# A NOTE ON ESTIMATION FROM RANDOMISED RESPONSE SURVEYS

*by*

JAGBIR SINGH*

*Temple University, Philadelphia, Pa. 19122*

(Received : March, 1980)

SUMMARY

**Following Warner's [1965] basic paper on randomised response survey techniques, many papers have appeared extending or modifying his model For all these models, estimators have beea developed for the proportion of population in the sensitive category. Most of these estimators of proportions could assume values greater than one or less than zero and yet have been erroneously called the maximum likelihood estimators (m. l. e.) In this note we examine the cause of this confusion and notice that the erroneously called m. l. e. are not even admissible when compared with the true m. l. e.'s. Using a recent model, alternative to that of Warner, we demonstrate that the erroneously called m. l. e. for this model also continue to have large probability of assuming values outside (0, 1) even when the sample size is as large as 210.**

## 1. INTRODUCTION

In surveys related to delicate questions, Warner [1965] introduced a randomised response technique for eliciting information, and thus estimating the proportion $\pi$ of the population in the sensitive category $A$. The technique consists in providing a spinner, or some suitable randomising device, with two outcomes $A$ or not $A$ with associated probabilities $p$ and $\bar{p}=(1-p)$ to each respondent. The respondent spins the spinner unobserved by the interviewer and answers *yes* if he has the characteristic indicated by the pointer and *no* otherwise. If the outcomes of the device are independent of the individual's responses, then obviously the probability of a *yes* response, say $\theta$ is,

$$\theta = p\pi + \bar{p}(1-\pi)$$
$$= p + (2p-1)\pi \qquad ...(1.1)$$

* This research was done while the author visited the Indian Agricultural Statistics Research Institute in New Delhi From August 1977 to May 1978.

If $m$ is the number of *yes* responses out of $n$ independent responses, and if the responses are truthful, Warner stated that the maximum likelihood estimator (m.l.e.) of $\theta$ is $\hat{\theta}=m/n$, and consequently that of $\hat{\pi}=(\hat{\theta}-\bar{p})/(2p-1)$ provided $p\neq\frac{1}{2}$. We may sssume without any loss of generality that $p>\frac{1}{2}$.

In a note [1976], this author pointed out that $\hat{\theta}$ and $\hat{\pi}$ are not the m.l.e.'s of $\theta$ and $\pi$ respectively. Indeed the note was motivated by a comment of Warner that his estimator $\hat{\pi}$ of $\pi$ can be negative or greater than 1. The comment led us to ask : If a value of $\pi$ is to be chosen from (0, 1) so that the likelihood function is maximised, then how is it possible to obtain an estimate $\hat{\pi}$ of $\pi$ outside the interval (0, 1) ? Obviously something was at a miss which was recognized by this author in [1976] note that the probability $\theta$ of a *yes* response can only assume values in the interval $[\bar{p}, p]$ and not (0, 1). Hence the true m.l.e. $\tilde{\theta}$ of $\theta$ instead is :

$$\tilde{\theta}=\begin{cases}\theta \text{ if } \bar{p}<\hat{\theta}<p\\ p \text{ if } \hat{\theta}\geqslant p\\ \bar{p} \text{ if } \hat{\theta}\leqslant\bar{p}\end{cases}$$

provided $p$ is known. From (1.1) it is now obvious that the m.l.e. of $\pi$ is

$$\hat{\pi}=\begin{cases}\hat{\pi} \text{ if } \bar{p}<\tilde{\theta}<p,\\ 1 \text{ if } \hat{\theta}\geqslant p\\ 0 \text{ if } \hat{\theta}\leqq\bar{p}\end{cases}$$

Indeed it was further shown that neither $\hat{\theta}$ nor $\bar{\pi}$ is admissible. The true m.l.e.'s $\tilde{\theta}$ and $\tilde{\pi}$ are uniformly better than $\hat{\theta}$ and $\hat{\pi}$ respectively with respect to the squared error loss function.

Following Warner's basic paper, many more papers have appeared extending, modifying and studying various other aspects of his model. Throughout in most of this literature, estimators developed for the proportion of population in the sensitive category are not truly the m.l.e.'s just as $\hat{\pi}$ is not the m.l.e. of $\pi$ for the Warner's model. In passing, we may quote the following two more from the literature.

Horvitz and his associates [1967] and Greenburg *et. al.* [1969] modified Warner's model by introducing a non-delicate question $Y$. Assume $\pi_y$, the proportion of people in $Y$, as known. If a randomly selected iudividual answers the sensitive question with probability $p$ and nonsensitive question $Y$ with probability $(1-p)=\bar{p}$, then the probability of *Yes* response is $\theta=p\pi+\bar{p}\pi_y$. It may be noted that $\theta$ lies in $[\bar{p}\pi_y,\ p+\bar{p}\pi_y]$. The authors having failed to notice this restriction, estimated $\theta$ by $n_1/n$, proportion of *Yes* response and in turn obtained an estimate, erroneously called m.l.e., of $\pi$ to be $\left[\frac{n_1}{n}-\bar{p}\pi_y\right]/p$, which could be negative or greater than one.

Latter Liu and Chow [1976], proposed a "Discrete Quantitative Response Model," for estimating $\pi_i$, the proportion of respondents who possess "$i$" quantitative measure $(\Sigma\pi_i=1)$. For example $\pi_i$ may be the proportion of individuals having "$i$" abortions. Because of the randomization, the probability that an individual will respond "$i$" is $\theta_i=\pi_i p+p_i$, where $p$ and $p_i$ are respectively the proportions of red balls in the randomizing device and white balls marked "$i$" in the device. Once again, it is because of the randomization that $p_i < \theta_i < p+p_i$. Having failed to recognize this restriction, $\hat{\pi}_i=\left(\frac{n_1}{n}-p_i\right)/p$ was erroneously called the m.l.e. of $\pi_i$. Hence $n_i/n$ is the observed proportion of respondents answering "$i$". Needless to say that $\hat{\pi}_i$ could be negative or greater than one.

These are but two examples from among many in this area of research.

In defense of the erroneously called m.l.e.'s, one may think of two arguments. First, since no one would really estimate a proportion by a negative number or a number greater than one, the estimates can be truncated to lie between zero and 1. For example, Warner's estimator $\hat{\pi}$ when truncated at zero or 1 coincides with $\tilde{\pi}$ This argumeat serves well as long as $\hat{\pi}$ is only used to estimate $\pi$. However, when one compares Warner's model with the alternative models, it is the variance of $\hat{\pi}$ which is used and not the mean square error of $\tilde{\pi}$. This is improper because as we showed in [1976] that the true m.l.e. $\tilde{\pi}$ is uniformly better than $\hat{\pi}$. Since such improper comparisons have, however, been made in the literature, for example Dowling [1975], it is difficult to accept the truncation argument.

The second argument is that $\hat{\pi}$ converges in distribution to $\tilde{\pi}$ and, therefore, the probability of $\hat{\pi}$ being negative or greater than one is essentially zero. This argument is no doubt theoretically sound and even practically valid for simple models, such as Warner's But as we shall see that for at least one randomised response model, proposed as an alternative to that of Warner, the so called m.l.e. has substantial probability of assuming values outside the admissible interval (0, 1) even for moderately large sample sizes.

In the remainder, we consider a recent randomised response model proposed by Takahasi and Sakasegawa (1977). We notice that some natural restrictions results on the parameters of this model as well because of the randomization imposed to maintain confidentiality. Recall the restriction on $\theta$ in the Warner's model that it must lie in $[\bar{p}, p]$ and not in (0, 1) as a result of the randomization It is these restrictions resulting from randomization which have gone mostly unnoticed and have been the root cause of erroneously believing the estimators to be the m.l.e. while they indeed are not. Takahasi and Sakasegawa also failed to account for such restrictions. They proceeded to derive an estimator of $\pi$, and called it the m.l.e. We demonstrate here that their esiimator is not really the m.l.e. and non-even admissible. In fact, without incorporating an additional assumptlon in their model, it is not possible to get the m.l.e. of $\pi$. In short, they can estimate $\pi$ using their model but the model needs modification in order to obtain the m.l.e. But for the modified model, their estimator is rendered inadmissible. The note ends with an example showing that even when the sample size is large, their estimator continues to have large amounts of probabilities associated with its negative and greater than one range.

## 2. Takahasi and Sakasegawa Model

The Warner's technique and most of its variants require some randomising device, the outcome of which is assumbed to be independent of any characteristic of the individual. Takahasi and Sakasegawa believed that the independence assumption is unrealistic. In an attempt to remove this drawback they proposed a model which does not require the use of any randomising device. For details, see their paper. Briefly the model is this :

The independent samples are drawn each with replacement. A respondent in each sample is asked to make a silent choice of one item from among the three items, say $B_1$, $B_2$, $B_3$ and answer 0 (no) or 1 (yes) according to a scheme summarized below.

| Attribute Item | 1st Sample | | 2nd Sample | | 3rd Sample | |
|---|---|---|---|---|---|---|
| | A | Not A | A | Not A | A | Not A |
| $B_1$ | 0 | 1 | 1 | 0 | 1 | 0 |
| $B_2$ | 1 | 0 | 0 | 1 | 1 | 0 |
| $B_3$ | 1 | 0 | 1 | 0 | 0 | 1 |

For example, if a respondent in the second sample has made a silent choice of $B_3$, then his response would be 1 if he has the attribute $A$ and 0 otherwise.

Following Takahasi and Sakasegawa's notation, let

$P(A, i)$=the proportion in the population of those who have $A$ and will choose $B_i$; $i=1, 2, 3$

$P(\bar{A}, i)$=the proportion in the population of those who do not have $A$ and will choose $B_i$; $i=1, 2, 3$

$q_i$=probability that a respondent in the $i$-th sample answers "$i$".

Since $\Sigma P(A, i)=\pi$, it follows that,

$$q_i=\pi+P(\bar{A}, i)-P(A, i)\,;\ i=1, 2, 3 \qquad ...(2.1)$$

and

$$\pi=\Sigma\, q_i-1. \qquad ...(2.2)$$

For the model to be meaningful, we must assume that $\Sigma\, q_i \neq 1$.

## 3. Estimation and an Example

For the purpose of this note, let $p_i$ denote the probability that any given individual would make a silent choice of $B_i$. Surely' $p_i$ is unknown for the model under consideration but we assume for the sake of discussion and without any loss that $p_i<1/2$; $i=1, 2, 3$. Now notice that if $\pi=1$, then $P(A, i)=p_i$. On the other hand, $P(A, i),=0$ if $\pi=0$. Likewise, $P(\bar{A}, i)=pi$ or zero as $\pi=0$, or 1. Now from (2.1), we may find the upper and lower bounds for $q_i$ as $\pi\to 1$, and $\pi\to 0$ respectively. That is :

$$p_i \leqslant q_i \leqslant 1-p_i;\ i=1, 2, 3$$

If $n_i$ is the size of the $i$-th sample and $y_i$ is the number of respondent answering '1' out of $n_i$, then $\hat{q}_i=y_i/n_i$ was suggested to be the m.l.e. of $q_i$ by Takahasi and Sakasegawa. Remember that $q_i$ is restricted to lie in $[p_i, \bar{p}_i]$, but $\hat{q}_i$ could take values outside this this admissible interval. Depending upon the magnitude of $p_i$, there is a positive probability that $\hat{q}_i$ will lie outside the interval $[p_i, \bar{p}_i]$ and, therefore, cannot be the m.l.e. of $q_i$. Further, because of (2.2) Takahasi and Sakasegawa claimed $\hat{\pi} = \Sigma\, \hat{q} - 1$ to be the m.l.e. of $\pi$. Once again it is obvious that $\hat{\pi}$ is not really the m.l.e. Indeed $\hat{\pi}$ can assume values outside the admissible interval (0, 1).

If $p_i$ is assumed to be known, then the m.l.e. of $q_i$ is :

$$\tilde{q}_i = \begin{cases} \hat{q}_i \text{ if } p_i < q_i < 1 - p_i \\ p_i \text{ if } q_i \leqslant p_i \\ (1 - p_i) \text{ if } q_i \geqslant 1 - p_i \end{cases} \qquad i = 1, 2, 3$$

Apparently the range of $\tilde{q}_i$ is contained in the range of $\hat{q}$. This fact can be used to show that $\hat{q}$ is not even admissible with respect to the squared error loss. On the other hand if $p_i$ is unknown then the m.l.e. does not even exist. In that case one may decide to use $\hat{\pi}$ anyway, but the fact remains that it is not the m.l.e, of $\pi$ regardless of other properties it may have.

We shall conclude the note by considering an example to demonstrate that even for fairly large sample sizes, there is a considerable amount of chance that $\hat{\pi}$ would be observed outside (0, 1). Let us consider Takahasi and Sakasegawa model with parameters $\pi = 0.2$, $q_1 = 0.3$, $q_2 = 0.4$, $q_3 = 0.5$ and examine the behaviour of $\hat{\pi}$ if each of the three samples is of equal size $n = 70$. That is, the randomised response experiment would have $3n = 210$ respondents in all. Remember that $\hat{\pi} = \sum \frac{y_i}{n} - 1$ is unbiased for $\pi = \Sigma\, q_i - 1$. Consider,

$$\begin{aligned} \hat{\pi} - \pi &= \sum \frac{y_i}{n} - 1 - \pi \\ &= \sum \frac{y_i}{n} - \Sigma\, q_i \\ &= \frac{1}{\sqrt{n}} \Sigma\, Z_i, \end{aligned}$$

where $Z_i=(y_i-n\,q_i)/\sqrt{n}$ is asymptotically $N(0, q_i\,\bar{q_i})$. We may use this fact to calculate $P(\hat{\pi}<0)+P(\hat{\pi}>1)$. To this end we notice that $\Sigma Z_i$ asymptotically $N(0, 0.70)$. It then follows from further simple calculations that $P(\hat{\pi}>1)\approx 0$, and $P(\hat{\pi}<0)\approx .421$. That is, the chance is nearly 42 per cent that the erroneously called m.l.e. $\hat{\pi}$ will be found negative even when the sample size is as large as 210. We emphasize that the values of the parameters chosen for demonstrations are very realistic and were not picked to bias the example at all.

## References

Dowling, T.A. and Shactman Richard, H. (1975). : "On the Relative Efficiency of Randomized Response Models," *J. Am. Sta. Assoc.*, 70, pp. 84-87.

Greenberg, Benard G., Abul-Ela, Abbel-Latif A., Simmons, Walt R., Horvitz, Daniel G. (1969). : "The unrelated question randomized response model : Theoretical framework," *J. Am. Statist. Assoc.*, 64, pp. 520-539.

Horvitz, Daniel G., Shah, B.V., Simmons, Walt R., (1967) : "The Unrelated Randomized Response Model, *Proceedings of the Social Statistics Section. American Statistical Association.*

Liu, P.T., Chow, L.P-, and Mosley, W,H. (1975). : "Use of the Randomized Response Technique with a New Randomizing Device", *J. Am. Statist. Assoc.*, 70, pp. 329-332.

Singh, Jagbir (1976) : "A note on the Randomised Response Technique," *Procedings of the Social Statistics Seetion, American Statistical Assoc.*, p. 772.

Takahasi, K, and Sakasegawa, H. (1977). : "A Randomised Response Technique without making use of ans Randomisine Device." *Ann. Inst. Statist. Math.* 29, Part Aw, 1-8.

Warner, S.L. (1965) : "Randomized Response : A survey technique for eliminating evasive answer bias," *J. Am. Stat. Assoc.* 60, pp. 63-69.