Eliciting Information on Sensitive Matters Without Inviting Respondents' Ire

Randomized Response Techniques

V R Padmawar

V R Padmawar teaches statistics at the Indian Statistical Institute, Bangalore. His research interests lie in the area of sampling theory.

It is now part of folklore that eliciting information on sensitive matters is riddled with nonresponse or untruthful response. Typically respondents fear that the traditional methods based on direct response do not protect their privacy. This fear results in nonresponse or untruthful response. We introduce in this article ingenious methods based on Randomized Response techniques. Such methods instill confidence in the respondent to furnish truthful answers to queries even on sensitive matters. We will also see how to carry out statistical inference using these methods.

Keywords
Sample surveys, simple random sampling, randomized response.

1. Introduction

Imagine a large college campus where the authorities have reasons to suspect that the young students indulge in cigarette smoking. They would like to get fairly accurate idea about the proportion π_A of students that indulge in smoking. One does not have to be a rocket scientist to realize that a young college student who indulges in smoking is very unlikely to own up to it if confronted by the authorities directly. In such a situation how do we estimate π_A? In this article we try to find answers to this question.

We are all aware of the need and importance of *survey sampling*. The advantages of survey sampling over *census* or *complete enumeration* are captured by the phrase 'better ESSAY'. Here the acronym ESSAY stands for Economy, Speed, Scope and Accuracy. Many of us are

familiar with at least the basic *sampling theory*. Interested reader may refer to [1, 2] for Sampling Techniques. In the present article, however, we deal only with the problem of nonresponse in certain sensitive survey situations.

The advantages of sample survey over *census* or *complete enumeration* are captured by the phrase 'better ESSAY'. Here the acronym ESSAY stands for Economy, Speed, Scope and Accuracy.

In socioeconomic surveys the investigators often need to collect data on matters that are either personal or sensitive. Respondents are averse to the idea of furnishing information on economic features like savings, tax evasions, illegal trade practises; social features like addiction to alcohol or drugs; health characteristics like suffering from certain types of diseases, sexual preferences, history of induced abortions, etc. Attempts to elicit direct information on such sensitive, personal or stigmatizing features invariably end up in nonresponse or evasive untruthful response. The college campus mentioned in the first paragraph is a case in point. Therefore ingenious methods are required to collect data in such situations, randomized response (RR) being one such device.

The setbacks that a survey statistician has to suffer while collecting data on some sensitive issues are well documented in the literature vide [3]. The book also discusses various randomized response models evolved over a period of two decades since the first and now well-known Warner [4] model. There is now a comprehensive book [5] on RR.

The main purpose of this article is to introduce the reader to the field of randomized response by way of describing a few models with examples. We first describe Warner's model and unrelated question model. We then consider models due to Yu, Tian and Tang [6], where randomization device is *not* required. We finally consider a simple example under Padmawar–Vijayan model. On the application front, we make an attempt to provide some numerical examples in real life situations where it

The main purpose of this article is to introduce the reader to the field of randomized response by way of describing a few models with examples.

Collecting data on sensitive or stigmatizing variables such as alcohol consumption or tax evasion, by the traditional direct response methods, is riddled with evasive untruthful response.

would be appropriate to use these models. We have kept the article as free of technical details as possible. The present article is, in fact, a *nontechnical avatar* of the article [7], where an initiated reader may find the immediate technical details as well as a discussion on performance and comparison of different models and estimators. Other useful references may be found in Suggested Reading.

2. Preliminaries and The Models

Let $U = \{1, 2, \cdots, N\}$ be a finite population under consideration. Let the study variate y take value y_i on unit i, $1 \leq i \leq N$. In particular, y may be a dichotomous variable taking values 1 and 0. Suppose we are interested in estimating the population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ of variable y. There are many strategies for estimating $\overline{Y}$ in traditional survey sampling theory. If, however, y happens to be stigmatizing or sensitive like extent of tax evasion or amount of alcohol consumption etc., the traditional direct response methods are riddled with nonresponse or evasive response. We resort to indirect methods of collecting data. Randomized response (RR) is an important such method.

S L Warner (1965) was the first one to advocate the use of randomized response techniques to elicit information on sensitive or stigmatizing variables like some kind of *addiction* or *tax evasion*, to avoid evasive untruthful response.

A typical traditional RR method may be described as a procedure in which the statistician chooses x, an innocuous but 'similar valued' variable whenever the variable y is sensitive. A respondent reports either the sensitive study variate value y or the innocuous variate value x with some prespecified probabilities. The statistician, does not, however, know whether the reported value corresponds to the variable x or the variable y. This may be viewed as the statistician receiving either the signal y or the noise x with some specified probabilities. Here randomization is used by the respondent to determine whether to report signal y or the noise x. *Unrelated question model,* which would be discussed later, is an example of traditional RR model.

Padmawar and Vijayan [8] suggested a different method of randomization. The respondent reports $y + x$, where y is the true value and x is a value taken by a suitably chosen random variable X whose probability distribution is known. Note that X is innocuous but it need not be 'similar valued'. In this method the statistician receives a 'mixture of signal y and noise x'. Here randomization is used to generate an observation or value x. The statistician does not ever get to know either the true value y or the randomly generated value x since the respondent reports only $y + x$. The statistician has to 'extract' the signals or 'eliminate' the noise to be able to carry out the estimation or inference.

The statistician does not ever get to know either the true stigmatizing value *y* (*signal*) or the randomly generated innocuous value *x* (*noise*) since the respondent reports only *y* + *x*. The statistician has to 'extract' the *signals* or 'eliminate' the *noise* to be able to carry out the inference.

The characteristic differences among direct response method, a typical traditional RR model and *Padmawar–Vijayan model* may be understood well by the following example.

Suppose a pediatrician has to treat a child. We all know that children hate bitter pills (for that matter who doesn't?). Direct response method is like the exercise of directly administering the bitter medicine to the child. We all know how difficult such an exercise is.

To make the exercise of administering the bitter pill less painful the doctor designs the following procedure. She takes m tiny identical boxes and keeps medicine tablets in some of them and chocolates in the rest in some known ratio $p : 1 - p$. The child is then asked to pick one of the boxes and advised to consume the contents of the box chosen. The child still has to swallow the bitter pill, which she still continues to detest, with probability p. She may not actually swallow the bitter pill if left entirely to herself. This is akin to a typical traditional RR method.

Consider now yet another scenario. The pediatrician offers chocolate coated bitter pill to the child and asks her to swallow that without chewing. The child would

Yu, Tian and Tang suggested methods that use *implicit randomization* therefore field workers do not require to carry any randomization devices to collect data.

oblige being blissfully unaware of the bitter kernel of the chocolate-like object. This is akin to the randomization method used in [8].

Randomized response methods, nonetheless, need to use randomization devices to carry out randomization. Yu, Tian and Tang [6], however, suggested methods that use implicit randomization therefore do not require any randomization devices.

In what follows we describe with illustrations some of the models used in literature that deal with sensitive variables.

2.1 *Warner's Model*

If y is a $1-0$ variable like, say, possessing an attribute $\mathcal{A}$ or not, then the population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ simplifies to proportion π_{A} say, of individuals possessing the attribute $\mathcal{A}$. If $\mathcal{A}$ is sensitive like some kind of *addiction* or *tax evasion,* to avoid nonresponse or evasive untruthful response, Warner [4] advocated the use of RR techniques.

Let $U = \{1,\ 2,\ \cdots,\ N\}$ be the population under consideration, for example all students on roll in a college. Let $A \subset U$ be the set of individuals possessing the stigmatizing attribute $\mathcal{A}$. Under Warner's model a respondent answers one of the following questions, without the knowledge of the interviewer, with preassigned probabilities p and $1-p$. The first question being: Do you belong to A? The other question being: Do you belong to A^C? Here $A^C = U - A$ is the set of individuals not possessing the attribute $\mathcal{A}$.

Such a randomization can be achieved using a deck of cards. Consider a deck that has two types of cards in the known ratio $p : 1-p$; $0 \leq p \leq 1$ but $p \neq \frac{1}{2}$. The first kind has the question *'Do you belong to A*?' written on it. The second kind has the question *'Do you belong to* A^C?' written on it. A respondent picks a card at random

from the well-shuffled deck and answers corresponding question in '*yes*' or '*no*' as the case may be, without the knowledge of the interviewer. In other words, the interviewer does not know to which question the respondent has furnished the answer. All respondents in the sample of size n say, furnish their answers likewise, albeit independently. Based on this information the survey statistician has to estimate the unknown proportion π_A.

One may use a bottle containing beads of two colours or an urn containing balls of two colours and so on, in place of the deck of cards for randomization.

One may use a bottle containing beads of two colours or an urn containing balls of two colours and so on, in place of the deck of cards for randomization.

Suppose we draw a sample of n respondents from the population U using Simple Random Sampling With Replacement (SRSWR). Under Simple Random Sampling each unit in the population has the same chance of being included in the sample. (One may again refer to the classic *Sampling Techniques* by W G Cochran [1] or Delampady and Padmawar [2] for the details of SRSWR). Let the number of '*yes*' responses be m.

Under *simple random sampling* each member of the population has the same chance of being included in the sample.

One can argue that an *unbiased* estimator for π_A is given by

$$\widehat{\pi}_{AW} = \frac{\frac{m}{n} - (1-p)}{(2p-1)}, \qquad (1)$$

$p \neq \frac{1}{2}$.

A statistician is also expected to report the *accuracy* of the estimator $\widehat{\pi}_{AW}$ in (1). We therefore obtain $\sqrt{\widehat{Var}\,(\widehat{\pi}_{AW})}$, estimated *standard error* of the estimator $\widehat{\pi}_{AW}$, using the formula

$$\widehat{Var}\,(\widehat{\pi}_{AW}) = \frac{\frac{m}{n}(1-\frac{m}{n})}{(n-1)(2p-1)^2}. \qquad (2)$$

Smaller the *standard error* more *accurate* the estimator.

Smaller the standard error more accurate the estimator.

See *Box* 1 for an example that illustrates Warner's model.

Box 1. Example illustrating Warner's Model

Consider a large college campus where the authorities suspect that the young students indulge in cigarette smoking. They would like to get a fairly accurate idea about the proportion π_A of such students.

Let A be the set of students in the college who indulge in smoking. Suppose we take a *random sample* of 50 students from the college. A student in the sample is presented with a deck of say 40 cards. The deck has 15 cards with the question *'Do you belong to* A*?'* written on them and 25 cards with the question *'Do you belong to* A^C*?'* written on them. The student shuffles the deck and picks one of the cards at random. The student furnishes answer '*yes*' or '*no*' as the case may be, to the question selected, without the knowledge of the interviewer. All 50 respondents in the sample, furnish their answers likewise, albeit independently. Based on this information the survey statistician has to estimate the unknown proportion π_A. Suppose the number of *yes* responses, among the total of 50 responses, is 27. We use the formula (1) to get an estimate of π_A. In our example we have $p = \frac{15}{40} = \frac{3}{8}$, p being the proportion of cards in the deck bearing the question *'Do you belong to* A*?'*; n, the sample size or the number of students interviewed, is 50; m, the number of *yes* responses among the total of 50 responses, is 27. Plugging in these values in (1), we get

$$\begin{aligned} \widehat{\pi}_{AW} &= \frac{\frac{m}{n} - (1-p)}{(2p-1)} \\ &= \frac{\frac{27}{50} - (1-\frac{3}{8})}{(2\times\frac{3}{8}-1)} \\ &= 0.34 \end{aligned}$$

Thus an estimated 34% of the college students indulge in smoking.

Using the formula (2),

$$\begin{aligned} \widehat{Var}\,(\widehat{\pi}_{AW}) &= \frac{\frac{m}{n}(1-\frac{m}{n})}{(n-1)(2p-1)^2} \\ &= \frac{\frac{27}{50}(1-\frac{27}{50})}{(50-1)(2\times\frac{3}{8}-1)^2} \\ &= 0.08111 \end{aligned}$$

Therefore the estimated standard error is $\sqrt{0.08111} = 0.2848$.

2.2 *Unrelated Question Model*

Warner's model definetly protects the privacy of the respondent yet the questions *'Do you belong to* A*?'* and *'Do you belong to* A^C*?'* both pertain to the same stigmatizing attribute $\mathcal{A}$. In unrelated question model, due

to Greenberg *et al* [9], however, the second question pertains to an innocuous attribute $\mathcal{B}$. Let $B \subset U$ be the set of individuals possessing attribute $\mathcal{B}$ and π_{B} be the proportion of individuals possessing attribute $\mathcal{B}$. Often in practice π_{B} is known. For instance, the attribute $\mathcal{B}$ may be 'being born in the month of January' or 'being fond of soccer'. We assume that the two attributes $\mathcal{A}$ and $\mathcal{B}$ are *independent.* The innocuous nature of attribute $\mathcal{B}$ seems to instill more confidence in the respondent to answer truthfully.

In *unrelated question model*, the second question pertains to an innocuous attribute $\mathcal{B}$ like 'being born in the month of January' or 'being fond of soccer'.

In unrelated question model the respondent answers one of the two questions, without the knowledge of the interviewer, with preassigned probabilities p and $1-p$.

Moreover, such a randomization can be achieved, as in Warner's model using a deck of cards with simple modifications. Consider a deck that has two types of cards in the known ratio $p : 1-p$. The first kind has the question *'Do you belong to A*?' written on it. The second kind has the question *'Were you born in the month of January*?' written on it. A respondent picks a card at random from the well-shuffled deck and answers corresponding question in '*yes*' or '*no*' without the knowledge of the interviewer. All respondents in the sample of size n say, furnish their answers likewise, albeit independently. Based on this information the survey statistician has to estimate the unknown proportion π_{A}. Here π_{B} is assumed to be available to the survey statistician. It is possible to deal with the situation when π_{B} is not known by taking two independent samples of respondents.

Here π_{B} is assumed to be known. It is possible to deal with the situation when π_{B} is not known by taking two independent samples of respondents.

Suppose we draw a sample of n respondents from the population U using SRSWR and that the number of '*yes*' responses is m. It may be argued that an unbiased estimator for π_{A} is given by

$$\widehat{\pi}_{\mathrm{AU}} = \frac{\frac{m}{n} - (1-p)\pi_{\mathrm{B}}}{p}, \qquad (3)$$

$p > 0$.

We are again expected to report the *accuracy* of our estimator $\widehat{\pi}_{\mathrm{AU}}$ in (3). We therefore obtain $\sqrt{\widehat{Var}\,(\widehat{\pi}_{\mathrm{AU}})}$, estimated *standard error* of our estimator $\widehat{\pi}_{\mathrm{AU}}$, using the formula

$$\widehat{Var}\,(\widehat{\pi}_{\mathrm{AU}}) = \frac{\frac{m}{n}(1-\frac{m}{n})}{(n-1)p^2}. \qquad (4)$$

Smaller the *standard error* more *accurate* the estimator. See *Box* 2 for an example.

Box 2. Example for Unrelated Question Model

Here we use the same basic setup as given in *Box* 1.

Let again A be the set of students possessing the stigmatizing attribute of smoking. Let B be the set of students possessing the innocuous attribute of being born in the month of January. Let π_{B} be the proportion of students in the college born in the month of January. We can find π_{B} looking into the records of the college or we can take $\pi_{\mathrm{B}} = \frac{1}{12}$ from probabilistic considerations. We assume that the two attributes $\mathcal{A}$ and $\mathcal{B}$ are *independent.* Suppose we take a random sample of 50 students from the college. We can again use a deck of cards for randomization, or acquaint ourselves with a different randomization device in the form of a corked bottle with opaque body and a transparent neck. The bottle contains certain number of balls. The neck of the bottle is just enough to hold exactly one ball when the bottle is inverted. The bottle used in the trial contains say total of 60 balls, 12 red and 48 green, but otherwise identical. When the bottle is shaken and inverted exactly one ball is seen in the neck of the bottle. The red colour corresponds to attribute $\mathcal{A}$ and the green colour corresponds to attribute $\mathcal{B}$.

A student in the sample is presented with the bottle. The student shakes the bottle well and inverts it. If the ball in the neck is red the student responds to the question *'Do you belong to A*?' If the ball in the neck is *green* the student responds to the question *'Were you born in the month of January?'* The student furnishes answer '*yes*' or '*no*' as the case may be, to the question corresponding to the *colour* seen in the neck, without the knowledge of the interviewer. All 50 respondents in the sample, furnish their answers likewise, albeit independently. Based on this information the survey statistician has to estimate the unknown proportion π_{A}. Suppose the number of *yes* responses, among the total of 50 responses, is 8. We use the formula (3) to get an estimate of π_{A}. In our example we have $p = \frac{12}{60} = \frac{1}{5}$, p being the proportion of red balls in the bottle; n, the sample size or the number of students interviewed, is 50; m, the number of *yes* responses among the total of 50 responses, is 6. Finally, $\pi_{\mathrm{B}} = \frac{1}{12}$. Plugging in these values in (3), we get

Box 2 continued...

Box 2 continued...

$$\begin{aligned}\widehat{\pi}_{\mathrm{AU}} &= \frac{\frac{m}{n} - (1-p)\pi_{\mathrm{B}}}{p} \\ &= \frac{\frac{6}{50} - (1-\frac{1}{5}) \times \frac{1}{12}}{\frac{1}{5}} \\ &= 0.2667\ .\end{aligned}$$

Thus an estimated 26.67% of the college students indulge in smoking.

We further, find the value of estimated *standard error* of $\widehat{\pi}_{\mathrm{AU}}$ using the formula (4).

$$\begin{aligned}\widehat{Var}\,(\widehat{\pi}_{\mathrm{AU}}) &= \frac{\frac{m}{n}(1-\frac{m}{n})}{(n-1)p^2} \\ &= \frac{\frac{6}{50}(1-\frac{6}{50})}{(50-1)(\frac{1}{5})^2} \\ &= 0.053878\ .\end{aligned}$$

Therefore the estimated standard error is $\sqrt{\widehat{Var}\,(\widehat{\pi}_{\mathrm{AU}})} = \sqrt{0.053878} = 0.23212$.

2.3 *Yu–Tian–Tang Models*

Subsequent to Warner's model and unrelated question model various RR models were proposed and studied. Interested readers may refer to [3] for details. Many of these models need appropriate randomization devices. Yu, Tian and Tang [6] suggested methods that use *implicit randomization*, which obviate the need to use randomization devices. They, however, continue to use innocuous attributes.

2.3.1 ***Circle and Triangle Model***: As in the Warner's model, let $A \subset U$ be the class of individuals possessing the sensitive attribute $\mathcal{A}$ under consideration, say addiction to drugs. The purpose again is to estimate the unknown proportion π_{A} of drug addicts. Let being born in the interval [*August* 1, *December* 31] be our innocuous attribute $\mathcal{B}$. We assume that the proportion π_{B} of individuals born in the interval [*August* 1, *December* 31] is known. We also assume that $\mathcal{A}$ and $\mathcal{B}$ are independent. Let $B \subset U$ be the class of individuals born in the interval [*August* 1, *December* 31] . Consider the following scheme of things.

Implicit randomization obviates the need to use randomization devices, however, we need to have innocuous attributes.

Figure 1. Circle and triangle model.

$\mathcal{B} \longrightarrow$ $\mathcal{A} \downarrow$	B^C	B
A^C	O	$\cdot$
A	$\cdot$	$\cdot$

In the personal interviews, the interviewee is asked to put a tick (✓) in the circle if s/he belongs to A^C(not a drug addict) and was born in the interval [*January* 1, *July* 31] else put a tick (✓) in the triangle formed by the dots in the remaining three cells (*Figure* 1). Note that the circle option is anyway innocuous and the triangle option is not a complete give-away. This, we believe, is encouraging enough for the respondents to report truthfully.

All respondents in the sample of size n say, furnish their answers likewise, albeit independently.

Suppose we draw a sample of n respondents from the population U using SRSWR and that the number of ticks (✓) in the circle is m. It may be argued that an unbiased estimator for π_{A} is given by

$$\widehat{\pi}_{\mathrm{AYTT1}} = 1 - \frac{\frac{m}{n}}{(1-\pi_{\mathrm{B}})}, \tag{5}$$

$\pi_{\mathrm{B}} < 1$.

We again report the *accuracy* of our estimator $\widehat{\pi}_{\mathrm{AYTT1}}$ in (5). We therefore obtain $\sqrt{\widehat{Var}\,(\widehat{\pi}_{\mathrm{AYTT1}})}$, estimated *standard error* of our estimator using the formula

The circle option is innocuous and the triangle option is not complete give-away. This instills enough confidence in the respondent to report truthfully.

$$\widehat{Var}\,(\widehat{\pi}_{\mathrm{AYTT1}}) = \frac{\frac{m}{n}(1-\frac{m}{n})}{(n-1)(1-\pi_{\mathrm{B}})^2}. \tag{6}$$

For an example illustrating *circle and triangle model,* see *Box* 3.

Box 3. Example for Circle and Triangle Model

Suppose there is an association of professionals. We would like to know the proportion π_{A} of its members that indulge in unlawful tax evasion. As in Warner's model, let A be the set of members possessing the sensitive attribute $\mathcal{A}$, namely indulging in unlawful tax evasion. The purpose again is to estimate the unknown proportion π_{A}. Let being born in the interval [*August* 1, *December* 31] be our innocuous attribute $\mathcal{B}$. Let B be the set of members born in this interval . We assume that the proportion π_{B} of such individuals born is known from the membership records or it may be taken as $\frac{5}{12}$ from probabilistic considerations. We assume that $\mathcal{A}$ and $\mathcal{B}$ are independent. Consider the following scheme of things.

For the personal interviews, suppose we take a random sample of 100 members. A respondent is given a sheet with *Figure* 1. The respondent is asked to put a tick (✓) in the circle if s/he belongs to A^C(does not indulge in tax evasion) and was born in the interval [*January* 1, *July* 31] , else put a tick (✓) in the triangle formed by the dots.

All 100 respondents in the sample, furnish their answers likewise, albeit independently. We use the formula (3) to get an unbiased estimate of π_{A}. Suppose the total number of ticks (✓) in the circle among the 100 respondents is 51. In our example $m = 51$, $n = 100$ and $\pi_{\mathrm{B}} = \frac{5}{12}$. We plug in these values in the formula (5) and get

$$\begin{aligned}\widehat{\pi}_{\mathrm{AYTT1}} &= 1 - \frac{\frac{m}{n}}{(1-\pi_{\mathrm{B}})} \\ &= 1 - \frac{\frac{51}{100}}{(1-\frac{5}{12})} \\ &= 0.1257\,.\end{aligned}$$

Thus an estimated 12.57% of the members of the association indulge in tax evasion. The estimated standard error is (using (6)) $\sqrt{\widehat{Var}\,(\widehat{\pi}_{\mathrm{AYTT1}})} = \sqrt{0.0071482} = 0.084547$.

2.3.2 ***Diagonal Model***: Yu, Tian and Tang [6] also suggested the following model which is similar, in spirit, to circle and triangle model.

In the personal interviews, the interviewee is asked to draw the diagonal containing her/his true value, i.e., either join the circles or dots depending on the respondent's true value (*Figure* 2). Neither diagonal is a complete give-away. This is conducive enough for the respondents to report truthfully. One wants to evolve, as before, a procedure to estimate π_{A}.

Neither diagonals are complete give-away. This instills enough confidence in the respondent to report truthfully.

Figure 2. Diagonal model.

$\mathcal{A}\downarrow$ \ $\mathcal{B}\longrightarrow$	B^C	B
A^C	○	·
A	·	○

Suppose we draw a sample of n respondents from the population U using SRSWR and that the number of diagonals joining the circles is m. An interested reader may be urged to establish that an unbiased estimator for π_{A} is given by

$$\widehat{\pi}_{\mathrm{AYTT2}} = \frac{\frac{m}{n} - (1 - \pi_{\mathrm{B}})}{2\pi_{\mathrm{B}} - 1}, \tag{7}$$

$\pi_{\mathrm{B}} \neq \frac{1}{2}$.

An example of *diagonal model* would be *similar* to that of *circle and triangle model.*

2.6 *Padmawar–Vijayan Model*

All the models considered so far can be used only for attributes, that is, when y is a $1-0$ variate. They cannot be used to estimate the population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ when, for instance, y is number of induced abortions or y is amount of tax evasion. Padmawar and Vijayan [8] suggested the following model.

Let y be the stigmatizing variable. Let X be a random variable with known distribution F say, that has finite *variance*. A respondent generates an observation on X without the knowledge of the interviewer and adds that to her/his y value and reports $Z = y + X$. Here adding noise to the signal is really like a sugar-coated pill. This encourages the respondent to respond with little or no inhibitions. All respondents in the sample of size n say, report likewise, albeit independently. Based on this information the survey statistician has to estimate the unknown population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Here y could be number of induced abortions and X may be normally distributed with known mean and variance, $\mathcal{N}(\nu, \xi^2)$ say.

In the Padmawar–Vijayan model adding *noise x* to the *signal y* is really like a sugar-coated pill which encourages the respondent to respond with little or no inhibitions.

Suppose we draw a sample of n respondents from the population U using SRSWR. Let $Z_1, Z_2, \cdots, Z_n$ be n responses from the selected individuals, where $Z_i = y_i + X_i$, 1 i n. (Here $X_1, X_2, \cdots, X_n$ are independent *and identically distributed* (iid) $N(\nu, \xi^2)$ random variables.) Let $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. It may be argued that an unbiased estimator for $\overline{Y}$ is given by

$$\widehat{\overline{Y}}_{\text{PV}} = \overline{Z} - \nu. \tag{8}$$

As before we may obtain $\sqrt{\widehat{Var}(\widehat{\overline{Y}}_{\text{PV}})}$, estimated *standard error* of our estimator, using the formula

$$\widehat{Var}(\widehat{\overline{Y}}_{\text{PV}}) = \frac{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2}{n(n-1)}. \tag{9}$$

We give an example of Padmawar–Vijayan model in *Box* 4.

Box 4. Example for Padmawar–Vijayan Model

We continue with the example in *Box* 3 so that we have the same basic set up.

In that example we had estimated the proportion π_{A} of members that indulge in unlawful tax evasion. Instead, we might be interested in estimating the population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, where y stands for amount of tax evaded by unlawful means during the last financial year. This estimate would give us an idea about per capita tax evaded by the association of professionals under consideration. From this we can also estimate the total amount of tax evaded by the members of the association. Note that none of the models considered hitherto can be used for this problem. This is because they are all useful only when y is a $1-0$ variable.

Suppose we take a random sample of 100 members of the association using SRSWR. As mentioned let y denote the amount of tax evaded by unlawful means (in thousand rupees) and x denote an observation on a $\mathcal{N}(-10, 1)$, *normal* random variable with known mean -10 and known variance 1, say. For this survey we need a portable device like a notebook equipped with a program that generates random numbers and more precisely generate observations from given distribution, here $\mathcal{N}(-10, 1)$. A respondent is asked to generate a value x using the portable equipment and add his/her true value y of amount of tax

Box 4 continued...

Box 4 continued...

evaded (in thousand rupees) and report $z = y + x$. Since the respondent is furnishing only z and not separate x and y values, the interviewer does not know the true value y, the amount of tax evaded, of the respondent. This ensures anonymity as well as maintains complete privacy. A typical z value might look like -6.923, say. All 100 respondents in the sample, furnish the data likewise, albeit independently. We use the formula (8) to get an unbiased estimate of $\overline{Y}$. Suppose $\overline{Z}$, the sample average of z values, is -2.249 and $s^2 = \frac{1}{99}\sum_{i=1}^{100}\left(Z_i - \overline{Z}\right)^2 = 82.739$. In our example $\nu = -10$. We plug in these values in the formula (8) and get

$$\begin{aligned}\widehat{\overline{Y}}_{\text{PV}} &= \overline{Z} - \nu \\ &= -2.249 - (-10) \\ &= 7.751\end{aligned}$$

Thus estimated average amount of tax evaded (in thousand rupees) for the association is 7.751. That is, on an average, a member of the association is estimated to save Rs $7,751$ on tax by unlawful means.

Finally, we estimate the *standard error* of our estimator $\widehat{\overline{Y}}_{\text{PV}}$. For this we use the formula (9). Here $n = 100$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2 = 82.739$

$$\begin{aligned}\widehat{Var}(\widehat{\overline{Y}}_{\text{PV}}) &= \frac{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2}{n(n-1)} \\ &= 0.82739\end{aligned}$$

Therefore estimated *standard error* $\sqrt{\widehat{Var}(\widehat{\overline{Y}}_{\text{PV}})}$ of our estimator $\widehat{\overline{Y}}_{\text{PV}}$ is given by

$$\sqrt{\widehat{Var}(\widehat{\overline{Y}}_{\text{PV}})} = \sqrt{.82739} \cong 0.91$$

that is, Rs 910.

Conclusions and Scope

We have thus far got a flavor of the topic randomized response. Warner was the pioneer of randomized response technique. However he used randomized response essentially pertaining to the same sensitive attribute. Unrelated question model truly exploited the idea pioneered by Warner by introducing an absolutely innocuous attribute. Unrelated question model randomizes between sensitive and innocuous attributes. This instills more

confidence in the respondent. Padmawar and Vijayan [8] used a different method of randomization. This has both psychological advantage over traditional RR models as well as theoretical advantage by way of providing an estimator with smaller variability, that is, better precision (vide [8]). We also noted that various RR methods, nonetheless, need to use randomization devices to carry out randomization. Yu, Tian and Tang [6], however, suggested methods that use implicit randomization therefore do not require any randomization devices. Finally, we noted that Warner's model, Unrelated question model as well as Yu-Tian-Tang models can be used only when the study variate y is $1-0$ valued, that is, when we have to deal with a sensitive attribute. Padmawar-Vijayan model, however, is not riddled with such a limitation and can be used for any real valued sensitive variable y.

Unrelated question model truly exploited the idea pioneered by Warner by introducing an absolutely innocuous attribute. It randomizes between sensitive and innocuous attributes. This instills more confidence in the respondent.

Although the various models that we considered have distinctive features, there are many more interesting models in the literature. It is beyond the scope of this introductory article even to list those models. An initiated reader may refer to [3] and [5] .

Although we considered examples based on Simple Random Sampling with Replacement (SRSWR) samples, that allow same individuals being interviewed more than once, we can easily study more general *sampling designs*. However, unlike direct response methods, when we use RR methods, same individual would report possibly different values on different trials precisely because of randomization. This is true of traditional RR models. Also under Padmawar–Vijayan model an individual interviewed more than once would *almost surely* report different values of z in different trials. Under Yu Tian Tang models, however, being implicit randomization models, response of an individual does not change from trial to trial.

Padmawar–Vijayan model is not confined to 1–0 variates hence has wider scope. It also uses a different method of randomization that has psychological advantage over traditional RR models to elicit information. Moreover it also results in better precision.

The models that we considered have distinctive features. There are many more interesting models in the literature. It is beyond the scope of this introductory article even to list them. An initiated reader may refer to [3] and [5] .

Finally, note that the *performance* of various estimators would depend on known as well as unknown *parameters. Performance* is often measured in terms of *variability.* It may therefore be argued that for given values of p, π_B and ξ^2 etc., the estimators may be compared based on their *performance.* An interested reader may refer to Suggested Reading for further details. The immediate technical details and related material may be found in [7].

Acknowledgement

The author is thankful to the referee and the editor for their suggestions that led to the refinement of the article.

Suggested Reading

[1] **W G Cochran, Sampling Techniques – Theory and Methods, Third Edition, John Wiley and Sons, New York, 1977.**

[2] **M Delampady and V R Padmawar, Sampling, probability models and statistical reasoning, *Resonance*, Vol.1, pp.49–58, 1996.**

[3] **A Chaudhuri and R Mukerjee, Randomized Response – Theory and Techniques, Marcel Dekker, New York, 1987.**

[4] **S L Warner, RR: a survey technique for eliminating evasive answer bias, *J. Amer. Statist. Assoc.*, Vol.60, pp.63–69, 1965.**

[5] **A Chaudhuri, Randomized Response and Indirect Questioning Techniques in Surveys, Chapman and Hall CRC, 2010.**

[6] **J -W Yu, G.-L Tian and M.-L Tang, Two new models for survey sampling with sensitive characteristic: design and analysis, *Metrika*, Vol.67, pp.251–263, 2008.**

[7] **V R Padmawar, Randomized response techniques. An unpublished article, Indian Statistical Institute, Bangalore, India. 2011.**

[8] **V R Padmawar and K Vijayan, Randomized response revisited, *J. Statist. Plann. Inference*, Vol.90, pp.293–304, 2000.**

[9] **B G Greenberg, A Abdel-Latif, Abul-Ela, W R Simmons and D G Horvitz, The unrelated question randomized response model for human surveys, *J. Amer. Statist. Assoc.*, Vol.64, pp.520–539, 1969.**

Address for Correspondence
V R Padmawar
Stat-Math. Division
Indian Statistical Institute
8th Mile, Mysore Road
Bangalore 560 059, India.
Email: vrp@isibang.ac.in