CORRELATIONS IN RANDOMIZED RESPONSE SURVEYS

Geun–Shik Han and William D. Warde, Oklahoma State University
William D. Warde, 301 MS, Department of Statistics

ABSTRACT

This paper examines methods to calculate the correlation between pairs of variables obtained in a simple random survey when one or both of the variables has been measured using the randomized response technique. Solutions are obtained for the Warner model and the unrelated question randomized response model in the case of binomial variables.

INTRODUCTION AND LITERATURE REVIEW

Sampling theory generally assumes that the data collected on units in the sample are accurate representations of the values associated with the units sampled. In many cases, this assumption is not valid, and a number of errors occur. These are variously called measurement errors, interviewer errors and bias, social desirability bias, etc. The randomized response models were introduced with the idea of minimizing social desirability bias.

An example of this type of bias is given by Sudman and Bradburn (1983) in which it was determined that over a third of the residents of Chicago who reported having voted in a primary election had in fact not voted. This was determined by an examination of the voting records for those individuals who had been interviewed in the survey. This type of bias is likely to be much greater when the question posed is of a socially undesirable or incriminating nature.

The randomized response technique first introduced by Warner (1965) is a method which can be used to help minimize social desirability bias. The technique involves a questionnaire on which two questions are given, but only one answer is requested. The respondent is asked to respond YES or NO to one of the two questions selected at random by the respondent. The method of selecting the question to be answered is designated and requires the use of a randomization device for which the researcher knows precisely the probability of selection of each of the two questions by the respondent. The questions are constructed in such a manner that the second question is the negative of the first.

For example, the respondent might be asked to respond to the question "Have you used cocaine within the past six months?" or to "Have you NOT used cocaine within the past six months?" The two questions might be placed on the survey instrument with instructions to the respondent to toss a fair die and respond truthfully to the first question if the die shows a one or a two, and to respond truthfully to the second question if the die shows a three, four, five or a six. The success of the method depends upon the interviewer being unaware of the result of the die roll, and merely recording a YES or a NO response without being aware of which question the respondent was answering. This is necessary in order to guarantee the confidentiality of the response given by the respondent.

The technique used to estimate the proportion of the population who had used cocaine in the past six months (S) is as follows. If the survey elicits responses from n individuals, n_1 of whom respond with a YES, and if we define p to be the probability of the respondent being asked the direct question (1/3 in our example), then

$$P(\text{YES answer}) = pS + (1-p)(1-S) \quad (1)$$

This can be estimated by the ratio n_1/n, and so by equating n_1/n to (1) and solving for S we obtain:

$$\hat{S} = (n_1/n - (1-p))/(1-2p) \quad (2)$$

and it can be shown that

$$V(\hat{S}) = S(1-S)/n + p(1-p)/(n(2p-1)^2)$$

Note that the divisor $(1 - 2p)$ in (2) requires that the randomization device selected may not be a fair coin for which $p = .5$, but must be such that p is different from .5.

Since Warner's original paper, a number of techniques have been developed to expand on or improve the method. These include the unrelated question model of Horvitz, Shah and Simmons (1967) which was extended by Greenberg, Abdul–Ela, Simmons and Horvitz (1967) to a multinomial model. Gould, Shah and Abernathy (1969) developed a two trials method to deal with some of the problems which were encountered in these methods.

The general framework for these methods is to replace the negated version of the sensitive question used by Warner with a totally unrelated question of a non sensitive nature for which the researcher knows a priori the probability of a YES response. An example of such a question might be "Does your social security number end in the number 2?" for which the probability of a YES response will be $P_N = .1$. As before, the respondent will be given some randomization device which asks them to respond to the sensitive question with probability p and the non sensitive question with probability $(1-p)$. Thus in this case

$$P(\text{YES response}) = pS + (1-p)P_N$$

which leads to an estimate of S given by

$$\hat{S} = (n_1/n - (1-p))/p$$

with

$$V(\hat{S}) = (pS + P_N)(1 - pS - P_N)/(np^2)$$

Methods have also been developed to deal with continuous responses in which the respondent selects a random number which is then either added to or multiplied by the true response and the resulting number reported to the interviewer. So long as the distribution from which the random number was selected is known to the researcher, then the necessary information about S can be deduced.

These methods have been shown to reduce the problem of bias in the estimation of the incidence the sensitive trait, and also to be practical in their administration. However, in many survey situations a goal of the investigation is to examine the relationship between several different questions and such techniques as correlation analysis and Chi–square contingency table analysis may be desired. The objective of this paper is to determine reasonable methods to compute a Pearson Product Moment Correlation coefficient (or Chi–square statistic) when one or both of the variables has been measured using one of the randomized response techniques.

WARNER'S METHOD

In the case of Warner's method, it can be shown that the sample correlation coefficient between the responses recorded by the interviewer will be identical to the correlation coefficient between two sensitive responses measured using the technique.

Let S_1 and S_2 denote the true

proportion of YES responses to the first and second question respectively, and let p_1 and p_2 designate the probability of each sensitive question being asked directly. If we define

$$r_i = \begin{cases} 1 & \text{if a respondent answers YES the i–th question} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$r_i = p_iS_i + (1 - P_i)(1 - S_i)$$

$$E(r_i) = (2p_i - 1)\, E(S_i) + (1 - p_i)$$

$$V(r_i) = (2p_i - 1)^2\, V(S_i)$$

for i = 1 or 2.

$$\begin{aligned} E(r_1r_2) = {} & (2p_1 - 1)(2p_2 - 1)E(S_1S_2) \\ & + (2p_1 - 1)(1 - p_2)E(S_1) \\ & + (1 - p_1)(2p_2 - 1)E(S_2) \\ & + (1 - p_1)(1 - p_2) \\ = {} & (2p_1 - 1)(2p_2 - 1)E(S_1S_2) + K \end{aligned}$$

$$\begin{aligned} \mathrm{Corr}(S_1S_2) = {} & \frac{(E(S_1S_2 - E(S_1)\ E(S_2))}{\sqrt{V(S_1)\ V(S_2)}} \\ = {} & \mathrm{Corr}(r_1r_2). \end{aligned}$$

Note that for the binomial situation, the usual Chi–square statistic can be obtained as $n\mathrm{Corr}(S_1S_2)$.

UNRELATED QUESTION METHODS

For this section, it is convenient to use the two trial methodology given by Gould, Shah and Abernathy (1969). In this technique, the sample is divided into two groups, not necessarily of equal sizes, and the probability of the respondent being asked the sensitive question is changed from one group to the next. This technique allows the use of a nonsensitive question for which the probability of a YES answer is not precisely known a priori.

For this technique there will be four questions involved. Questions 1 and 2, with Question 1 designated as the sensitive question, will form the pair which is used for the first Sensitive question pair, and questions 3 and 4 with question 3 designated sensitive will form the second sensitive pair. For convenience, we will designate the sample size as 2n with the probability of selection of question 1 denoted by p_1 and for question 3 by p_3 in the first group of n interviews. These probabilities will be changed to p_2 and p_4 respectively in the second group of n interviews.

We will designate the incidence of a YES answer to the i–th non sensitive question by Y_1.

The response equations are given by

$$\begin{aligned} r_1 &= p_1S_1 + (1 - p_1)Y_1 \\ r_2 &= p_2S_2 + (1 - p_2)Y_2 \\ r_3 &= p_3S_1 + (1 - p_3)Y_1 \\ r_4 &= p_4S_2 + (1 - p_4)Y_2 \end{aligned}$$

Since the sensitive variables are assumed to be independent of the two non sensitive variables, then the correlations can be written as

$$\begin{aligned} \mathrm{Corr}(r_1r_2) = {} & p_1p_2\mathrm{Corr}(S_1S_2) \\ & + (1 - p_1)(1 - p_2)\mathrm{Corr}(Y_1Y_2) \end{aligned}$$

$$\begin{aligned} \mathrm{Corr}(r_3r_4) = {} & p_3p_4\mathrm{Corr}(S_1S_2) \\ & + (1 - p_3)(1 - p_4)\mathrm{Corr}(Y_1Y_2) \end{aligned}$$

These two equations may be solved to give $\mathrm{Corr}(S_1S_2)$ independent of

$\mathrm{Corr}(Y_1Y_2)$ as

$$\frac{\mathrm{Corr}(r_1r_2) - \dfrac{(1-p_1)(1-p_2)\mathrm{Corr}(r_3r_4)}{(1-p_3)(1-p_4)}}{p_1p_2 - \dfrac{(1-p_1)(1-p_2)(p_3p_4)}{(1-p_3)(1-p_4)}}$$

which can be estimated by replacing $\mathrm{Corr}(r_1r_2)$ and $\mathrm{Corr}(r_3r_4)$ by their estimates based on the sample data.

Note that for the binomial situation, the usual Chi-square statistic can be obtained in this case as $2n\mathrm{Corr}(S_1S_2)$.

A Monte Carlo study of this situation by Han (1993) indicates that although the estimates of $\mathrm{Corr}(S_1S_2)$ are independent of the choice of p_i; $i=1,2,3,4$ which are made by the researcher, the standard deviation of $\mathrm{Corr}(S_1S_2)$ decreases as $|p_1 - p_3|$, and also as $|p_2 - p_4|$, increase. Thus our selection should be to have both p_1 and p_3 as far away from .5 as we can justify without compromising the confidentiality of the respondents, which would be non existent of any p_i were either 0 or 1. The results of this Monte Carlo study are given in Tables 1 and 2.

REFERENCES

Gould, A.L., Shah, B.V. and Abernathy, J.R. (1967) "Unrelated Question Randomized Response Techniques with Two Trials per Respondent" Proceedings of the Social Statistics Section, American Statistical Association, 351–359.

Greenberg, D.G., Abdul–Ela, A.A., Simmons, W.R. and Horvitz, D.G. (1969) "The Unrelated Question Randomized Response Model Theoretical Framework" Journal of the American Statistical Association, 64:520–539.

Han, G–S. (1993) "Correlation Analysis for the Randomized Response Models" Unpublished Ph.D. dissertation, Oklahoma State University, Stillwater, OK.

Horvitz, D.G., Shah, B.V. and Simmond, W.R. (1967) "The Unrelated Question Randomized Response Model" Proceedings of the Social Statistics Section, American Statistical Association, 65–72.

Sudman, S. and Bradburn, N. (1983) "Asking Questions: A Practical Guide to Questionnaire Design" Jossey–Bass Publishers, San Francisco.

Warner, S.L. (1965) "Randomized Response: A Survey Technique for Eliminating Evasive Answers" Journal of the American Statistical Association, 60:63–69.

TABLE 1

ESTIMATED CORRELATION FOR THE UNRELATED RANDOMIZED RESPONSE MODEL WITH n=100.

P_1	P_3	$\rho_{Y_1Y_2}$				
		0.22	0.3	0.4	0.5	0.6
0.3	0.4	0.64216 (1.21164)	0.62861 (1.22923)	0.63454 (1.22557)	0.62885 (1.23004)	0.64619 (1.21609)
0.3	0.6	0.60406 (0.32265)	0.59795 (0.32537)	0.59967 (0.32591)	0.60064 (0.32467)	0.60062 (0.32148)
0.3	0.7	0.60163 (0.22093)	0.59346 (0.22292)	0.59420 (0.22315)	0.59436 (0.22254)	0.59453 (0.22173)
0.3	0.8	0.59836 (0.16567)	0.59379 (0.16165)	0.59369 (0.16295)	0.59381 (0.16287)	0.59412 (0.16361)
0.4	0.6	0.60206 (0.37488)	0.59554 (0.37324)	0.59725 (0.37469)	0.59651 (0.37422)	0.59639 (0.37281)
0.4	0.7	0.60078 (0.23157)	0.59236 (0.23345)	0.593038 (0.23353)	0.59255 (0.23298)	0.59269 (0.23273)
0.4	0.8	0.59804 (0.16754)	0.59344 (0.16329)	0.59330 (0.16458)	0.59321 (0.16451)	0.59352 (0.16544)
0.6	0.7	0.61012 (0.37164)	0.60409 (0.37912)	0.60582 (0.37793)	0.60025 (0.37439)	0.59901 (0.37393)
0.6	0.8	0.59984 (0.18359)	0.59630 (0.17904)	0.59631 (0.18032)	0.59509 (0.17975)	0.59510 (0.18096)
0.7	0.8	0.60127 (0.23289)	0.59882 (0.22893)	0.59849 (0.22993)	0.59563 (0.22820)	0.59452 (0.22960)

Inside values of ( ) are standard deviations.
Simulation includes 1000 trials.
$\rho_{S_1S_2} = 0.6$, n = 100.

TABLE 2

ESTIMATED CORRELATION FOR THE UNRELATED RANDOMIZED RESPONSE MODEL WITH n=200.

P_1	P_3	$\rho_{Y_1Y_2}$				
		0.22	0.3	0.4	0.5	0.6
0.3	0.4	0.61612 (0.83676)	0.60314 (0.85359)	0.60058 (0.85901)	0.60042 (0.85139)	0.61474 (0.84824)
0.3	0.6	0.59779 (0.22317)	0.59062 (0.22043)	0.59204 (0.22062)	0.59265 (0.22263)	0.59319 (0.22387)
0.3	0.7	0.60015 (0.01545)	0.59615 (0.15507)	0.59669 (0.15542)	0.59618 (0.15339)	0.59661 (0.15338)
0.3	0.8	0.59935 (0.12118)	0.59743 (0.11063)	0.59692 (0.11145)	0.59634 (0.11006)	0.59681 (0.11015)
0.4	0.6	0.59577 (0.25985)	0.58614 (0.25549)	0.58768 (0.25577)	0.58760 (0.25998)	0.58791 (0.26149)
0.4	0.7	0.59954 (0.16211)	0.59482 (0.16305)	0.59536 (0.16334)	0.59454 (0.16172)	0.59488 (0.16208)
0.4	0.8	0.59914 (0.11331)	0.59702 (0.11197)	0.59649 (0.11275)	0.59582 (0.11146)	0.59625 (0.11166)
0.6	0.7	0.60591 (0.26218)	0.60090 (0.26366)	0.60251 (0.26445)	0.59894 (0.26050)	0.59900 (0.26259)
0.6	0.8	0.60057 (0.12483)	0.59872 (0.12335)	0.59830 (0.12402)	0.59700 (0.12220)	0.59739 (0.12274)
0.7	0.8	0.60015 (0.16035)	0.59923 (0.16105)	0.59868 (0.16094)	0.59650 (0.15794)	0.59640 (0.15837)

Inside values of ( ) are standard deviations.
Simulation includes 1000 trials.
$\rho_{S_1S_2} = 0.6$, n = 200.