A Sensitive Survey Model for HIV Seroprevalence Research

*[1]A. Usman; [1]M. Baba; [2]S. A. Damisa

[1]Department of Mathematics and Statistics,
Kaduna Polytechnic,
Kaduna, Nigeria.
aliyusman2007@gmail.com

[2]Department of Statistics,
Ahmadu Bello University,
Zaria, Nigeria.
asdamisa@abu.edu.ng

***Abstract*— Sensitive questions like HIV status may cause biased estimation of unknown population parameters as well as increase in the variance of the estimates due to evasive responses. The randomized response techniques (RRT) can be used to avoid the concealment of information or evasive answers. The RRT guarantees the anonymity of respondents in surveys aimed at determining the frequency of stigmatic, embarrassing or criminal behaviour where direct techniques for data collection may induce respondents to refuse to answer or give false responses. Different randomized response models (RRMs) have been devised in the past decades for dealing with sensitive items; which usually involve the use of random devices, such as dice or cards to collect reliable data on sensitive issues. Most of these RRMs have been proposed without some specific applications to HIV seroprevalence surveys. The motivation was to improve upon the existing RRMs as well as to apply them to estimate HIV seroprevalence rates. The objectives were to use research frontier to devise a mixed-stratified RRMs and use same to estimate HIV seroprevalence rates in a given population and compare results with the existing seroprevalence rates. [1]Proposed the pioneering RRM for estimating the proportion of persons bearing a socially disapproved character. [2]Produced unified criteria for all RRTs, [3] proposed a stratified RRM and so many others. Furthermore, the procedure of the field work and sampling design were well coordinated for the target population of 3,740 people aged 18 years and above using a sample size of 550. Furthermore, the model was used to estimate the HIV seroprevalence rate in a small population of adults attending a clinic in Kaduna, Nigeria. The Model estimated the HIV seroprevalence rate as 8.74% with a standard error of 0.0134 and a 95% confidence interval of [6.1%, 11.4%]. Accordingly, the sentinel projected seroprevalence rate, using the Epidemic Projection Package (EPP), for the next ten years (2013) was 9.7%; very consistent with the 95% confidence interval. Hence, the RRTs herein can serve as new viable methods for HIV seroprevalence surveys.**

Keywords- *Randomized response techniques, randomized response models, seroprevalence rates, mixed-stratified, design parameter, efficiency, sentinel surveys, stratified random sampling.*

Abbreviations-AIDS-Acquired Immune Deficiency Syndrome, CDC-Centre for Disease Control, EPP-Epidemic Projection Package, HIV-Human Immunodeficiency Virus, NACA-National Agency for the Eradication of AIDS, RR-Randomized Response, RRM-Randomized Response Model, RRT-Randomized Response Technique, USAID-United States Agency for International Development.

I. INTRODUCTION

Sensitive questions like HIV status may cause biased estimation of unknown population parameters as well as increase in the variance of the estimates due to evasive responses. The randomized response techniques (RRTs) were especially developed to improve the accuracy of answers to sensitive questions. Socially sensitive questions are thought to be threatening to respondents [14]. When sensitive topics are studied, respondents often react in ways that negatively affect the validity of the data. Such a threat to the validity of the results is the respondents' tendency to give socially desirable answers to avoid social embarrassment and to project a positive self-image [15]. Some scholars reasoned that the reluctance of the respondents to reveal sensitive or probably harmful information would diminish when respondents could be convinced that their anonymity was guaranteed [1].

Following this assumption, the first randomized response model(RRM) was designed by [1].

The crux of his method and all other RRTs that followed is that the meaning of the respondents' answers is hidden by a deliberate contamination of the data. Studies with RRTs have been conducted in the areas of health care [4], on alcohol, drug abuse and sexual behavior[5], on child molestation [6], on tax evasion [7], among others. Meta-analysis on 42 comparative studies showed that RRTs resulted in more valid population estimates than direct question–answer techniques [8].

The advantage of using RRTs to question sensitive topics is that the results are less distorted than when direct question–answer designs are used, making the RRM more effective. A second advantage of using RRT when conducting sensitive research is that, the individual 'yes'-answer becomes meaningless as it is only a 'yes-answer' to the random device [9]. However, the disadvantage of using RR methods is that they are less efficient than direct question designs. Since the RRTs work by adding random noise to the data, they all suffer from larger standard errors, leading to reduced power which makes it necessary to use larger samples than in question–answer designs. Unfortunately, larger samples are associated with prolonged completion time and higher research costs, making RRTs less attractive to applied researchers. This leads to the topic of efficiency versus effectiveness. [10]Defined HIV seroprevalence as the study of the number of cases where HIV is present in a specific population at a designated time. The presence of HIV in a specific individual is determined by the finding of HIV antibodies in the serum (HIV seropositivity).

This study is set to develop an efficient mixed-stratified RRM particularly for HIV seroprevalence surveys and to use the new Model for estimating the seroprevalence rate in a small population.

II. Materials and Methods

The procedure of the field work and sampling design were well coordinated for the target population of *3,740* adults aged 18 years and above attending Gwamna Awan Hospital in Kaduna, Nigeria using a sample size of *550*. Furthermore, the model was used to estimate the HIV seroprevalence rate in the same population. [2], [11] and[12] both theoretically and empirically analyzed the effect of different design parameters on the performance of RRTs using different levels of privacy protection. [2]Concluded that 0.7 approximately works well for every mixed RRM where the questions are regarded as highly sensitive. Hence, we adopt 0.7 as our design parameter and deck of 50 cards as our random device throughout. In stratified sampling, the population of N units is first divided into subpopulations (strata) of $N_1, N_2, ..., N_L$ units, respectively. These subpopulations are non-overlapping and together they comprise the whole of the population so that $N_1 + N_2 ... + N_L = N$. The sample sizes within the strata are denoted by $n_1, n_2, ..., n_L$, respectively. If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling. The marital status is used to form three strata for this study.

The Proposed RRT Model

The HIV seroprevalence surveys Model requires that a sample respondent in stratum h to answer an innocuous direct question and asked to use the random device R_{h1} if his/her answer to direct question is "yes". If answer to the direct question is "no", he/she is requested to use another random device R_{h2}. The random device R_{h1} consists of two statements (i) "I am HIV positive" and (ii) "I am HIV negative", presented with probabilities P_{h1} and $(1-P_{h1})$ respectively. Similarly, the random device R_{h2} consists of the two statements (i) "I am HIV positive" and (ii) "I am HIV negative", presented with probabilities and P_{h2} and $(1-P_{h2})$ respectively. The probabilities of a 'yes' response from the respondents using R_{h1} and R_{h2} are respectively given by:

$$\lambda_{h1} = P_{h1}\pi_h + (1-P_{h1})\pi_{hy} = P_{h1}\pi_h + (1-P_{h1}) \quad (1)$$

And

$$\lambda_{h2} = P_{h2}\pi_h + (1-P_{h2}) \quad (2)$$

On the other hand, the probabilities of a 'no' response from the respondents using R_{h1} and R_{h2} are respectively given by:

$$\lambda'_{h1} = P_{h1}(1-\pi_h) + (1-P_{h1})(1-\pi_{hy}) = P_{h1}(1-\pi_h) \quad (3)$$

and

$$\lambda'_{h2} = P_{h2}(1-\pi_h) \quad (4)$$

Since the respondent using R_{h1} has already answered yes to the direct question, $\pi_{hy} = 1$.

Among those that answered 'yes' to the innocuous questions in stratum h; suppose that n_{h1} report 'yes' and $(n_h - n_{h1})$ report 'no', the likelihood of the sample in the same stratum is as follows:

$$\xi = \left[P_{h1}\pi_h + (1-P_{h1})\right]^{n_{h1}} \times \left[P_{h1}(1-\pi_h)\right]^{n_h - n_{h1}} \quad (5)$$

The natural log of the likelihood is given below:

$$\log \xi = n_{h1} \log[P_{h1}\pi_h + (1-P_{h1})] + (n_h - n_{h1}) \log[P_{h1}(1- \quad (6)$$

To obtain the value of π_h, differentiate $\log \xi$ w.r.t. π_h and equate to zero as follows:

$$\frac{\partial \log \xi}{\partial \pi_h} = \frac{n_{h1}P_{h1}}{P_{h1}\pi_h + (1-P_{h1})} - \frac{(n_h - n_{h1})P_{h1}}{P_{h1}(1-\pi_h)} = 0 \quad (7)$$

$$\frac{n_{h1}P_{h1}}{P_{h1}\pi_h + (1-P_{h1})} = \frac{(n_h - n_{h1})P_{h1}}{P_{h1}(1-\pi_h)}$$

$$n_{h1}P_{h1}(1-\pi_h) = (n_h - n_{h1})[P_{h1}\pi_h + (1-P_{h1})]$$

$$n_{h1}P_{h1} - n_{h1}P_{h1}\pi_h = n_h P_{h1}\pi_h + n_h - n_h P_{h1} - n_{h1}P_{h1}\pi_h - n_{h1} + n_{h1}P_{h1}$$

$$n_h P_{h1}\pi_h = n_h P_{h1} - n_h + n_{h1}$$

$$\pi_h = \frac{n_h P_{h1} - n_h + n_{h1}}{n_h P_{h1}}$$

$$\pi_h = \frac{n_h P_{h1} - n_h + n_{h1}}{n_h P_{h1}}$$

Hence, the unbiased estimators in terms of the responses of the respondents using R_{h1} is given by:

$$\hat{\pi}_{h1} = \frac{\hat{\lambda}_{h1} - (1-P_{h1})}{P_{h1}} \quad (8)$$

Where the proportion of 'yes' answers from R_{h1} in the sample is $\hat{\lambda}_{h1} = n_{h1}/n_h$. The variance of $\hat{\pi}_{h1}$ is obtained as follows:

$$Var(\hat{\pi}_{h1}) = \left[\frac{1}{P_{h1}}\right]^2 Var(\hat{\lambda}_{h1}) \quad (9)$$

$$= \left[\frac{1}{P_{h1}}\right]^2 \left(\frac{\hat{\lambda}_{h1}(1-\hat{\lambda}_{h1})}{n_{h1}}\right)$$

$$= \left[\frac{1}{P_{h1}}\right]^2 \frac{[P_{h1}\pi_h + (1-P_{h1})][P_{h1}(1-\pi_h)]}{n_{h1}}$$

$$= \frac{[P_{h1}\pi_h + (1-P_{h1})](1-\pi_h)}{n_{h1}P_{h1}}$$

Hence;

$$Var(\hat{\pi}_{h1}) = \frac{(1-\pi_{h1})(P_{h1}\pi_{h1} + 1 - P_{h1})}{n_{h1}P_{h1}}$$

Similarly, the unbiased estimators in terms of the responses of the respondents using R_{h2} is given by:

$$\hat{\pi}_{h2} = \frac{\hat{\lambda}_{h2} - (1-P_{h2})}{P_{h2}} \quad (10)$$

Where the proportion of 'yes' answers from R_{h2} in the sample is $\hat{\lambda}_{h2} = n_{h2}/n_h$. The variance of $\hat{\pi}_{h2}$ is obtained as follows:

$$Var(\hat{\pi}_{h2}) = \frac{(1-\pi_{h2})(P_{h2}\pi_{h2} + 1 - P_{h2})}{n_{h2}P_{h2}} \quad (7)$$

In stratum h two randomization devices R_{h1} and R_{h2} are equally protective against the privacy of the respondents if $P_{h1} = P_{h2} = P_h$. Under this setting, the variances of the two unbiased estimators $\hat{\pi}_{h1}$ and $\hat{\pi}_{h2}$ become the same. We can also propose an estimator based on all the information collected in stratum h which we can use to estimate seroprevalence rates in stratum h as follows:

$$\hat{\pi}_h = \frac{n_{h1}}{n_h}\hat{\pi}_{h1} + \frac{n_{h2}}{n_h}\hat{\pi}_{h2} \quad (11)$$

Its variance is given by:

$$Var(\hat{\pi}_h) = \left(\frac{n_{h1}}{n_h}\right)^2 Var(\hat{\pi}_{h1}) + \left(\frac{n_{h2}}{n_h}\right)^2 Var(\hat{\pi}_{h2}) \quad (12)$$

$$= \left(\frac{n_{h1}}{n_h}\right)^2 \left[\frac{(1-\pi_h)(P_{h1}\pi_h + 1 - P_{h1})}{n_{h1}P_{h1}}\right] + \left(\frac{n_{h2}}{n_h}\right)^2 \left[\frac{(1-\pi_h)(P_{h2}\pi_h + 1 - P_{h2})}{n_{h2}P_{h2}}\right]$$

$$= \left(\frac{n_{h1}}{n_h^2}\right)\left[\frac{(1-\pi_h)(P_{h1}\pi_h + 1 - P_{h1})}{P_{h1}}\right] + \left(\frac{n_{h2}}{n_h^2}\right)\left[\frac{(1-\pi_h)(P_{h2}\pi_h + 1 - P_{h2})}{P_{h2}}\right]$$

If we decide that $P_{h1} = P_{h2} = P_h$ thus we get:

$$Var(\hat{\pi}_h) = \left(\frac{n_{h1} + n_{h2}}{n_h^2}\right)\left[\frac{(1-\pi_h)(P_h\pi_h + 1 - P_h)}{P_h}\right] \quad (13)$$

$$= \left(\frac{1}{n_h}\right)\left[\frac{(1-\pi_h)[P_h\pi_h + (1-P_h)]}{P_h}\right]$$

Hence;

$$Var(\hat{\pi}_h) = \frac{\pi_h(1-\pi_h)}{n_h} + \frac{(1-P_h)(1-\pi_h)}{n_h P_h}$$

An unbiased stratified seroprevalence rates estimator is given by:

$$\hat{\pi}_{sero} = \sum_{h=1}^{L} W_h \hat{\pi}_h \quad (14)$$

where; $W_h = N_h / N$ for is $h = 1,2,...,L$

N_h is the total number of individuals in the stratum h

N is the total number of individuals in the population

obviously $\sum_{h=1}^{L} W_h = 1$

Its variance is given by:

$$Var\ (\hat{\pi}_{Sero}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left[\pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}\right]^2 \quad (15)$$

Cochran (1977) established that the sampling fraction n_h / n is ignorable, then $Var(\hat{\pi}_{Sero})$ is minimized for a fixed total sample size n if:

$$n_h = \frac{nW_h\left[\pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}\right]^{\frac{1}{2}}}{\sum_{h=1}^{L} W_h\left[\pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}\right]^{\frac{1}{2}}} \quad (16)$$

where;

$$n_h = n_{h1} + n_{h2}$$

$$\sum_{h=1}^{L} n_h = n$$

Thus substituting the optimum value of n_h in (15) we get:

$$Var(\hat{\pi}_{Sero}) = \frac{1}{n}\left[\sum_{h=1}^{L} W_h\left[\pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}\right]^{\frac{1}{2}}\right]^2 \quad (17)$$

Relative Efficiency of the RRT Model

One of the most important ways of assessing any sample survey model is through its efficiency relative to the existing models. We hereby compare the relative efficiency of the proposed for HIV seroprevalence Model with Kim and Warde (2005) stratified estimator. Hence, the proposed Model is more efficient for a fixed sample size if and only if:

$$Var(\hat{\pi}_{SK}) - Var(\hat{\pi}_{Sero}) \geq 0 \quad (18)$$

$$\frac{1}{n}\left[\sum_{h=1}^{L} W_h\left\{\pi_h(1-\pi_h) + \frac{(1-P_h)\{\lambda_h P_h(1-\pi_h)+1-\lambda_h\}}{P_h^2}\right\}^{\frac{1}{2}}\right]^2 - \frac{1}{n}\left[\sum_{h=1}^{L} W_h\left[\pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}\right]^{\frac{1}{2}}\right]^2 \geq 0$$

The above inequality will be true if for each stratum h, $h = 1,2,...,L$ we have:

$$\left\{\pi_h(1-\pi_h) + \frac{(1-P_h)\{\lambda_h P_h(1-\pi_h)+1-\lambda_h\}}{P_h^2}\right\}^{\frac{1}{2}} - \left\{\pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}\right\}^{\frac{1}{2}} \geq 0$$

$$\frac{(1-P_h)\{\lambda_h P_h(1-\pi_h)+1-\lambda_h\}}{P_h^2} - \frac{(1-P_h)(1-\pi_h)}{P_h} \geq 0$$

$$1 - P_h(1-\pi_h) \geq 0 \quad (19)$$

The LHS of (19) is always nonnegative, hence the proposed model is more efficient than Kim and Warde (2005) stratified estimator.

The LHS of (19) is always nonnegative, hence the proposed model is more efficient than Kim and Warde (2005) stratified estimator.

III. Analysis

The analysis was maually computed to arrive at the following results

IV. Results

Recall that the unbiased mixed-stratified seroprevalence Model is given by:

$$\hat{\pi}_{sero} = \sum_{h=1}^{L} W_h \hat{\pi}_h \quad \text{where; } W_h = N_h / N \text{ for is } h = 1,2,..., L$$

Its variance is given by:

$$Var(\hat{\pi}_{Sero}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left[\pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}\right]^2$$

Where;

$$\hat{\pi}_h = \frac{n_{h1}}{n_h}\hat{\pi}_{h1} + \frac{n_{h2}}{n_h}\hat{\pi}_{h2}$$

$$Var\ (\hat{\pi}_h) = \frac{\pi_h(1-\pi_h)}{n_h} + \frac{(1-P_h)(1-\pi_h)}{n_h P_h}$$

Table 1:Samples and Strata sizes

Strata	Strata Description	N_h	n_h	n_{h1}	n_{h2}	W_h
1	Married (Men/ Women)	1,285	189	35	38	0.344
2	Unmarried (Men/ Women)	2,020	297	57	58	0.540
3	Divorced/Separated/Widowed	435	64	11	9	0.116
Total		3,740	550	103	105	1.000

Table 2:Summary of result of the Random Devices

Strata	$\hat{\lambda}_{h1}$	$\hat{\pi}_{h1}$	$V(\hat{\pi}_{h1})$	$\hat{\lambda}_{h2}$
1	0.365	0.093	0.0135	0.409
2	0.383	0.119	0.0085	0.392
3	0.324	0.034	0.0406	0.300

Table 3:Summary of result of the Random Devices (Contd)

$\hat{\pi}_{h2}$	$V(\hat{\pi}_{h2})$	$\hat{\pi}_h$	$V(\hat{\pi}_h)$
0.156	0.0130	0.098	0.0052
0.131	0.0838	0.097	0.0033
0.000	0.0476	0.011	0.0156

Table 5:Summary of computations

Strata	W_h	$\hat{\pi}_h$	$W_h\hat{\pi}_h$
1	0.344	0.098	0.0337
2	0.540	0.097	0.0524
3	0.116	0.011	0.0013
Total	1.000		0.0874

Table 5:Summary of computations (Contd)

W_h^2/n_h	$\hat{\pi}_h(1-\hat{\pi}_h)$	$\sum_{h=1}^{L}\frac{W_h^2}{n_h}\phi^2$
0.00063	0.156	0.000056
0.00098	0.131	0.000088
0.00021	0.000	0.000036
		0.000180

The other computations are summarized below:

$$\phi = \pi_h(1-\pi_h) + (1-P_h)(1-\pi_h)/P_h$$

$$\hat{\pi}_{sero} = \sum_{h=1}^{L} W_h\hat{\pi}_h = 0.0874$$

$$Var(\hat{\pi}_{Sero}) = \sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}\right]^2 = 0.00018$$

$$SE(\hat{\pi}_{Sero}) = \sqrt{Var(\hat{\pi}_{Sero})} = 0.0134$$

The *95%* confidence interval for HIV seroprevalence rate is given by:

$$\hat{\pi}_{sero} \pm 1.96 \times SE(\hat{\pi}_{Sero}) = 0.0874 \pm 1.96 \times 0.0134 = [0.061, 0.114]$$

V. Discussions

This study has helped to avoid evasive answer on HIV surveys. It was motivated by the fact that conventional data collection techniques usually cause evasive or untruthful responses when people are asked sensitive questions like their HIV serostatus. As a result, it is difficult to make accurate inferences from such unreliable data. This study has devised a mixed-stratified RRM using the work of [1], [13], [2], among others particularly for HIV seroprevalence surveys. The proposed model was proved to be more efficient than a frontier similar model by [3].

VI. Conclusion

We have been able to develop a sensitive survey model for HIV seroprevalence. The model was used to estimate HIV seroprevalence rate in a small adult population using a sample size of *550* and a design parameter of *0.7*. Using the survey data, the model estimated the HIV seroprevalence rate as *8.74%* with a standard error of *0.0134* and 95% confidence bands of [*6.1%*, *11.4%*]. These estimates are for adults who are 18 years and above who attend a hospital. These results are consistent with that of Nigerian sentinel survey (2003) conducted by NACA, USAID and CDC which estimated the HIV seroprevalence in Kaduna State as *6.0%*. Accordingly, the sentinel projected seroprevalence rate, using the EPP Package, for the next ten years (2013) was 9.7%; very consistent with the 95% confidence interval. Hence, the RRTs herein can serve as new viable methods for HIV seroprevalence surveys.

Acknowledgment

The authors wish to thank the anonymous reviewers for their time and constructive reviewing.

References

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

[2] Quatember, A. (2009). A standardization of randomized response strategies. *Survey Methodology, 35(2), 143*-152.

[3] Kim, J.M. and Warde, W.D. (2005). A stratified Warner's randomized response model. *Journal of Statistical* Planning and Inference 120(2), 155-165.

[4] Volicer, B. J. and Volicer, L. (1982). Randomised response technique for estimating alcohol use and non compliance in hypertensives. Journal of Studies in Alcohol 43, 739-750.

[5] Jarman, B. J. (1997). The Prevalence and Precedence of Socially Condoned Sexual aggression Within a Dating Context as Measured by Direct Questioning and Randomized response Technique. *Journal of Statistical Planning and Inference 76, 229-238.*

[6] Fox, J.A. and Tracy, P.E. (1986). *Randomized response: a method for sensitive surveys*. Beverly Hills, CA: Sage Publications Inc.

[7] Houston, J. and Tran, Alfred. (2008). A survey of tax evasion using randomized response technique. *Journal of Statistical Planning and Inference 148, 225-233*

[8] Lensvelt-Mulders, G.J.L.M.; Hox, J.J.; van der Heijden, P.G.M. and Maas, C.J.M. (2005). Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation, Sociological Methods & Research 33, 319-348.

[9] Van der Hout, A., Van der Heijden, P.G.M and Gilchrist, R. (2002). A multivariate logistic regression model for randomized response data. *Quantitative Methods31, 25-41.*

[10] Brookmeyer, R. & Gail, M.H. (2004). *AIDS epidemiology a quantitative approach.* London: Oxford.

[11] Blair, G., Imai, K. & Zhou, Y. (2015). Design and analysis of the randomized response. *Journal of the American Statistical Association*. 110 (511), 1304-1319.

[12] Saha, A. (2016). Simple randomized response technique in complex surveys. *International Journal of Statistics*. 59-66 A. 65 (5), 48-55.

[13] Arnab, R. (2004). Optional randomized response techniques for complex survey designs. *Biometrical Journal 46(1), 114-124*

[14] Lee, R.M. (1993). *Doing research on sensitive topics*. Newbury Park, CA: Sage.

[15] Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W. and Lee, L. (1999). Methods of data collection, perception of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology 22, 465–484.*