RANDOMIZED CONDITIONAL RESPONSE

Peter W. DeLacy, Cornell University[1]

1. Introduction

Consider a population of size N which can be divided into three subpopulations of size N_1, N_2, and N_3, where $N_1 + N_2 + N_3 = N$. Suppose that we are interested in estimating $N_1/(N_1 + N_2)$, as we have no direct interest in the third group and only secondary interest in N_1/N and $(N_1 + N_2)/N$. Procedures are well defined in Cochran [1963] for point and interval estimation of ratios. What is proposed here are methods of accomplishing these goals in situations requiring use of a randomized response technique to eliminate untruthful responses to questions which might cause embarrassment or attach stigma.

We employ an extension of the unrelated question technique in Greenberg et al. [1967], where a second question is asked of those persons responding "yes" to the first question. To illustrate, suppose we are interested in the proportion of persons involved in extramarital sexual experiences whose experiences are homosexual. We might ask the following pairs of questions:

Set A
1. Are you involved in any extramarital sexual experiences?
2. If so, are these experiences homosexual?

Set B
1. Is the last digit of your SSAN 0, 1, 2, 3, 4, 5, or 6?
2. If so, is the last digit of your SSAN 5 or 6?

A respondent (randomly selected from a suitable population) would then choose either set A with probability p or set B with probability 1-p. He would then answer the questions with the responses "Yes, Yes", "Yes, No", or "No".

We shall define the following notation:

Π_1 = probability of answering "yes" to question 1 set A.
Π_2 = probability of answering "yes" to question 2 set A given that question 1 set A was answered "yes".
θ_1 = probability of answering "yes" to question 1 set B.
θ_2 = probability of answering "yes" to question 2 set B given that question 1 set B was answered "yes".
n_1 = number of people in a sample of size n who answer "yes, yes".
n_2 = number of people in a sample of size n who answer "yes, no".
Λ_1 = probability of answering "yes" to question 1 (either set).
Λ_2 = probability of answering "yes" to question 2 (either set).
$\lambda_1 = (n_1 + n_2)/n$.
$\lambda_2 = n_1/n$.

We thus have the following equalities:

$$\Lambda_1 = p\Pi_1 + (1-p)\theta_1, \tag{1}$$

$$\Lambda_2 = p\Pi_1\Pi_2 + (1-p)\theta_1\theta_2. \tag{2}$$

From (1) we have

$$\Pi_1 = \frac{\Lambda_1 - (1-p)\theta_1}{p}, \tag{3}$$

and from (2) we have

$$\Pi_2 = \frac{\Lambda_2 - (1-p)\theta_1\theta_2}{p\Pi_1}. \tag{4}$$

Thus,

$$\Pi_2 = \frac{\Lambda_2 - (1-p)\theta_1\theta_2}{\Lambda_1 - (1-p)\theta_1}.$$

2. Estimation

Since $n_1 + n_2$ is distributed binomially (n,Λ_1), similar arguments are used to show that n_1 given $n_1 + n_2$ is distributed binomially $(n_1 + n_2, \Lambda_2/\Lambda_1)$. Also, since the joint probability density of n_1 and $n_1 + n_2$ is the product of the density of $n_1 + n_2$ and of n_1 given $n_1 + n_2$, we have the joint likelihood function:

$$L \propto \Lambda_1^{n_1+n_2}(1-\Lambda_1)^{n-(n_1+n_2)}\left(\frac{\Lambda_2}{\Lambda_1}\right)^{n_1}\left(1 - \frac{\Lambda_2}{\Lambda_1}\right)^{n_1+n_2-n_1}$$

$$= (1-\Lambda_1)^{n-n_1-n_2}(\Lambda_2)^{n_1}(\Lambda_1 - \Lambda_2)^{n_2}.$$

Solving for Λ_1 and Λ_2 we have:

$$\tilde{\Lambda}_1 = (n_1 + n_2)/n$$

and

$$\tilde{\Lambda}_2 = n_1/n$$

as our maximum likelihood estimators. But since

$$\Lambda_1 = p\Pi_1 + (1-p)\theta_1$$

and

$$\Lambda_2 = p\Pi_1\Pi_2 + (1-p)\theta_1\theta_2,$$

then

$$\hat{\Pi}_1 = \left(\frac{n_1 + n_2}{n} - (1-p)\theta_1\right)/p$$

and

$$\hat{\Pi}_2 = \frac{n_1/n - (1-p)\theta_1\theta_2}{\frac{n_1 + n_2}{n} - (1-p)\theta_1}$$

are maximum likelihood estimators for Π_1 and Π_2, respectively, by the invariance property.

It may be easily shown that $\hat{\Pi}_2$ is not unbiased.

3. Accuracy of $\hat{\Pi}_2$

Recalling the literature of ratio estimation we recognize $\hat{\Pi}_2$ as a ratio estimate. The approximate variance of $\hat{\Pi}_2$, then, can be arrived at by

direct application of procedures well defined in the literature. There are various equivalent expressions for Var $\hat{\Pi}_2$. Among the simplest is

$$\mathrm{Var}(\Pi_2) \doteq \frac{[\Lambda_2-(1-p)\theta_1\theta_2]^2}{[\Lambda_1-(1-p)\theta_1]^2\, n}\left\{\frac{\Lambda_2(1-\Lambda_2)}{[\Lambda_2-(1-p)\theta_1\theta_2]^2}\right.$$

$$\left. + \frac{\Lambda_1(1-\Lambda_1)}{[\Lambda_1-(1-p)\theta_1]^2} - \frac{2\Lambda_2(1-\Lambda_1)}{[\Lambda_1-(1-p)\theta_1][\Lambda_2-(1-p)\theta_1\theta_2]}\right\}.$$

This is equivalent to

$$\Pi_2^2\left(C_{\lambda_2\lambda_2} + C_{\lambda_1\lambda_1} - 2C_{\lambda_1\lambda_2}\right),$$

where

$$C_{\lambda_2\lambda_2} = \frac{\Lambda_2(1-\Lambda_2)}{n[\Lambda_2 - (1-p)\theta_1\theta_2]^2}$$

and

$$C_{\lambda_1\lambda_1} = \frac{\Lambda_1(1-\Lambda_1)}{n[\Lambda_1 - (1-p)\theta_1]^2}$$

are the squared coefficients of variation of $\lambda_2 - (1-p)\theta_1\theta_2$ and $\lambda_1 - (1-p)\theta_1$, respectively, and

$$C_{\lambda_1\lambda_2} = \frac{\Lambda_2(1-\Lambda_1)}{n[\Lambda_1 - (1-p)\theta_1][\Lambda_2 - (1-p)\theta_1\theta_2]}$$

is the relative covariance of $\lambda_2 - (1-p)\theta_1\theta_2$ and $\lambda_1 - (1-p)\theta_1$. If, however, λ_1 and λ_2 follow a bivariate normal distribution (which they will asymptotically), Sukhatme [1954] has shown that to terms of order $1/n^2$

$$E(\hat{\Pi}_2-\Pi_2)^2 \doteq (\Pi_2)^2\left(C_{\lambda_1\lambda_1} + C_{\lambda_2\lambda_2} - 2C_{\lambda_1\lambda_2}\right)\left(1+3C_{\lambda_1\lambda_1}\right) + \frac{6C_{\lambda_1\lambda_1}\left(\rho^2 C_{\lambda_2\lambda_2} + C_{\lambda_1\lambda_1} - 2C_{\lambda_1\lambda_2}\right)}{n}$$

$$= (\Pi_2)^2\left(C_{\lambda_1\lambda_1} + C_{\lambda_2\lambda_2} - 2C_{\lambda_1\lambda_2}\right)\left(1 + 3C_{\lambda_1\lambda_1} + 6C_{\lambda_1\lambda_1}\left(\frac{\rho^2 C_{\lambda_2\lambda_2} + C_{\lambda_1\lambda_1} - 2C_{\lambda_1\lambda_2}}{C_{\lambda_2\lambda_2} + C_{\lambda_1\lambda_1} - 2C_{\lambda_1\lambda_2}}\right)\right).$$

Since the last term inside parentheses is less than $6C_{\lambda_1\lambda_1}$,

$$E(\hat{\Pi}_2-\Pi_2)^2 < (\Pi_2^2)\left(C_{\lambda_1\lambda_1} + C_{\lambda_2\lambda_2} - 2C_{\lambda_1\lambda_2}\right)\left(1 + 9C_{\lambda_1\lambda_1}\right),$$

to terms $O(n^{-2})$.

This leads us to conclude that if we make n large enough to keep $C_{\lambda_1\lambda_1} < .01$, we will underestimate by less than 9%, the true mean squared error (MSE). Let us examine the idea that we need to keep $C_{\lambda_1\lambda_1} \leq .01$ to be within 9% of the true MSE.

Since

$$C_{\lambda_1\lambda_1} = \frac{(1-\Lambda_1)\Lambda_1}{n[\Lambda_1 - (1-p)\theta_1]^2},$$

this is equivalent to

$$n \geq \frac{100\Lambda_1(1-\Lambda_1)}{(p\Pi_1)^2}.$$

From this expression it is easy to see that we need to keep p as large as possible to keep the sample size small. In order to get a better feeling for what we mean by "large" and "small", the following tables of n for selected values of Π_1, θ_1 are presented:

$\Pi_1 = .5$		p = .9	.7	.5	.3
	.9	123	193	336	763
	.7	124	201	384	1024
θ_1	.5	124	**204**	400	1112
	.3	124	202	384	1024
	.1	123	193	336	763

$\Pi_1 = .3$		p = .9	.7	.5	.3
	.9	316	566	1066	2489
	.7	308	552	1112	3008
θ_1	.5	299	**522**	1067	3042
	.3	288	477	934	2596
	.1	277	414	712	1660

$\Pi_1 = .1$		p = .9	.7	.5	.3
	.9	1823	4580	10,000	24,934
	.7	1660	4115	9600	27,733
θ_1	.5	1487	**3502**	8400	26,178
	.3	1304	2743	6400	20,267
	.1	1112	1838	3600	10,000

$\Pi_1 = .05$		p = .9	.7	.5	.3
	.9	5767	17,304	39,900	
	.7	5026	15,100	37,500	
θ_1	.5	4246	**12,308**	31,900	etc.
	.3	3426	8929	23,100	
	.1	2567	4962	11,100	

To see exactly in terms of variances how these values perform, we selectively choose values for Π_2 and θ_2 also. To reduce the effect of these parameters we choose $\Pi_2 = \theta_2$ and arbitrarily

choose $\Pi_2 = .5$. The variances for the n-values in boxes above are:

	p = .7	p = .5
$\Pi_1 = .5$	Var $\hat{\Pi}_2 = .0050$	Var $\hat{\Pi}_2 = .0050$
$= .3$	$= .0039$	$= .0042$
$= .1$	$= .0032$	$= .0036$
$= .05$	$= .0031$	$= .0034$

$\Pi_1 = \theta_1 = .5$, p = .7, n = 2044

Π_2 \ θ_2	.1	.3	.5	.7	.9
.1	.00180	.00273	.00358	.00436	.00507
.3	.00369	.00420	.00465	.00502	.00532
.5	.00486	.00497	.00500	.00497	.00486
.7	.00532	.00502	.00465	.00420	.00369
.9	.00507	.00436	.00358	.00273	.00180

$\Pi_1 = .1$, $\theta_1 = .5$, p = .7, n = 3502

Π_2 \ θ_2	.1	.3	.5	.7	.9
.1	.00115	.00250	.00374	.00488	.00694
.3	.00194	.00269	.00334	.00388	.00432
.5	.00299	.00315	.00320	.00315	.00299
.7	.00432	.00388	.00334	.00269	.00194
.9	.00694	.00488	.00374	.00250	.00115

From these tables it can be concluded that to keep the MSE low, we need to have p as large as possible and θ_1 as small as possible without compromising the anonymity of the respondents. It is also clear that for fixed Π_1, θ_1, p, if $\theta_2 > \Pi_2 > \frac{1}{2}$ or if $\theta_2 < \Pi_2 < \frac{1}{2}$, the MSE will be lower than for the case $\theta_2 = \Pi_2$. For $\Pi_2 = \frac{1}{2}$, however, all $\theta_2 \neq \Pi_2$ result in smaller MSE, for the case $\Pi_1 = .5$ and larger MSE for $\Pi_1 = .1$.

However, Π_2 is never known, and if we assume that it can achieve any value between 0 and 1, then we must look at each column of these tables for the maximal MSE. A quick inspection leads to the conclusion that for every value of $\theta_2 \neq \frac{1}{2}$, the maximum MSE is greater than the maximum MSE for $\theta_2 = \frac{1}{2}$. We have numerically derived then, a minimax rule for choice of θ_2: always choose $\theta_2 = \frac{1}{2}$. If we use different assumptions about the range of Π_2, for example $\Pi_2 > \frac{1}{2}$ or $\Pi_2 < \frac{1}{2}$, we would use $\theta_2 > \frac{1}{2}$ and $\theta_2 < \frac{1}{2}$, respectively, as minimax rules.

4. Extensions

This entire process can be extended to a series of k questions each conditioned on a "yes" response to the previous question. Thus

$$\hat{\Pi}_k = \frac{\lambda_k - (1-p)\theta_1\theta_2 \cdots \theta_k}{\lambda_{k-1} - (1-p)\theta_1\theta_2 \cdots \theta_{k-1}},$$

$$\text{Var } \hat{\Pi}_k \doteq (\Pi_k)^2\left(C_{\lambda_k\lambda_k} + C_{\lambda_{k-1}\lambda_{k-1}} - 2C_{\lambda_k\lambda_{k-1}}\right),$$

and we require the coefficient of variation of $\lambda_{k-1} - (1-p)\theta_1 \cdots \theta_{k-1}$ to be less than .1 in order to have negligible bias in estimation.

However, one can quickly see that with this repeated subsampling, Λ_{k-1} can become very small, and

$$\frac{\Lambda_k(1-\Lambda_{k-1})}{[\Lambda_{k-1} - (1-p)\theta_1 \cdots \theta_{k-1}]^2}$$

very large. It then becomes necessary to have extremely large samples to attain any precision on estimates of Π_k, as well as accuracy.

5. Summary

Randomized response techniques can be used, in a census of human populations, for obtaining information on a sensitive characteristic. In sample surveys of human populations, it might be of interest to measure the proportion of individuals belonging to group A, the members of which are associated with a characteristic that is stigmatic in the opinion of the population in general. Hence a member of such a group might suffer embarrassment in conceding explicitly his association with the group. The randomized response technique is devised to mask the respondent's answer so that he can feel assured that his anonymity as to the response is preserved. In certain surveys it might be of interest to obtain an estimate of the membership of a subgroup of A. In such a case, the following procedure, which is called randomized conditional response model, can be applied.

Two sets of two questions each are given as a part of the questionnaire. One set of questions are designed to elicit information on a sensitive characteristic, and the other set of questions are innocuous. The respondent chooses any of the two sets, assisted by a chance mechanism, and answers the first question of the set. If the answer is affirmative then he answers the second question. If the answer to the first question is negative then he ignores the second question and reports a "no" response. Thus the response to the second question is dependent on the response to the first question. In this sense, the procedure is called randomized conditional response model.

In this paper, the maximum likelihood estimator of the conditional probability (treated as a parameter) of answering "yes" to the second question is obtained. The properties of such an estimator are studied in terms of mean squared error. Some guidelines for reducing the mean squared error and sample size by manipulation of parameters are given.

6. Bibliography

Cochran, W. G. Sampling Techniques, John Wiley and Sons, Inc., New York, 1963.

Fieller, E. C. The distribution of the index in a normal bivariate population. Biometrika 46(1932), pp. 477-480.

Greenberg, B. G., Abul-Ela, Abdel-Latif, A., Simmons, W. R., and Horvitz, D. G. The unrelated question randomized response model: theoretical framework. JASA 64(1969), pp. 520-539.

Greenberg, B. G., Horvitz, D. G., and Abernathy, J. R. A comparison of randomized response designs. Reliability and Biometry, Statistical Analysis of Lifelength Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1974, pp. 787-815.

Hartley, H. O. and Ross, A. Unbiased ratio estimates. Nature 174(1954), pp. 270-271.

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

Moors, J. J. A. Optimization of the unrelated question randomized response model. JASA 66(1971), pp. 627-629.

Sukhatme, P. V. Sampling Theory of Surveys with Applications, Iowa State College Press, Ames, Iowa, 1954.

Warner, S. L. Randomized response: a survey technique for eliminating evasive answer bias. JASA 60(1965), pp. 63-69.

---

[1] Current address: Captain Peter W. DeLacy, Infantry School, Ft. Benning, Columbus, Georgia.