	I have never been hospitalized	
2	Substatement 2a: I do have hay fever	
	Substatement 2b: I do have eczema	
3	I have taken antibiotics during the last two years	
4	Substatement 4a: I do not make use of marijuana	
	Substatement 4b: During my childhood I did not have asthma	
5	At least one of my parents was a smoker during my childhood	
6	I consider smoking no less harmful than the use of marijuana	
	Total Score	

Acknowledgements

The authors thank the referees for their comments on the paper. Especially, they would like to thank one of the referees whose suggestions led to an improved version of the manuscript.

References

Chaudhuri, A., Pal, S., 2002. On certain alternative mean square error estimators in complex survey sampling. J. Statist. Plann. Inference. 104, 363–375.

Droitcour, J.A., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W., Ezzati, T.M., 1991. The Item Count Technique as a method of indirect questioning: a review of its development and a case study application. In: Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N., Sudman, S. (Eds.), Measurement Errors in Surveys. Wiley, New York.

Droitcour, J.A., Larson, E.M., 2002. An innovative technique for asking sensitive questions: the three card method. Bull. Methodologie Sociologie 75, 5–23.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47, 663–685.

Miller, J.D., 1985. The nominative technique: a new method of estimating heroin prevalence. NIDA Research Monograph, No. 57, pp. 104–124.

Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. J. Amer. Statist. Assoc. 60, 63–69.