RANDOMIZED RESPONSE MODELS : A RATIO ESTIMATOR

Abdel-Latif A. Abul-Ela, and Hala M. Dakrouri
Mansoura University, and Cairo University, Egypt.

1. Introduction

The randomized response technique was suggested by Warner{6}as an interviewing method to get information from individuals possessing some stigmatizing characteristic. Different models have been proposed, mainly to eliminate or at least to reduce the bias resulting from refusal to respond or intentional untruthful reporting. Some few schemes have treated the application of the technique to the problem of related characteristics.

Barksdale { 1} presented a model to interview and estimate the proportion of individuals who belong to group A characterized by a sensitive trait and another group B possessing another related trait that may or may not be sensitive. Consequently, the respondent will not react independently when questioned about A , especially when the trait of B is sensitive.

De Lacy { 4} proposed his conditional response model as an extension of the unrelated question model of Greenberg et al. { 5 } . According to his model, the respondent is asked two consecutive questions on his membership of group A, and its subgroup a, both of whom are associated with a stigmatizing characteristic. The response to the second question is then dependent on that to the first.

This paper applies the ratio estimate method to the randomized response technique { 3 { as an attempt to improve theprecision of the RR estimates of two related characteristies.

2. A Randomized Ratio Estimate

Suppose there are two correlated charcteristics, A and B, where A is sensitive, and B may or may not be sensitive. In order to improve the estimate of the proportion of the population in group A, say π_1 , an estimate of those in group B, say π_2 , may be used, and the randomized response technique applied. The ratio of π_1 to π_2 is in itself a relevant quantity, and will be our main interest. Knowing this ratio, the number of individuals in group A could be estimated.

In the application of the ratio estimation procedure to two correlated attributes using the RR technique, we have used the Warner dichotomous RR Model { 6 } twice, to obtain separate estimates of each of the two proportions. Thus, it is assumed that there are two independent samples, of size n_1 and n_2, each drawn with replacement, from the same population. The probabilities of drawing the question of interest, Q(A) in the first sample, and Q(B) in the second are P_1 and P_2, respectively. Thus, applying Warner's RR model { 6 } , the first sample is used to estimate π_1 , where the number of "Yes" answers obtained is n_1'. Similarly, the second sample is used to estimate π_2 and the number of "Yes" answers will be n_2' . Assuming 100% truthful reporting in both samples. The estimates take the form :

$$\hat{\pi}_i = \frac{P_i-1}{2p_i-1} + \frac{n_i'}{n_i(2p_i-1)} \quad \text{for } i = 1,2 \qquad (1)$$

As for the variance, it is of the form :

$$\mathrm{Var}(\hat{\pi}_i) = \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i} + \frac{P_i(1-p_i)}{n_i(2p_i-1)^2} \quad \text{for } i = 1,2 \qquad (2)$$

The randomizing device to be used could be one of several of those mentioned in the literature, such as the spinner, cards, or dice.

As an example, the sensitive question Q(A) could be concerning tax evasion, where the respondent could be queried about his having filled out the appropriate tax forms and having handed them in to the authorities. As for the question Q(B), it could either be concerning a sensitive topic, such as income, where the popultion can be divided into " high income" and "low income" groups, using a suitable criterion to distinguish one from the other; or else Q(B) could be concerning a nonsensitive topic, such as type of occupation, also divided dichotomously into "white collar" and " other..

3. The Randomized Ratio Estimate and its Variance

Generally, if an estimate is required for the ratio of a variable X, to the value of another correlated variable Y, it takes the form :

$$\hat{R} = \frac{x}{y} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

where x and y are the sample values of X and Y respectively. Applying this notion to randomized response, it is required to derive the ratio of π_1 to π_2 . Thus the randomized ratio estimate would take the form :

$$\hat{R}_{ran} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \left\{\frac{p_1-1}{2p_1-1} + \frac{n_1'}{n_1(2p_1-1)}\right\} \Big/ \left\{\frac{p_2-1}{2p_2-1} + \frac{n_2'}{n_2(2p_2-1)}\right\} \quad \ldots\ldots\ldots\ldots(4)$$

In order to simplify the sample ratio estimate form (4), since the two samples are independent and estimates of π_1 and π_2 are obtained independently, it is assumed that $n_1 = n_2 = n$, and that $p_1 = p_2 = p$.

Thus, the randomized ratio estimate reduces to :

$$\hat{R}_{ran} = \frac{n(p-1)+n_1'}{n(2p-1)} \cdot \frac{n(2p-1)}{n(p-1)+n_2'}$$

$$= \frac{(p-1)+\frac{n_1'}{n}}{(p-1)+\frac{n_2'}{n}}$$

$$\therefore \hat{R}_{ran} = \frac{(p-1)+\ {}_1}{(p-1)+} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

where $\lambda_1 = \frac{(n_1')^2}{n}$ and $\lambda_2 = \frac{(n_2')}{n}$ are the proportions of "Yes" answers in the first and second sample, respectively.

As for the variance of the regular ratio estimate, and ignoring the f.p.c. Cochran { 2 } showed that :

$$\mathrm{Var}\ (R) = \frac{1}{n\bar{y}^2}\left(S_x^2 + R^2 S_y^2 - 2\rho R\ S_x S_y\right) \ldots\ldots\ldots(6)$$

where R is defined as following :

$$\hat{R} = \frac{x}{\bar{y}} = \frac{\bar{x}}{y} = \frac{n_1\hat{\pi}_1}{n_2\hat{\pi}_2} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \hat{R}_{ran} \text{ for } n_1=n_2=n.$$

Therefore, $Var(\hat{R}_{ran})$ is given by :

$$Var(\hat{R}_{ran}) = \frac{1}{\hat{\pi}_2^2}\left(Var(\hat{\pi}_1) + \hat{R}^2_{ran}Var(\hat{\pi}_2) - 2\rho\hat{R}_{ran}\sqrt{Var(\hat{\pi}_1)\,Var(\hat{\pi}_2)}\right) \quad \text{.......(7)}$$

If the population value of the proportion π_2 can be obtained from another source, it it used in (7), and the variance of the sample ratio estimate $\hat{R}_{ran}$ should decrease, as will be shown later, since we will be using the two samples combined to estimate π_1. If not, the sample estimate $\hat{\pi}_2$ may be uses instead.

4. Precision of the Randomized Ratio Estimate

In this section, the precision of the randomized ratio estimate, will be numerically investigated for different values of the parameters p, π_2 and ρ. Using the variance formula (7) the following combinations of parameter values were used to calculate this variance:

n = 1000
π_1 = 0.1,0.2.
π_2 = 0.2,0.4,0.6,0.8.
p = 0.6,0.7,0.8,0.9.
ρ = -1.0, -0.5, -0.2,0.0, 0.5,1.0.

Tables (1) and (2) present a summary of the results and will be used to detect some general trends which may suggest some optimum parameter values.

From the results it may be noted :

a. Except for p = 0.5, in which case the var($\hat{R}_{ran}$) is undefined, the values of this variance for any value of p will be almost identical to that for (1-p) because of the presence of the term $\frac{p(1-p)}{n(2p-1)^2}$ in var($\hat{\pi}_i$) resulting from the use of the randomizing device. However, too high or low values of p may affect the degree of confidence of the respondent in the interviewing process and respectively his cooperation.

b. For fixed values of π_1, π_2 and p, the variance decreases as ρ varies from - 1.0 to 1.0, with a minimum at ρ=1. Table(1) shows some of the values which corroborate this point.

c. For all fixed values of p, ρ, and π_1, Var($\hat{R}_{ran}$) decreases as π_2 increases from 0.2 to 0.8. when $\pi_1 = \pi_2$ and ρ =1 in which case Var ($\hat{R}_{ran}$) = 0, however it is a trivial case.

d. Var(R) decreased as p goes from 0.6 to 0.9. This is seen in Table (2). Table(1) also shows the gradual decrease in variance as p increases, for fixed values of π_1 and ρ, and, although π_2 varies, the trend is obvious.

TABLE(2) THE Variance of the Randomized Ratio Estimate for π_2 from 0.2 to 0.8, For some Values of π_1, and ρ.

π_1		π_2	p= 0.6 Var($\hat{R}_{ran}$)	p = 0.9 Var($\hat{R}_{ran}$)
0.1	-1.0	0.2	0.34387	0.01423
		0.4	0.05976	0.00252
		0.6	0.02311	0.00094
		0.8	0.01206	0.00047
	-0.2	0.2	0.22137	0.00896
		0.4	0.04435	0.00178
		0.6	0.01854	0.00072
		0.8	0.01014	0.00039
	0.5	0.2	0.01419	0.00436
		0.4	0.03087	0.00113
		0.6	0.01454	0.00053
		0.8	0.00847	0.00032
	1.0	0.2	0.03763	0.00106
		0.4	0.02124	0.00066
		0.6	0.01169	0.00040
		0.8	0.00727	0.00026
0.2	-1.0	0.2	0.61600	0.03006
		0.4	0.08700	0.00459
		0.6	0.03052	0.00158
		0.8	0.01504	0.00073
	-0.2	0.2	0.36960	0.01804
		0.4	0.05600	0.00290
		0.6	0.02133	0.00108
		0.8	0.01119	0.00055
	0.5	0.2	0.15400	0.00752
		0.4	0.02888	0.00142
		0.6	0.01330	0.00064
		0.8	0.00782	0.00038
	1.0	0.2	0.00000	0.00000
		0.4	0.00950	0.00036
		0.6	0.00756	0.00033
		0.8	0.00541	0.00026

On the basis of the previous results it can be concluded that an optimal choice of parameter values for π_2, p and ρ, which increase the precision of the estimate may be achieved when one or more of the following rules are used :

1. Choose the characteristic B to be more frequent than A. However, π_1 is usually unknown, so that the choice of a target value for π_2 may be based upon some previous knowledge or an approximation of π_1. However, the effect of this choice on the bias and of the impact of this procedure on respondents in terms of respondent cooperation should be further investigated.

TABLE (1) THE Variance of the Randomized Ratio Estimate for ρ from - 1.0 to 1.0 for some Values of p, π_1 and π_2

P	π_1	π_2		Var($\hat{R}_{ran}$)	P	π_1	π_2		Var($\hat{R}_{ran}$)
0.6	0.1	0.2	-1.0	0.34387	0.8	0.1	0.6	-1.0	0.00210
			-0.5	0.26731				-0.5	0.00162
			0.0	0.19075				0.0	0.00154
			0.5	0.11419				0.5	0.00115
			1.0	0.03763				1.0	0.00098
0.7	0.2	0.8	-1.0	0.00359	0.9	0.2	0.4	-1.0	0.00459
			-0.5	0.00301				-0.5	0.00353
			0.0	0.00244				0.0	0.00247
			0.5	0.00187				0.5	0.00142
			1.0	0.00129				1.0	0.00036

2.Chose p as high (or as low) as possible,without endangering respondent cooperation, since, as we have observed, the lowest variances were at p=0.9, or its equivalent p = 0.1. This choice of p is in agreement with most of the other recommendations concerning p in the literature.

3. As for the choice of ρ , the nearer it is to+1, the better,or in other words the two attributes should be as positively correlated as possible. This was also the recommendation of Barksdale.

5.The Randomized Ratio Estimate when π_2 is known:

Assuming π_2 is known in advance or could be known from the survey or from some other source, then we could use both samples to estimate π_1 , the proportion of the stigmatizing characteristic. For example if π_1 denotes the proportion of illegitmate births to be estimated by the RR techique from the present survey, then π_2 could be that same proportion for the same population at some previous time as shown in the vital statistics reports. Also it could be the proportion of some other non-sensitive characteristic B in the present population that is highly and positively correlated with the senitive charcteristic A. In such a case the estimate of π_1 using Warner's model {6} is given by :

$$\hat{\pi}_1 = \frac{P-1}{2p-1} + \frac{(2n')}{2n(2p-1)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(8)$$

and the sample ratio estimate will be :

$$\hat{R}_{ran} \mid \pi_2 = \frac{\hat{\pi}_1}{\pi_2} = \frac{2n(p-1)+(2n')}{2n(2p-1)\,\pi_2} = \frac{p-1+\lambda}{\pi_2\,(2p-1)} \cdots(9)$$

where $\lambda = \frac{2n'}{2n}$ = the proportion of "Yes"answers.

Therefore, Var $(\hat{R})_{ran}$ given that π_2 is known in advance will be :

$$\mathrm{Var}\,(\hat{R}_{ran} \mid \pi_2) = \frac{1}{\pi_2^2}\,\mathrm{Var}\,(\hat{\pi}_1) + R^2\,\mathrm{Var}\,(\pi_2) - 2\,\rho\,R\,\sqrt{\mathrm{Var}\,(\hat{\pi}_1)\,\mathrm{var}\,(\pi_2)} \qquad (10)$$

where $\mathrm{Var}\,(\pi_2) = \frac{\pi_2(1-\pi_2)}{2n}$

As noted before, the variance of the sample ratio estimate is expected to decrease and sizeable gains in precision will be obtained . This has been shown by the short numerical investigation presented in table (3).

6. Conclusion :

Using the randomized ratio estimator, it was found that greater precision is achieved the more the variables are positively correlated, and for large values of π_2 and p.

Our parameter of interest was the ratio of π_1 to π_2, denoted $\hat{R}_{ran}$, but knowing it, we could also easily estimate the number of indivduals possessing the sensitive characteristic A in the sample or in the population, where :

$$\hat{R} = \frac{\hat{\pi}_1}{\hat{\pi}_2}$$

$$A = n\,\hat{R}\,\hat{\pi}_2 = n\,\hat{\pi}_1$$

$$\text{or}\quad A = N\,\hat{R}\,\hat{\pi}_2 = N\,\hat{\pi}_1$$

It should be noted however, that cov (π_1, π_2) or equivalently ρ should be calculated to estimate $\mathrm{Var}(\hat{R}_{ran})$, a point which is not treated in this paper. It is obvious that we need to consider the joint distsibution of π_1 and π_2 and applying Barksdale's or Delacy's models partially on a subsample from the two samples used to estimate π_1 and π_2 separately to calculate the value of $\hat{\rho}$. This raises the question : How much practical is this ratio estimate compared with Barksdale's estimate or Delacy's ? However, one should consider the expected gain in efficiency when respondents are questioned about each characteristic independently and applying the ratio estimate suggested in this paper.

Finally, our study of the efficiency of the randomized ratio estimate was based solely upon the variance as criterion. Nevertheless, the bias is equally important and should also be taken into consideration in a study of this nature. As the bias depends mostly upon respondent cooperation and trust in the technique, the extent to which our randomized ratio estimate affects the bias may be determined empirically.

***

TABLE (3) Relative Precision of RR Ratio Estimate with π_2 known to RR Ratio Estimate with π_2 Unknown*

P	π_1	π_2		Relative Precision	P	π_1	π_2		Relative Precision
0.6	0.1	0.2	-1.0	3.86	0.8	0.1	0.6	-1.0	2.28
			-0.5	3.22				-0.5	1.93
			0	2.49				0	2.03
			0.5	1.62				0.5	1.69
			1.0	0.59				1.0	1.69
0.7	0.2	0.8	-1.0	2.66	0.9	0.2	0.4	-1.0	3.28
			-0.5	2.38				-0.5	2.80
			0	2.10				0	2.21
			0.5	1.75				0.5	1.42
			1.0	1.32				1.0	0.42

* Relative precision defined by $\dfrac{V(\hat{R}_{ran})}{V(\hat{R}_{ran} \mid \pi_2)}$

REFERENCES

{ 1 } Barksdale, W.B. New Randomized Response Techniques For Control Of Non-sampling Errors In Surveys, The Institute of Statistics, The Consolidated University of North Carolina, Ph. D. thesis, University of North Carolina, Chapel Hill, 1971.

{ 2 } Cochran, W.G. Sampling Techniques, 3 rd ed., John Wiley and Sons, Inc., New York, 1977.

{ 3 } Dakrouri, H,M. Randomized Response Models. The case of Related Characteristics, unpublished MSc. thesis, Cairo University, 1980.

{ 4 } Delacy, P.W. Randomized Conditional Response, Proceedings of the Social Statistics Section, American Statistical Association (1975), pp. 383-386.

{ 5 } Greenberg, B.G., Abul-Ela, A.A., Simmons, W.R., and Horvitz, D G. The Unrelated Question Randomized Response Model: Theoretical Framework. JASA 64 (1969), pp. 520-539.

{ 6 } Warner, S.L. Randomized Response : A Survey Technique For Eliminating Evasive Answer Bias, JASA 60 (1965), pp. 63-69.
