STRATIFIED RANDOMIZED RESPONSE TECHNIQUE FOR MODELLING HIV SEROPREVALENCE RATES IN NIGERIA

Aliyu Usman & Mamuda Ibrahim Kukasheka

Department of Mathematic and Statistics, Kaduna Polytechnic, Nigeria.

Abstract

Warner (1965) proposed the pioneering RRT for estimating the proportion of persons bearing a socially disapproved character. The RRT are used to avoid the concealment of sensitive information from respondents. Similarly, Su (2021) proposed a randomized response techniques (RRT) for tracking drug usage. 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 techniques have been devised in the past decades. Most of these RRTs have been proposed without some specific applications to HIV seroprevalence surveys. The objective here is to apply the RRT to estimate HIV seroprevalence rates. Quatember (2009) produced unified criteria for all RRTs, Kim and Warde (2005) proposed a stratified randomized response model (RRM) and so many others. The proposed RRM for HIV seroprevalence surveys was relatively more efficient than the Kim and Warde (2005) stratified estimator for a fixed sample size. Using the criteria of Quatember (2009) who derived the statistical properties of the standardized estimator for general probability sampling and privacy protection, the chosen design parameter was $\pi_h = 0.7$. The procedure of the field work and sampling design were well coordinated for the target population using a sample size of 400. Furthermore, the model was used to estimate the HIV seroprevalence rate of adults attending a clinic in Kaduna, Nigeria. Using the survey data, the model estimated the HIV seroprevalence rate is *1.1%* with a standard error of *0.0024* and 95% confidence bands of [*0.6%, 1.6%*]. 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 (2018) conducted by Nigeria HIV/AIDS Indicator and Impact survey (NAIIS) and United Nations Programme on HIV/AIDS (UNAIDS) which estimated the HIV seroprevalence in Nigeria as *1.4%*. Accordingly, this is within the 95% confidence interval. Hence, the RRT can serve as cheaper and faster viable methods for HIV seroprevalence surveys.

Key words
Randomized response techniques, seroprevalence, design parameter, stratified random sampling.

Abbreviations
AIDS-Acquired Immune Deficiency Syndrome, CDC-Centre for Disease Control, HIV-Human Immunodeficiency Virus, MICS-Multiple Indicators Cluster Surveys, NACA-National Agency for the Eradication of AIDS, NDHS-National Demographic and Health Survey, RR-Randomized Response, RRM-Randomized Response Model, RRT-Randomized Response Technique, USAID-United States Agency for International Development.

Introduction

Usman and Oshungade (2012) proposed a stratified randomized response model (RRM) and used same to estimate the HIV seroprevalence in Nigeria. The stratification then was by

marital status. A similar model is hereby proposed to estimate same on a different stratification domain. This estimation is stratified by hospital location. Brookmeyer and Gail (2004) 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 has applied an efficient randomized response model (RRM for HIV seroprevalence surveys in Nigeria.

Socially sensitive questions are thought to be threatening to respondents. Hence, randomized response techniques (RRTs) were particularly developed to improve the response rates as well as the accuracy of responses to sensitive questions. For surveys with sensitive topics, 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 (Rasinski, 1999). Warner (1965) 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. Following this assumption, Warner (1965) designed the first RRM. The idea 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.

Furthermore, positive effect on the validity of the results was seen when the estimates of RRTs were compared to known population estimates and when the results of RRTs were compared to other data collection methods. It also appeared that the results of the RRTs became more valid when the topic under investigation became more sensitive. Therefore, an 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 (Van der Hout, et al., 2002).

However, the shortcoming of using RRT 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. Effectiveness is related to the validity of research results in the same way that efficiency is related to reliability. The randomized response design is more effective than the direct question-answer design (Lensvelt-Mulders et al., 2005). The loss of efficiency in RR designs could be compensated when the results prove to be more valid (Kuk, 1990). When the loss in efficiency can be kept as small as possible the use of a RR design to study sensitive questions will become more profitable.

Objectives

1. Use the RRM to estimate HIV seroprevalence rates in Kaduna State, Nigeria.
2. Compare the estimated HIV seroprevalence by RRM with other clinical sources.

Methodology

The procedure of the field work and sampling design were well coordinated for the target population of adults aged 18 years and above attending four selected clinics in Kaduna, Nigeria using a sample size of 400. The sampling strategy is to consider the respondents until the sample size is achieved in each stratum. Furthermore, the model was devised to estimate the HIV seroprevalence rate in the same population. Quatember (2009) has theoretically and empirically analyzed the effect of different design parameters, π_h, on the performance of RRTs using different levels of privacy protection. He concluded that the design parameters $\pi_h = 0.7$ approximately works well for every mixed RRM where the questions are regarded as highly sensitive. Hence, $\pi_h = 0.7$ is hereby adopted as the design parameter for the electronic random device throughout.

In stratified sampling, the population of *N* units is first divided into subpopulations (strata) of $N_1, N_2, \cdots, N_L$ units, respectively. These subpopulations are non-overlapping and together they comprise the whole of the population such that $N_1 + N_2 + \cdots + N_L = N$. The sample sizes within the strata are denoted by $n_1, n_2, \cdots, n_L$ such that $n_1 + n_2 + \cdots + n_L = n$. 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.

Results

The HIV seroprevalence RRM 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 (1) and (2):

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

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

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

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

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:

$$\therefore V(\hat{\pi}_{h1}) = \frac{(1-\pi_h)(P_{h1}\pi_h + 1 - P_{h1})}{n_{h1}P_{h1}}$$

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

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

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:

$$V(\hat{\pi}_{h2}) = \frac{(1-\pi_h)(P_{h2}\pi_h + 1 - P_{h2})}{n_{h2}P_{h2}}$$

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. An estimator based on all the information collected in stratum *h* is hereby proposed which can be used to estimate seroprevalence rates in stratum *h* given by the following equation:

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

Its variance is given by the following equation:

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

An unbiased stratified seroprevalence rates estimator is given by the following equation:

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

Where:

N is the total number of individuals in the population

N_h is the total number of individuals in the stratum *h*

$W_h = N_h/N$ for is $h = 1,2,\cdots,L$

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

The variance is given by (15):

$$V(\hat{\pi}_{sero}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\phi^2$$

Where,

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

Results

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

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

Its variance is given by the following equation:

$$V(\hat{\pi}_{sero}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \phi^2$$

Where,

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

Table 1: Samples and Strata Sizes

Strata	Strata Names	N_h	n_h	W_h
1	Gwamna Awan General Hospital	1,953	100	0.239
2	Barau Dikko Teaching Hospital	2,045	100	0.250
3	Yusuf Dantsoho Memorial Hospital	2,311	100	0.283
4	Kawo General Hospital	1,862	100	0.228
	Total	8,171	400	1.000

Table 3: Summary of Results of Random Devices

Strata	n_{h1}	$\hat{\lambda}_{h1}$	$\hat{\pi}_{h1}$	$V(\hat{\pi}_{h1})$	n_{h2}	$\hat{\lambda}_{h2}$	$\hat{\pi}_{h2}$	$V(\hat{\pi}_{h2})$
1	16	0.356	0.079	0.0291	21	0.382	0.117	0.0232
2	19	0.333	0.048	0.0258	21	0.488	0.269	0.0267
3	21	0.420	0.171	0.0229	19	0.380	0.114	0.0280
4	22	0.431	0.188	0.0213	15	0.306	0.009	0.0334
Total	78				76			

Table 5: Summary of Computations

Strata	$\hat{\pi}_h$	$V(\hat{\pi}_h)$	W_h^2/n_h	$\hat{\pi}_h(1-\hat{\pi}_h)$	$W_h\hat{\pi}_h$	$\sum_{h=1}^{L} \frac{W_h^2}{n_h}\phi^2$
1	0.0073	0.0046	0.00057	0.0675	0.001745	0.00000123
2	0.0147	0.0049	0.00063	0.1257	0.003675	0.00000152
3	0.0115	0.0048	0.00080	0.1019	0.003255	0.00000185
4	0.0084	0.0047	0.00052	0.0768	0.001915	0.00000115
Total					0.010589	0.00000575

The other computations are summarized below:

$$\phi = \pi_h(1-\pi_h) + \frac{(1-P_h)(1-\pi_h)}{P_h}$$

$$\hat{\pi}_{sero} = \sum_{h=1}^{L} W_h\hat{\pi}_h = 0.011$$

$$V(\hat{\pi}_{sero}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \phi^2 = 0.00000575$$

$$SE(\hat{\pi}_{sero}) = \sqrt{V(\hat{\pi}_{sero})} = 0.0024$$

Hence, the *95%* confidence interval for HIV seroprevalence rate is given by:

$$\hat{\pi}_{sero} \pm 1.96 \times SE(\hat{\pi}_{sero})$$

Table 7: Summary of Seroprevalence Results

n	$\hat{\pi}_{sero}$	$SE(\hat{\pi}_{sero})$	95% confidence interval	
			Lower limit	Upper limit
400	0.011	0.0024	0.006	0.016

Conclusion

The RRM was used to estimate HIV seroprevalence rate in a small adult population using a sample size of 400 and a design parameter of *0.7*. Using the survey data, the model estimated the HIV seroprevalence rate is *1.1%* with a standard error of *0.0024* and 95% confidence bands of [*0.6%, 1.6%*]. 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 (2018) conducted by Nigeria HIV/AIDS Indicator and Impact survey (NAIIS) and United Nations Programme on HIV/AIDS (UNAIDS) which estimated the HIV seroprevalence in Nigeria as *1.4%*. Accordingly, this is within the 95% confidence interval. Hence, the RRT is hereby recommended to serve as cheaper and faster viable methods for HIV seroprevalence surveys. Further research projects can also take advantage of this RRT for HIV seroprevalence surveys in other places as well all apply same to other sensitive surveys.

References

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

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.*

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

Su, S., Salinas, V.I., Zamora, M.L., Sedory, S.A., and Singh, S. (2021). Randomized response sampling with applications to tracking drugs for better life. *Statistica Sinica, 31, 1-20.*

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

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.*

Usman, A. and Oshungade, I. O. (2012a). A Mixed-Stratified Randomized Response Technique Model for HIV Seroprevalence Surveys. *Research Journal of Mathematics and Statistics 4(3), 70-75*.

Usman, A. and Oshungade, I. O. (2012b). A Two-way Randomized Response Technique in Stratification for Tracking HIV Seroprevalence. *Mathematical Theory and Modeling 2(7), 86-97.*

Van der Hout, A., Van der Heijden, P.G.M and Gilchrist, R. (2002). A multivariate logistic regression model for randomized response data. *Quantitative Methods 31, 25-41.*

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