Ann. Inst. Statist. Math.
29 (1977), Part A, 1-8

A RANDOMIZED RESPONSE TECHNIQUE WITHOUT MAKING USE OF ANY RANDOMIZING DEVICE

KOITI TAKAHASI AND HIROTAKA SAKASEGAWA

(Received July 19, 1976)

1. Introduction

A randomized response technique for eliminating evasive answer bias has been developed by Warner [6] and several variants of the Warner technique have been suggested [1], [3].

These randomized response techniques need some randomizing device; for example, a deck of cards, a sealed plastic box with colored beads [3], or a flask containing different colored balls (the Hopkins' randomizing device III) [4].

In this paper we propose a randomized response technique without making use of any randomizing device. Such a technique will be applicable not only to face-to-face interview survey, but also easily to self-administrated, telephone interview and mail surveys.

There are two essential properties of randomizing devices which are used in randomized response techniques;

(i) the results of trials with a randomizing device by a respondent are independent of any characteristic of him;

(ii) the probabilities of the results of trials (for example, in [6] the probability that the spinner points the letter A or in [3] the probability that a bead of some color appears in a small window) are known (and also can be controlled).

After quick consideration it might come into our mind to use some auxiliary question instead of randomizing devices. An auxiliary question is shown to the respondent. For example, it may be the question, "Which do you prefer 'spring' or 'autumn'?" After having made a silent answer to this auxiliary question the respondent is asked to say 0 or 1 according to the following list:

The List;

If you prefer 'spring' and
- you have A, then your answer is 0.
- you do not have A, then your answer is 1.

If you prefer 'autumn' and
{ you have A, then your answer is 1.
{ you do not have A, then your answer is 0.

In the list A denotes the sensitive attribute that we want to estimate the proportion of subgroup having it in the population.

If the preference of one season to the other is independent of whether the respondent has A or not and the proportion p of the preference of 'spring' to 'autumn' in the population is known, then this technique has the same mathematical model as the Warner's technique. If the independency still holds, then it does not raise any difficult problem that p is unknown, because another independent sample will give an estimate of p. However, if the independency can not be assumed, then the estimation of the proportion of subgroup having the attribute A in the population becomes impossible.

As shown in [5] it is very difficult or might be impossible to find any auxiliary question which is independent of some attribute.

On the other hand, the technique which we are going to propose in this paper does not assume the independency of the auxiliary question and the attribute.

2. Description of our technique

Let $\sharp$ be the proportion of subgroup having a sensitive attribute A in the population. The purpose is to estimate $\sharp$. Three independent samples are required with this technique. For simplicity we assume that each sample is drawn with replacement from the population. Each respondent is asked to make a silent choice of one item from among the three items, say $B1$, $B2$ and $B3$, for example, one color from among the three colors of violet, blue and green, but not to tell the color to the interviewer and to answer 0 or 1 according to the following lists:

The List for the First Sample;

If you have chosen 'violet' and
{ you have A, then your answer is "0".
{ you do not have A, then your answer is "1".
If you have chosen 'blue' and
{ you have A, then your answer is "1".
{ you do not have A, then your answer is "0".
If you have chosen 'green' and
{ you have A, then your answer is "1".
{ you do not have A, then your answer is "0".

The List for the Second Sample;

If you have chosen 'violet' and
{ you have A, then your answer is "1".
{ you do not have A, then your answer is "0".
If you have chosen 'blue' and
{ you have A, then your answer is "0".
{ you do not have A, then your answer is "1".
If you have chosen 'green' and
{ you have A, then your answer is "1".
{ you do not have A, then your answer is "0".

The List for the Third Sample;

If you have chosen 'violet' and
{ you have A, then your answer is "1".
{ you do not have A, then your answer is "0".
If you have chosen 'blue' and
{ you have A, then your answer is "1".
{ you do not have A, then your answer is "0".
If you have chosen 'green' and
{ you have A, then your answer is "0".
{ you do not have A, then your answer is "1".

These are summarized as follows:

attribute / color	1st sample		2nd sample		3rd sample	
	A	not A	A	not A	A	not A
violet	0	1	1	0	1	0
blue	1	0	0	1	1	0
green	1	0	1	0	0	1

Instead of the three colors of violet, blue and green we may, of course, use other kinds of items: i.e., the three animals of leopard, tiger and lion; the three countries of France, Canada and England; the three numbers of 3, 5 and 7, etc.

3. Estimation of

Let

$p(A, i)$=the proportion in the population of the persons who have the attribute A and will choose Bi $(i=1, 2, 3)$,

$p(\bar{A}, i)$=the proportion in the population of the persons who do not have the attribute A and will choose Bi $(i=1, 2, 3)$,

$n(i)$=the size of the ith sample $(i=1, 2, 3)$,
$y(i)$=the number of the respondents answering "1" in the ith sample $(i=1, 2, 3)$

and

$$q(i)=\#+\bar{p}(A, i)-p(A, i),\ (i=1, 2, 3),$$

or $q(i)$ is a probability that a respondent in the ith sample answers "1".

We have

$$\#=\sum_{i=1}^{3} q(i)-1$$

and

$$\mathrm{E}(y(i))=n(i)q(i)\ ,\qquad (i=1, 2, 3)\ .$$

The maximum likelihood estimator, say T, of the proportion $\#$ is given by

$$T=\sum_{i=1}^{3} y(i)/n(i)-1\ . \tag{0}$$

This is an unbiased estimator of $\#$ with the variance

$$\mathrm{Var}(T)=\sum_{i=1}^{3} q(i)(1-q(i))/n(i)\ . \tag{1}$$

If we put $n(i)=n/3$, $(i=1, 2, 3)$, then (1) is reduced to

$$\mathrm{Var}(T)=\frac{3}{n}\sum_{i=1}^{3} q(i)(1-q(i))=\frac{3}{n}\left(1-\#\bar{\#}-\sum_{i=1}^{3} d(i)^2\right)\ , \tag{2}$$

where $\bar{\#}=1-\#$ and $d(i)=p(A, i)-p(\bar{A}, i)$, $(i=1, 2, 3)$.

As noticed above we do not assume the independence of the attribute A and the choice of Bi's, and it involves a lot of risk to assume the independence without careful considerations [5].

Throughout the remaining part of this section we assume that $n(i)=n/3$, $(i=1, 2, 3)$. From $\sum_{i=1}^{3} d(i)=\#-\bar{\#}$, we have

$$\begin{aligned}\sum_{i=1}^{3} d(i)^2&=\frac{1}{3}\left[\left(\sum_{i=1}^{3} d(i)\right)^2+\sum\sum_{i<j}(d(i)-d(j))^2\right]\\&=\frac{1}{3}(\#-\bar{\#})^2+\frac{1}{3}\sum\sum_{i<j}(d(i)-d(j))^2\ .\end{aligned}$$

Thus we have

$$\sum_{i=1}^{3} d(i)^2\geqq\frac{1}{3}(\#-\bar{\#})^2 \tag{3}$$

and the equality holds when

$$d(1)=d(2)=d(3)=(\#-\bar{\#})/3\ . \tag{4}$$

On the other hand, since

$$\begin{aligned}\sum_{i=1}^{3} d(i)^2&=\sum_{i=1}^{3}(p(A,i)-p(\bar{A},i))^2\\&=\sum_{i=1}^{3}p(A,i)^2+\sum_{i=1}^{3}p(\bar{A},i)^2-2\sum_{i=1}^{3}p(A,i)p(\bar{A},i)\\&\leqq\sum_{i=1}^{3}p(A,i)^2+\sum_{i=1}^{3}p(\bar{A},i)^2\\&\leqq\Big(\sum_{i=1}^{3}p(A,i)\Big)^2+\Big(\sum_{i=1}^{3}p(\bar{A},i)\Big)^2\\&=\#^2+\bar{\#}^2\ ,\end{aligned}$$

we have

$$\sum_{i=1}^{3} d(i)^2\leqq\#^2+\bar{\#}^2 \tag{5}$$

and the equality holds when

$$p(A,i)=\# \qquad \text{for some } i$$

and

$$p(\bar{A},j)=\bar{\#} \qquad \text{for some } j\neq i\ . \tag{6}$$

From (2), (3) and (5), we get the following inequality for the variance of T:

$$\frac{3}{n}\#\bar{\#}\leqq\mathrm{Var}\,(T)\leqq\frac{1}{n}(2+\#\bar{\#}) \tag{7}$$

and the first equality holds when (4) holds and the second equality holds when (6) holds.

In [1] if we put p (=the probability that the spinner points to A) $=1/3$, then the variance of Warner's estimator is given by $(2+\#\bar{\#})/n$. From (7) the variance of T is not greater than the variance of Warner's estimator in the case of $p=1/3$.

If the attribute A and the choice of Bi is independent, we have

$$\sum_{i=1}^{3} d(i)^2=\sum_{i=1}^{3}(p(A,i)-p(\bar{A},i))^2=(\#-\bar{\#})^2\sum_{i=1}^{3}a_i^2\ , \tag{8}$$

where $a_i=p(A,i)/\#=p(\bar{A},i)/\bar{\#}$ and therefore $\sum_{i=1}^{3}a_i=1$. In this case, we have

$$\frac{1}{3}(\#-\bar{\#})^2\leqq\sum_{i=1}^{3}d(i)^2\leqq(\#-\bar{\#})^2\ , \tag{9}$$

and the equality of the right-hand side holds when

$$a_i=1 \qquad \text{for some } i\,.$$

In this case we have from (2)

$$\mathrm{Var}\,(T)=\frac{3}{n}\left(1-\#\bar{\#}-(\#-\bar{\#})^2\sum_{i=1}^{3}a_i^2\right)$$

and from (9)

$$\frac{9\#\bar{\#}}{n}\leqq\mathrm{Var}\,(T)\leqq\frac{1}{n}(2+\#\bar{\#})\,.$$

4. Similar techniques

There may be many similar randomized response techniques without making use of any randomizing device to the one of Section 2. We present here two of them.

In the first one two independent samples are required. Each respondent is asked to make a silent choice of one out of two possibilities, for example, one out of "autumn" or "spring", but not to tell which one to the interviewer and to answer 0 or 1 according to the following lists:

The List for the First Sample;

If you have chosen 'autumn' and
- you have A, then your answer is "1".
- you do not have A, then your answer is "0".

If you have chosen 'spring', then your answer is "1" whether you have A or not.

The List for the Second Sample;

If you have chosen 'autumn', then your answer is "1" whether you have A or not.

If you have chosen 'spring' and
- you have A, then your answer is "1".
- you do not have A, then your answer is "0".

Let $n(i)$ be the size of the ith sample and $y(i)$ the number of individuals answering 1 in the ith sample ($i=1, 2$). The maximum likelihood estimator $T1$ of the proportion $\#$ of A in the population is given by

$$T1=\frac{y(1)}{n(1)}+\frac{y(2)}{n(2)}-1\,.$$

We have

$$\mathrm{E}(T1)=\#$$

and

$$\mathrm{Var}(T1)=\sum_{i=1}^{2}\frac{q(i)(1-q(i))}{n(i)}\ ,$$

where $q(i)=1-p(\bar{A}, i)$, $p(\bar{A}, 1)$ is the proportion in the population of the persons who do not have the attribute A and will choose 'autumn' and $p(\bar{A}, 2)$ is the proportion in the population of the persons who do not have the attribute A and will choose 'spring'. In the case of $n(1)=n(2)=n/2$, we have

$$\mathrm{Var}(T1)=\frac{2}{n}(\#\bar{\#}+2p(\bar{A}, 1)p(\bar{A}, 2))$$

and

$$\frac{2\#\bar{\#}}{n}\leqq\mathrm{Var}(T1)\leqq\frac{1}{n}(1+2\#\bar{\#})\ .$$

In this technique it should be noticed that the answer "0" includes only the possibility of 'not A', though the answer "1" includes both possibilities of 'A' and 'not A'. Therefore if both to have and not to have the attribute A are sensitive, this technique is not applicable.

The second similar technique is the same as the one of Section 2 except the lists. The Lists are summarized as follows:

attribute / item	1st sample		2nd sample		3rd sample	
	A	not A	A	not A	A	not A
$B1$	1	0	0	0	1	1
$B2$	1	1	1	0	0	0
$B3$	0	0	1	1	1	0

Each respondent is asked to make a silent choice of one item from among the three items of $B1$, $B2$ and $B3$ and then answer "0" or "1" according to the above lists.

The maximum likelihood estimator $T2$ of $\#$ is also given by (0);

$$T2=\sum_{i=1}^{3}y(i)/n(i)-1$$

and we have

$$\mathrm{E}(T2)=\#$$

and

$$\mathrm{Var}\,(T2)=\sum_{i=1}^{3} q(i)(1-q(i))/n(i)\ ,$$

where $q(i)$ is the sum of the proportions in the population which correspond to the cells including "1" of the List for the ith sample ($i=1$, 2, 3).

5. Comments

This technique is based on the implicit assumptions that

(i) the respondents do not change the choice of Bi's after having seen their lists, and

(ii) they answer honestly and correctly according to their lists.

If we use the items as Bi's in which most people will choose one of them or most people will not choose one of them, or the choice of which will be highly correlated with the attribute, then the privacies of the respondents can not be maintained. It will be advisable to use the items as Bi's each of which has approximately equal chance being chosen and which are not so highly correlated with the attribute.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Abdul-Ela, Abdel-Latif A., Greenberg, Bernard G. and Horvitz, Daniel G. (1967). A multi-proportions randomized response model, *J. Amer. Statist. Ass.*, **62**, 990-1008.

[2] Gould, A. L., Shah, B. V. and Abernathy, J. R. (1969). Unrelated question randomized response techniques with two trials per respondent, *Proceedings of the Social Statistics Section, American Statistical Association*, 351-359.

[3] Horvitz, Daniel G., Shah, B. V. and Simmons, Walt R. (1967). The unrelated question randomized response model, *Proceedings of the Social Statistics Section, American Statistical Association*, 65-72.

[4] Liu, P. T., Chow, L. P. and Mosley, W. H. (1975). Use of the randomized response technique with a new randomizing device, *J. Amer. Statist. Ass.*, **70**, 329-332.

[5] Suzuki, T., Takahasi, K. and Sakasegawa, H. (1976). Some notes on randomized response techniques, *Proc. Inst. Statist. Math.*, **24**, 1-13.

[6] Warner, Stanley L. (1965). Randomized response: A survey technique for eliminating evasive answer bias, *J. Amer. Statist. Ass.*, **60**, 63-69.