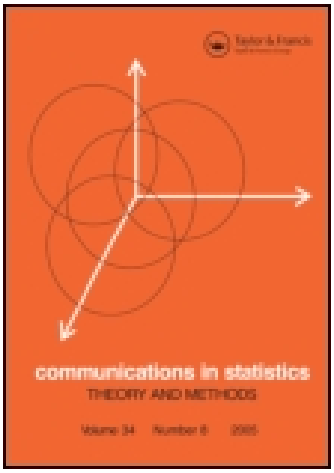

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Confidentiality guaranteed: a noninvasive procedure for collecting sensitive information

Samprit Chatterjee [a] & Gary Simon [a]
[a] Stern School of Business , New York University , 44 West Fourth Street, New York, NY, 10012-1126
Published online: 27 Jun 2007.

COMMUN. STATIST.—THEORY METH., 22(6), 1629-1651 (1993)

CONFIDENTIALITY GUARANTEED: A NONINVASIVE PROCEDURE FOR COLLECTING SENSITIVE INFORMATION

Samprit Chatterjee
Gary Simon

Stern School of Business
New York University
44 West Fourth Street
New York NY 10012-1126

KEYWORDS: *confidentiality; randomized response; sensitive information*

ABSTRACT

We present a scheme in which subjects select balls from each of two urns, reporting only the result from the urn appropriate to their response to a sensitive question. Subjects can thus be induced to respond to the sensitive question while preserving confidentiality. Design aspects are discussed.

1. INTRODUCTION

The technique known as randomized response is designed to allow researchers to ask sensitive questions by creating an environment in which each subject believes that his response to the sensitive question cannot be linked to his identity. It is often important to estimate population parameters on socially-disapproved items about which individuals are reluctant to

respond. This problem is especially important in dealing with AIDS, in which it is important to obtain estimates of prevalences of various sexual practices. Mere assurances that "your responses will remain confidential" are not sufficient to convince respondents to deal honestly with the questions.

A typical version of the randomized response involves a setup question followed by two substantive questions. For example, the setup question could be this:

> Examine your social security number. Observe whether the last digit is an even number or an odd number. If this last digit is an even number, answer question A below and place your response on the answer sheet. If this last digit is an odd number, answer question B below and place your response on the answer sheet.

Question A will deal with the sensitive matter (Have you used drugs?) and question B will deal with a population proportion which is known. For example, question B might ask "Were you born in the first six months of the year (January through June)?" Since we know the population proportions for the setup question (0.5 for odd/even social security numbers), and since we know the population proportion for question B (here 0.5), we can estimate the proportion in the population who answer yes to question A.

This technique is generally attributed to Horvitz, Shah, and Simmons (1967), who modified a simpler technique proposed by Warner (1965) in which no unrelated question was asked. For a recent bibliography, see Nathan (1988). In many situations, the

technique does not provide sufficient confidentiality assurance to the respondents to make them respond correctly; see Warner (1986), Fox and Tracy (1986), and Chaudhuri and Mukerjee (1988). Kuk (1990) suggested a further modification in the hope of improving respondent cooperation. In Kuk's version, there are two urns, and the respondent is instructed to draw a number of balls from each of the urns with replacement. The respondent is asked to supply (only) the result of the random drawing from urn 1 if he or she would respond "yes" to the sensitive question and to supply (only) the result from urn 2 if the response is "no." Since the contents of the two urns will be known to us, we will be able to estimate the population proportion which would answer "yes" to the sensitive question.

From a probabilistic point of view, we could equivalently tell the respondent to use urn 1 if the sensitive response is "yes" and to use urn 2 otherwise. It will probably help compliance, however, if we actually make the respondent go through the act of using both urns.

The main advantage of the modified procedure is that in no case does any respondent have to answer directly the sensitive question, and hence this removes respondent jeopardy. We also do not have to deal with distracting setup questions.

A practical method of implementing this system might, of course, use decks of cards rather than urns.

This paper will consider Kuk's scheme further. We consider both sampling with and without replacement. In sections 2 and 3

we obtain moment estimators of the proportion of people having a particular characteristic. The variances of these estimators and unbiased estimators of their variances are also obtained. In section 4 we obtain the maximum likelihood solutions for this problem. In section 5 we discuss the efficiency of the maximum likelihood estimators. This discussion leads to section 6 in which we consider the problems of optimal design. The final section is a comparative discussion of the new method with previously proposed procedures.

2. NOTATION AND DEVELOPMENT: PROBABILISTIC MODEL

We have two mutually exclusive groups of people, G_1 and G_2. In the population of interest, the proportion of people in G_1 is π, which we would like to estimate.

Accordingly, we select n people at random from this population. It is convenient to create a random variable Γ, so that $\Gamma_i = 1$ if the i^{th} individual is in G_1 and $\Gamma_i = 2$ if in G_2. We confront each person with two urns. There are red balls and black balls in each urn. We ask each person to select k balls from each urn (with-replacement and without-replacement will be discussed later), mentally noting the number of red balls obtained from each. We do not observe the drawing of the balls, and the subjects understand that we will not know the results of the separate draws. Then, each person is told to reveal the number of red balls obtained from the urn corresponding to his group. Persons from G_1 will tell us the number of red balls

obtained from urn 1, and the persons from G_2 will tell us the number of red balls obtained from urn 2. By such a mechanism, confidentiality of individuals is preserved.

The data consist of a single value reported by each subject.

Suppose that k = 1, meaning that each subject makes a single draw. Let θ_1 and θ_2 be the proportions of red balls in the two urns. We control the values of θ_1 and θ_2, and we would never consider the case $\theta_1 = \theta_2$. The probability that a single subject will report a red ball is $\phi = \theta_1 \pi + \theta_2 (1-\pi)$; this defines ϕ. Suppose that r is the proportion of the n subjects who report getting a red ball. Certainly

$$\hat{\phi} = \theta_1 \hat{\pi} + \theta_2 (1-\hat{\pi}) = r \qquad (1)$$

is the maximum likelihood estimate of ϕ, and

$$\hat{\pi}_1 = \frac{r - \theta_2}{\theta_1 - \theta_2} \quad . \qquad (2)$$

The subscript 1 is a reminder that k=1 ball is taken from each urn.

We should note in (1) that r is a weighted average of θ_1 and θ_2. If it happens in the data that r is between θ_1 and θ_2, we will have $0 < \hat{\pi}_1 < 1$. If r is not between θ_1 and θ_2, then $\hat{\pi}_1$ is outside the interval (0,1). From a practical point of view, getting r outside the interval is some evidence that there is malingering — the subjects are not cooperating. Since $\hat{\pi}_1$ can sometimes be outside (0,1), it is certainly not the maximum likelihood estimate, nor is it admissible in the decision-theory sense.

Since $\mathrm{Var}(r) = \phi(1-\phi)/n$, it is easy to show that

$$V_1 = \mathrm{Var}(\hat{\pi}_1) = \frac{\phi\ (1 - \phi)}{n\ (\theta_1 - \theta_2)^2} \quad . \tag{3}$$

Since $E\ \hat{\phi} = \phi$ and $E\ \hat{\phi}^2 = \phi(1-\phi)/n + \phi^2$, it can be verified that

$$E\ [\ \hat{\phi}\ (1-\hat{\phi})\ ] = E\ \hat{\phi} - E\ \hat{\phi}^2 = \phi(1-\phi)\ \frac{n-1}{n} \quad .$$

It is noted that an unbiased estimate $\hat{v}_1$ of $\mathrm{Var}(\hat{\pi}_1)$ is

$$\hat{v}_1 = \frac{\hat{\phi}\ (1 - \hat{\phi})}{(n-1)\ (\theta_1 - \theta_2)^2} \quad . \tag{4}$$

Consider next the case in which each individual draws k balls. We assume that there are N balls in each urn. Let $r[i]$ be the proportion of red balls reported by the i^{th} individual. Then

$$E\ r[i] = \phi$$

and, as shown in the Appendix,

$$\mathrm{Var}\ r[i] = \frac{k - c}{k} \left[\theta_1^2\ \pi + \theta_2^2\ (1-\pi) \right] - \frac{\phi}{k}\ (k\ \phi - c) \quad . \tag{5}$$

If the k balls are drawn with replacement, then $c = 1$. If the balls are drawn without replacement, $c = (N-k)/(N-1)$. Note further that when $k = 1$, $\mathrm{Var}\ r[i] = \phi(1-\phi)$.

Now let $r_k = \dfrac{r[1] + r[2] + \ldots + r[n]}{n}$, the proportion of red balls among all n subjects. It is clear that $E\ r_k =$

E r[i] = $\phi = \pi\ \theta_1 + (1-\pi)\ \theta_2$, so that we can find an unbiased estimate of π as

$$\hat{\pi}_k = \frac{r_k - \theta_2}{\theta_1 - \theta_2} \quad . \tag{6}$$

This $\hat{\pi}_k$ is not the maximum likelihood estimate, since it will lie outside (0,1) whenever r_k is not betweeen θ_1 and θ_2.

Also, Var r_k = n^{-1} Var r[i] and

$$V_k = \text{Var}(\hat{\pi}_k) = \frac{\text{Var } r_k}{(\theta_1 - \theta_2)^2} = \frac{\text{Var r[i]}}{n\ (\theta_1 - \theta_2)^2} \quad .$$

In the case of sampling *with* replacement, this expression reduces to

$$V_k = \frac{V_1}{k} + \left(1 - \frac{1}{k}\right) \frac{\pi(1-\pi)}{n} \quad . \tag{7}$$

A technique shown in section 3 will enable us to get an unbiased estimate of Var r_k and hence of V_k = Var($\hat{\pi}_k$) for any general drawing scheme.

For the simple scheme of drawing without replacement, an unbiased estimate of Var($\hat{\pi}_k$) is easy to obtain. An unbiased estimate of Var($\hat{\pi}_k$) is $\hat{v}_k$ where

$$\hat{v}_k = \frac{1}{k}\hat{v}_1 + \left(\frac{k\ \hat{\pi}_k\ (1 - \hat{\pi}_k) + \hat{v}_1}{k(n+1) - 1}\right) \frac{k-1}{k} \quad . \tag{8}$$

where

$$\hat{v}_1 = \frac{\hat{\phi}\ (1 - \hat{\phi})}{(n-1)\ (\theta_1 - \theta_2)^2} \quad .$$

This uses $\hat{\phi} = r_k$.

It is easy to verify that

$$E[\ \hat{\pi}_k\ (1-\hat{\pi}_k)\ ]\ =\ \pi(1-\pi)\left\{1\ -\ \frac{k-1}{kn}\right\}\ -\ \frac{1}{k}\,V_1\quad .$$

It follows that

$$E\left\{\ \hat{\pi}_k(1-\hat{\pi}_k)\ +\ \frac{1}{k}\,\hat{v}_1\ \right\}\ =\ \pi(1-\pi)\ \frac{kn\ -\ k\ +\ 1}{n}$$

and hence

$$E\left\{\ \frac{k\ \hat{\pi}_k\ (1-\hat{\pi}_k)\ +\ \hat{v}_1}{kn\ -\ k\ +\ 1}\ \right\}\ =\ \frac{\pi(1-\pi)}{n}\quad .$$

An unbiased estimate of $V_k = \mathrm{Var}(\ \hat{\pi}_k\ )$ is the $\hat{v}_k$ given above.

3. A GENERAL DRAWING SCHEME

Suppose that we consider the problem in general. This will allow us to derive results for any randomized scheme, not exclusively the binomial or hypergeometric. The subjects will be reporting a value from either urn 1 or urn 2. Let us assume that the probabilities are α_i = P[i red balls from urn 1] and β_i = P[i red balls from urn 2]. Let μ_j and σ_j^2 (j=1,2) be the means and variances of these two distributions.

The μ's and σ's are based only on the contents of the urns. It is assumed that M is the greatest number of red balls that can be reported (from either urn). We need not specify the actual distributions, but the most practically relevant distribution would be the hypergeometric distribution (sampling without replacement).

We assume that the α's and β's are not identical; that is, the contents of the urns are not identical. The differential compositions of the urns is what makes this method work. We will comment on the composition of the urns in sections 5 and 6.

Let R[i] be the number of red balls reported by the i^{th} individual.

Clearly $E\ R[i] = \pi\ \mu_1 + (1-\pi)\ \mu_2 = \mu$. This defines μ. Also,

$$\begin{aligned} \text{Var } R[i] &= \pi\ (\mu_1 - \mu)^2 + (1-\pi)\ (\mu_2 - \mu)^2 \\ &\quad + \pi\ \sigma_1^2 + (1-\pi)\ \sigma_2^2 \\ &= \pi(1-\pi)\ (\mu_1 - \mu_2)^2 + \pi\ \sigma_1^2 + (1-\pi)\ \sigma_2^2 \ . \end{aligned}$$

Now let $\bar{R} = (\ R[1] + R[2] + \ldots + R[n]\ )/n$.

Hence $E\ \bar{R} = \mu = \pi\ \mu_1 + (1-\pi)\ \mu_2 = \mu_2 + \pi(\mu_1 - \mu_2)$ and

$$\text{Var } \bar{R} = \frac{\pi(1-\pi)\ (\mu_1 - \mu_2)^2 + \pi\ \sigma_1^2 + (1-\pi)\ \sigma_2^2}{n} \ .$$

Since we know μ_1 and μ_2, we can get an unbiased estimator of π as

$$\hat{\pi} = \frac{\bar{R} - \mu_2}{\mu_1 - \mu_2} \ .$$

Note (as before) that this estimator is outside of (0,1) whenever $\bar{R}$ is not between μ_1 and μ_2.

We would like to find an unbiased estimator of Var $\bar{R}$. Observe that $E\ \bar{R}^2 = \text{Var } \bar{R} + [\ E\ \bar{R}\ ]^2$

$$= \frac{\pi(1-\pi)\ (\mu_1 - \mu_2)^2 + \pi\ \sigma_1^2 + (1-\pi)\ \sigma_2^2}{n} + \mu^2$$

$$= A\pi^2 + B\pi + C ,$$

where

$$A = \frac{n-1}{n}(\mu_1-\mu_2)^2 ,$$

$$B = \frac{(\mu_1-\mu_2)^2}{n} + \frac{\sigma_1^2-\sigma_2^2}{n} + 2\mu_2(\mu_1-\mu_2)$$

and

$$C = \frac{\sigma_2^2}{n} + \mu_2^2 .$$

Note that A, B, and C are all functions of known quantities.

Now let $D = \mu_1 - \mu_2$ and let $E = \mu_2$. We showed above that $E\bar{R} = D\pi + E$. We also noted previously that

$$\text{Var}\,\bar{R} = E(\bar{R}^2) - [E\bar{R}]^2 = A\pi^2 + B\pi + C - [D\pi + E]^2$$

$$= (A - D^2)\pi^2 + (B - 2ED)\pi + (C - E^2)$$

$$= F\pi^2 + G\pi + H, \text{ where}$$

$$F = -\frac{(\mu_1 - \mu_2)^2}{n} ,$$

$$G = \frac{(\mu_1 - \mu_2)^2}{n} + \frac{\sigma_1^2 - \sigma_2^2}{n}$$

and

$$H = \frac{\sigma^2}{n} .$$

For an unbiased estimator of $\text{Var}\,\bar{R}$ consider $U\bar{R}^2 + V\bar{R} + W$. This estimator will be unbiased if for each value of π, we have

$$U\left(A\pi^2 + B\pi + C\right) + V\left(D\pi + E\right) + W = F\pi^2 + G\pi + H$$

and this is ensured when

$$U = F/A = -1/(n-1) \quad ,$$

$$V = (G - UB)/D = (AG - BF)/(AD)$$

$$= \frac{\mu_1 + \mu_2}{n-1} + \frac{1}{n-1}\,\frac{\sigma_1^2 - \sigma_2^2}{\mu_1 - \mu_2} \quad , \text{ and}$$

$$W = H - UC - VE = H - CF/A - E(AG - BF)/(AD)$$

$$= \frac{\sigma_2^2}{n-1} - \frac{\mu_2}{n-1}\left(1 + \frac{\sigma_1^2 - \sigma_2^2}{\mu_1 - \mu_2}\right)$$

Since we have an unbiased estimator of $\mathrm{Var}\,\bar{R}$, we can easily construct an unbiased estimator of $\mathrm{Var}(\hat{\pi})$. In particular cases, such as the binomial or hypergeometric models, as shown below, we can make the values of U, V, and W explicit.

Sampling With Replacement

Consider the simple binomial case in which the proportions are θ_1 and θ_2, with k draws taken from each urn. Then

$$\mu_1 = k\,\theta_1 \quad , \quad \sigma_1^2 = k\,\theta_1\,(1-\theta_1) \quad ,$$

$$\mu_2 = k\,\theta_2 \quad , \quad \sigma_2^2 = k\,\theta_2\,(1-\theta_2) \quad .$$

For sampling with replacement, the binomial case, the values of U, V, and W simplify as

$$U = -\frac{1}{n-1} \quad ,$$

$$V = \frac{1}{n-1}\left[(k-1)(\theta_1 + \theta_2) + 1\right] \quad , \text{ and}$$

$$W = \frac{k\theta_2(\theta_1 - 1)}{n - 1} .$$

An unbiased estimator of the variance of $\hat{\pi}_k$ is

$$-\frac{\bar{R}^2}{n-1} + \frac{\bar{R}}{n-1}\left((k-1)(\theta_1 + \theta_2) + 1 \right) + \frac{k\theta_2(\theta_1 - 1)}{n - 1} . \quad (9)$$

Sampling Without Replacement

Consider next the hypergeometric case in which the sampling is done without replacement. As before the urns contain N balls, and the proportions of red balls are θ_1 and θ_2, with k draws taken from each urn. Then

$$\mu_1 = k\,\theta_1 \quad , \quad \sigma_1^2 = c\,k\,\theta_1\,(1-\theta_1) ,$$

$$\mu_2 = k\,\theta_2 \quad , \quad \sigma_2^2 = c\,k\,\theta_2\,(1-\theta_2) .$$

The value of c is $\dfrac{N - k}{N - 1}$.

Then,

$$U = -\frac{1}{n-1} \quad \text{as before,}$$

$$V = \frac{1}{(n-1)(N-1)}\left[N(k-1)(\theta_1 + \theta_2) + N - k \right] , \quad \text{and}$$

$$W = \frac{k\theta_2}{n-1}\left[c\theta_1 - 1 \right] .$$

It follows that an unbiased estimator for $\mathrm{Var}(\hat{\pi}_k)$ is

$$\frac{\bar{R}}{(n-1)(N-1)}[N(k-1)(\theta_1+\theta_2)+N-k] + \frac{k\theta_2}{n-1}\left(\frac{N-k}{N-1}\theta_1 - 1 \right) - \frac{\bar{R}^2}{n-1} . \quad (10)$$

Note that (10) reduces to (9) when $N/(N-1)$ and $(N-k)/(N-1)$ are replaced by 1; these are the conditions for sampling with replacement.

An Example:

Consider a survey in which a sample of size 100 is drawn from the population of interest. Suppose that the urn associated with group 1 has θ_1 = proportion of red balls = 0.3 and that with group 2 has θ_2 = proportion of red balls = 0.4. Each urn has 10 balls, and 4 are drawn out by each respondent. The proportion of red reported by the subjects is 0.38. In our notation $n = 100$, $\theta_1 = 0.3$, $\theta_2 = 0.4$, $k = 4$, and $N = 10$.

An estimator of the proportion π belonging to group 1 is

$$\hat{\pi}_4 = \frac{0.38 - 0.4}{0.3 - 0.4} = 0.2 .$$

An unbiased estimator of the variance of the estimator is obtained easily. For sampling with replacement we have from (9) $v(\hat{\pi}_4) = 0.0047$, and the standard error is 0.068. If these results were obtained for sampling without replacement, we have from (10) $v(\hat{\pi}_4) = 0.0044$, with a standard error of 0.066.

These are the method of moments estimators. They can be easily calculated and their variances have closed form expressions.

4. MAXIMUM LIKELIHOOD ESTIMATION

We now derive the maximum likelihood estimator of π for the proposed sampling scheme. The probability of getting response j

from any individual is $\pi \alpha_j + (1-\pi) \beta_j$. Let n_j be the number of subjects (out of n) who report the selection of j red balls. The likelihood is the multinomial

$$L = \frac{n!}{n_0!\, n_1!\, n_2! \ldots n_M!} \prod_{j=0}^{M} [\, \pi \alpha_j + (1-\pi) \beta_j \,]^{n_j} .$$

It follows that

$$\frac{d}{d \pi} \log L = \sum_{j=0}^{M} n_j \frac{\alpha_j - \beta_j}{\pi \alpha_j + (1-\pi) \beta_j} .$$

Setting this equation to zero does not provide an explicit solution for π.

Let P be the set of indices j for which $\alpha_j > \beta_j$, and let N be the set of indices j for which $\alpha_j < \beta_j$. Thus

$$P = \{\, j \mid \alpha_j > \beta_j \,\} \text{ and } N = \{\, j \mid \alpha_j < \beta_j \,\} .$$

The set of indices j where $\alpha_j = \beta_j$ does not contribute to the likelihood. Since the α's and β's both sum to 1, and since we have assumed that the α's and β's are not identical, both sets P and N must be nonempty.

It is easy to see that $\frac{d}{d \pi} \log L = 0$ is equivalent to the condition

$$\sum_{j \in P} n_j \frac{\alpha_j - \beta_j}{\pi \alpha_j + (1-\pi) \beta_j} = \sum_{j \in N} n_j \frac{\beta_j - \alpha_j}{\pi \alpha_j + (1-\pi) \beta_j} .$$

In this form, we ask for the equality of two sums involving positive terms. In the left sum, increasing π causes the sum

to decrease (since $\alpha_j > \beta_j$ in this sum). In the right sum, increasing π causes the sum to increase (since $\alpha_j < \beta_j$ in this sum). The left side decreases with π and the right side increases with π, so there can be only one solution. (It is possible that the two sums cannot be made equal. In such a case, we would have $\hat{\pi} = 1$ or $\hat{\pi} = 0$ as the maximum likelihood estimator.) The solution can be found (or its nonexistence demonstrated) by simple bisection.

Using standard statistical theory, we can get the limiting variance from Fisher's information. Here we use

$$I(\pi) = - E \frac{d^2}{d\pi^2} \log L = n \sum_{j=0}^{M} \frac{(\alpha_j - \beta_j)^2}{\pi \alpha_j + (1-\pi) \beta_j} . \qquad (13)$$

The estimated information is of course

$$I(\hat{\pi}) = n \sum_{j=0}^{M} \frac{(\alpha_j - \beta_j)^2}{\hat{\pi} \alpha_j + (1-\hat{\pi}) \beta_j} . \qquad (14)$$

A large sample 95% confidence interval for π is

$$\hat{\pi} \pm 1.96/[\ I(\hat{\pi})\ ]^{1/2} .$$

An investigation of $I(\pi)$ is considered in the next section.

5. PROPERTIES OF INFORMATION $I(\pi)$

The results of the previous section raise many interesting questions and they do not seem to have been addressed in the literature. An obvious question is how the α's and β's should be designed to maximize $I(\pi)$. We will investigate this issue for the case in which the α's and β's correspond to the hypergeometric model. If the α's and β's are very far apart,

$I(\pi)$ would become large, but that would also make the scheme transparent to the subjects. What restrictions on the α's and β's are appropriate?

As a simple first step, we will calculate some values for $I(\pi)/n$. We consider hypergeometric models, so that each urn contains a number of balls, and the subjects are asked to take a sample of these without replacement. In the cases considered (Table I), the urns are given *opposite* distributions; an example of such a situation is one in which urn 1 contains 14 red balls and 6 black balls, while urn 2 contains 6 red balls and 14 black balls. Many such situations were investigated. Since they were qualitatively very similar, only two examples are shown in Table I.

The following conclusions can be drawn from all the cases we have considered:

* When the two urns are identical (the case used in the last line of each part of Table I), the information $I(\pi)$ is zero.

* $I(\pi)$ is considerably larger when π is small.

* $I(\pi)$ is larger when the urn contents are very different (corresponding to small values of C in Table I).

* Given the same sample size and same proportional division of the colors in the urns, the small-urn case is more informative. As an example, consider the case of taking a sample of size 4. The urns with 10 balls (2 red/8 black or 8 red/2 black) give more information than the urns with 20 balls (4 red/16 black or 16 red/4 black).

TABLE I (10,4)
10 balls per urn
Sample 4 balls from selected urn
Urn 1 has C red balls and 10 - C black balls
Urn 2 has 10 - C red balls and C black balls

Values of π

	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
C= 1	21.05	11.11	7.84	6.25	5.33	4.76	4.40	4.17	4.04	4.00
C= 2	18.25	9.63	6.80	5.42	4.62	4.13	3.81	3.61	3.50	3.47
C= 3	7.81	5.05	3.92	3.30	2.91	2.65	2.49	2.38	2.32	2.30
C= 4	1.36	1.16	1.04	0.96	0.91	0.87	0.84	0.82	0.81	0.81
C= 5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE I (20,4)
20 balls per urn
Sample 4 balls from selected urn
Urn 1 has C red balls and 20 - C black balls
Urn 2 has 20 - C red balls and C black balls

Values of π

	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
C= 1	21.05	11.11	7.84	6.25	5.33	4.76	4.40	4.17	4.04	4.00
C= 2	20.39	10.76	7.60	6.05	5.16	4.61	4.26	4.04	3.91	3.87
C= 3	17.93	9.77	6.98	5.59	4.79	4.28	3.96	3.76	3.64	3.61
C= 4	13.94	8.10	5.95	4.84	4.18	3.76	3.49	3.32	3.23	3.20
C= 5	9.59	6.04	4.62	3.85	3.38	3.08	2.87	2.75	2.67	2.65
C= 6	5.68	3.95	3.18	2.74	2.46	2.27	2.15	2.07	2.02	2.00
C= 7	2.71	2.15	1.85	1.66	1.53	1.44	1.38	1.34	1.32	1.31
C= 8	0.97	0.87	0.81	0.76	0.73	0.71	0.69	0.68	0.67	0.67
C= 9	0.20	0.20	0.19	0.19	0.19	0.19	0.19	0.18	0.18	0.18
C=10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6. OPTIMUM DESIGN ASPECTS

We consider here design choices which will lead to superior estimators. In the binomial case, in which we are sampling *with* replacement, recall that

$$\mathrm{Var}(\hat{\pi}_k) =$$

$$= \frac{\mathrm{Var}(\hat{\pi}_1)}{k} + \left(1 - \frac{1}{k}\right) \frac{\pi(1-\pi)}{n} \quad .$$

where $\mathrm{Var}(\hat{\pi}_1) = \dfrac{\phi(1-\phi)}{n\,(\theta_1-\theta_2)^2}$.

As can be seen, $\mathrm{Var}(\hat{\pi}_k)$ decreases as the number of subjects, n, increases.

To investigate the dependence on k, observe that

$$\mathrm{Var}(\hat{\pi}_{k+1}) - \mathrm{Var}(\hat{\pi}_k) = \frac{-\pi\,\theta_1(1-\theta_1) - (1-\pi)\theta_2(1-\theta_2)}{nk(k+1)\,(\theta_1 - \theta_2)^2} \quad .$$

The numerator is negative. Hence, $\mathrm{Var}(\hat{\pi}_{k+1}) - \mathrm{Var}(\hat{\pi}_k) < 0$; that is, $\mathrm{Var}(\hat{\pi}_k)$ is a decreasing function of k, the number of draws per subject. This variance decreases to $\pi(1-\pi)/n$ as k becomes large.

Observe that $\mathrm{Var}(\hat{\pi}_k)$ depends on θ_1 and θ_2 only through $\mathrm{Var}(\hat{\pi}_1)$, and it decreases as $|\theta_1-\theta_2|$ becomes large. We cannot let θ_1 and θ_2 get too far apart, lest the scheme become transparent to the subjects and lead to non-response or malingering.

We will minimize $\mathrm{Var}(\hat{\pi}_k)$ subject to the constraint $\theta_1 - \theta_2 = \delta > 0$, where δ is a fixed specified positive quantity. Here δ represents the maximum on the differences between the compositions of the urns that will not be objected to by the respondents.

The motivating idea here is to achieve sampling design efficiency without unduly increasing respondent jeopardy. Sampling efficiency and respondent jeopardy are in conflict with each other (similar to the conflict between cost and precision of a survey) and they have to be balanced. We will maximize efficiency (reducing the variance of the estimator) while ensuring that the respondent jeopardy is held at an acceptable threshold level. This analysis is along the lines of Lenke (1976) and Flinger et. al. (1977).

Here we use the form

$$\mathrm{Var}(\hat{\pi}_k) = \frac{\pi(1-\pi)}{n} + \frac{\pi\,\sigma_1^2 + (1-\pi)\,\sigma_2^2}{n\,(\mu_1 - \mu_2)^2}. \qquad (15)$$

Since $\mu_1 - \mu_2 = k(\theta_1 - \theta_2) = k\delta$, we need only minimize $\pi\,\sigma_1^2 + (1-\pi)\,\sigma_2^2$. In the case of sampling either with or without replacement from urns with the same total number of balls, we can write $\sigma_i^2 = C\,\theta_i\,(1-\theta_i)$. For sampling with replacement, $C = 1$, and for sampling without replacement, $C = (N-k)/(N-1)$, where N is the number of balls in the urn. Thus we have to minimize

$$\pi\,C\,\theta_1\,(1-\theta_1) + (1-\pi)\,C\,\theta_2\,(1-\theta_2).$$

Substituting $\theta_2 = \theta_1 - \delta$ and minimizing over θ_1 we find the minimizing θ_1 to be $0.5 + \delta(1-\pi)$, and the corresponding θ_2 is $0.5 - \pi\delta$. Substituting these minimizing values into (15), the minimum variance is

$$\min(\ \mathrm{Var}(\hat{\pi}_k)\ ) = \frac{\pi(1-\pi)}{n}\left(1 - \frac{C}{k}\right) + \frac{C}{4nk\delta^2} \quad . \qquad (16)$$

The resulting minimized $\mathrm{Var}(\hat{\pi}_k)$ from (16) for sampling with and without replacement, respectively, are given by

$$V_{\text{with}} = \frac{1}{4\ nk\ \delta^2} + \left(1 - \frac{1}{k}\right)\frac{\pi(1-\pi)}{n}$$

and

$$V_{\text{without}} = \frac{N-k}{4\ nk(k-1)\ \delta^2} + \left(1 - \frac{N-k}{k(k-1)}\right)\frac{\pi(1-\pi)}{n} \quad .$$

7. CONCLUSION

Collecting sensitive information in a survey often results in a large non-response. The responses are often distorted because the respondents desire not to be identified with a negative trait. We have presented a survey method which collects sensitive information without compromising the identities of the respondents. In randomized techniques proposed earlier there was always a proportion of people who had to answer "yes" to the sensitive question with the concern that their answers might be decoded. This factor often causes a distorted response. In the method we have proposed and developed no individual directly answers the sensitive question. We have provided a non-invasive sampling scheme which guarantees complete confidentiality. For this scheme we have provided both the unbiased and maximum

likelihood estimators. We have also given unbiased estimators for the variance of the unbiased estimators. The estimators have higher efficiencies than those provided by other randomized response methods. Optimum sampling design issues for the proposed scheme have also been outlined. It is hoped that the scheme will be useful and can be implemented for collecting personally sensitive data, related to the epidemics such as AIDS. The proposed scheme accommodates two conflicting goals, the society's need to know (for effective public policy) and the individual's right to privacy for human dignity.

APPENDIX

In this appendix, we derive Var r[i], the proportion of red balls obtained by subject i. It is assumed that each subject makes k drawings with replacement from the appropriate urn. The notation is that of section 2.

$$\text{Var } r[i] = \text{Var}\{ E(r[i] \mid \Gamma_i) \} + E\{ \text{Var}(r[i] \mid \Gamma_i) \}$$

$$= \pi(\theta_1 - \phi)^2 + (1-\pi)(\theta_2-\phi)^2 + \frac{\pi\, c\theta_1(1-\theta_1) + (1-\pi)\, c\theta_2(1-\theta_2)}{k}$$

$$= \pi(\theta_1 - \phi)^2 + (1-\pi)(\theta_2-\phi)^2 + c\,\frac{\phi - \pi\,\theta_1^2 - (1-\pi)\,\theta_2^2}{k}$$

$$= \frac{k\pi(\theta_1 - \phi)^2 + k(1-\pi)(\theta_2-\phi)^2 + c\,[\,\phi - \pi\,\theta_1^2 - (1-\pi)\,\theta_2^2\,]}{k}$$

$$= \frac{\theta_1^2\,(k-c)\pi + \theta_2^2\,(k-c)(1-\pi) - 2k\phi\,[\pi\theta_1+(1-\pi)\theta_2] + c\phi + \phi^2 k}{k}$$

$$= \frac{\theta_1^2 (k-c)\pi + \theta_2^2 (k-c)(1-\pi) - 2k\phi^2 + c\phi + \phi^2 k}{\cdot\ k}$$

$$= \frac{\theta_1^2 (k-c)\pi + \theta_2^2 (k-c)(1-\pi) - k\phi^2 + c\phi}{k}$$

$$= \frac{k-c}{k}\left[\theta_1^2 \pi + \theta_2^2 (1-\pi)\right] - \frac{\phi}{k}(k\phi - c) \quad .$$

BIBLIOGRAPHY

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response*, New York: Marcel Dekker.

Flinger, M.A., Policello, G.F. and Singh, J. (1977). "A comparison of two RR survey methods with consideration for the level of respondent protection," *Communications in Statistics: Theory and Methods*, 6, 1511-1526.

Fox, J.A. and Tracy, P.E. (1986). *Randomized Response*, Beverly Hills: Sage Publishers.

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

Kuk, A.Y.C. (1990). "Asking sensitive questions indirectly," *Biometrika*, 77, 436-438.

Lenke, J. (1976). "On the degree of protection in randomized interviews," *International Statistical Review*, 44, 197-203.

Nathan, G. (1988). "A bibliography of papers on randomized response techniques," *Survey Methodology*, vol 14, no. 2.

Warner, S.L. (1965). "Randomized response: a survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 60, 63-69.

Warner, S.L. (1986). "The omitted digit randomized response model for telephone applications," *Proceedings of Survey Research Methods Section*, American Statistical Association, 441-443.

Received May 1992; Revised February 1993