

**CORRELATION ANALYSIS FOR THE RANDOMIZED
RESPONSE MODELS**

By

GEUN-SHIK HAN

**Bachelor of Science in Statistics
Korea University
Seoul, Korea
1984**

**Master of Science in Statistics
Iowa State University
Ames, Iowa
1988**

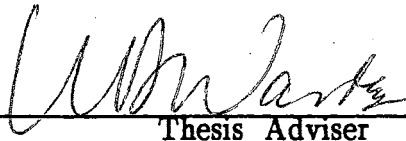
**Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfilment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 1993**

THE UNIVERSITY OF MICHIGAN

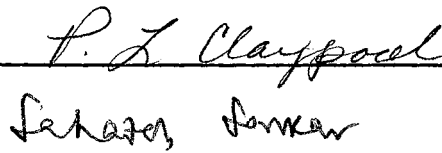
Thesis
1993D
H233C

CORRELATION ANALYSIS FOR THE RANDOMIZED
RESPONSE MODELS

Thesis Approved:

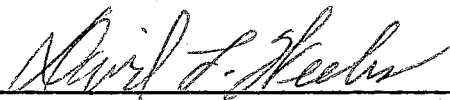


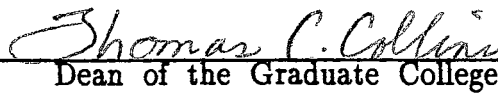
Thesis Adviser



Sethares, Jonkaw







Dean of the Graduate College

ACKNOWLEDGMENTS

I wish to express my sincere thanks and appreciations to my advisor, Dr. William D. Warde, for the constant encouragement, assistance and guidance during the course of this study. I am grateful to Dr. L. Claypool, Dr. S. Sarkar, and Dr. D. Schreiner for serving on my graduate committee. Their suggestions and support were very helpful throughout the study.

I wish to thank my wife, Saim Woo, for her sacrifice and understanding, and my daughters, Ji-Hee, and Ji-Young for their patience.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION.	1
Sources of Error in Surveys.	2
The Effect of Refusal on the Estimates	3
Sensitive Question Bias	4
Scope of the Study	4
II. CORRELATION ANALYSIS FOR THE DICHOTOMOS RANDOMIZED RESPONSE TECHNIQUE	6
Literature Review	6
The Warner Model	7
Unrelated Randomized Response Model	11
Bivariate Binomial Data Analysis Collected by Warner's Randomized Response Technique.	18
Correlation Analysis	20
Cell proportion Estimation	26
Test of Hypothesis	26
Sample Size Estimation	27
A New Randomized Response Technique for Bivariate Binomial Data	29
Correlation Analysis for the Warner's Model Versus Direct Survey	34
Bivariate Binomial Data Analysis Collected by the Unrelated Randomized Response Models.	35
Correlation Analysis	36
Test of Hypothesis	41
Sample Size Estimation	41
Unrelated Randomized Response Technique versus Direct Survey Technique	43
III. RANDOMIZED RESPONSE TECHNIQUE FOR MULTIPLE ATTRIBUTES	51
Additive Randomized Response Technique	51
Correlation Analysis for Another version of Additive Model	55
Correlation Analysis for the Additive Model.	62
Scrambled Randomized Response Technique	65

Chapter	Page
Correlation Analysis for the Multivariate version of the Scrambled Randomized Response Model	69
Multiproportional Randomized Response Technique	73
Test of Hypothesis	79
Multiproportional Randomized Response Technique with Reordering Cell Numbers	81
IV. RANDOMIZED RESPONSE TECHNIQUE FOR THE QUANTITATIVE ATTRIBUTES	84
Additive Randomized Response Models	84
Scrambled Randomized Response Models	94
LITERATURE CITED	108

LIST OF TABLES

Table	Page
1. Reduced Correlations	19
2. Estimated Correlation (n=100) for the Unrelated Randomized Response Model	39
3. Estimated Correlation (n=200) for the Unrelated Randomized Response Model	40
4. Estimated Correlation for the Unrelated Randomized Model versus Direct Survey.	50
5. Reordered Contingency Table	56
6. Transformed Response for T = 5	67
7. Reordered Population Categories	70
8. Transformed Response for T = 9	71
9. Estimated Means and Standard Deviations for the Additive Randomized Response Model	90
10. Estimated Correlations and Biases for the Additive Randomized Response Models	91
11. Conditions for the Positive Bias for the Additive Randomized Response Model	93
12. Estimated Correlations and Biases for the Scrambled Randomized Response Models (n=100)	100
13. Estimated Correlations and Biases for the Scrambled Randomized Response Models (n=200)	101
14. Estimated Means and Standard Deviations for the Scrambled Randomized Response Models	102
15. Conditins for the Positive Biases of Scrambled Randomized Response Models	103

LIST OF FIGURES

Figure	Page
1. Warner's Randomizing Device	8
2. The Unrelated Randomizing Device	13
3. Hopkins' Randomizing Device	75

CHAPTER I

INTRODUCTION

In most of sampling theory, it has been assumed that the data collected on the units in the sample are always accurate or true values of the characteristics observed, and that the estimates of the population values obtained from the data are subject only to sampling errors. In practice, the situation is rarely as simple.

The nonsampling errors that arise from the method of measurement or interviewing, and other sources of errors in surveys are present in a census. These nonsampling errors may be equally as important as sampling error, or perhaps more important for surveys of human populations. For voting questions in Chicago, approximately one third of all residents who reported voting in the primary election were found not to have voted when the record was checked (Sudman and Bradburn, 1983). This problem becomes more serious when respondents are questioned about sensitive matters, especially when truthful answers may place them in an unfavorable light. The question dealing with acceptance of racial intermarriage produced a difference by race of interviewer of over 45 percent (Hatchett and Schuman, 1975). For the socially undesirable questions, direct measurement of valid information on human populations is difficult because of untruthful reporting and refusal to respond.

The randomized response methodology of survey technique is designed to encourage cooperation and truthful replies to questions involving socially undesirable activities.

1.1 Sources of Error in Surveys

The theory of survey sampling assumes throughout that some kind of probability sampling is used and that the observation, say y_i , on the i -th unit is the correct value for that unit. The error of estimation arises solely from the random sampling variation that is present when n of the units are measured instead of the complete population of N units. This makes up what is termed sampling error.

Many sampling techniques and estimation techniques have been developed to collect data and use methods of estimation so as to minimize sampling error, and improve the efficiency of survey estimates.

Even though the various survey operations carried out strictly according to the rules laid down are expected to yield the true value, x_i , which is the characteristic under study, this can rarely be achieved in practice. The discrepancy between the value actually obtained, y_i , to be called the survey value and the true value is called the observational or response error (Hansen, Hurwitz, and Madow, 1951) and arises primarily from the variable performance of enumerators and lack of precision in measurement techniques. Hence even when the sampling fraction is unity, that is $n = N$, the value of any population parameter obtained from a census will differ from the true value of the parameter. The discrepancy between the survey value and the true value also arises due to several other causes, such as incomplete coverage, faulty methods of selection, faulty methods of estimation, and so on. Together with the observational errors these make up what are termed as non-sampling errors (Sukhatme, Sukhatme, Sukhatme, & Asok, 1984). Deming (1960) and Cochran (1977) have discussed in detail the sources of non-sampling error and its effects on sampling estimates.

The main source of non-sampling error in any survey comes as the result of

non-response. Non-response occurs when an element of the sample fails to provide data to the researcher. In effect, this keeps the sample from truly being a random sample from the survey population. This can often lead to a considerable bias in the survey results and hence distort the conclusions regarding the population of interest. Four distinct types of non-response are generally recognized.

These are as follows.

- i) non-coverage
- ii) unable to answer
- iii) not at home
- iv) hard core refusal

Among all these problems, we will study two types of non-sampling errors:

i) non-response error resulting from the respondents who adamantly refuse to be interviewed.

ii) response error resulting from giving incorrect answers.

Systematic distortion of the respondent's true status jeopardizes the validity of survey measurements. Unlike random error, response bias does not cancel out over repeated measurement.

1.1.1 The Effect of Refusal on the Estimates

The answers from a survey are heavily weighted with people who are willing to respond, and many characteristics of these people are different from the characteristics of people who are not willing to respond. Deming (1960, pp. 67) states "People sometimes enquire whether 50 % response is good enough, or whether 80 %, or 90 % is good enough, or just what do we consider to be good ? The answer depends on the characteristic and how it is distributed. If half the

people or firms with very high incomes, sales, or inventories are nonrespondents, the error may be large, even though the response over all classes combined be only 5 %".

1.1.2 Sensitive Question Bias

When the survey is about sensitive matters, the non-response and response error becomes more serious because the respondent will tend to give incorrect answers, the interviewer may hesitate to ask such questions, and sometimes even omit or alter them. The people who have high incomes will try to underreport, and the people who have low incomes will try to overreport. Worse cases are where the subject is asked to respond to questions about sensitive issues such as: abortion, drunken driving, or marijuana smoking. The respondents often prefer to give an answer that is socially acceptable.

1.2 Scope of the Study

After Warner's (1965) proposal, many other researchers improved and developed the theory and techniques of the randomized response models. Their main discussion was the estimation of population proportions or population means. Here we are studying the covariance and correlation between two sensitive variables. Although Kraemer (1980), Fox and Tracy (1984), and Edgell, Himmelfarb, and Cira (1986) discussed estimation of correlations, their assumptions are not practical.

A review of Warner's model and the unrelated randomized response models are given in chapter II. And the analysis of correlation for Warner's model and the unrelated randomized response model is also given in chapter II.

Chapter III contains a review of the additive models, the scrambled models,

and multiproportion models, and the estimation of the product moment correlation for each model is also given.

In chapter IV, the correlation analysis for the continuous sensitive variables is given for the additive models and the scrambled randomized response models. The correlation between the response variables is expressed in terms of the correlation between the two sensitive variables and bias due to random device. The bias due to random device is estimated for the additive and scrambled randomized response models.

CHAPTER II

CORRELATION ANALYSIS FOR THE DICHOTOMOUS RANDOMIZED RESPONSE TECHNIQUE

Literature Review

In surveys of human populations, respondents are not likely to participate or tell the truth when the reply may tend to stigmatize them in the eyes of the surveyer or the reply represents a socially undesirable behavior.

Sample surveys of human populations have established the fact that refusal to respond and intentional giving of incorrect answers are two main sources of non sampling error. The bias produced by these two sources of error can sometimes make the sample estimates seriously misleading. This problem becomes more serious when respondents are questioned about sensitive matters, especially those questions for which truthful answers may place them in an unfavorable light. For example, questions about the number of times that a woman has had an abortion, incidence of drunken driving, use of marijuana, sexual activity and child abuse will create biases of these types.

In surveys on these topics, the respondents may refuse to answer or give incorrect answers. This will lead to response bias, and these sources of bias persist no matter how much effort is put into completeness of returns or into the improvement of sampling techniques.

A survey technique for eliminating or reducing this bias was introduced by Warner (1965) and is generally called the randomized response technique.

The technique was designed to eliminate or reduce response bias for

sensitive questions in estimation of the proportion of a population belonging to a sensitive group. In other words, this technique reduced the frequency of false (incorrect) answers by giving the respondent a randomization device.

2.1 The Warner Model

Suppose that every person in a population belongs to either group S or the complementary group \bar{S} , and it is necessary to estimate the proportion of persons who belong to group S from a sample survey. A simple random sample of n people is drawn with replacement from the population and provisions are made for each person selected to be interviewed. Before the interviews, each interviewer is furnished with an identical spinner (random device, see figure 1) which points to the question Q with probability P , and to the question \bar{Q} with probability $1-P$. A die, a container with marbles, and a deck of cards, each can be used as randomization devices for Warner's model. In each interview, the respondent is asked to spin the spinner unobserved by the interviewer and report only 'yes' or 'no' according to the question to which the spinner points. The interviewer is told not to make any attempt to identify the group to which the spinner points. Thus the interviewer does not know whether the respondent's answer is for the sensitive question or the nonsensitive question, and all that the interviewer records is the respondent's answer (yes or no). Let

π_S = the true population proportion of respondents belonging to group S

P = the probability that the spinner points to S , and

$$r_i = \begin{cases} 1 & \text{if the } i\text{-th respondent reports 'yes',} \\ 0 & \text{if the } i\text{-th respondent reports 'no'}. \end{cases}$$

Each respondent is provided with a randomization device by which he or

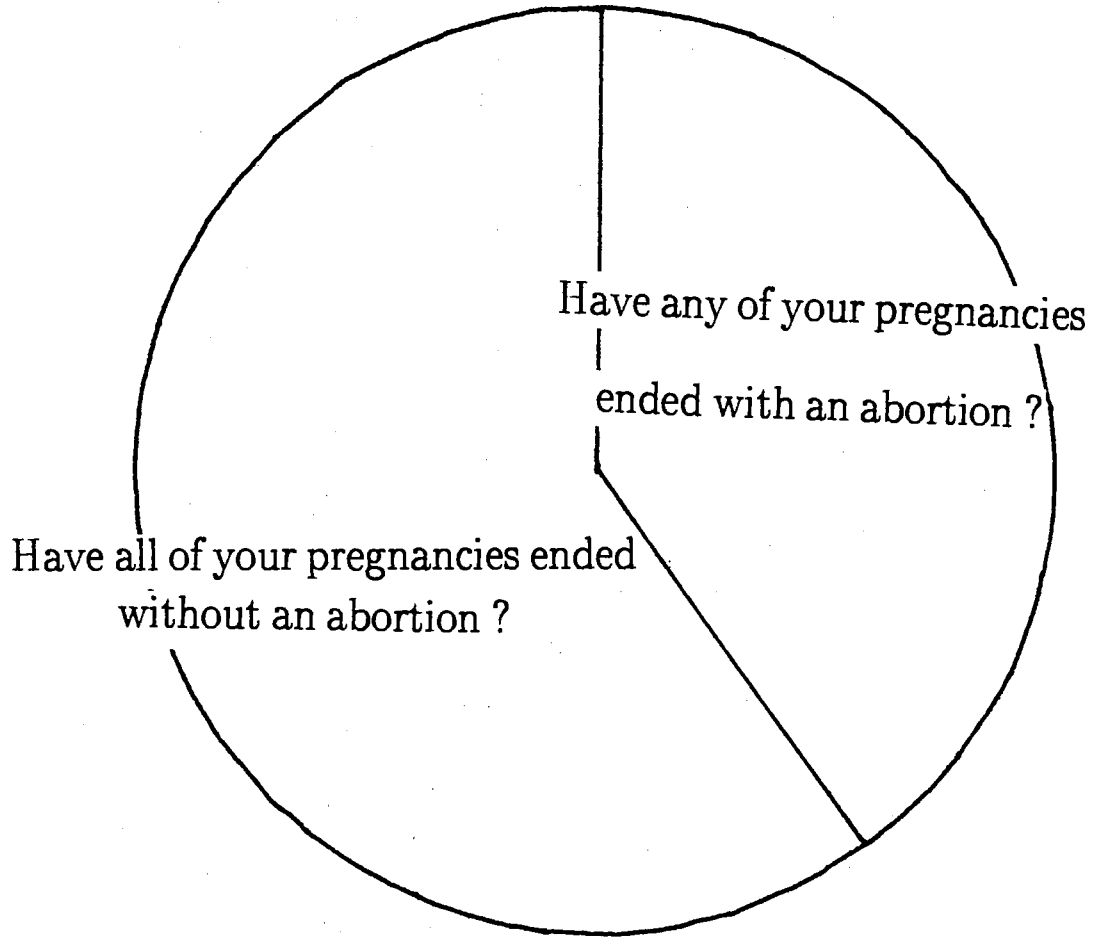


Figure 1. Warner's Randomizing Device

she chooses one of the two questions.

For example the two questions might be

Q ; Have any of your pregnancies ended with an abortion ?

\bar{Q} ; Have all of your pregnancies ended without an abortion ?

The randomization device is designed to ask question Q with probability P, where $0 < P < 1$ as in figure 1. The proportion of the sensitive question (P) is preassigned and question \bar{Q} is the complement of question Q.

The probability of getting a 'yes' response (λ) is

$$\begin{aligned}\lambda &= P(r_i = 1) = P(\text{yes} \mid Q) P(Q) + P(\text{yes} \mid \bar{Q}) P(\bar{Q}) \\ &= \pi_S P + (1 - \pi_S) (1 - P) \\ &= (2P - 1) \pi_S + (1 - P).\end{aligned}\tag{2.1.1}$$

Denoting the number of 'yes' responses in the sample by n_1 , the sample proportion of 'yes' responses is $\frac{n_1}{n}$ ($\hat{\lambda} = \frac{n_1}{n}$), and since n_1 follows a binomial distribution with parameters n and λ ,

$$E \hat{\lambda} = E \frac{n_1}{n} = \frac{1}{n} n \lambda = \lambda, \text{ so } \frac{n_1}{n} \text{ is an unbiased estimator of } \lambda.$$

Therefore an unbiased estimate of π_S is

$$\begin{aligned}\hat{\pi}_S &= \frac{P - 1}{2P - 1} + \frac{\hat{\lambda}}{2P - 1} \\ &= \frac{P - 1}{2P - 1} + \frac{n_1}{(2P - 1)n}\end{aligned}\tag{2.1.2}$$

and since n_1 follows binomial distribution with parameters n and λ , Eq.(2.1.2) is an unbiased estimate of π_S when $P \neq \frac{1}{2}$.

If $P = \frac{1}{2}$, the probability of getting a 'yes' response does not even depend on π . If $\frac{1}{2} < P < 1$ (or $0 < P < \frac{1}{2}$), the person interviewed provides useful but not absolute information as to exactly which group he (she) is in. In this context the P can be thought of as describing the nature of the cooperation between the interviewer and the respondent.

From Eq.(2.1.2) the variance of an unbiased estimate of π_S is given by

$$\begin{aligned} \text{Var}(\hat{\pi}_S) &= \frac{n\lambda(1-\lambda)}{n^2(2P-1)^2} \\ &= \frac{[(1-P) + (2P-1)\pi_S] [1 - \{(1-P) + (2P-1)\pi_S\}]}{n(2P-1)^2} \\ &= \frac{\pi_S(1-\pi_S)}{n} + \frac{1}{n} \left[\frac{P(1-P)}{(2P-1)^2} \right], \end{aligned} \quad (2.1.3)$$

where the second term on the right hand side of Eq.(2.1.3) is the variance due to the random device. This bias is symmetric about $P = \frac{1}{2}$, and as P increases to 1 (P decreases to 0) the bias decreases.

$$\text{Since } E(n_1) = n\lambda, \quad E(n_1)^2 = n\lambda + n(n-1)\lambda^2$$

$$\begin{aligned} E \left[\frac{\hat{\lambda}(1-\hat{\lambda})}{n-1} \right] &= E \left[(n-1)^{-1} \frac{n_1}{n} \left(1 - \frac{n_1}{n} \right) \right] \\ &= E \left[\frac{n n_1 - n_1^2}{n^2(n-1)} \right] \\ &= \frac{1}{n^2(n-1)} [n^2\lambda - (n\lambda + n(n-1)\lambda^2)] \\ &= \frac{\lambda(1-\lambda)}{n}. \end{aligned}$$

Thus an unbiased estimate of $\text{Var}(\hat{\pi}_S)$ is given by

$$\begin{aligned}\widehat{\text{Var}}(\hat{\pi}_S) &= \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)(2P-1)^2} \\ &= \frac{\hat{\pi}_S(1-\hat{\pi}_S)}{n-1} + \frac{1}{n-1} \left[\frac{P(1-P)}{(2P-1)^2} \right].\end{aligned}\quad (2.1.4)$$

The first term of Eq.(2.1.4) is the variance of $\hat{\pi}_S$ as in the direct survey procedure, therefore our variance consists of the variance due to sampling plus the variance due to the random device. When the selection probability, P close to 0.5, the variance due to the random device increases.

2.2 Unrelated Randomized Response Model

The unrelated question randomized response technique was developed by Horvitz, Shah, and Simmons (1967) and its theoretical framework has been discussed by Greenberg, Abul-Ela, Simmons, and Horvitz (1969). Abul-Ela, Greenberg, and Horvitz (1967) extended the unrelated randomized response technique to a multiproportions model. Gould, Shah, and Abernathy (1969) considered two trials per person for the unrelated randomized response technique, and Moors (1971) compared Warner's model and the unrelated model. The first major field trial of the unrelated randomized response technique conducted by the Research Triangle Institute for the National Center for Health Statistics (1965-1966).

This technique requires the respondents to randomly select one of two unrelated questions (sensitive question or unrelated nonsensitive question). In the Warner's randomized response technique, the interviewer asked the respondent

whether he or she belongs to the sensitive group S or to the complementary group \bar{S} . If two unrelated questions (including one nonsensitive question) are used, the respondent may have more confidence that his or her response is confidential and so this will increase the cooperation of the respondent. This possibility leads to the unrelated question model.

Two independent, non-overlapping simple random samples of size n_1 and n_2 are drawn from the population. The size of n_1 and n_2 are not necessarily equal.

Every respondent in the samples is asked to reply with only a 'yes' or 'no' answer to the specific single question which turns up in his case. The selection of the question is made by a randomization device on probability basis. In this way, the respondent's status is not revealed to the interviewer provided that the interviewer cannot observe the randomization process in the device.

Suppose the randomization device consists of a wheel of two parts (Figure 2). In this model, two randomization devices need to be used. Wheel 1 is used for the respondents in the first sample, and wheel 2 is used for respondents in the second sample. If more than one interviewer is used in either sample, every interviewer in sample 1 has a randomization device identical to wheel 1, and every interviewer in sample 2 has a randomization device identical to wheel 2. The two wheels, 1 and 2, must also be different with respect to the probability that the sensitive question, Q_1 , will be selected.

Let the randomization devices be such that the sensitive question Q_1 is represented on a probability basis by P_1 on wheel 1 and P_2 on wheel 2, and $P_1 \neq P_2$. Similarly, let the unrelated non-sensitive question, Q_2 be represented on a probability basis by $(1 - P_1)$ on wheel 1, and $(1 - P_2)$ on wheel 2.

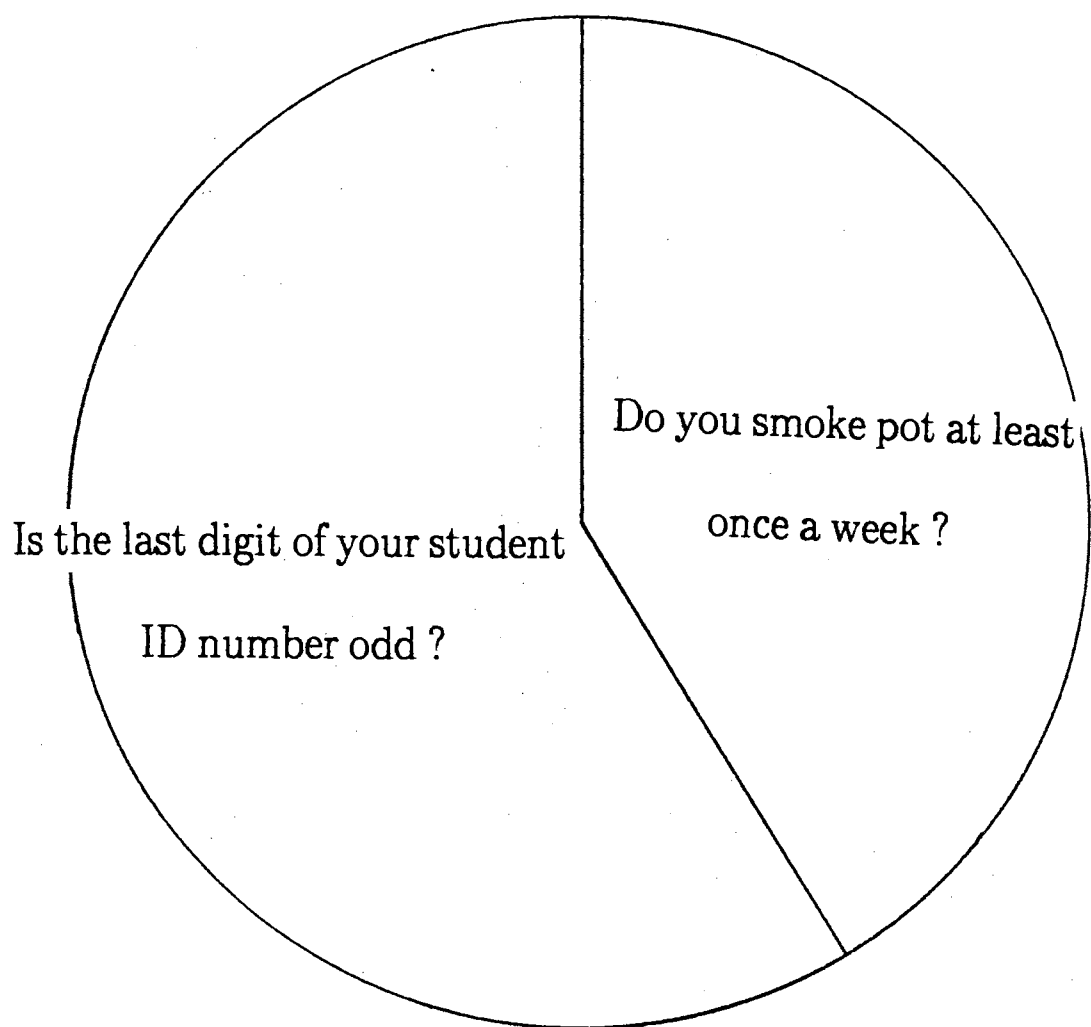


Figure 2. The Unrelated Randomizing Device

The unrelated question model uses two questions such as

Q1 ; Do you smoke pot at least once a week ?

Q2 ; Is the last digit of your student ID number odd ?

Let

π_S = the true population proportion of belonging to the group S.

π_Y = the true population probability of belonging to the nonsensitive group

Y.

P_1 = the probability that the sensitive question is selected by the random device in the first sample.

P_2 = the probability that the sensitive question is selected by the random device in the second sample.

n_1 ; size of the first sample

n_2 ; size of the second sample

$$r_{ij} = \begin{cases} 1 & \text{if the } i\text{-th respondent in the } j\text{-th sample reports 'yes' to the} \\ & \text{selected question.} \\ 0 & \text{otherwise,} \end{cases}$$

where $i=1,2, \dots, n_j$, $j=1, 2$.

The probability of getting a 'yes' response is

$$\begin{aligned} \lambda_1 &= P(r_{i1} = 1) = P(\text{yes} \mid Q_1) P(Q_1) + P(\text{yes} \mid Q_2) P(Q_2) \\ &= \pi_S P_1 + \pi_Y (1-P_1), \end{aligned} \quad (2.2.1)$$

$$\begin{aligned} \lambda_2 &= P(r_{i2} = 1) = P(\text{yes} \mid Q_1) P(Q_1) + P(\text{yes} \mid Q_2) P(Q_2) \\ &= \pi_S P_2 + \pi_Y (1-P_2). \end{aligned} \quad (2.2.2)$$

Denote the number of 'yes' responses in the first sample as n_{11} and in the second sample as n_{12} , where $n_{11} = \sum_{i=1}^{n_1} r_{i1}$ and $n_{12} = \sum_{i=1}^{n_2} r_{i2}$.

The sample proportion of 'yes' responses are $\hat{\lambda}_1 (= \frac{n_{11}}{n_1})$ and $\hat{\lambda}_2 (= \frac{n_{12}}{n_2})$ for each sample, and since n_{1j} follows a binomial distribution with parameters n_j and λ_j , $\frac{n_{11}}{n_1}$ and $\frac{n_{12}}{n_2}$ are unbiased estimates of $\hat{\lambda}_1$ and $\hat{\lambda}_2$, and from Eq.(2.2.1) and Eq.(2.2.2)

$$\pi_S = \frac{1}{P_1 - P_2} \left[(1 - P_2) \lambda_1 - (1 - P_1) \lambda_2 \right], \quad (2.2.3)$$

provided $P_1 \neq P_2$.

Therefore an estimate of π_S is given by

$$\begin{aligned} \hat{\pi}_S &= \frac{1}{P_1 - P_2} \left[(1 - P_2) \hat{\lambda}_1 - (1 - P_1) \hat{\lambda}_2 \right] \\ &= \frac{1}{P_1 - P_2} \left[(1 - P_2) \frac{n_{11}}{n_1} - (1 - P_1) \frac{n_{12}}{n_2} \right]. \end{aligned} \quad (2.2.4)$$

The denominator in Eq.(2.2.4) can become quite small by choosing P_2 too close to P_1 with the result that the point estimate of π_S might be greater than 1 in the unrelated question model. Thus, a first general rule is that P_2 should be selected as far from P_1 as possible without jeopardizing the likelihood of a respondent's cooperation. Obviously, when $P_2 = 0$, or 1, this would not be a randomization device at all.

The estimate of π_Y is

$$\begin{aligned}\hat{\pi}_Y &= \frac{1}{P_2 - P_1} \left[P_2 \hat{\lambda}_1 - P_1 \hat{\lambda}_2 \right] \\ &= \frac{1}{P_2 - P_1} \left[P_2 \frac{n_{11}}{n_1} - P_1 \frac{n_{12}}{n_2} \right].\end{aligned}$$

Since n_{1j} follows a binomial distribution with parameters n_j and λ_j , for $j=1, 2$, $\hat{\pi}_S$ and $\hat{\pi}_Y$ are unbiased estimates of π_S and π_Y .

From Eq.(2.2.4) the variance of $\hat{\pi}_S$ is

$$\text{Var}(\hat{\pi}_S) = (P_1 - P_2)^{-2} \left[\frac{(1-P_2)^2 \lambda_1 (1-\lambda_1)}{n_1} + \frac{(1-P_1)^2 \lambda_2 (1-\lambda_2)}{n_2} \right]. \quad (2.2.5)$$

Since $\text{Var}(\hat{\lambda}_i) = \frac{\lambda_i(1-\lambda_i)}{n_i}$, and $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are independent, then using Eq.(2.2.5),

an unbiased estimate of $\text{Var}(\hat{\pi}_S)$ is

$$\hat{\text{Var}}(\hat{\pi}_S) = \frac{1}{(P_1 - P_2)^2} \left[\frac{(1-P_2)^2 \hat{\lambda}_1 (1-\hat{\lambda}_1)}{n_1 - 1} + \frac{(1-P_1)^2 \hat{\lambda}_2 (1-\hat{\lambda}_2)}{n_2 - 1} \right],$$

provided $P_1 \neq P_2$.

Since $\lambda_i = \pi_S P_i + \pi_Y (1 - P_i)$, if π_Y is close enough to zero, the variance of the estimate, $\hat{\pi}_S$, is smaller, but if $\pi_Y = 0$, this technique reduces to the direct survey. Therefore the respondent reporting sensitive information is not protected by the method whenever $\pi_Y = 0$.

To reduce the variance, $\text{Var}(\hat{\pi}_S)$, it is desirable to choose P_1 as far away from P_2 as possible and to keep the respondent's confidence, P_1 and P_2 should be as large as can be efficiently afforded, and π_Y should be large.

But in the case when the true probability of a 'yes' answer to the unrelated non-sensitive question in the population is known in advance, one sample is enough to estimate $\hat{\pi}_S$. When the distribution of π_Y is known in advance, from Eq.(2.2.1), an estimate of π_S is

$$\begin{aligned}\hat{\pi}_S &= \frac{\hat{\lambda}_1 - (1-P_1) \pi_Y}{P_1} \\ &= \frac{\frac{n_{11}}{n_1} - (1-P_1) \pi_Y}{P_1}.\end{aligned}\tag{2.2.6}$$

Since n_{11} follows a binomial distribution with parameters n_1 and λ_1 , $\hat{\pi}_S$ is an unbiased estimate of π_S .

From Eq.(2.2.6) the variance of $\hat{\pi}_S$ is

$$\text{Var}(\hat{\pi}_S) = \frac{\lambda_1(1-\lambda_1)}{n_1 P_1^2}.$$

Since n_{11} follows a binomial distribution with parameters n_1 and λ_1 and $E(n_{11}) = n_1 \lambda_1$ and $E(n_{11}^2) = n_1 \lambda_1 + n_1 (n_1-1) \lambda_1^2$, therefore

$$E\left[\frac{\hat{\lambda}_1(1-\hat{\lambda}_1)}{n_1-1}\right] = \frac{\lambda_1(1-\lambda_1)}{n_1}.$$

Hence, an unbiased estimate of $\text{Var}(\hat{\pi}_S)$ is given by

$$\text{Var}(\hat{\pi}_S) = \frac{\hat{\lambda}(1-\hat{\lambda}_1)}{(n_1-1)P_1^2} \quad (2.2.7)$$

Greenberg, Abul-Ela, Simmon, and Horvitz (1969) showed that the unrelated randomized response model with known π_Y is better than that with unknown π_Y and both the unrelated randomized response model with known π_Y and with unknown π_Y are better than Warner's model despite the fact that Warner's model is always asking about the sensitive group, S either the complementary group of S.

2.3 Bivariate Binomial Data Analysis Collected by Warner's Randomized Response Technique

Using randomized response models, we can estimate the proportion or mean of a population, but we cannot observe individual level data. Therefore direct computation of correlation procedures are not possible.

Consider two sensitive variables S_1 and S_2 with dichotomized qualitative groups ($S_1 \bar{S}_1$) and ($S_2 \bar{S}_2$), along with a sample of size n drawn from a bivariate binomial distribution with correlation ρ . Using the randomized response technique, we may estimate the marginal parameters π_1 and π_2 for a 2×2 contingency table, but we may not observe cell proportions, n_{00} , n_{01} , n_{10} , n_{11} . Denoting "yes" response = "1" and "no" response = "0" for each group, we get

		S_2		
		yes	no	
S_1	yes	n_{11}	n_{10}	n_{1+}
	no	n_{01}	n_{00}	n_{0+}
		n_{+1}	n_{+0}	

Therefore we have analytic limits to analyse the data collected by randomized response models, and as we showed in the introduction, the estimated variance will be inflated by the random device bias. If we estimate the covariance and correlation using the observed response data, we may have reduced (inflated) estimates and tests of hypothesis will also give misleading results as we can see in table 1. The estimated correlation between the two reported data obtained by the unrelated randomized response technique is shown table 1.

The estimated correlation between the two reported variables decrease as the selection probabilities (P_1 and P_3) for the sensitive variables decrease.

TABLE 1
REDUCED CORRELATIONS

P_1	P_3	Estimated Correlation Unrelated model	Estimated True Correlation
0.4	0.4	0.17772	0.5954
0.6	0.4	0.24777	0.6007
0.7	0.7	0.31478	0.5996
0.8	0.8	0.38185	0.6000
0.9	0.9	0.48053	0.5991

The true correlation is 0.6
 $n = 100$

Our concerns are how do we correct these correlations, covariances, and test statistics.

Kraemer (1980) considered estimation of the correlation coefficient between two sensitive groups each surveyed by Warner's technique and the unrelated randomized response models when the population parameters of the nonsensitive variables are known.

2.3.1 Correlation Analysis

Here we propose a correlation analysis for Warner's model.

We will show that the correlation between two unknown sensitive variables is the same as that of the two observed response variables for the Warner's model.

i.e., $\rho_{S_1 S_2} = \rho_{r_1 r_2}$, for Warner's model.

As we showed in Warner's model, the response variables can be expressed by

$$r_1 = (2P_1 - 1)S_1 + (1 - P_1)$$

$$r_2 = (2P_2 - 1)S_2 + (1 - P_2).$$

The means and variances can be expressed as

$$E r_1 = (2P_1 - 1) E S_1 + (1 - P_1) \tag{2.3.1}$$

$$E r_2 = (2P_2 - 1) E S_2 + (1 - P_2) \tag{2.3.2}$$

$$V(r_1) = (2P_1 - 1)^2 V(S_1) \tag{2.3.3}$$

$$V(r_2) = (2P_2 - 1)^2 V(S_2) \quad (2.3.4)$$

$$\begin{aligned} E r_1 r_2 &= (2P_1 - 1)(2P_2 - 1) E S_1 S_2 + (2P_1 - 1)(1 - P_2) E S_1 \\ &\quad + (1 - P_1)(2P_2 - 1) E S_2 + (1 - P_1)(1 - P_2) \\ &= (2P_1 - 1)(2P_2 - 1) E S_1 S_2 + K, \end{aligned} \quad (2.3.5)$$

where $K = (2P_1 - 1)(1 - P_2) E S_1 + (1 - P_1)(2P_2 - 1) E S_2 + (1 - P_1)(1 - P_2)$.

Now the formula for the correlation between S_1 and S_2 is

$$\rho_{S_1 S_2} = \frac{\sigma_{S_1 S_2}}{\sigma_{S_1} \sigma_{S_2}} = \frac{E S_1 S_2 - E S_1 E S_2}{\sigma_{S_1} \sigma_{S_2}}. \quad (2.3.6)$$

Substituting Eq.(2.3.3) and Eq.(2.3.5) into Eq.(2.3.6), then we have

$$\rho_{S_1 S_2} = \frac{E r_1 r_2 - E r_1 E r_2 + (1 - P_2) E r_1 + (1 - P_1) E r_2 - (1 - P_1)(1 - P_2) - K}{\sigma_{r_1} \sigma_{r_2}}$$

Substituting for $E S_1$ and $E S_2$ yields

$$K = (1 - P_2) E r_1 - (1 - P_2)(1 - P_1) + (1 - P_1) E r_2.$$

Therefore

$$\rho_{S_1 S_2} = \frac{E r_1 r_2 - E r_1 E r_2}{\sigma_{r_1} \sigma_{r_2}} = \rho_{r_1 r_2'} \quad (2.3.7)$$

and hence $\text{Cov}(S_1, S_2) = \rho_{S_1 S_2} \sigma_{S_1} \sigma_{S_2}$.

The bivariate binomial density function of (S_1, S_2) is

$$\pi_{S_1 S_2} = \pi_{1+}^{s_1} (1-\pi_{1+})^{1-s_1} \pi_{+1}^{s_2} (1-\pi_{+1})^{1-s_2} \left[1 + c \frac{(s_1 - \pi_{1+})(s_2 - \pi_{+1})}{\pi_{1+}(1-\pi_{1+})\pi_{+1}(1-\pi_{+1})} \right],$$

where $c = \rho \sqrt{\pi_{1+}(1-\pi_{1+})\pi_{+1}(1-\pi_{+1})}$.

From the density,

$$\pi_{11} = \pi_{1+} \pi_{+1} + c$$

$$\pi_{10} = \pi_{1+} (1-\pi_{+1}) - c$$

$$\pi_{01} = (1-\pi_{1+}) \pi_{+1} - c$$

$$\pi_{00} = (1-\pi_{1+})(1-\pi_{+1}) + c.$$

This can be displayed in a 2 x 2 table as follow:

		S_2		
		1	0	
S_1	1	π_{11}	π_{10}	π_{1+}
	0	π_{01}	π_{00}	π_{0+}
		π_{+1}	π_{+0}	

and observed cells are given by

		1	S_2	0					
		<table style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; border-right: 1px solid black; padding: 5px; text-align: center;">n_{11}</td> <td style="padding: 5px; text-align: center;">n_{10}</td> </tr> <tr> <td style="border-right: 1px solid black; padding: 5px; text-align: center;">n_{01}</td> <td style="padding: 5px; text-align: center;">n_{00}</td> </tr> </table>		n_{11}	n_{10}	n_{01}	n_{00}	n_{1+}	n_{0+}
n_{11}	n_{10}								
n_{01}	n_{00}								
S_1	1								
	0								
		n_{+1}		n_{+0}					

The likelihood function is

$$L \propto \pi_{11}^{n_{11}} \pi_{10}^{n_{10}} \pi_{01}^{n_{01}} \pi_{00}^{n_{00}}$$

The log likelihood function is

$$\log L \propto n_{11} \log \pi_{11} + n_{10} \log \pi_{10} + n_{01} \log \pi_{01} + n_{00} \log \pi_{00}$$

By taking derivatives of $\log L$ with respect to π_{1+} , π_{+1} , and c , and then equating to 0 we obtain :

$$\frac{\partial \log L}{\partial \pi_{1+}} = \frac{n_{11} \pi_{+1}}{\pi_{+1} \pi_{1+} + c} - \frac{n_{01} \pi_{+1}}{\pi_{+1} (1 - \pi_{1+}) - c}$$

$$+ \frac{n_{10} (1 - \pi_{+1})}{(1 - \pi_{+1}) \pi_{1+} - c} - \frac{n_{00} (1 - \pi_{+1})}{(1 - \pi_{+1}) (1 - \pi_{1+}) + c} = 0,$$

$$\frac{\partial \log L}{\partial \pi_{+1}} = \frac{n_{11} \pi_{1+}}{\pi_{+1} \pi_{1+} + c} - \frac{n_{01} (1 - \pi_{1+})}{\pi_{+1} (1 - \pi_{1+}) - c}$$

$$+ \frac{n_{10} \pi_{1+}}{(1 - \pi_{+1}) \pi_{1+} - c} - \frac{n_{00} (1 - \pi_{1+})}{(1 - \pi_{+1}) (1 - \pi_{1+}) + c} = 0,$$

$$\frac{\partial \log L}{\partial c} = \frac{n_{11}}{\pi_{+1}\pi_{1+} + c} - \frac{n_{01}}{\pi_{+1}(1-\pi_{1+}) - c} + \frac{n_{10}}{(1-\pi_{+1})\pi_{1+} - c} - \frac{n_{00}}{(1-\pi_{+1})(1-\pi_{1+}) + c} = 0.$$

Solving these equations we may find the maximum likelihood estimators (m.l.e)

$$\hat{\pi}_{1+} = \frac{n_{1+}}{n}, \quad \hat{\pi}_{+1} = \frac{n_{+1}}{n}, \quad \hat{c} = \frac{n_{11}n_{00} - n_{10}n_{01}}{n^2}$$

and since $c = \rho \sqrt{\pi_{1+}(1-\pi_{1+})\pi_{+1}(1-\pi_{+1})}$

$$\hat{\rho} = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{1+}n_{0+}n_{+1}n_{+0}}, \text{ and is the m.l.e of } \rho \text{ by the invariance}$$

property of m.l.e.

To see the properties of the estimated correlation, we introduce the following Theorems.

Theorem (Muirhead 1982) Multivariate central limit theorem

Let the 2-component vectors $\mathbf{r}_1, \mathbf{r}_2, \dots$ be independently and identically distributed vectors with $E(\mathbf{r}_i) = \boldsymbol{\mu} = (\mu_1, \mu_2)'$ and covariance matrices

$$\text{Cov}(\mathbf{r}_{i1}, \mathbf{r}_{i2}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Let $\bar{r}_n = \frac{1}{n} \sum_{i=1}^n r_i$, then $\sqrt{n}(\bar{r}_n - \mu) \longrightarrow N_2(0, \Sigma)$ as $n \longrightarrow \infty$.

Proof : See Muirhead (1982).

Now since $\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$, Σ can be rewritten as follows

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12} \sigma_1 \sigma_2 \\ \rho_{12} \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}. \quad (2.3.8)$$

Theorem (Anderson 1984)

Let $(r_{1t}, r_{2t})'$, $t = 1, 2, \dots, n$, be i.i.d $N_2(\mu, \Sigma)$, where Σ is the same as Eq.(2.3.8), $\sigma_i^2 > 0$, $|\rho| \leq 1$. Let γ be the sample correlation coefficient. Then

$$\sqrt{n}(\gamma - \rho) \longrightarrow N[0, (1 - \rho^2)^2] \text{ as } n \longrightarrow \infty.$$

whereas $\gamma \longrightarrow N\left[\rho, \frac{(1 - \rho^2)^2}{n}\right]$.

Proof : See Anderson pp 122.

Hence

$$\hat{\text{Var}}(\gamma) = \frac{(1 - \gamma^2)^2}{n},$$

where $\gamma = \frac{n_{11} n_{00} - n_{10} n_{01}}{n_{1+} n_{0+} n_{+1} n_{+0}}$.

2.3.2 Cell Proportion Estimation

Here we propose a method to estimate the cell proportions, using the estimated correlation $\hat{\rho}_{S_1 S_2}$.

$$\hat{\pi}_{ij} = \hat{\pi}_{i.}^{s_1} (1 - \hat{\pi}_{i.})^{1-s_1} \hat{\pi}_{.j}^{s_2} (1 - \hat{\pi}_{.j})^{1-s_2} \left[1 + \hat{c} \frac{(s_1 - \hat{\pi}_{i.})(s_2 - \hat{\pi}_{.j})}{\hat{\pi}_{i.}(1 - \hat{\pi}_{i.})\hat{\pi}_{.j}(1 - \hat{\pi}_{.j})} \right],$$

where $\hat{c} = \hat{\rho} \sqrt{\hat{\pi}_{i.}(1 - \hat{\pi}_{i.})\hat{\pi}_{.j}(1 - \hat{\pi}_{.j})}$, $s_1 = 0, 1$, $s_2 = 0, 1$.

2.3.3 Tests of Hypothesis

Using the relationship between the correlation and chi-square statistics, we can perform a test of hypothesis.

In a 2 x 2 contingency table, by coding each level "0" and "1", we may derive a relationship between correlation and chi-square test statistics as

$$\begin{aligned} \chi^2 &= \sum_{i=0}^1 \sum_{j=0}^1 \frac{(\hat{\pi}_{ij} - \hat{\pi}_{i+} \hat{\pi}_{+j})^2}{\hat{\pi}_{i+} \hat{\pi}_{+j}} \\ &= \frac{(\hat{\pi}_{00} \hat{\pi}_{11} - \hat{\pi}_{01} \hat{\pi}_{10})^2}{\hat{\pi}_{0+} \hat{\pi}_{1+} \hat{\pi}_{+0} \hat{\pi}_{+1}} = \gamma^2. \end{aligned}$$

Thus in a 2 x 2 tables, γ^2 is simply $\frac{\chi^2}{n}$, i.e., the Pearson chi-square test statistic for independence of rows and columns divided by the sample size (Bishop, Fienberg, and Holland, 1975). But this relationship is not true in general for r x c contingency table (r > 2, c > 2).

The null hypothesis is $H_0 : \rho_{S_1 S_2} = 0$

The alternative is $H_A : \rho_{S_1 S_2} \neq 0$

To test this, we just need the value of $n \gamma_{S_1 S_2}^2$ with critical values of χ^2 ,

i.e., if $n \gamma^2 > \chi_\alpha^2$ we may say that there is evidence against H_0 at significance level α .

2.3.4 Sample Size Estimation

Now consider the sample size estimation procedure.

i. Some margin of error d in the estimated proportion π_S of units in the population has been agreed on, and there is a small risk α that we are willing to incur that the actual error is larger than d . So how many elements are needed to satisfied these conditions ?

We want

$$\Pr [| \hat{\pi}_S - \pi_S | \geq d] = \alpha.$$

We use simple random sampling, and since $\hat{\pi}_S$ is a maximum likelihood estimate, $\hat{\pi}_S$ is approximately normally distributed,

$$\hat{\pi}_S \sim AN \left[\pi_S, \frac{\pi_S(1-\pi_S)}{n} + \frac{P(1-P)}{n(2P-1)^2} \right].$$

Hence the formula that connects n with the desired degree of precision is

$$d = z_{\alpha/2} \sqrt{ \frac{\pi_S(1-\pi_S)}{n} + \frac{P(1-P)}{n(2P-1)^2} },$$

where z is the abscissa of the normal curve that cuts off an area of $\frac{\alpha}{2}$ in the upper tail.

Solving for n , we find

$$n = \frac{z^2 \alpha/2}{d^2} \left[\pi_S(1-\pi_S) + \frac{P(1-P)}{(2P-1)^2} \right].$$

At this point a difficulty appear that is common to all problems in the estimation of sample size. The above formula for n depends on the parameter of the population that is to be sampled.

The parameter is the quantity π_S that we would like to measure. Estimates of π_S for the purpose of estimating n may be obtained in a number of ways (Warde, 1991).

We try to obtain an estimate of π_S which is as close to the real value of π_S as possible, but which, it is not correct, will obtain a conservative value for n . This concept is referred to as a conservative assumption in that it is made to ensure that the specified tolerances are met or exceeded. In this situation, with no prior knowledge of π_S , using the value $\pi_S = 0.5$ will yield a conservative value for n . If some information on π_S is known, the value of π_S closest to 0.5 in the range of values of π_S that we believe to be reasonable a priori will yield a conservative value for n .

- ii. How many more elements are needed to get the same variance as the direct survey ?

We know the variance of $\hat{\pi}_S$ for the direct survey is given by

$$V(\hat{\pi}_S)_d = \frac{\pi_S(1-\pi_S)}{n_d},$$

and

$$V(\hat{\pi}_S)_R = \frac{\pi_S(1-\pi_S)}{n_R} + \frac{P(1-P)}{n_R(2P-1)^2}.$$

In order to have the same variance as the direct survey, the sample size (n_d) should be increased. The number of additional sample units needed can be found by reexpressing $V(\hat{\pi}_S)_R$ and $\text{Var}(\pi)_d$, respectively, as

$$V(\hat{\pi}_S)_R = \frac{\pi_S(1-\pi_S)(2P-1)^2 + P(1-P)}{n_R(2P-1)^2}$$

and

$$\text{Var}(\hat{\pi}_S)_d = \frac{\pi_S(1-\pi_S)(2P-1)^2}{n_d(2P-1)^2}$$

If we equate both variances and solve for n_R , then

$$n_R = n_d + \frac{P(1-P)}{(2P-1)^2 V(\hat{\pi}_S)_d}$$

The formula for n has been obtained, but n depends on the quantity $V(\hat{\pi}_S)_d$ that we would like to measure. Here we also use a conservative assumption that is made to ensure that the specified tolerances are met or exceeded.

2.3.5 A New Randomized Response Technique for Bivariate Binomial Data

Now we propose a new data collection technique and hence the cell proportions for a 2 x 2 contingency table can now be estimated.

Consider two sensitive variables S_1 and S_2 with dichotomized groups (S_1 \bar{S}_1) and (S_2 \bar{S}_2), which follow bivariate binomial distributions with correlation ρ . By applying Warner's technique, each interviewer is furnished with two spinners (random devices). In each interview, the respondent is asked to spin spinner 1 unobserved by the interviewer, and pointer 1 gives a question from the first two

statements. Without reporting the answer to the interviewer he (she) spins spinner 2, and pointer 2 gives a question from the second two statements. Then the respondent reports a pair of answers: yes yes; yes no; no yes; or no no; or a pair of coded answer 11; 10; 01; or 00.

The bivariate binomial density function of (x,y) is given by Hamdan and Martinson (1971) and Kocherlakota and Kocherlakota (1992) as

$$f_{x,y} = p_1^x q_1^{1-x} p_2^y q_2^{1-y} [1 + d(x-p_1)(y-p_2)] / (p_1 q_1 p_2 q_2)$$

where $c = \rho \sqrt{p_1 q_1 p_2 q_2}$, $x, y = 0, 1$, $q_1 = 1-p_1$ and $q_2 = 1-p_2$.

An example of possible questions to be used are

Q_1 ; Have you ever had an abortion ?

\bar{Q}_1 ; Have you ever not had an abortion ?

Q_2 ; Have you ever smoked marijuana ?

\bar{Q}_2 ; Have you never smoked marijuana ?

Let π_{S_j} be the population proportion of the respondent belong to group S_j

P_j be the probability that spinner j points to S_j ($j=1, 2$)

r_i be the reported response value from the i -th respondent and

$P[A_1 A_2 | Q_1 Q_2]$ be the conditional probability of response given questions $(Q_1 Q_2)$

$$A_j = \begin{cases} 1 & \text{if response is 'yes'} \\ 0 & \text{if response is 'no'} \end{cases} \quad \text{for } j = 1, 2$$

where Q_1 is the first question which is selected by the first device and Q_2 is the second question which is selected by the second device and a question is randomly selected one at a time by the random device, hence $P[Q_1 Q_2] = P[Q_1]P[Q_2]$, then the probability of getting a response (1 1, 1 0, 0 1, or 0 0) is

$$\begin{aligned}\lambda_{11} &= P[r_i = (1\ 1)] = P[1\ 1|Q_1\ Q_2] P(Q_1)P(Q_2) + P[1\ 0|Q_1\ \bar{Q}_2]P(Q_1)P(\bar{Q}_2) \\ &\quad + P[0\ 1|\bar{Q}_1\ Q_2] P(\bar{Q}_1)P(Q_2) + P[0\ 0|\bar{Q}_1\ \bar{Q}_2] P(\bar{Q}_1)P(\bar{Q}_2) \\ &= \pi_{11} P_1 P_2 + \pi_{10} P_1(1-P_2) + \pi_{01} (1-P_1)P_2 + \pi_{00} (1-P_1)(1-P_2)\end{aligned}$$

$$\begin{aligned}\lambda_{10} &= P[r_i = (1\ 0)] = P[0\ 1|Q_1\ Q_2] P(Q_1)P(Q_2) + P[0\ 0|Q_1\ \bar{Q}_2] P(Q_1)P(\bar{Q}_2) \\ &\quad + P[1\ 1|\bar{Q}_1\ Q_2] P(\bar{Q}_1)P(Q_2) + P[1\ 0|\bar{Q}_1\ \bar{Q}_2] P(\bar{Q}_1)P(\bar{Q}_2) \\ &= \pi_{01} (1-P_1)P_2 + \pi_{00} P_1(1-P_2) + \pi_{11} (1-P_1)P_2 + \pi_{10} (1-P_1)(1-P_2)\end{aligned}$$

$$\begin{aligned}\lambda_{01} &= P[r_i = (0\ 1)] = P[1\ 0|Q_1\ Q_2] P(Q_1)P(Q_2) + P[1\ 1|Q_1\ \bar{Q}_2] P(Q_1)P(\bar{Q}_2) \\ &\quad + P[0\ 0|\bar{Q}_1\ Q_2] P(\bar{Q}_1)P(Q_2) + P[0\ 1|\bar{Q}_1\ \bar{Q}_2] P(\bar{Q}_1)P(\bar{Q}_2) \\ &= \pi_{10} P_1 P_2 + \pi_{11} P_1(1-P_2) + \pi_{00} (1-P_1)P_2 + \pi_{01} (1-P_1)(1-P_2)\end{aligned}$$

$$\begin{aligned}\lambda_{00} &= P[r_i = (0\ 0)] = P[0\ 0|Q_1\ Q_2] P(Q_1)P(Q_2) + P[0\ 1|Q_1\ \bar{Q}_2] P(Q_1)P(\bar{Q}_2) \\ &\quad + P[1\ 0|\bar{Q}_1\ Q_2] P(\bar{Q}_1)P(Q_2) + P[1\ 1|\bar{Q}_1\ \bar{Q}_2] P(\bar{Q}_1)P(\bar{Q}_2) \\ &= \pi_{00} P_1 P_2 + \pi_{01} P_1(1-P_2) + \pi_{10} (1-P_1)P_2 + \pi_{11} (1-P_1)(1-P_2).\end{aligned}$$

In matrix form we can express those probabilities as

$$\begin{bmatrix} \lambda_{11} \\ \lambda_{10} \\ \lambda_{01} \\ \lambda_{00} \end{bmatrix} = \begin{bmatrix} P_1 P_2 & P_1(1-P_2) & (1-P_1)P_2 & (1-P_1)(1-P_2) \\ (1-P_1)P_2 & (1-P_1)(1-P_2) & P_1 P_2 & P_1(1-P_2) \\ P_1(1-P_2) & P_1 P_2 & (1-P_1)(1-P_2) & (1-P_1)P_2 \\ (1-P_1)(1-P_2) & (1-P_1)P_2 & P_1(1-P_2) & P_1 P_2 \end{bmatrix} \begin{bmatrix} \pi_{11} \\ \pi_{10} \\ \pi_{01} \\ \pi_{00} \end{bmatrix}$$

In matrix notation this becomes

$$\Lambda = P \Pi$$

and an unbiased estimator is

$$\hat{\Lambda} = P \hat{\Pi}.$$

The variance of $\hat{\Pi}$ is given by

$$\text{Var}(\hat{\Pi}) = P^{-1} \text{Var}(\hat{\Lambda}) (P^{-1})'$$

$$= \frac{1}{n} P^{-1} \begin{bmatrix} \lambda_{11}(1-\lambda_{11}) & -\lambda_{11}\lambda_{10} & -\lambda_{11}\lambda_{01} & -\lambda_{11}\lambda_{00} \\ -\lambda_{10}\lambda_{11} & \lambda_{10}(1-\lambda_{10}) & -\lambda_{10}\lambda_{01} & -\lambda_{10}\lambda_{00} \\ -\lambda_{01}\lambda_{11} & -\lambda_{01}\lambda_{10} & \lambda_{01}(1-\lambda_{01}) & -\lambda_{01}\lambda_{00} \\ -\lambda_{00}\lambda_{11} & -\lambda_{00}\lambda_{10} & -\lambda_{00}\lambda_{01} & \lambda_{00}(1-\lambda_{00}) \end{bmatrix} (P^{-1})'$$

$$= n^{-1} [\text{diag } \Lambda - \Lambda \Lambda'],$$

$$\text{where } \text{diag}(\Lambda) = \begin{bmatrix} \lambda_{11} & 0 & 0 & 0 \\ 0 & \lambda_{10} & 0 & 0 \\ 0 & 0 & \lambda_{01} & 0 \\ 0 & 0 & 0 & \lambda_{00} \end{bmatrix}.$$

To compare this model with the direct survey method, if we apply the usual

direct survey method, the outcome of the n independent repetitions of that trial follows the multinomial distribution with probability density function defined by

$$f(x_{11}, x_{10}, x_{01}, x_{00}) = n! \prod_{i=0}^1 \prod_{j=0}^1 \frac{\pi_{ij}^{x_{ij}}}{x_{ij}!}$$

For the vector of observed counts $\mathbf{x} = (x_{11}, x_{10}, x_{01}, x_{00})'$, $0 \leq x_{ij} \leq n$
for $i, j = 0, 1$ and $\sum_{i=0}^1 \sum_{j=0}^1 x_{ij} = n$.

The direct survey estimates of π_{11}^d , π_{10}^d , π_{01}^d , and π_{00}^d are

$$\hat{\pi}_{11}^d = \frac{x_{11}}{n}, \hat{\pi}_{10}^d = \frac{x_{10}}{n}, \hat{\pi}_{01}^d = \frac{x_{01}}{n}, \hat{\pi}_{00}^d = \frac{n - \hat{\pi}_{11}^d - \hat{\pi}_{10}^d - \hat{\pi}_{01}^d}{n}$$

These estimates are unbiased and the Covariance matrix is

$$\Sigma_d = \frac{1}{n} \begin{bmatrix} \pi_{11}^d(1-\pi_{11}^d) & -\pi_{11}^d\pi_{10}^d & -\pi_{11}^d\pi_{01}^d & -\pi_{11}^d\pi_{00}^d \\ -\pi_{10}^d\pi_{11}^d & \pi_{10}^d(1-\pi_{10}^d) & -\pi_{10}^d\pi_{01}^d & -\pi_{10}^d\pi_{00}^d \\ -\pi_{01}^d\pi_{11}^d & -\pi_{01}^d\pi_{10}^d & \pi_{01}^d(1-\pi_{01}^d) & -\pi_{01}^d\pi_{00}^d \\ -\pi_{00}^d\pi_{11}^d & -\pi_{00}^d\pi_{10}^d & -\pi_{00}^d\pi_{01}^d & \pi_{00}^d(1-\pi_{00}^d) \end{bmatrix}$$

Comparing the estimates under both models, randomized and direct, we observed that :

1. The estimates of π_{ij} and π_{ij}^d are unbiased under both models.
2. The direct survey estimates are expected to be of higher precision (i.e. lower variance) than the randomized response estimates. This is because the use of a

random device in interviewing introduces an additional source of variability to sample variation. Also the variance of the direct multinomial estimators, $\text{Var}(\hat{\pi}_{ij}^d)$, ($i, j = 0, 1$) as a function of n decreases faster than the variance of the randomized multinomial estimators, $\text{Var}(\hat{\pi}_{ij})$ which is function of n and P_j .

$\text{Var}(\hat{\pi}_{ij})$ can be minimized by choosing $P_j = 1$, but in that case the model is no longer a randomized response model. Therefore P_j are to be determined to increase the cooperation of the respondents and at the same time minimize the variances of the randomized response estimators. Since each cell of the 2×2 contingency table is known, by applying $\gamma^2 = \frac{1}{n} \chi^2$, we can estimate the correlation between two sensitive variables. And to test $H_0 : \rho_{S_1 S_2} = 0$, and $H_1 : \rho_{S_1 S_2} \neq 0$, we just need the value of $n \gamma_{S_1 S_2}^2$ with critical value of χ^2 .

2.3.6 Correlation Analysis for the Warner's Model versus Direct Survey

If a researcher wants to estimate the correlation between two variables where one variable is sensitive and another is nonsensitive, one possibility is to collect data on the sensitive variable using Warner's model, and on the nonsensitive variable data by a direct survey.

As we showed in Warner's model, the response variable can be expressed as

$$r_1 = (2P - 1) S + (1 - P)$$

and the response variable for the nonsensitive variable (Y) using a direct survey can be expressed as

$$r_2 = Y.$$

Let

$$S = \begin{cases} 1 & \text{if the individual says 'yes' for the Warner model} \\ 0 & \text{otherwise.} \end{cases}$$

$$Y = \begin{cases} 1 & \text{if the individual says 'yes' for the direct survey} \\ 0 & \text{otherwise.} \end{cases}$$

The outcome of each trial can be displayed in a 2 x 2 table as follows:

		S		
		yes	no	
Y	yes	S = 1 Y = 1	S = 0 Y = 1	Y = 1
	no	S = 1 Y = 0	S = 0 Y = 0	Y = 0
		S = 1	S = 0	

The correlation between the two variables (S, Y) is

$$\rho_{r_1 r_2} = \frac{(2P-1) \text{Cov}(S, Y)}{\sqrt{(2P-1)^2 V(S) V(Y)}} = \frac{\text{Cov}(S, Y)}{\sqrt{V(S) V(Y)}} = \rho_{SY}.$$

We observe that this result is the same as the Warner model versus Warner model given in section 2.3.1.

2.4 Bivariate Binomial Data Analysis Collected by the Unrelated Randomized Response Models

For the unrelated randomized response technique, like Warner's model, we cannot estimate cell proportions. To analyse the correlation between two sensitive variables, Kraemer (1980), Fox and Tracy (1984), and Edgel, Himmelfarb, and Cira (1986) assumed that two sensitive variables (S_1 and S_2) are independent of the two unrelated variables (Y_1 and Y_2) and that the two unrelated variables are also independent. These assumptions are not practical

(these assumptions are too strong), because S_i and Y_j ($i=1,2$ $j=1,2$) can be independent, but Y_1 and Y_2 may not be independent.

Gould, Shah, and Abernathy (1969) tried to use the unrelated randomized response techniques with two trials per respondent to get the covariance, but their model contains forty two parameters and not all of these parameters are simultaneously estimable, therefore they failed to estimate the covariance between the two sensitive variables.

2.4.1 Correlation Analysis

Here we propose a method to estimate the correlation between two sensitive variables. To estimate the correlation between the sensitive variables, the interviewer has to prepare two sets of questions such as.

Q_1 : Do you smoke pot at least once a week ?

Q_2 : Is the last digit of your student ID number odd ?

Q_1 : Have you ever had abortion ?

Q_2 : Were you born in Oklahoma ?

(These types of questions were used by several authors)

Suppose we have a sample of size $2n$ drawn from a population. The first n respondents are asked to answer "yes" or "no" to one of two questions from the first set. The probability of selecting the sensitive question is predetermined as P_1 , and the question to be answered will be selected by a random device. After completing the first question, the respondents are asked to answer one of two questions from the second set of questions. The probability of selecting the sensitive question is predetermined as P_2 , and by a random device a question will be selected.

For the next n respondents, the interviewer will change the probability of selecting the sensitive questions for both question sets. Thus, the probability of the first sensitive variable changes from P_1 to P_3 and the probability of the second sensitive variable changes from P_2 to P_4 . With these probabilities, the next n respondents will answer the questions like the first n respondents.

Then the response equations are :

$$r_1 = P_1 S_1 + (1 - P_1) Y_1$$

$$r_2 = P_2 S_2 + (1 - P_2) Y_2$$

$$r_3 = P_3 S_1 + (1 - P_3) Y_1$$

$$r_4 = P_4 S_2 + (1 - P_4) Y_2$$

Since the sensitive variables are independent of the two unrelated variables, the correlation equations can be written as

$$\rho_{r_1 r_2} = P_1 P_2 \rho_{S_1 S_2} + (1 - P_1)(1 - P_2) \rho_{Y_1 Y_2} \quad (2.4.1)$$

$$\rho_{r_3 r_4} = P_3 P_4 \rho_{S_1 S_2} + (1 - P_3)(1 - P_4) \rho_{Y_1 Y_2}$$

where P_i ($i=1,2,3,4$) are predetermined, and $\rho_{r_1 r_2}$ and $\rho_{r_3 r_4}$ can be estimated from the observed data.

Solving Eq(2.4.1) for $\rho_{S_1 S_2}$, we have

$$\rho_{S_1 S_2} = \frac{\rho_{r_1 r_2} - [(1 - P_1)(1 - P_2) \rho_{r_3 r_4}] [(1 - P_3)(1 - P_4)]^{-1}}{P_1 P_2 - [(1 - P_1)(1 - P_2) P_3 P_4] [(1 - P_3)(1 - P_4)]^{-1}}$$

Therefore we may estimate $\rho_{S_1 S_2}$ by

$$\hat{\rho}_{S_1 S_2} = \frac{\hat{\rho}_{r_1 r_2} - [(1 - P_1)(1 - P_2) \hat{\rho}_{r_3 r_4}] [(1 - P_3)(1 - P_4)]^{-1}}{P_1 P_2 - [(1 - P_1)(1 - P_2) P_3 P_4] [(1 - P_3)(1 - P_4)]^{-1}}$$

where $\hat{\rho}_{r_1 r_2}$ and $\hat{\rho}_{r_3 r_4}$ are provided by the observed data, and the selection probabilities P_1 , P_2 , P_3 , and P_4 are known, and hence we can estimate the correlation between the two sensitive variables.

To illustrate this procedure, we have simulated randomized response data for estimating the correlation between the two sensitive variables. The true correlation between the two sensitive variables S_1 and S_2 was set at 0.6. The true correlation between the two unrelated variables Y_1 and Y_2 was set at 0.2, 0.3, 0.4, 0.5 and 0.6 for fixed correlation between the sensitive variables. Means of each of the variables were set at $\mu_{S_1} = 0.2$, $\mu_{Y_1} = 0.2$, $\mu_{S_2} = 0.3$, and $\mu_{Y_2} = 0.3$. In the simulation, the probabilities of selecting the sensitive question were set to be various values. The results of the simulations are presented in table 2 and table 3 for $n = 100$ and $n = 200$ respectively. Each table gives the estimated correlations and standard deviations of the sampling distribution of the correlation coefficient obtained by using the unrelated question model under the assumptions stated previously. The standard deviations decrease as $|P_1 - P_3|$ increases. The estimated correlation $\hat{\rho}_{S_1 S_2}$, does not depend on the correlation between unrelated variables for each set of (P_1, P_2, P_3, P_4) .

TABLE 2
ESTIMATED CORRELATION FOR THE UNRELATED
RANDOMIZED RESPONSE MODEL

P_1	P_3	$\rho_{Y_1 Y_2}$				
		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.

$\rho_{S_1 S_2} = 0.6, n = 100.$

TABLE 3
ESTIMATED CORRELATION FOR THE UNRELATED
RANDOMIZED RESPONSE MODEL

P_1	P_3	$\rho_{Y_1 Y_2}$				
		0.22	0.3	0.4	0.5	0.6
0.4	0.4	0.61612 (0.83676)	0.60314 (0.85359)	0.60058 (0.85901)	0.60042 (0.85139)	0.61474 (0.84824)
0.6	0.6	0.659779 (0.22317)	0.59062 (0.22043)	0.59204 (0.22062)	0.59265 (0.22263)	0.59319 (0.22387)
0.7	0.7	0.60015 (0.01545)	0.59615 (0.15507)	0.59669 (0.15542)	0.59618 (0.15339)	0.59661 (0.15338)
0.8	0.8	0.59935 (0.12118)	0.59743 (0.11063)	0.59692 (0.11145)	0.59634 (0.11006)	0.59681 (0.11015)
0.6	0.6	0.59577 (0.25985)	0.58614 (0.25549)	0.58768 (0.25577)	0.58760 (0.25998)	0.58791 (0.26149)
0.7	0.7	0.59954 (0.16211)	0.59482 (0.16305)	0.59536 (0.16334)	0.59454 (0.16172)	0.59488 (0.16208)
0.8	0.8	0.59914 (0.11331)	0.59702 (0.11197)	0.59649 (0.11275)	0.59582 (0.11146)	0.59625 (0.11166)
0.7	0.7	0.60591 (0.26218)	0.60090 (0.26366)	0.60251 (0.26445)	0.59894 (0.26050)	0.59900 (0.26259)
0.8	0.8	0.60057 (0.12483)	0.59872 (0.12335)	0.59830 (0.12402)	0.59700 (0.12220)	0.59739 (0.12274)
0.8	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.
 $\rho_{S_1 S_2} = 0.6, n = 200.$

2.4.2 Test of Hypothesis

Here we propose a method to perform a test of hypothesis. For 2 x 2 tables (not I x J tables in general) $\gamma^2 = \chi^2/2n$. Therefore we may conduct a test of independence directly from the estimate of ρ with critical value of χ^2 .

2.4.3 Sample Size Estimation

Here we propose sample size estimation.

i. Some margin of error d in the estimated proportion π of units in the population has been agreed on, and there is a small risk α that we are willing to incur that the actual error is larger than d . So how many elements are needed to satisfied these conditions ?

We want

$$\Pr [| \hat{\pi}_S - \pi_S | \geq d] = \alpha.$$

We use simple random sampling, and since $\hat{\pi}_S$ is a maximum likelihood estimate, $\hat{\pi}_S$ is approximately normally distributed (for the case where π_Y is known),

$$\hat{\pi}_S \sim \text{AN} \left[\pi_S, \frac{\lambda(1-\lambda)}{n P^2} \right],$$

where $\lambda = \pi_S P + \pi_Y(1 - P)$.

Hence the formula that connects n with the desired degree of precision is

$$d = z_{\alpha/2} \sqrt{\frac{\lambda(1-\lambda)}{n P^2}},$$

where z is the abscissa of the normal curve that cuts off an area of $\frac{\alpha}{2}$ in the upper tail.

Solving for n, we find

$$n = \frac{z_{\alpha/2}^2}{d^2} \left[\frac{\lambda(1-\lambda)}{P^2} \right].$$

For practical use, an estimate $\hat{\lambda}$ of λ is substituted in the above formula. But λ depends on the parameter π_S of the population that is to be sampled. Hence we may use a conservative assumption to ensure that the specified tolerances are met or exceeded.

ii. How many more elements are needed to get the same variance as the direct survey. We know the variance of $\hat{\pi}_S$ for the direct survey is given by

$$V(\hat{\pi}_S)_d = \frac{\pi_S(1-\pi_S)}{n_d},$$

and

$$V(\hat{\pi}_S)_R = \frac{\lambda(1-\lambda)}{n P^2}.$$

In order to have the same variance as for the direct survey, the sample size (n_d) must be increased. The number of extra sample units can be found by equating both variances and solving for n_R , then

$$n_R = n_d \left[\frac{1}{P} + \frac{\pi_S(1-P)}{P(1-\pi_S)} + \frac{\pi_Y(1-P)[1-\pi_Y(1-P)] - 2\pi_S\pi_Y P(1-P)}{P^2\pi_S(1-\pi_S)} \right].$$

The formula for n depends on the population parameters π_S and π_Y which are to be estimated.

Hence we may use a conservative assumption in that it is made to ensure that the specified tolerances are met or exceeded.

2.4.5 Unrelated Randomized Response Technique versus Direct

Survey technique

The purpose of the present model is to estimate the correlation between two variables; one variable is sensitive (S) and the other variable is nonsensitive (Y).

To estimate the proportion of the sensitive variable we may use the unrelated randomized response model with an alternative nonsensitive variable (Y_1) which is unrelated to the sensitive variable (S) but can be related to the nonsensitive variable (Y). Since Y is a nonsensitive variable, we may use the usual direct survey methodology.

Suppose we have a sample of size $2n$ drawn from a population. In this particular model, two randomization devices need to be used. The first one is used for the first n respondents, and the second one is used for the next n respondents.

Let the randomization devices be the two wheels. One side of wheel designates sensitive question S and the other side designates nonsensitive unrelated question, Y_1 . The selection probability of the sensitive question S is predetermined as P_1 for the first wheel and as P_2 for the second wheel. P_1 should not be the same as P_2 .

An example of possible questions to be used are:

S : Have you ever smoked marijuana ?

Y_1 : Have you watched any sports game on TV in the past week ?

and let the question for the usual direct survey methodology be:

Y : Have you had a beer in the past week ?

The first n respondents are asked to answer 'yes' or 'no' to one of two questions by using the first wheel and also answer the direct question, Y , on a nonsensitive topic. The next n respondents will answer the question like the first n respondents but using the second wheel. Then the interviewer will observe ;

For the first n respondents

$$\lambda_1 = P_1 \pi_S + (1 - P_1)\pi_{Y_1} \quad (2.4.1)$$

$$\lambda_2 = \pi_Y. \quad (2.4.2)$$

For the second n respondents

$$\lambda_3 = P_2 \pi_S + (1 - P_2)\pi_{Y_1} \quad (2.4.3)$$

$$\lambda_4 = \pi_Y. \quad (2.4.4)$$

From Eq(2.4.1) and Eq(2.4.3)

$$\pi_S = \frac{(1 - P_2)\lambda_1 - (1 - P_1)\lambda_3}{(1 - P_2)P_1 - (1 - P_1)P_2}, \quad \text{provided } P_1 \neq P_2.$$

Therefore an estimate of π_S is

$$\begin{aligned} \hat{\pi}_S &= \frac{(1 - P_2)\hat{\lambda}_1 - (1 - P_1)\hat{\lambda}_3}{(1 - P_2)P_1 - (1 - P_1)P_2} \\ &= \frac{(1 - P_2)\frac{n_1}{n} - (1 - P_1)\frac{n_2}{n}}{(1 - P_2)P_1 - (1 - P_1)P_2}, \end{aligned} \quad (2.4.5)$$

where n_1 is the number of 'yes' responses from the first n respondents, and n_2 is the number of 'yes' responses from the second n respondents.

Since n_1 and n_2 follow binomial distributions with parameters n and λ_i ,
 $i = 1, 2$, $\hat{\pi}_S$ is an unbiased estimate of π_S .

From Eq(2.4.5), the variance of $\hat{\pi}_S$ is

$$\text{Var}(\hat{\pi}_S) = \frac{[(1 - P_2)^2 \text{Var}(\hat{\lambda}_1) + (1 - P_1)^2 \text{Var}(\hat{\lambda}_3)]}{(P_1 - P_2)^2}$$

where $\text{Var}(\hat{\lambda}_1) = n^{-2} \text{Var}(n_1) = n^{-1} \lambda_1(1 - \lambda_1)$

$$\text{Var}(\hat{\lambda}_2) = n^{-2} \text{Var}(n_2) = n^{-1} \lambda_3(1 - \lambda_3).$$

Hence $\text{Var}(\hat{\pi}_S)$ is given by

$$\text{Var}(\hat{\pi}_S) = \frac{[(1 - P_2)^2 \lambda_1(1 - \lambda_1) + (1 - P_1)^2 \lambda_3(1 - \lambda_3)]}{n (P_1 - P_2)^2}.$$

Since n_i follow binomial distributions with parameters n and λ_i for $i=1,2$
 an unbiased estimate of $\text{Var}(\hat{\pi}_S)$ is given by

$$\hat{\text{Var}}(\hat{\pi}_S) = \frac{[(1 - P_2)^2 \hat{\lambda}_1(1 - \hat{\lambda}_1) + (1 - P_1)^2 \hat{\lambda}_3(1 - \hat{\lambda}_3)]}{(n-1)(P_1 - P_2)^2}.$$

and from Eq(2.4.1) and Eq(2.4.3)

$$\pi_{Y_1} = \frac{[P_2 \lambda_1 - P_1 \lambda_3]}{P_2 - P_1}.$$

Since $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are unbiased, an unbiased estimate of π_{Y_1} is given by

$$\begin{aligned}\hat{\pi}_{Y_1} &= \frac{[P_2 \hat{\lambda}_1 - P_1 \hat{\lambda}_2]}{P_2 - P_1} \\ &= \frac{[P_2 \frac{n_1}{n} - P_1 \frac{n_2}{n}]}{P_2 - P_1},\end{aligned}$$

and the variance of π_{Y_1} is give by

$$\text{Var}(\hat{\pi}_{Y_1}) = \frac{[P_2^2 \lambda_1(1 - \lambda_1) + P_1^2 \lambda_2(1 - \lambda_2)]}{n(P_2 - P_1)^2}.$$

Hence an unbiased estimate of $\text{Var}(\hat{\pi}_{Y_1})$ is given by

$$\hat{\text{Var}}(\hat{\pi}_{Y_1}) = \frac{[P_2^2 \hat{\lambda}_1(1 - \hat{\lambda}_1) + P_1^2 \hat{\lambda}_2(1 - \hat{\lambda}_2)]}{(n-1)(P_2 - P_1)^2}.$$

However, we also have an estimate of π_Y by direct observation, and hence we have two estimates of π_Y , each of them is unbiased. To get the best linear unbiased estimate, let

$\hat{\pi}_Y^1$ be the first estimate of π_Y

$\hat{\pi}_Y^2$ be the second estimate of π_Y

σ_1^2 be the variance of $\hat{\pi}_Y^1$

σ_2^2 be the variance of $\hat{\pi}_Y^2$

σ_{12} be the covariance between $\hat{\pi}_Y^1$ and $\hat{\pi}_Y^2$.

$\hat{\pi}_Y = \xi \hat{\pi}_Y^1 + (1-\xi) \hat{\pi}_Y^2$ is a linear unbiased for π_Y for any value of ξ ,

where $0 \leq \xi \leq 1$.

The variance of $\hat{\pi}_Y$ is given by

$$\text{Var}(\hat{\pi}_Y) = \xi^2 \text{Var}(\hat{\pi}_Y^1) + (1-\xi)^2 \text{Var}(\hat{\pi}_Y^2) + 2 \xi (1-\xi) \text{Cov}(\hat{\pi}_Y^1, \hat{\pi}_Y^2). \quad (2.4.6)$$

To minimize this variance, we take the first derivative with respect to ξ , and set the resulting equation equal to 0.

That is

$$\frac{\partial \text{Var}(\hat{\pi}_Y)}{\partial \xi} = \xi(2\sigma_1^2 + 2\sigma_2^2 - 2\sigma_{12}) - 2\sigma_2^2 + 2\sigma_{12} = 0.$$

Solving for ξ , we get ξ_0 which is given by

$$\xi_0 = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}. \quad (2.4.7)$$

Therefore $\hat{\pi}_Y$ is the best linear unbiased estimate whenever we use $\xi = \xi_0$.

Substituting Eq.(2.4.7) into Eq.(2.4.6), $\text{Var}(\hat{\pi}_Y)$ is given by

$$\text{Var}(\hat{\pi}_Y) = \frac{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}.$$

To estimate the correlation between two variables (S and Y),

we may rewrite Eq.(2.4.1) and Eq.(2.4.4)

$$r_1 = P_1 S + (1 - P_1) Y_1$$

$$r_2 = Y$$

$$r_3 = P_2 S + (1 - P_2) Y_1$$

$$r_4 = Y.$$

Since S and Y_1 are independent, the correlation equations can be written as

$$\rho_{r_1 r_2} = P_1 \rho_{SY} + (1 - P_1) \rho_{Y_1 Y}$$

$$\rho_{r_3 r_4} = P_2 \rho_{SY} + (1 - P_2) \rho_{Y_1 Y}.$$

From these two equations we obtain

$$\rho_{SY} = \frac{\left[(1 - P_2) \rho_{r_1 r_2} - (1 - P_1) \rho_{r_3 r_4} \right]}{P_1 - P_2}.$$

Hence from the observed data we may estimate $\rho_{r_1 r_2}$ and $\rho_{r_3 r_4}$, and the selection probabilities P_1 and P_2 are known hence we may estimate the correlation and so obtain

$$\hat{\rho}_{SY} = \frac{\left[(1 - P_2) \hat{\rho}_{r_1 r_2} - (1 - P_1) \hat{\rho}_{r_3 r_4} \right]}{P_1 - P_2}.$$

Since $\hat{\rho}_{r_1 r_2}$ and $\hat{\rho}_{r_3 r_4}$ are m.l.e., $\hat{\rho}_{SY}$ is also m.l.e..

To illustrate this procedure, we have simulated randomized response data for estimating the correlation between the two variables. The true correlation between the sensitive variable S and the non-sensitive variable Y was set at 0.6. The true correlation between the unrelated variable Y_1 and the non-sensitive variable which is conducted by direct survey was set at 0.35. In simulation, the probabilities of selecting the sensitive question were set to be various values. Means of each of the variables were set at $\mu_S = 0.2$, $\mu_Y = 0.3$, and $\mu_{Y_1} = 0.3$.

The results of the simulations are presented in table 4. Table 4 gives the estimated correlations and standard deviations (third column for $n = 100$, fourth column for $n = 200$) and the effective sample sizes. To minimize the variance of ρ_{SY} , for fixed P_1 (or P_2), we should choose $P_2 = 1$ (or $P_1 = 1$), but in that case, this is no longer randomized response model. Therefore, P_2 (or P_1) have to be chosen as far from P_1 (or P_2) as possible. As we can see in table 4, as $|P_1 - P_2|$ increases the standard deviation decreases.

TABLE 4

ESTIMATED CORRELATION FOR THE UNRELATED RANDOMIZED
MODEL VERSUS DIRECT SURVEY

P ₁	P ₂	$\rho_{S_1 S_2}$	$\rho_{S_1 S_2}$	Effective Sample Size
		n=100	n=200	
0.2	0.3	0.61285 (0.54785)	0.61260 (0.37662)	50
0.2	0.4	0.61628 (0.26637)	0.61849 (0.18950)	60
0.2	0.6	0.60490 (0.13498)	0.60544 (0.10093)	80
0.2	0.7	0.60386 (0.11986)	0.60322 (0.08428)	90
0.2	0.8	0.60424 (0.10299)	0.60310 (0.07360)	100
0.3	0.4	0.60702 (0.49167)	0.60492 (0.35092)	70
0.3	0.6	0.60850 (0.16269)	0.60964 (0.11073)	90
0.3	0.7	0.59858 (0.12600)	0.60707 (0.08882)	100
0.3	0.8	0.60578 (0.10472)	0.60141 (0.07591)	110
0.4	0.6	0.60269 (0.20378)	0.60155 (0.14389)	100
0.4	0.7	0.60091 (0.13875)	0.60385 (0.09875)	110
0.4	0.8	0.60143 (0.11183)	0.60082 (0.07695)	120
0.6	0.7	0.60020 (0.24899)	0.60046 (0.17894)	130
0.6	0.8	0.59811 (0.13076)	0.59811 (0.09026)	140

Inside values of () are standard deviations.
Simulation includes 1000 trials.

CHAPTER III

RANDOMIZED RESPONSE TECHNIQUE FOR MULTIPLE ATTRIBUTES

3.1 Additive Randomized Response Technique.

An additive randomized response technique was proposed by Kim and Flueck (1978). The additive randomized response technique will be explained briefly.

Let C_j be the true category for the j -th respondent, where the C_j have T mutually exclusive and exhaustive categories with population proportions $\pi_1, \pi_2, \pi_3, \dots, \pi_T$, respectively, and $\sum_{t=1}^T \pi_t = 1$. Let $Y_j (1, 2, \dots, T)$ be a randomly selected augmentation value for the j -th respondent, with selection probability $P(Y_j = t) = P_t, t = 1, 2, \dots, T$, and $\sum_{t=1}^T P_t = 1$. The selection probability (P_t) of the augmentation value is preassigned and the distribution of augmentation variable (Y) is known. Each respondent is asked to select his own category but to maintain confidentiality, they are instructed to add the augmented value selected to their own category number. Then the j -th respondent's added response whose true group is C_j , is

$$C_j + Y_j, j = 1, 2, \dots, n; C_j = 1, 2, \dots, T; \text{ and } Y_j = 1, 2, \dots, T.$$

To provide further confidentiality to the respondent, the interviewer asks the respondent to transform the added value and report the value r_j .

$$r_j = \begin{cases} C_j + Y_j & \text{if } C_j + Y_j \leq T \\ C_j + Y_j - T & \text{if } C_j + Y_j > T. \end{cases}$$

For the case $T = 3$, the questions (Kim and Flueck, 1978) are :

Q_1 : I have never cheated

Q_2 : I was prepared to cheat before the test but did not actually cheat

Q_3 : I cheated.

Then, the probability λ_r that a respondent reports value r (1, 2 or 3) is

$$\lambda_1 = P(r=1) = P(C_j=1, Y_j=3) + P(C_j=2, Y_j=2) + P(C_j=3, Y_j=1)$$

$$= \pi_1 P_3 + \pi_2 P_2 + \pi_3 P_1,$$

$$\lambda_2 = P(r=2) = P(C_j=1, Y_j=1) + P(C_j=2, Y_j=3) + P(C_j=3, Y_j=2)$$

$$= \pi_1 P_1 + \pi_2 P_3 + \pi_3 P_2,$$

$$\lambda_3 = P(r=3) = P(C_j=1, Y_j=2) + P(C_j=2, Y_j=1) + P(C_j=3, Y_j=3)$$

$$= \pi_1 P_2 + \pi_2 P_1 + \pi_3 P_3. \quad (3.1.1)$$

Since $\lambda_3 = 1 - \lambda_1 - \lambda_2$ and $\pi_3 = 1 - \pi_1 - \pi_2$, these equations reduce to

$$\lambda_1 = P_1 + (P_3 - P_1)\pi_1 + (P_2 - P_1)\pi_2,$$

$$\lambda_2 = P_2 + (P_1 - P_2)\pi_1 + (P_3 - P_2)\pi_2.$$

We may rewrite these equations in matrix form

$$\begin{bmatrix} \lambda_1 - P_1 \\ \lambda_2 - P_2 \end{bmatrix} = \begin{bmatrix} P_3 - P_1 & P_2 - P_1 \\ P_1 - P_2 & P_3 - P_2 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}.$$

From the observed data, the sample proportion for $r = 1$ is $\frac{n_1}{n}$, the sample proportion for $r = 2$ is $\frac{n_2}{n}$, and the sample proportion for $r=3$ is $(1 - \frac{n_1}{n} - \frac{n_2}{n})$.

$$\text{Since } \hat{\lambda}_1 = \frac{n_1}{n}, \hat{\lambda}_2 = \frac{n_2}{n}, \hat{\lambda}_3 = 1 - \hat{\lambda}_1 - \hat{\lambda}_2,$$

we may estimate $\hat{\Pi}$, by $\hat{\Pi} = P^{-1} \hat{\Lambda}^*$, provided $P_1 \neq P_2 \neq P_3$

$$\text{where } \hat{\Lambda}^* = \begin{bmatrix} \lambda_1 - P_1 \\ \lambda_2 - P_2 \end{bmatrix}, P = \begin{bmatrix} P_3 - P_1 & P_2 - P_1 \\ P_1 - P_2 & P_3 - P_2 \end{bmatrix}, \text{ and } \hat{\Pi} = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}.$$

The unbiased estimate of $\hat{\Pi}$ is

$$\hat{\pi}_1 = \frac{1}{|P|} [(P_3 - P_2)(\hat{\lambda}_1 - P_1) + (P_1 - P_2)(\hat{\lambda}_2 - P_2)],$$

$$\hat{\pi}_2 = \frac{1}{|P|} [(P_2 - P_1)(\hat{\lambda}_1 - P_1) + (P_3 - P_1)(\hat{\lambda}_2 - P_2)],$$

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2. \quad (3.1.2)$$

The variances of $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3$ and $\text{Cov}(\hat{\pi}_1, \hat{\pi}_2)$ are

$$\text{Var}(\hat{\pi}_1) = \frac{(P_3 - P_2)^2 \lambda_1(1-\lambda_1) + (P_1 - P_2)^2 \lambda_2(1-\lambda_2) - 2(P_3 - P_2)(P_1 - P_2)\lambda_1\lambda_2}{n|P|^2},$$

$$\text{Var}(\hat{\pi}_2) = \frac{(P_2 - P_1)^2 \lambda_1(1-\lambda_1) + (P_3 - P_1)^2 \lambda_2(1-\lambda_2) - 2(P_2 - P_1)(P_3 - P_1)\lambda_1\lambda_2}{n|P|^2},$$

$$\text{Var}(\hat{\pi}_3) = \text{Var}(\hat{\pi}_1) + \text{Var}(\hat{\pi}_2) + 2 \text{Cov}(\hat{\pi}_1, \hat{\pi}_2),$$

$$\begin{aligned} \text{Cov}(\hat{\pi}_1, \hat{\pi}_2) &= \frac{(P_3 - P_2)(P_2 - P_1)\lambda_1(1 - \lambda_1) + (P_2 - P_1)^2\lambda_1\lambda_2}{n|P|^2} \\ &\quad + \frac{(P_3 - P_2)(P_3 - P_1)\lambda_1\lambda_2 + (P_1 - P_2)(P_3 - P_1)\lambda_2(1 - \lambda_2)}{n|P|^2}. \end{aligned}$$

Since n_i follows a binomial distribution with parameters n and λ_i for $i = 1, 2$,

$$\begin{aligned} E \left[\frac{\hat{\lambda}_1(1 - \hat{\lambda}_1)}{n-1} \right] &= E \left[\frac{\frac{n_1}{n} \left(1 - \frac{n_1}{n}\right)}{n-1} \right] \\ &= \frac{1}{n-1} E \left(\frac{n_1}{n} - \frac{n_1^2}{n^2} \right) = \frac{\lambda_1(1 - \lambda_1)}{n}. \end{aligned}$$

$$\begin{aligned} \text{and } E \left[\frac{\hat{\lambda}_1 \hat{\lambda}_2}{n-1} \right] &= E \left[\frac{\frac{n_1}{n} \frac{n_2}{n}}{n-1} \right] \\ &= \frac{1}{n^2(n-1)} E n_1 n_2 \\ &= \frac{\lambda_1 \lambda_2}{n}, \end{aligned}$$

since $E n_1 n_2 = \text{Cov}(n_1, n_2) + E n_1 E n_2$.

The unbiased estimate of $\text{Var}(\hat{\pi}_1)$, $\text{Var}(\hat{\pi}_2)$, and $\text{Var}(\hat{\pi}_3)$ is given by

$$\hat{\text{Var}}(\hat{\pi}_1) = \frac{(P_3 - P_2)^2 \hat{\lambda}_1(1 - \hat{\lambda}_1) + (P_1 - P_2)^2 \hat{\lambda}_2(1 - \hat{\lambda}_2) - 2(P_3 - P_2)(P_1 - P_2)\hat{\lambda}_1\hat{\lambda}_2}{(n-1)|P|^2}$$

$$\widehat{\text{Var}}(\hat{\pi}_2) = \frac{(P_2 - P_1)^2 \hat{\lambda}_1(1 - \hat{\lambda}_1) + (P_3 - P_1)^2 \hat{\lambda}_2(1 - \hat{\lambda}_2) - 2(P_2 - P_1)(P_3 - P_1)\hat{\lambda}_1\hat{\lambda}_2}{(n-1)|P|^2}$$

$$\widehat{\text{Var}}(\hat{\pi}_3) = \widehat{\text{Var}}(\hat{\pi}_1) + \widehat{\text{Var}}(\hat{\pi}_2) + 2 \widehat{\text{Cov}}(\hat{\pi}_1, \hat{\pi}_2),$$

where

$$\widehat{\text{Cov}}(\hat{\pi}_1, \hat{\pi}_2) = \frac{(P_3 - P_2)(P_2 - P_1)\hat{\lambda}_1(1 - \hat{\lambda}_1) + (P_2 - P_1)^2\hat{\lambda}_1\hat{\lambda}_2}{(n-1)|P|^2} + \frac{(P_3 - P_2)(P_3 - P_1)\hat{\lambda}_1\hat{\lambda}_2 + (P_1 - P_2)(P_3 - P_1)\hat{\lambda}_2(1 - \hat{\lambda}_2)}{(n-1)|P|^2}$$

3.2 Correlation Analysis for Another Version of the Additive Model

Suppose we have two sensitive characters and each character has more than two subcategories (S_1 has r subcategories, S_2 has c subcategories). Thus the population can be tabulated as an $r \times c$ contingency table, and we need to estimate the corresponding cell proportions $\pi_{11}, \pi_{12}, \dots, \pi_{rc}$, where $0 < \pi_{ij} < 1$,

($i = 1, 2, \dots, r, j = 1, 2, \dots, c$) and $\sum_{i,j} \pi_{ij} = 1$.

Let S_1 have 3 subcategories and S_2 have 3 subcategories, then to estimate each cell proportion, we may apply the additive randomized response technique by reordering each cell number as follow:

Let $(1\ 1) = 1, (1\ 2) = 2, (1\ 3) = 3, (2\ 1) = 4, (2\ 2) = 5, (2\ 3) = 6,$
 $(3\ 1) = 7, (3\ 2) = 8, (3\ 3) = 9.$

The first number of each pair is the row subcategory and the second number is the column subcategory, then our contingency table will be as follows:

TABLE 5
REORDERED CONTINGENCY TABLE

	S_2		
	1	2	3
1	π_{11} 1	π_{12} 2	π_{13} 3
2	π_{21} 4	π_{22} 5	π_{23} 6
3	π_{31} 7	π_{32} 8	π_{33} 9

As with the usual additive model, let Y_j be the j -th respondent's randomly selected augmentation value ($Y_j = 1, 2, \dots, 9$), and the selection probabilities P_t ($t = 1, 2, \dots, 9$) are known, then this system is the same as the Kim and Flueck's Additive randomized response model. Hence we may apply the additive model for estimating the cell proportions of the above contingency table.

Then the j -th respondent's added response whose true group is C_j is

$$C_j + Y_j \quad C_j = 1, 2, \dots, 9$$

$$Y_j = 1, 2, \dots, 9.$$

The possible added responses are 2, 3, ..., 16, 17, 18, the added responses 2 and 18 have only one possibility to be that added values, hence the respondents will hesitate to release their information.

To provide further confidentiality to the respondent, the j -th respondent's added value $C_j + Y_j$ is transformed by the respondent to the reported value,

$$r_j = \begin{cases} C_j + Y_j & \text{if } C_j + Y_j \leq 9 \\ C_j + Y_j - 9 & \text{if } C_j + Y_j > 9. \end{cases}$$

Then the possible reported values and their sources are :

Observed number	source ($C_j + Y_j$)
1	1+9 2+8 3+7 4+6 5+5 6+4 7+3 8+2 9+1
2	1+1 2+9 3+8 4+7 5+6 6+5 7+4 8+3 9+2
3	1+2 2+1 3+9 4+8 5+7 6+6 7+5 8+4 9+3
4	1+3 2+2 3+1 4+9 5+8 6+7 7+6 8+5 9+4
5	1+4 2+3 3+2 4+1 5+9 6+8 7+7 8+6 9+5
6	1+5 2+4 3+3 4+2 5+1 6+9 7+8 8+7 9+6
7	1+6 2+5 3+4 4+3 5+2 6+1 7+9 8+8 9+7
8	1+7 2+6 3+5 4+4 5+3 6+2 7+1 8+9 9+8
9	1+8 2+7 3+6 4+5 5+4 6+3 7+2 8+1 9+9

Let π_j be the proportion of j -th category for that population (see table 5).

Let P_j be the selection probability of the augmentation values.

From the above transformed response values the probability (λ_r) that a respondent reports value r ($r = 1, 2, 3, 4, \dots, 9$) is :

$$\lambda_1 = \pi_1 P_9 + \pi_2 P_8 + \pi_3 P_7 + \pi_4 P_6 + \pi_5 P_5 + \pi_6 P_4 + \pi_7 P_3 + \pi_8 P_2 + \pi_9 P_1,$$

$$\lambda_2 = \pi_1 P_1 + \pi_2 P_9 + \pi_3 P_8 + \pi_4 P_7 + \pi_5 P_6 + \pi_6 P_5 + \pi_7 P_4 + \pi_8 P_3 + \pi_9 P_2,$$

$$\lambda_3 = \pi_1 P_2 + \pi_2 P_1 + \pi_3 P_9 + \pi_4 P_8 + \pi_5 P_7 + \pi_6 P_6 + \pi_7 P_5 + \pi_8 P_4 + \pi_9 P_3,$$

$$\lambda_4 = \pi_1 P_3 + \pi_2 P_2 + \pi_3 P_1 + \pi_4 P_9 + \pi_5 P_8 + \pi_6 P_7 + \pi_7 P_6 + \pi_8 P_5 + \pi_9 P_4,$$

$$\lambda_5 = \pi_1 P_4 + \pi_2 P_3 + \pi_3 P_2 + \pi_4 P_1 + \pi_5 P_9 + \pi_6 P_8 + \pi_7 P_7 + \pi_8 P_6 + \pi_9 P_5,$$

$$\lambda_6 = \pi_1 P_5 + \pi_2 P_4 + \pi_3 P_3 + \pi_4 P_2 + \pi_5 P_1 + \pi_6 P_9 + \pi_7 P_8 + \pi_8 P_7 + \pi_9 P_6,$$

$$\lambda_7 = \pi_1 P_6 + \pi_2 P_5 + \pi_3 P_4 + \pi_4 P_3 + \pi_5 P_2 + \pi_6 P_1 + \pi_7 P_9 + \pi_8 P_8 + \pi_9 P_7,$$

$$\lambda_8 = \pi_1 P_7 + \pi_2 P_6 + \pi_3 P_5 + \pi_4 P_4 + \pi_5 P_3 + \pi_6 P_2 + \pi_7 P_1 + \pi_8 P_9 + \pi_9 P_8,$$

$$\lambda_9 = \pi_1 P_8 + \pi_2 P_7 + \pi_3 P_6 + \pi_4 P_5 + \pi_5 P_4 + \pi_6 P_3 + \pi_7 P_2 + \pi_8 P_1 + \pi_9 P_9.$$

Since $\lambda_9 = 1 - \sum_{l=1}^8 \lambda_l$, and $\pi_9 = 1 - \sum_{l=1}^8 \pi_l$, these equations can be reduced as follows

$$\lambda_1 = P_1 + (P_9 - P_1)\pi_1 + (P_8 - P_1)\pi_2 + (P_7 - P_1)\pi_3 + (P_6 - P_1)\pi_4$$

$$+ (P_5 - P_1)\pi_5 + (P_4 - P_1)\pi_6 + (P_3 - P_1)\pi_7 + (P_2 - P_1)\pi_8$$

$$\lambda_2 = P_2 + (P_1 - P_2)\pi_1 + (P_9 - P_2)\pi_2 + (P_8 - P_2)\pi_3 + (P_7 - P_2)\pi_4$$

$$+ (P_6 - P_2)\pi_5 + (P_5 - P_2)\pi_6 + (P_4 - P_2)\pi_7 + (P_3 - P_2)\pi_8$$

$$\lambda_3 = P_3 + (P_2 - P_3)\pi_1 + (P_1 - P_3)\pi_2 + (P_9 - P_3)\pi_3 + (P_8 - P_3)\pi_4$$

$$+ (P_7 - P_3)\pi_5 + (P_6 - P_3)\pi_6 + (P_5 - P_3)\pi_7 + (P_4 - P_3)\pi_8$$

$$\lambda_4 = P_4 + (P_3 - P_4)\pi_1 + (P_2 - P_4)\pi_2 + (P_1 - P_4)\pi_3 + (P_9 - P_4)\pi_4$$

$$+ (P_8 - P_4)\pi_5 + (P_7 - P_4)\pi_6 + (P_6 - P_4)\pi_7 + (P_5 - P_4)\pi_8$$

$$\begin{aligned}\lambda_5 = & P_5 + (P_4 - P_5)\pi_1 + (P_3 - P_5)\pi_2 + (P_2 - P_5)\pi_3 + (P_1 - P_5)\pi_4 \\ & + (P_9 - P_5)\pi_5 + (P_8 - P_5)\pi_6 + (P_7 - P_5)\pi_7 + (P_6 - P_5)\pi_8\end{aligned}$$

$$\begin{aligned}\lambda_6 = & P_6 + (P_5 - P_6)\pi_1 + (P_4 - P_6)\pi_2 + (P_3 - P_6)\pi_3 + (P_2 - P_6)\pi_4 \\ & + (P_1 - P_6)\pi_5 + (P_9 - P_6)\pi_6 + (P_8 - P_6)\pi_7 + (P_7 - P_6)\pi_8\end{aligned}$$

$$\begin{aligned}\lambda_7 = & P_7 + (P_6 - P_7)\pi_1 + (P_5 - P_7)\pi_2 + (P_4 - P_7)\pi_3 + (P_3 - P_7)\pi_4 \\ & + (P_2 - P_7)\pi_5 + (P_1 - P_7)\pi_6 + (P_9 - P_7)\pi_7 + (P_8 - P_7)\pi_8\end{aligned}$$

$$\begin{aligned}\lambda_8 = & P_8 + (P_7 - P_8)\pi_1 + (P_6 - P_8)\pi_2 + (P_5 - P_8)\pi_3 + (P_4 - P_8)\pi_4 \\ & + (P_3 - P_8)\pi_5 + (P_2 - P_8)\pi_6 + (P_1 - P_8)\pi_7 + (P_9 - P_8)\pi_8.\end{aligned}$$

In matrix notation,

$$\Lambda^* = P \Pi, \tag{3.2.1}$$

where P is

$$\begin{bmatrix} P_9 - P_1 & P_8 - P_1 & P_7 - P_1 & P_6 - P_1 & P_5 - P_1 & P_4 - P_1 & P_3 - P_1 & P_2 - P_1 \\ P_1 - P_2 & P_9 - P_2 & P_8 - P_2 & P_7 - P_2 & P_6 - P_2 & P_5 - P_2 & P_4 - P_2 & P_3 - P_2 \\ P_2 - P_3 & P_1 - P_3 & P_9 - P_3 & P_8 - P_3 & P_7 - P_3 & P_6 - P_3 & P_5 - P_3 & P_4 - P_3 \\ P_3 - P_4 & P_2 - P_4 & P_1 - P_4 & P_9 - P_4 & P_8 - P_4 & P_7 - P_4 & P_6 - P_4 & P_5 - P_4 \\ P_4 - P_5 & P_3 - P_5 & P_2 - P_5 & P_1 - P_5 & P_9 - P_5 & P_8 - P_5 & P_7 - P_5 & P_6 - P_5 \\ P_5 - P_6 & P_4 - P_6 & P_3 - P_6 & P_2 - P_6 & P_1 - P_6 & P_9 - P_6 & P_8 - P_6 & P_7 - P_6 \\ P_6 - P_7 & P_5 - P_7 & P_4 - P_7 & P_3 - P_7 & P_2 - P_7 & P_1 - P_7 & P_9 - P_7 & P_8 - P_7 \\ P_7 - P_8 & P_6 - P_8 & P_5 - P_8 & P_4 - P_8 & P_3 - P_8 & P_2 - P_8 & P_1 - P_8 & P_9 - P_8 \end{bmatrix}$$

Λ^* is

$$[\lambda_1^{-P_1}, \lambda_2^{-P_2}, \lambda_3^{-P_3}, \lambda_4^{-P_4}, \lambda_5^{-P_5}, \lambda_6^{-P_6}, \lambda_7^{-P_7}, \lambda_8^{-P_8}, \lambda_9^{-P_9}]$$

and

$$\Pi' = [\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7, \pi_8].$$

From Eq.(3.2.1)

$$\hat{\Pi} = P^{-1} \hat{\Lambda}^* \text{ provided } |P| \neq 0,$$

and

$$\text{Var}(\hat{\Pi}) = P^{-1} \text{Var}(\hat{\Lambda}^*) P^{-1},$$

$$\text{where } \text{Var}(\hat{\Lambda}^*) = \frac{1}{n} \begin{bmatrix} \lambda_1(1-\lambda_1) & -\lambda_1\lambda_2 & \dots & -\lambda_1\lambda_8 & -\lambda_1\lambda_9 \\ -\lambda_2\lambda_1 & \lambda_2(1-\lambda_2) & \dots & -\lambda_2\lambda_8 & -\lambda_2\lambda_9 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ -\lambda_8\lambda_1 & -\lambda_8\lambda_2 & \dots & \lambda_8(1-\lambda_8) & -\lambda_8\lambda_9 \\ -\lambda_9\lambda_1 & -\lambda_9\lambda_2 & \dots & -\lambda_9\lambda_8 & \lambda_9(1-\lambda_9) \end{bmatrix},$$

By decoding $\pi_1 = \pi_{11}$, $\pi_2 = \pi_{12}$, , $\pi_9 = \pi_{33}$, we may estimate each cell proportion. Using these estimated cell proportions, we may estimate the product moment correlation between two sensitive variables.

For I x J contingency table

$$\rho = \frac{\sum_i \sum_j \pi_{ij} (a_i - \bar{a}_+)(b_j - \bar{b}_+)}{\sqrt{[\sum_i \pi_{i+} (a_i - \bar{a}_+)^2] [\sum_j \pi_{+j} (b_j - \bar{b}_+)^2]}}$$

where a_i is a value assigned to the i -th row category, and b_j is a value assigned to the j -th column category, and $\bar{a}_+ = \frac{1}{\sum_{i=1}^I \pi_{i+}} \sum_{i=1}^I \pi_{i+} a_i$ and $\bar{b}_+ = \frac{1}{\sum_{j=1}^J \pi_{+j}} \sum_{j=1}^J \pi_{+j} b_j$.

The estimator is

$$\gamma = \frac{\sum_i^I \sum_j^J \hat{\pi}_{ij} (a_i - \hat{\bar{a}}_+) (b_j - \hat{\bar{b}}_+)}{\sqrt{[\sum_i \pi_{i+} (a_i - \hat{\bar{a}}_+)^2] [\sum_j \pi_{+j} (b_j - \hat{\bar{b}}_+)^2]}}$$

$$\text{where } \hat{\bar{a}}_+ = \sum_i \left[\frac{\hat{\pi}_{i+}}{\hat{\pi}_{++}} \right] a_i, \text{ and } \hat{\bar{b}}_+ = \sum_j \left[\frac{\hat{\pi}_{+j}}{\hat{\pi}_{++}} \right] b_j.$$

For the general case, S_1 has r subcategories and S_2 has c subcategories.

The above procedure is extendable, by renumbering the $r \times c$ contingency table,

$$(1\ 1) = 1, (1\ 2) = 2, \dots, (1\ c) = i, (2\ 1) = i+1, \dots, (r\ c) = m.$$

The probability of getting each renumbered cell is :

$$\lambda_1 = P_1 + (P_m - P_1)\pi_1 + (P_{m-1} - P_1)\pi_2 + \dots + (P_2 - P_1)\pi_{m-1},$$

$$\lambda_2 = P_2 + (P_1 - P_2)\pi_1 + (P_m - P_1)\pi_2 + \dots + (P_3 - P_2)\pi_{m-1},$$

$$\lambda_3 = P_3 + (P_2 - P_3)\pi_1 + (P_1 - P_3)\pi_2 + \dots + (P_4 - P_3)\pi_{m-1},$$

.

.

$$\lambda_{m-1} = P_{m-1} + (P_{m-1} - P_{m-2})\pi_1 + \dots + (P_m - P_{m-1})\pi_{m-1}.$$

We can express these equations in matrix notation, $\Lambda^* = P \Pi$, and by solving these equations for Π , we may estimate the product moment correlation between two sensitive variables.

3.3 Correlation Analysis for the Additive Model

To estimate the correlation between two sensitive variables which have more than two subcategories, we may apply the additive model.

Suppose each sensitive variable has 3 subcategories, then by applying the additive model, the respondents are asked to select their own category for the first sensitive variable and add their augmented value (number) to their own selected category. By using this procedure for the second sensitive variable in the same manner as with the first variable, the respondent reports a pair of answers (transformed responses) to the interviewer. To give further confidence to the respondent, the reported value r is reduced modulo 3 if r is greater than 3, and the interviewer records a pair of responses.

The transformed response table is given by

1+3=4	1+1=2	1+2=3
2+2=4	2+3=5	2+1=3
3+1=4	3+2=5	3+3=6
1	2	3

1+3=4	4	4	4	2	4	3
2+2=4	4	4	4	5	4	3
3+1=4	4	4	4	5	4	6
1+1=2	2	4	2	2	2	3
2+3=5	5	4	5	5	5	3
3+2=5	5	4	5	5	5	6
1+2=3	3	4	3	2	3	3
2+2=3	3	4	3	5	3	3
3+3=6	6	4	6	5	6	6
	π_1		π_2		π_3	

From this table, the probabilities (λ_{ij} , $i = 1, 2, 3$; $j = 1, 2, 3$) of getting each cell response is

$$\begin{aligned}\lambda_{11} = & P_3 P_3 \pi_{11} + P_3 P_2 \pi_{12} + P_3 P_1 \pi_{13} + P_2 P_3 \pi_{21} + P_2 P_2 \pi_{22} + P_2 P_1 \pi_{23} \\ & + P_1 P_3 \pi_{31} + P_1 P_2 \pi_{32} + P_1 P_1 \pi_{33}\end{aligned}$$

$$\begin{aligned}\lambda_{12} = & P_3 P_1 \pi_{11} + P_3 P_3 \pi_{12} + P_3 P_2 \pi_{13} + P_2 P_1 \pi_{21} + P_2 P_3 \pi_{22} + P_2 P_2 \pi_{23} \\ & + P_1 P_1 \pi_{31} + P_1 P_3 \pi_{32} + P_1 P_2 \pi_{33}\end{aligned}$$

$$\begin{aligned}\lambda_{13} = & P_3 P_2 \pi_{11} + P_3 P_1 \pi_{12} + P_3 P_3 \pi_{13} + P_2 P_2 \pi_{21} + P_2 P_1 \pi_{22} + P_2 P_3 \pi_{23} \\ & + P_1 P_2 \pi_{31} + P_1 P_1 \pi_{32} + P_1 P_3 \pi_{33}\end{aligned}$$

$$\begin{aligned}\lambda_{21} = & P_1 P_3 \pi_{11} + P_1 P_2 \pi_{12} + P_1 P_1 \pi_{13} + P_3 P_3 \pi_{21} + P_3 P_2 \pi_{22} + P_3 P_1 \pi_{23} \\ & + P_3 P_3 \pi_{31} + P_3 P_2 \pi_{32} + P_2 P_1 \pi_{33}\end{aligned}$$

$$\begin{aligned}\lambda_{22} = & P_1 P_1 \pi_{11} + P_1 P_3 \pi_{12} + P_1 P_2 \pi_{13} + P_3 P_1 \pi_{21} + P_3 P_3 \pi_{22} + P_3 P_2 \pi_{23} \\ & + P_2 P_1 \pi_{31} + P_2 P_3 \pi_{32} + P_2 P_2 \pi_{33}\end{aligned}$$

$$\begin{aligned}\lambda_{23} = & P_1 P_2 \pi_{11} + P_1 P_1 \pi_{12} + P_1 P_3 \pi_{13} + P_3 P_2 \pi_{21} + P_3 P_1 \pi_{22} + P_3 P_3 \pi_{23} \\ & + P_2 P_2 \pi_{31} + P_2 P_1 \pi_{32} + P_2 P_3 \pi_{33}\end{aligned}$$

$$\begin{aligned}\lambda_{31} = & P_2 P_3 \pi_{11} + P_2 P_2 \pi_{12} + P_2 P_1 \pi_{13} + P_1 P_3 \pi_{21} + P_1 P_2 \pi_{22} + P_1 P_1 \pi_{23} \\ & + P_3 P_3 \pi_{31} + P_3 P_2 \pi_{32} + P_3 P_1 \pi_{33}\end{aligned}$$

$$\begin{aligned}\lambda_{32} = & P_2 P_1 \pi_{11} + P_2 P_3 \pi_{12} + P_2 P_2 \pi_{13} + P_1 P_1 \pi_{21} + P_1 P_3 \pi_{22} + P_1 P_2 \pi_{23} \\ & + P_3 P_1 \pi_{31} + P_3 P_3 \pi_{32} + P_3 P_2 \pi_{33}\end{aligned}$$

$$\begin{aligned}\lambda_{33} = & P_2 P_2 \pi_{11} + P_2 P_1 \pi_{12} + P_2 P_3 \pi_{13} + P_1 P_2 \pi_{21} + P_1 P_1 \pi_{22} + P_1 P_3 \pi_{23} \\ & + P_3 P_2 \pi_{31} + P_3 P_1 \pi_{32} + P_3 P_3 \pi_{33}.\end{aligned}$$

These equations can be written in matrix form,

$$\begin{bmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{21} \\ \lambda_{22} \\ \lambda_{23} \\ \lambda_{31} \\ \lambda_{32} \\ \lambda_{33} \end{bmatrix} = \begin{bmatrix} P_3 P_3 & P_3 P_2 & P_3 P_1 & P_2 P_3 & P_2 P_2 & P_2 P_1 & P_1 P_3 & P_1 P_2 & P_1 P_1 \\ P_3 P_1 & P_3 P_3 & P_3 P_2 & P_2 P_1 & P_2 P_3 & P_2 P_2 & P_1 P_1 & P_1 P_3 & P_1 P_2 \\ P_3 P_2 & P_3 P_1 & P_3 P_3 & P_2 P_2 & P_2 P_1 & P_2 P_3 & P_1 P_2 & P_1 P_1 & P_1 P_3 \\ P_1 P_3 & P_1 P_2 & P_1 P_1 & P_3 P_3 & P_3 P_2 & P_3 P_1 & P_2 P_3 & P_2 P_2 & P_2 P_1 \\ P_1 P_1 & P_1 P_3 & P_1 P_2 & P_3 P_1 & P_3 P_3 & P_3 P_2 & P_2 P_1 & P_2 P_3 & P_2 P_2 \\ P_1 P_2 & P_1 P_1 & P_1 P_3 & P_3 P_2 & P_3 P_1 & P_3 P_3 & P_2 P_2 & P_2 P_1 & P_2 P_3 \\ P_2 P_3 & P_2 P_2 & P_2 P_1 & P_1 P_3 & P_1 P_2 & P_1 P_1 & P_3 P_3 & P_3 P_2 & P_3 P_1 \\ P_2 P_1 & P_2 P_3 & P_2 P_2 & P_1 P_1 & P_1 P_3 & P_1 P_2 & P_3 P_1 & P_3 P_3 & P_3 P_2 \\ P_2 P_2 & P_2 P_1 & P_2 P_3 & P_1 P_2 & P_1 P_1 & P_1 P_3 & P_3 P_2 & P_3 P_1 & P_3 P_3 \end{bmatrix} \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{21} \\ \pi_{22} \\ \pi_{23} \\ \pi_{31} \\ \pi_{32} \\ \pi_{33} \end{bmatrix}.$$

In matrix notation

$$\Lambda = P \Pi$$

$$\hat{\Pi} = P^{-1} \hat{\Lambda} \text{ provided } |P| \neq 0,$$

then $E \hat{\Pi} = P^{-1} E \hat{\Lambda}$, and $\text{Var}(\hat{\Pi}) = P^{-1} \text{Var}(\hat{\Lambda}) P^{-1}$,

where $\text{Var}(\hat{\Lambda}) = \{ \sigma_{ij} \}$, $\sigma_{ii} = \frac{\lambda_i (1 - \lambda_i)}{n}$, and $\sigma_{ij} = -\frac{\lambda_i \lambda_j}{n}$ for $i = 1, 2, 3$;
 $j = 1, 2, 3$.

Using these estimated cell proportions, we may estimate the product moment correlation between two sensitive variables. The formula is given by

$$\gamma = \frac{\sum_i^I \sum_j^J \hat{\pi}_{ij} (a_i - \hat{\bar{a}}_+) (b_j - \hat{\bar{b}}_+)}{\sqrt{[\sum_i \hat{\pi}_{i+} (a_i - \hat{\bar{a}}_+)^2] [\sum_j \hat{\pi}_{+j} (b_j - \hat{\bar{b}}_+)^2]}}$$

where $\hat{\bar{a}}_+ = \sum_i \left[\frac{\hat{\pi}_{i+}}{\hat{\pi}_{++}} \right] a_i$, and $\hat{\bar{b}}_+ = \sum_j \left[\frac{\hat{\pi}_{+j}}{\hat{\pi}_{++}} \right] b_j$.

3.4 Scrambled Randomized Response Technique.

Here, instead of adding a random number which is generated by a random device, the respondent is asked to multiply a random number by his true category number. The product is given to the interviewer, who does not know the value of the random number. This technique is called "the scrambled randomized response technique" (Eichhorn and Hayre, 1983). Eichhorn and Hayre showed how to generate the values of multiplier variable. Pollock and Bek (1976) compared the additive and the scrambled models. The scrambled randomized response model will be explained briefly.

Let C_j be the true category number for the j -th respondent ($C_j = 0, 1, 2, 3, \dots, T-1$, and $j = 1, 2, 3, \dots, n$) and m_j be the randomly selected

multiplier number for the j -th respondent, ($m_j = 0, 1, 2, \dots, T-1$). The selection probabilities $P(m_j = t) = P_t$, ($t = 0, 1, 2, \dots, T-1$, and $\sum_{t=0}^{T-1} P_t = 1$) are preassigned, and the distribution of the multiplier variable is known. T is the number of category. Each respondent is asked to select his own category but to keep their response confidential, they then multiply the multiplier value by their own category number. Then the j -th respondent's scrambled response whose true category is C_j is

$$\begin{aligned} C_j * m_j, & \quad C_j = 0, 1, 2, \dots, T-1. \\ m_j & = 0, 1, 2, \dots, T-1. \\ j & = 1, 2, \dots, n. \end{aligned}$$

The possible scrambled responses are $0, 1, 2, \dots, (T-1)*(T-1)$, hence some responses like 1 or $(T-1)*(T-1)$ have only one possibility to be that number. It is therefore probable the respondents will hesitate to release their information.

To provide further confidence to the respondent, the interviewer asks the respondent to transform the scrambled value and report the transformed value r_j , where

$$r_j = \begin{cases} C_j * m_j & \text{if } C_j * m_j \leq T-1 \\ C_j * m_j \text{ mod}(T) & \text{if } C_j * m_j > T-1 \end{cases}$$

where T is a prime number.

For the case $T = 5$, the transformed response for the scrambled model is given in table 6.

TABLE 6
 TRANSFORMED RESPONSE FOR T = 5

		m_j				
		0	1	2	3	4
C_j	0	0	0	0	0	0
	1	0	1	2	3	4
	2	0	2	4	1	3
	3	0	3	1	4	2
	4	0	4	3	2	1

The reported value r_j satisfies;

$$r_j = \begin{cases} C_j * m_j & \text{if } C_j * m_j \leq 4 \\ C_j * m_j - r_j = k \pmod{5} & \text{if } C_j * m_j > 4 \end{cases}$$

where k is a number less than 5. As we can see in table 6, the first category is not protected by multiplying by a random number. If the respondent's answer 0 is a nonsensitive response, the fact that this answer is not protected by the randomization technique will not be problem.

Define π_{C_j} : true population proportion in category C_j

P_{m_j} : the probability of selecting a multiplier number m_j .

Then, for $T = 5$ the probability (λ_r) that a respondent reports a value $r_j(0,1, 2, 3, \text{ or } 4)$ is :

$$\begin{aligned}
\lambda_0 &= P(r=0) = P(C_j=0, m_j=0) + P(C_j=0, m_j=1) + P(C_j=0, m_j=2) + P(C_j=0, m_j=3) + \\
& P(C_j=0, m_j=4) + P(C_j=1, m_j=0) + P(C_j=2, m_j=0) + P(C_j=3, m_j=0) + P(C_j=4, m_j=0) \\
& = P_0 + (1 - P_0) \pi_0
\end{aligned}$$

$$\begin{aligned}
\lambda_1 &= P(r=1) = P(C_j=1, m_j=1) + P(C_j=2, m_j=3) + P(C_j=3, m_j=2) + P(C_j=4, m_j=4) \\
& = \pi_1 P_1 + \pi_2 P_3 + \pi_3 P_2 + \pi_4 P_4
\end{aligned}$$

$$\begin{aligned}
\lambda_2 &= P(r=2) = P(C_j=1, m_j=2) + P(C_j=2, m_j=1) + P(C_j=3, m_j=4) + P(C_j=4, m_j=3) \\
& = \pi_1 P_2 + \pi_2 P_1 + \pi_3 P_4 + \pi_4 P_3
\end{aligned}$$

$$\begin{aligned}
\lambda_3 &= P(r=3) = P(C_j=1, m_j=3) + P(C_j=2, m_j=4) + P(C_j=3, m_j=1) + P(C_j=4, m_j=2) \\
& = \pi_1 P_3 + \pi_2 P_4 + \pi_3 P_1 + \pi_4 P_2
\end{aligned}$$

$$\begin{aligned}
\lambda_4 &= P(r=4) = P(C_j=1, m_j=4) + P(C_j=2, m_j=2) + P(C_j=3, m_j=3) + P(C_j=4, m_j=1) \\
& = \pi_1 P_4 + \pi_2 P_2 + \pi_3 P_3 + \pi_4 P_1
\end{aligned}$$

We may rewrite these equation in matrix form.

$$\Lambda = P \Pi,$$

where P is

$$P = \begin{bmatrix} 1 - P_0 & 0 & 0 & 0 & 0 \\ 0 & P_1 & P_3 & P_2 & P_4 \\ 0 & P_2 & P_1 & P_4 & P_3 \\ 0 & P_3 & P_4 & P_1 & P_2 \\ 0 & P_4 & P_2 & P_3 & P_1 \end{bmatrix},$$

$$\Lambda' = (\lambda_0 - P_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4),$$

and

$$\Pi' = (\pi_0, \pi_1, \pi_2, \pi_3, \pi_4).$$

Hence $\hat{\Pi} = P^{-1} \hat{\Lambda}$ provided $|P| \neq 0$.

$$E \hat{\Pi} = P^{-1} E \hat{\Lambda} = P^{-1} \Lambda, \text{ and}$$

$$\text{Var}(\hat{\Pi}) = P^{-1} \text{Var}(\hat{\Lambda}) (P^{-1})',$$

$$\text{where } \text{Var}(\hat{\Lambda}) = \frac{1}{n} \begin{bmatrix} \lambda_0(1-\lambda_0) & -\lambda_0\lambda_1 & -\lambda_0\lambda_2 & -\lambda_0\lambda_3 & -\lambda_0\lambda_4 \\ -\lambda_1\lambda_0 & \lambda_1(1-\lambda_1) & -\lambda_1\lambda_2 & -\lambda_1\lambda_3 & -\lambda_1\lambda_4 \\ -\lambda_2\lambda_0 & -\lambda_2\lambda_1 & \lambda_2(1-\lambda_2) & -\lambda_2\lambda_3 & -\lambda_2\lambda_4 \\ -\lambda_3\lambda_0 & -\lambda_3\lambda_1 & -\lambda_3\lambda_2 & \lambda_3(1-\lambda_3) & -\lambda_3\lambda_4 \\ -\lambda_4\lambda_0 & -\lambda_4\lambda_1 & -\lambda_4\lambda_2 & -\lambda_4\lambda_3 & \lambda_4(1-\lambda_4) \end{bmatrix},$$

$$\hat{\Pi}' = (\hat{\pi}_0, \hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\pi}_4),$$

$$\hat{\Lambda}' = (\hat{\lambda}_0 - P_0, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4),$$

and $\hat{\lambda}_t = \frac{n_t}{n}$, and n_t is the number of respondents who reported value t .

3.4.1 Correlation Analysis for the Multivariate version of the Scrambled Randomized Response Model

Suppose we have two sensitive variables with r and c subcategories respectively. The population can be tabulated as an $r \times c$ contingency table, and we need to estimate the corresponding cell proportions, $\pi_{00}, \pi_{01}, \dots, \pi_{0 c-1}, \pi_{10}, \pi_{11}, \dots, \pi_{1 c-1}, \dots, \pi_{r-1 0}, \pi_{r-1 1}, \dots, \pi_{r-1 c-1}$,

where $0 < \pi_{ij} < 1$ ($i=0,1,2, \dots, r-1$. $c=0,1,2, \dots, c-1$) and $\sum_{i=0}^{r-1} \sum_{j=0}^{c-1} \pi_{ij} = 1$.

We will explain the $r = 3$, and $c = 3$ case detail, since the $r > 3$, and $c > 3$ case is an extension of this procedure.

Let C_j be the true category number for the j -th respondent. To estimate each cell proportion, we may apply the scrambled randomized response models by reordering each cell number as follows: $(0, 0) = 0$, $(0, 1) = 1$, $(0, 2) = 2$, $(1, 0) = 3$, $(1, 1) = 4$, $(1, 2) = 5$, $(2, 0) = 6$, $(2, 1) = 7$, $(2, 2) = 8$, where the first number of each pair is the row category and the second number is the column category. Then the population will be tabulated as shown in Table 7:

TABLE 7
REORDERED POPULATION CATEGORIES

		S_2		
		0	1	2
S_1	0	π_0	π_1	π_2
	1	π_3	π_4	π_5
	2	π_6	π_7	π_8

Using the same steps as were explained earlier in this chapter, the j -th respondent has a scrambled value $C_j^*m_j$. By scrambling, the respondent's answer cannot be protected for some scrambled values consequently, in order to give more confidence, the respondent is asked to transform the scrambled value and report the transformed value r_j ,

$$r_j = \begin{cases} C_j * m_j & \text{if } C_j * m_j \leq 8 \\ C_j * m_j \pmod{9} & \text{if } C_j * m_j > 8 \end{cases}$$

Then the transformed response for the scrambled model is given by table 8.

TABLE 8
TRANSFORMED RESPONSE FOR T=9

		m_j								
		0	1	2	3	4	5	6	7	8
C_j	0	0	0	0	0	0	0	0	0	0
	1	0	1	2	3	6	4	5	7	8
	2	0	2	1	6	3	8	7	5	4
	3	0	3	6	2	1	5	8	4	7
	4	0	6	3	1	2	7	4	8	5
	5	0	4	8	5	7	6	1	2	3
	6	0	5	7	8	4	1	3	6	2
	7	0	7	5	4	8	2	6	3	1
	8	0	8	4	7	5	3	2	1	6

Then the probability (λ_r) that a respondent reports a value r_j (0, 1, 2, 3, 4, 5, 6, 7, 8) is :

$$\begin{aligned} \lambda_0 &= \pi_0 P_0 + \pi_0 P_1 + \pi_0 P_2 + \pi_0 P_3 + \pi_0 P_4 + \pi_0 P_5 + \pi_0 P_6 + \pi_0 P_7 + \pi_0 P_8 \\ &= \pi_1 P_0 + \pi_2 P_0 + \pi_3 P_0 + \pi_4 P_0 + \pi_5 P_0 + \pi_6 P_0 + \pi_7 P_0 + \pi_8 P_0 \\ &= P_0 + (1 - P_0) \pi_0 \end{aligned}$$

$$\lambda_1 = \pi_1 P_1 + \pi_2 P_2 + \pi_3 P_4 + \pi_4 P_3 + \pi_5 P_6 + \pi_6 P_5 + \pi_7 P_8 + \pi_8 P_7$$

$$\lambda_2 = \pi_1 P_2 + \pi_2 P_1 + \pi_3 P_3 + \pi_4 P_4 + \pi_5 P_7 + \pi_6 P_8 + \pi_7 P_5 + \pi_8 P_6$$

$$\lambda_3 = \pi_1 P_3 + \pi_2 P_4 + \pi_3 P_1 + \pi_4 P_2 + \pi_5 P_8 + \pi_6 P_6 + \pi_7 P_7 + \pi_8 P_5$$

$$\lambda_4 = \pi_1 P_5 + \pi_2 P_8 + \pi_3 P_7 + \pi_4 P_6 + \pi_5 P_1 + \pi_6 P_4 + \pi_7 P_3 + \pi_8 P_2$$

$$\lambda_5 = \pi_1 P_6 + \pi_2 P_7 + \pi_3 P_5 + \pi_4 P_8 + \pi_5 P_3 + \pi_6 P_1 + \pi_7 P_2 + \pi_8 P_4$$

$$\lambda_6 = \pi_1 P_4 + \pi_2 P_3 + \pi_3 P_2 + \pi_4 P_1 + \pi_5 P_5 + \pi_6 P_7 + \pi_7 P_6 + \pi_8 P_8$$

$$\lambda_7 = \pi_1 P_7 + \pi_2 P_6 + \pi_3 P_8 + \pi_4 P_5 + \pi_5 P_4 + \pi_6 P_2 + \pi_7 P_1 + \pi_8 P_3$$

$$\lambda_8 = \pi_1 P_8 + \pi_2 P_5 + \pi_3 P_6 + \pi_4 P_7 + \pi_5 P_2 + \pi_6 P_3 + \pi_7 P_4 + \pi_8 P_1.$$

By rewriting these equations in matrix form, we get

$$\Lambda = P \Pi,$$

where P is

$$\begin{bmatrix} 1-P_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & P_1 & P_2 & P_4 & P_3 & P_6 & P_5 & P_8 & P_7 \\ 0 & P_2 & P_1 & P_3 & P_4 & P_7 & P_8 & P_5 & P_6 \\ 0 & P_3 & P_4 & P_1 & P_2 & P_8 & P_6 & P_7 & P_5 \\ 0 & P_5 & P_8 & P_7 & P_6 & P_1 & P_4 & P_3 & P_2 \\ 0 & P_6 & P_7 & P_5 & P_8 & P_3 & P_1 & P_2 & P_4 \\ 0 & P_4 & P_3 & P_2 & P_1 & P_5 & P_7 & P_6 & P_8 \\ 0 & P_7 & P_6 & P_8 & P_5 & P_4 & P_2 & P_1 & P_3 \\ 0 & P_8 & P_5 & P_6 & P_7 & P_2 & P_3 & P_4 & P_1 \end{bmatrix}$$

$$\Pi' = (\pi_0, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7, \pi_8),$$

and

$$\Lambda' = (\lambda_0 - P_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_8).$$

$$\text{Hence } \hat{\Pi} = P^{-1} \hat{\Lambda},$$

$$\text{Var}(\hat{\Pi}) = P^{-1} \text{Var}(\hat{\Lambda}) (P^{-1})',$$

$$\text{where } \text{Var}(\hat{\Lambda}) = \frac{1}{n} \begin{bmatrix} \lambda_0(1-\lambda_0) & -\lambda_0\lambda_1 & \dots & -\lambda_0\lambda_7 & -\lambda_0\lambda_8 \\ -\lambda_1\lambda_0 & \lambda_1(1-\lambda_1) & \dots & -\lambda_1\lambda_7 & -\lambda_1\lambda_8 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ -\lambda_7\lambda_0 & -\lambda_7\lambda_1 & \dots & \lambda_7(1-\lambda_7) & -\lambda_7\lambda_8 \\ -\lambda_8\lambda_0 & -\lambda_8\lambda_1 & \dots & -\lambda_8\lambda_7 & \lambda_8(1-\lambda_8) \end{bmatrix}.$$

Using these estimated cell proportions we can estimate the product moment correlation between the two sensitive variables. The formula for the estimated product moment correlation is given by

$$\gamma = \frac{\sum_i \sum_j \hat{\pi}_{ij} (a_i - \hat{\bar{a}}_+) (b_j - \hat{\bar{b}}_+)}{\sqrt{[\sum_i \pi_{i+} (a_i - \hat{\bar{a}}_+)^2] [\sum_j \pi_{+j} (b_j - \hat{\bar{b}}_+)^2]}}$$

$$\text{where } \hat{\bar{a}}_+ = \sum_i \left[\frac{\hat{\pi}_{i+}}{\pi_{++}} \right] a_i, \text{ and } \hat{\bar{b}}_+ = \sum_j \left[\frac{\hat{\pi}_{+j}}{\pi_{++}} \right] b_j.$$

3.5 Multiproportional Randomized Response Technique

Suppose we have two sensitive variables with r and c subcategories respectively. The population can be tabulated as an $r \times c$ contingency table, and

we need to estimate the corresponding cell proportions, $\pi_{11}, \pi_{12}, \dots, \pi_{1c}, \pi_{21}, \pi_{22}, \dots, \pi_{2c}, \dots, \pi_{r1}, \pi_{r2}, \dots, \pi_{rc}$, where $0 < \pi_{ij} < 1$ ($i = 1, 2, \dots, r$; $c = 1, 2, \dots, c$) and $\sum_{i=1}^r \sum_{j=1}^c \pi_{ij} = 1$.

We will explain the $r = 3$, and $c = 3$ case detail, since the $r > 3$, and $c > 3$ case is an extension of this procedure.

A simple random sample of size n is drawn with replacement from that population. Random devices are used to obtain, from the respondents in the sample, information concerning the category to which they belong on a probability basis, and in such a way that the respondent's status will not be revealed to the interviewer. Suppose that the random device is the Hopkins' Randomizing Device (which was developed by Liu and Chow 1976 a, b). A number of balls of two different colors, e.g., green and white, will be placed in the body of the device (see figure 3). A discrete number, such as 1, 2, 3, will be marked on the surface of each of the white balls. The proportion of green to white balls, and of white balls with different figures, will be predetermined. The respondent is asked to turn the device upside down, shake the device thoroughly, and turn it right side up to allow one of the balls to appear in the window of the device.

The ball in the window will either be green or white. If it is a green ball, the respondent will be asked to answer the sensitive question (e.g., the number of abortions she has had). If the ball is white, there will be a number marked on its surface, and the respondent simply tells the number. The answers will again be 1, 2, 3, depending on the number marked on the surface of the white ball.

Interviewers stand on the opposite side of the window of the device, and therefore do not know whether the respondents have been asked to respond to the sensitive question or whether the respondents are responding with the number on a white ball.

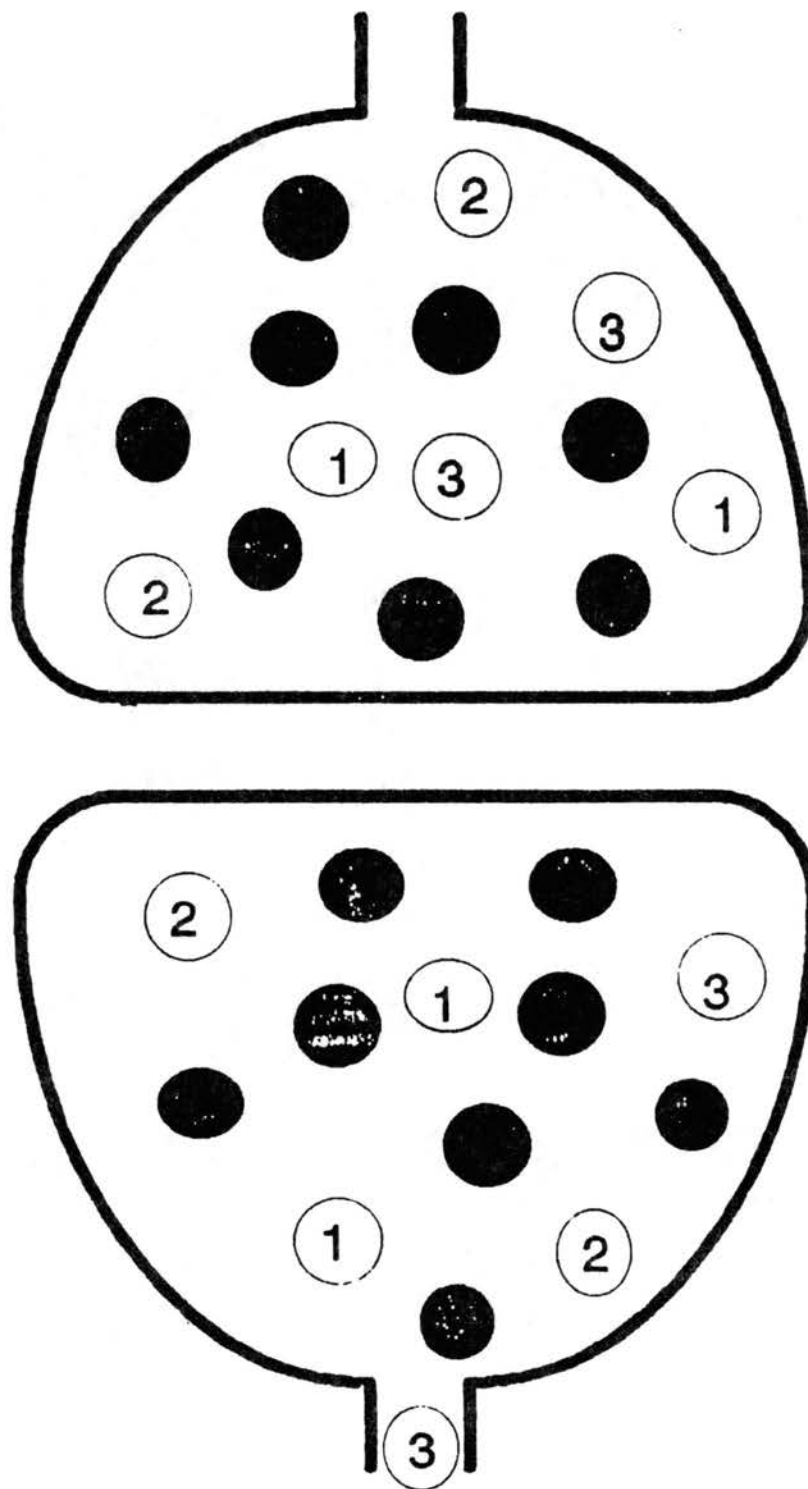


Figure 3. Hopkins' Randomizing Device

For two sensitive variables with 3 subcategories, the interviewer prepares a device for each question (or prepares one device and asks the respondent to use the same device for each question), and the respondent will be asked to use the first device for the first question and so on. Obviously, the respondent will return the ball into the device after answering the question.

Let w_i represent the number of white balls marked i (where $i = 1, 2, 3$) and where g represents the number of green balls (unmarked), then the total number of balls in the device is

$$g + w = g + \sum_{i=1}^3 w_i, \text{ where } w = \sum_{i=1}^3 w_i.$$

For the first question, the respondent will shake the device and will get a ball. If it is a green ball and he (she) really belongs to subcategory 2, then he (she) will report 2, and replace the ball into the device. For the second question, if it is a white ball marked 3, then he will report 3 whichever category he really belongs to.

Let π_{ij} represent the true proportion of respondents who possess i -th category for the first question, and j -th category for the second question.

Let $P[i' j' | i j]$ represent the conditional probability that the respondents give the responses $(i' j')$, given that they are in category $(i j)$.

Now the probability that the respondent gives the response $(2 3)$ is

$$\begin{aligned} \lambda_{23} = & P[2 3 | 1 1] \pi_{11} + P[2 3 | 1 2] \pi_{12} + P[2 3 | 1 3] \pi_{13} \\ & + P[2 3 | 2 1] \pi_{21} + P[2 3 | 2 2] \pi_{22} + P[2 3 | 2 3] \pi_{23} \\ & + P[2 3 | 3 1] \pi_{31} + P[2 3 | 3 2] \pi_{32} + P[2 3 | 3 3] \pi_{33} \end{aligned}$$

Since before reporting a pair of answers, the respondent uses the random device two times independently and each time the device selects a question, We can rewrite $P[i|j|j'] = P[i|i']P[j|j']$. Therefore λ_{23} can be rewritten as follows

$$\begin{aligned}\lambda_{23} = & P[2|1] P[3|1] \pi_{11} + P[2|1] P[3|2] \pi_{12} + P[2|1] P[3|3] \pi_{13} \\ & + P[2|2] P[3|1] \pi_{21} + P[2|2] P[3|2] \pi_{22} + P[2|2] P[3|3] \pi_{23} \\ & + P[2|3] P[3|1] \pi_{31} + P[2|3] P[3|2] \pi_{32} + P[2|3] P[3|3] \pi_{33}.\end{aligned}$$

Similarly, we may have λ_{ij} for all i,j , and we can express the probability (λ_{ij}) in matrix form using the Kronecker product (\otimes), as

$$\begin{bmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{21} \\ \lambda_{22} \\ \lambda_{23} \\ \lambda_{31} \\ \lambda_{32} \\ \lambda_{33} \end{bmatrix} = \begin{bmatrix} (P[1|1] P[1|2] P[1|3]) \otimes (P[1|1] P[1|2] P[1|3]) \\ (P[1|1] P[1|2] P[1|3]) \otimes (P[2|1] P[2|2] P[2|3]) \\ (P[1|1] P[1|2] P[1|3]) \otimes (P[3|1] P[3|2] P[3|3]) \\ (P[2|1] P[2|2] P[2|3]) \otimes (P[1|1] P[1|2] P[1|3]) \\ (P[2|1] P[2|2] P[2|3]) \otimes (P[2|1] P[2|2] P[2|3]) \\ (P[2|1] P[2|2] P[2|3]) \otimes (P[3|1] P[3|2] P[3|3]) \\ (P[3|1] P[3|2] P[3|3]) \otimes (P[1|1] P[1|2] P[1|3]) \\ (P[3|1] P[3|2] P[3|3]) \otimes (P[1|1] P[1|2] P[1|3]) \\ (P[3|1] P[3|2] P[3|3]) \otimes (P[1|1] P[1|2] P[1|3]) \end{bmatrix} \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{21} \\ \pi_{22} \\ \pi_{23} \\ \pi_{31} \\ \pi_{32} \\ \pi_{33} \end{bmatrix} \quad (3.5.1)$$

By rewriting these equations in matrix form, we get

$$\Lambda = \mathbf{P} \Pi,$$

where $\Lambda' = (\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{21}, \lambda_{22}, \lambda_{23}, \lambda_{31}, \lambda_{32}, \lambda_{33})$,

$$\Pi' = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{31}, \pi_{32}, \pi_{33}),$$

and \mathbf{P} can be rewritten as

$$\mathbf{P} = \begin{pmatrix} (\mathbf{P}[1|1] \ \mathbf{P}[1|2] \ \mathbf{P}[1|3]) \\ (\mathbf{P}[2|1] \ \mathbf{P}[2|2] \ \mathbf{P}[2|3]) \\ (\mathbf{P}[3|1] \ \mathbf{P}[3|2] \ \mathbf{P}[3|3]) \end{pmatrix} \otimes \begin{pmatrix} (\mathbf{P}[1|1] \ \mathbf{P}[1|2] \ \mathbf{P}[1|3]) \\ (\mathbf{P}[2|1] \ \mathbf{P}[2|2] \ \mathbf{P}[2|3]) \\ (\mathbf{P}[3|1] \ \mathbf{P}[3|2] \ \mathbf{P}[3|3]) \end{pmatrix}$$

Hence $\hat{\Pi} = \mathbf{P}^{-1} \hat{\Lambda}$, provided \mathbf{P} is nonsingular.

$$\text{Var}(\hat{\Pi}) = \mathbf{P}^{-1} \text{Var}(\hat{\Lambda}) (\mathbf{P}^{-1})'$$

$$\text{where } \text{Var}(\hat{\Lambda}) = \frac{1}{n} \begin{pmatrix} \lambda_{11}(1-\lambda_{11}) & -\lambda_{11}\lambda_{12} & \dots & -\lambda_{11}\lambda_{32} & -\lambda_{11}\lambda_{33} \\ -\lambda_{12}\lambda_{11} & \lambda_{12}(1-\lambda_{12}) & \dots & -\lambda_{12}\lambda_{32} & -\lambda_{12}\lambda_{33} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ -\lambda_{32}\lambda_{11} & -\lambda_{32}\lambda_{12} & \dots & \lambda_{32}(1-\lambda_{32}) & -\lambda_{32}\lambda_{33} \\ -\lambda_{33}\lambda_{11} & -\lambda_{33}\lambda_{12} & \dots & -\lambda_{33}\lambda_{32} & \lambda_{33}(1-\lambda_{33}) \end{pmatrix}$$

Using these estimated cell proportions we can estimate the product moment correlation between the two sensitive variables. The formula for the estimated product moment correlation is given by

$$\gamma = \frac{\sum_i \sum_j \hat{\pi}_{ij} (a_i - \hat{\bar{a}}_+) (b_j - \hat{\bar{b}}_+)}{\sqrt{[\sum_i \hat{\pi}_{i+} (a_i - \hat{\bar{a}}_+)^2] [\sum_j \hat{\pi}_{+j} (b_j - \hat{\bar{b}}_+)^2]}}$$

$$\text{where } \hat{\bar{a}}_+ = \sum_i \left[\frac{\hat{\pi}_{i+}}{\hat{\pi}_{++}} \right] a_i, \text{ and } \hat{\bar{b}}_+ = \sum_j \left[\frac{\hat{\pi}_{+j}}{\hat{\pi}_{++}} \right] b_j$$

3.5.1 Test of Hypothesis

Before discussing the test of hypothesis we will show the relationship between λ_{ij} and π_{ij} . i.e., we will show π_{ij} 's are independent if and only if the λ_{ij} 's are independent.

In a 3 x 3 contingency table, if the π_{ij} 's are independent then $\pi_{ij} = \pi_{i.} \pi_{.j}$.
From Eq(3.4.1),

$$\begin{aligned} \lambda_{11} = & P[1|1]P[1|1] \pi_{11} + P[1|1]P[1|2] \pi_{12} + P[1|1]P[1|3] \pi_{13} \\ & + P[1|2]P[1|1] \pi_{21} + P[1|2]P[1|2] \pi_{22} + P[1|2]P[1|3] \pi_{23} \\ & + P[1|3]P[1|1] \pi_{31} + P[1|3]P[1|2] \pi_{32} + P[1|3]P[1|3] \pi_{33}. \end{aligned}$$

Now assumes that the π_{ij} 's are independent, i.e., $\pi_{ij} = \pi_{i.} \pi_{.j}$, then

$$\begin{aligned} \lambda_{11} = & P[1|1]P[1|1] \pi_{1.} \pi_{.1} + P[1|1]P[1|2] \pi_{1.} \pi_{.2} + P[1|1]P[1|3] \pi_{1.} \pi_{.3} \\ & + P[1|2]P[1|1] \pi_{2.} \pi_{.1} + P[1|2]P[1|2] \pi_{2.} \pi_{.2} + P[1|2]P[1|3] \pi_{2.} \pi_{.3} \\ & + P[1|3]P[1|1] \pi_{3.} \pi_{.1} + P[1|3]P[1|2] \pi_{3.} \pi_{.2} + P[1|3]P[1|3] \pi_{3.} \pi_{.3} \end{aligned}$$

$$= (P[1|1] \ P[1|2] \ P[1|3]) \otimes (P[1|1] \ P[1|2] \ P[1|3]) \begin{bmatrix} \pi_{1.} \pi_{.1} \\ \pi_{1.} \pi_{.2} \\ \pi_{1.} \pi_{.3} \\ \pi_{2.} \pi_{.1} \\ \pi_{2.} \pi_{.2} \\ \pi_{2.} \pi_{.3} \\ \pi_{3.} \pi_{.1} \\ \pi_{3.} \pi_{.2} \\ \pi_{3.} \pi_{.3} \end{bmatrix}$$

Since $\lambda_{1.} = \lambda_{11} + \lambda_{12} + \lambda_{13}$ and $\lambda_{.1} = \lambda_{11} + \lambda_{21} + \lambda_{31}$,
and if the λ_{ij} 's are independent, then $\lambda_{ij} = \lambda_{i.} \lambda_{.j}$.

Now we will relate $\lambda_{i.} \lambda_{.j}$ to $\pi_{i.} \pi_{.j}$.

From Eq(3.5.1),

$$\begin{aligned} \lambda_{1.} = & P[1|1](P[1|1]+P[2|1]+P[3|1])\pi_{11} + P[1|1](P[1|2]+P[2|2]+P[3|2])\pi_{12} \\ & + P[1|1](P[1|3]+P[2|3]+P[3|3])\pi_{13} + P[1|2](P[1|1]+P[2|1]+P[3|1])\pi_{21} \\ & + P[1|2](P[1|2]+P[2|2]+P[3|2])\pi_{22} + P[1|2](P[1|3]+P[2|3]+P[3|3])\pi_{23} \\ & + P[1|3](P[1|1]+P[2|1]+P[3|1])\pi_{31} + P[1|3](P[1|2]+P[2|2]+P[3|2])\pi_{32} \\ & + P[1|3](P[1|3]+P[2|3]+P[3|3])\pi_{33}. \end{aligned}$$

Since $P[1|j]+P[2|j]+P[3|j] = 1$, then

$$\lambda_{1.} = P[1|1]\pi_{1.} + P[1|2]\pi_{2.} + P[1|3]\pi_{3.} \text{ and similarly}$$

$$\begin{aligned} \lambda_{.1} = & P[1|1](P[1|1]+P[2|1]+P[3|1])\pi_{11} + P[1|2](P[1|2]+P[2|2]+P[3|2])\pi_{12} \\ & + P[1|3](P[1|3]+P[2|3]+P[3|3])\pi_{13} + P[1|1](P[1|1]+P[2|1]+P[3|1])\pi_{21} \\ & + P[1|2](P[1|2]+P[2|2]+P[3|2])\pi_{22} + P[1|3](P[1|3]+P[2|3]+P[3|3])\pi_{23} \\ & + P[1|1](P[1|1]+P[2|1]+P[3|1])\pi_{31} + P[1|2](P[1|2]+P[2|2]+P[3|2])\pi_{32} \\ & + P[1|3](P[1|3]+P[2|3]+P[3|3])\pi_{33}. \end{aligned}$$

Since $P[1|j]+P[2|j]+P[3|j]=1$,

$$\begin{aligned}\lambda_{.1} &= P[1|1](\pi_{11}+\pi_{21}+\pi_{31}) + P[1|2](\pi_{12}+\pi_{22}+\pi_{32}) + P[1|3](\pi_{13}+\pi_{23}+\pi_{33}) \\ &= P[1|1] \pi_{.1} + P[1|2] \pi_{.2} + P[1|3] \pi_{.3}.\end{aligned}$$

Therefore

$$\lambda_{1.} \lambda_{.1} = (P[1|1] \ P[1|2] \ P[1|3]) \otimes (P[1|1] \ P[1|2] \ P[1|3]) \begin{pmatrix} \pi_{1.} \pi_{.1} \\ \pi_{1.} \pi_{.2} \\ \pi_{1.} \pi_{.3} \\ \pi_{2.} \pi_{.1} \\ \pi_{2.} \pi_{.2} \\ \pi_{2.} \pi_{.3} \\ \pi_{3.} \pi_{.1} \\ \pi_{3.} \pi_{.2} \\ \pi_{3.} \pi_{.3} \end{pmatrix}.$$

Therefore we showed $\pi_{ij} = \pi_{i.} \pi_{.j}$ if and only if $\lambda_{ij} = \lambda_{i.} \lambda_{.j}$.

Using this relationship we may discuss a test of independence,

$$\chi^2 = \frac{\sum_i^3 \sum_j^3 (n_{ij} - n \hat{\lambda}_{i.} \hat{\lambda}_{.j})^2}{n \hat{\lambda}_{i.} \hat{\lambda}_{.j}}$$

where n_{ij} is the observed ij -th cell count.

3.6 Multiproportional Randomized Response Technique

With Reordering Cell Numbers

For the multiproportional data, as we explained in the additive and scrambled randomized response models, by reordering each cell number (table 5) and using a random device (wheel, multifaced dice, or Hopkins' device).

We can estimate each cell proportion π_{ij} ($i = 1, 2, \dots, r, j = 1, 2, \dots, c$).

Suppose a sample was drawn from a bivariate discrete population. For instance if each variable has 3 subcategories, then the population can be tabulated as a 3 x 3 contingency table (table 5). By reordering each cell number as in table 5, the contingency table can be express as a 9 x 1 vector, $\Pi = (\pi_1, \pi_2, \pi_3, \dots, \pi_9)'$.

As a random device, we can use the Hopkins' Randomizing Device (Figure 3) then following the same steps section 3.4 we can estimate each cell proportion.

Let w_i represent the number of white balls marked i (where $i = 1, 2, \dots, 9$), and g represents the number of green balls (unmarked), then the total number of balls in the device is $g + \sum_{i=1}^9 w_i = g + w$ (where $w = \sum_{i=1}^9 w_i$). If π_i represents the true proportion of respondents who belong to i -th category (where $\sum_{i=1}^9 \pi_i = 1$), then the probability (λ_i) that a respondent reports i is:

$$\lambda_i = \pi_i \left(\frac{g}{g+w} \right) + \frac{w_i}{g+w}, \quad (3.6.1)$$

where $i = 1, 2, 3, \dots, 9$.

Let n_i be the number of respondents reporting i , then the proportion of respondents reporting i is $\frac{n_i}{n}$ ($= \hat{\lambda}_i$). Substituting this into Eq.(3.6.1), and solving for π_i , then the estimate of π_i becomes

$$\hat{\pi}_i = \frac{(g+w) \frac{n_i}{n} - w_i}{g}.$$

The estimated variance of $\hat{\pi}_i$ becomes

$$\hat{\text{Var}}(\hat{\pi}_i) = \left[\frac{g+w}{g} \right]^2 \frac{1}{n} \frac{n_i}{n} \left(1 - \frac{n_i}{n} \right). \quad (3.6.2)$$

The estimated covariance between $\hat{\pi}_i$ and $\hat{\pi}_j$ is

$$\hat{\text{Cov}}(\hat{\pi}_i, \hat{\pi}_j) = -\left[\frac{g+w}{g}\right]^2 \frac{\frac{n_i}{n} \frac{n_j}{n}}{n}. \quad (3.6.3)$$

Liu and Chow (1976 a) indicate that "the ratio of green balls to the total number of balls in the device is the major component which affects the efficiency of estimate." From Eq.(3.6.1) and Eq.(3.6.3), we can see that for fixed total number of balls, if we increases the number of green balls, $\text{Var}(\hat{\pi}_i)$ and $\hat{\text{Cov}}(\hat{\pi}_i, \hat{\pi}_j)$ decreases. But if the ratio of green balls to the white balls is large, the respondent's cooperation will decrease.

Now by decoding the reordered cell number, $\pi_1 = \pi_{11}, \pi_2 = \pi_{12}, \dots, \pi_9 = \pi_{33}$. we can estimate the correlation between two sensitive variables.

CHAPTER IV

RANDOMIZED RESPONSE TECHNIQUE FOR THE QUANTITATIVE ATTRIBUTES

4.1 Additive Randomized Response Models

Kim and Flueck's (1978) additive randomized response models can be used to obtain responses for sensitive questions when the answers are quantitative. The respondent is asked to sum his (her) sensitive attribute (S) and an augmented value (Y). The augmented value is generated from a random device and is not known to the interviewer, but the distribution of the augmented variable is completely known and the augmented variable is independent of the sensitive variable. Suppose a simple random sample is drawn from a bivariate continuous population. The respondent is asked to generate a random number from a random device and add it to his (her) own sensitive attribute for each question. The value of the random number which is generated by the respondent is unobserved and unknown to the interviewer.

Let r_i = observed response variable

S_i = unknown sensitive variable

Y = augmented variable

then the response value for each question can be written

$$r_1 = S_1 + Y$$

$$r_2 = S_2 + Y$$

where S_i ($i = 1, 2$) and Y are independent.

The estimated mean of the observed random variable is

$$\hat{\mu}_{r_i} = \hat{\mu}_{S_i} + \mu_Y \quad (4.1.1)$$

and hence the unbiased estimate of μ_{S_i} is

$$\hat{\mu}_{S_i} = \hat{\mu}_{r_i} - \mu_Y.$$

And $\text{Var}(r_i) = \text{Var}(S_i) + \text{Var}(Y)$, since S_i and Y are independent and hence the estimated variance, $\sigma_{S_i}^2$, is given by

$$\hat{\sigma}_{S_i}^2 = \hat{\sigma}_{r_i}^2 - \sigma_Y^2, \quad (4.1.2)$$

where $\hat{\sigma}_{r_i}^2 = \frac{1}{n-1} \sum_{j=1}^n (r_{ij} - \bar{r}_i)^2$, $\bar{r}_i = \frac{1}{n} \sum_{j=1}^n r_{ij}$.

Now From Eq.(4.1.1), since S_i and Y are independent, the estimated variance of $\hat{\mu}_{r_i}$ is

$$\text{Var}(\hat{\mu}_{r_i}) = \text{Var}(\hat{\mu}_{S_i}) + \text{Var}(\mu_Y).$$

Since the distribution of Y is completely known, the estimated variance of $\hat{\mu}_{S_i}$ is given by

$$\begin{aligned} \text{Var}(\hat{\mu}_{S_i}) &= \text{Var}(\hat{\mu}_{r_i}) - \text{Var}(\mu_Y) \\ &= \frac{\text{Var}(r_i)}{n} = \frac{1}{n} (\sigma_{S_i}^2 + \sigma_Y^2), \end{aligned} \quad (4.1.3)$$

where $\text{Cov}(S_i, Y) = 0$, because S_i and Y are independent.

Hence to reduce the estimated variance of $\hat{\mu}_{S_1}$, we need to choose an augmented variable with small variance. This result is shown in table 9, in other words, by choosing an augmented variable with small variance we may have a short confidence interval for $\hat{\mu}_{S_1}$.

Now since $\text{Cov}(r_1, r_2) = \text{Cov}(S_1, S_2) + \text{Var}(Y)$, the correlation between two sensitive variables is given by

$$\begin{aligned} \rho_{S_1 S_2} &= \frac{\text{Cov}(S_1, S_2)}{\sqrt{\text{Var}(S_1) \text{Var}(S_2)}} \\ &= \frac{\sigma_{r_1 r_2} - \sigma_Y^2}{\sqrt{\sigma_{S_1}^2 \sigma_{S_2}^2}} \end{aligned}$$

Divide both the numerator and denominator by $\sigma_{r_1} \sigma_{r_2}$,

$$\rho_{S_1 S_2} = \frac{\rho_{r_1 r_2} - \frac{\sigma_Y^2}{\sigma_{r_1} \sigma_{r_2}}}{\sqrt{\frac{\sigma_{S_1}^2 \sigma_{S_2}^2}{\sigma_{r_1}^2 \sigma_{r_2}^2}}}$$

Let $X = \sqrt{\frac{\sigma_{S_1}^2 \sigma_{S_2}^2}{\sigma_{r_1}^2 \sigma_{r_2}^2}}$, then since $\frac{1}{X} = \left(1 + \frac{1 - X}{X}\right)$, $\rho_{S_1 S_2}$ can be written

TABLE 9
ESTIMATED MEANS & STANDARD DEVIATIONS FOR THE
ADDITIVE RANDOMIZED RESPONSE MODELS

α	β	μ_{S_1}	STD	μ_{S_2}	STD
2	2	29.99325	1.76233	37.06723	1.86817
2	3	29.99117	1.78970	37.06521	1.89342
2	4	29.98909	1.82773	37.06313	1.92884
2	5	29.98700	1.87576	37.06104	1.97386
2	6	29.98492	1.93306	37.05896	2.02785
2	7	29.98284	1.99883	37.05688	2.09011
2	8	29.98076	2.07225	37.05480	2.15993
2	9	29.97868	2.15256	37.05272	2.23660
2	10	29.97659	2.23900	37.05060	2.31945
2	11	29.97451	2.33090	37.04855	2.40782
2	12	29.97243	2.42764	37.04647	2.50115
3	2	30.03489	1.80234	37.12420	1.96584
3	3	30.03519	1.83979	37.12451	2.00595
3	4	30.03550	1.89292	37.12482	2.06036
3	5	30.03580	1.96046	37.12512	2.12798
3	6	30.03611	2.04098	37.12543	2.20759
3	7	29.98284	1.99883	37.05688	2.09011
3	8	30.03672	2.23514	37.12604	2.39783
3	9	30.03703	2.34604	37.12635	2.50611
3	10	30.03733	2.46453	37.12665	2.62174
3	11	30.03764	2.58957	37.12696	2.74379

$$\begin{aligned}
\rho_{S_1 S_2} &= \left[\rho_{r_1 r_2} - \frac{\sigma_Y^2}{\sigma_{r_1} \sigma_{r_2}} \right] \left(1 + \frac{1-X}{X} \right) \\
&= \rho_{r_1 r_2} \left(1 + \frac{1-X}{X} \right) - \frac{\sigma_Y^2}{\sigma_{r_1} \sigma_{r_2}} \left(1 + \frac{1-X}{X} \right) \\
&= \rho_{r_1 r_2} + \rho_{r_1 r_2} \frac{1-X}{X} - \frac{\sigma_Y^2}{\sigma_{r_1} \sigma_{r_2}} \frac{1}{X}
\end{aligned}$$

Since $\frac{1-X}{X} = \frac{\sigma_{r_1} \sigma_{r_2}}{\sigma_{S_1} \sigma_{S_2}} - 1$, and $\frac{\sigma_Y^2}{\sigma_{r_1} \sigma_{r_2}} \frac{1}{X} = \frac{\sigma_Y^2}{\sigma_{S_1} \sigma_{S_2}}$,

$$\rho_{S_1 S_2} = \rho_{r_1 r_2} + \rho_{r_1 r_2} \left(\frac{\sigma_{r_1}}{\sigma_{S_1}} \frac{\sigma_{r_2}}{\sigma_{S_2}} - 1 \right) - \frac{\sigma_Y^2}{\sigma_{S_1} \sigma_{S_2}}, \quad (4.1.4)$$

where the last two terms are due to the random device. Since $\rho_{r_1 r_2}$, σ_{r_1} , and σ_{r_2} can be calculated from the observed data, and σ_{S_1} and σ_{S_2} can be estimated from Eq.(4.1.2), and σ_Y^2 is known, the bias can be estimated.

To estimate the correlation between two sensitive variables, first we need to estimate the variances of the sensitive variables, and then we find the estimated variance, $\sigma_{S_i}^2$ which is given by

$$\hat{\sigma}_{S_i}^2 = \hat{\sigma}_{r_i}^2 - \sigma_Y^2.$$

If we use an augmented variable which has a wide range and so σ_Y^2 is close to $\sigma_{S_i}^2$, or σ_Y^2 is greater than $\sigma_{S_i}^2$, then we may not be able to estimate the

correlation between two sensitive variables since we observe a negative estimated variance for the sensitive variables. To illustrate this procedure, we simulated a bivariate gamma distribution (Mardia, 1970, Ong, 1992) with the true correlation set at 0.6, means (μ_{S_1}, μ_{S_2}) equal to (30, 37.037) and variances $(\sigma_{S_1}^2, \sigma_{S_2}^2)$ equal to (300, 370.37).

The results of the simulations are presented in table 9, 10, and 11. Table 9 gives the expected means and standard deviations of the sampling distribution for population correlation values of 0.6 and for a sample size of 100. The standard deviations in table 9 are strictly increasing as the variance of the augmented variable increases. We expected this result from Eq.(4.1.2). Table 10 presents the results for the estimated correlations, standard deviations and biases. As we explained early, it is possible to obtain a negative estimate of the variance, $\hat{\sigma}_{S_1}^2$. A correlation cannot be calculated when this occurs. The table also indicates when the negative variance occurs. We observed negative variances of the sensitive variable (S), for augmented distributions, Gam(2,14), Gam(3,11), Gam(4,10), Gam(5,8), Gam(6,7). Where the variances of augmented variables are greater than $\text{Min}(\sigma_{S_1}^2, \sigma_{S_2}^2)$, only Gam(6,7) has less variance than $\text{Min}(\sigma_{S_1}^2, \sigma_{S_2}^2)$. Table 11 gives the conditions for having a positive bias. Each parameter, σ_Y^2 , $\sigma_{S_1}^2$, and $\sigma_{S_2}^2$ is changed from 1 to 20 by increasing by one unit.

For $\sigma_Y^2 = 1$, and $\text{Min}(\sigma_{S_1}^2, \sigma_{S_2}^2) \leq 7$, we observed a positive bias with relatively high $\rho_{r_1 r_2}$. For $\sigma_Y^2 = 1$, the smallest $\rho_{r_1 r_2}$ to give a positive bias was 0.6. For $\sigma_Y^2 = 2$, and $2 \leq \text{Min}(\sigma_{S_1}^2, \sigma_{S_2}^2) \leq 7$, we observed a positive bias with relatively high $\rho_{r_1 r_2}$. For $\sigma_Y^2 = 1$, the smallest $\rho_{r_1 r_2}$ to give a positive bias was 0.7. As σ_Y^2 increase, we have less chance to have a positive bias.

TABLE 10
 ESTIMATED CORRELATIONS & BIASES FOR THE
 ADDITIVE RANDOMIZED RESPONSE MODELS

α	β	$\rho_{r_1 r_2}$	Bias	$\rho_{S_1 S_2}$	STD	negative variance S_1	S_2
2	2	0.60825	-0.00971	0.59853	0.08594		
2	3	0.61910	-0.02130	0.59780	0.08723		
2	4	0.63340	-0.03655	0.59684	0.08910		
2	5	0.65028	-0.05469	0.59558	0.09174		
2	6	0.66889	-0.07490	0.59392	0.09536		
2	7	0.68843	-0.09669	0.59174	0.10026		
2	8	0.70824	-0.11938	0.58886	0.10682		
2	9	0.72780	-0.14273	0.58506	0.11559		
2	10	0.74672	-0.16669	0.58002	0.12744		
2	11	0.76474	-0.19149	0.57324	0.14411		
2	12	0.78170	-0.21797	0.56373	0.17076		
2	13	0.79754	-0.25375	0.54378	0.33975		
2	14	0.81222	-0.27826	0.53424	0.27100	1,	1
3	2	0.61599	-0.01425	0.60174	0.08474		
3	3	0.63161	-0.03085	0.60076	0.08655		
3	4	0.65151	-0.05208	0.59942	0.08946		
3	5	0.67412	-0.07654	0.59757	0.09381		
3	6	0.69802	-0.10298	0.59504	0.10012		
3	7	0.68843	-0.09669	0.59174	0.10026		
3	8	0.74538	-0.15877	0.58661	0.12238		
3	9	0.76745	-0.18795	0.57950	0.14341		

TABLE 10 (Continue)

α	β	$\rho_{r_1 r_2}$	Bias	$\rho_{S_1 S_2}$	STD	negative variance S_1 S_2
3	10	0.78795	-0.21714	0.57095	0.16737	1, 0
3	11	0.80675	-0.25299	0.55390	0.24146	1, 0
4	2	0.62040	-0.01884	0.60156	0.08727	
4	3	0.64063	-0.04027	0.60036	0.08980	
4	4	0.66560	-0.06696	0.59864	0.09373	
4	5	0.69299	-0.09681	0.59618	0.09963	
4	6	0.72086	-0.12820	0.59266	0.10835	
4	7	0.74785	-0.16027	0.58757	0.12151	
4	8	0.77311	-0.19306	0.58004	0.14313	
4	9	0.79620	-0.22856	0.56763	0.19599	
4	10	0.81698	-0.26887	0.54821	0.352333	1, 0
5	2	0.62302	-0.02345	0.59957	0.08709	
5	3	0.64790	-0.04945	0.59844	0.09021	
5	4	0.67765	-0.08092	0.59673	0.09509	
5	5	0.70917	-0.11505	0.59411	0.10244	
5	6	0.74014	-0.15001	0.59012	0.11347	
5	7	0.76911	-0.18524	0.58387	0.13130	
5	8	0.79538	-0.21959	0.57584	0.15182	1, 0
6	2	0.62665	-0.02792	0.59873	0.08806	
6	3	0.65536	-0.05829	0.59706	0.09218	
6	4	0.68882	-0.09437	0.59444	0.09912	
6	5	0.72326	-0.13294	0.59032	0.11048	
6	6	0.75615	-0.17252	0.58363	0.13039	
6	7	0.78615	-0.21349	0.57272	0.18067	1, 0

TABLE 11

CONDITIONS FOR THE POSITIVE BIAS OF THE ADDITIVE
RANDOMIZED RESPONSE MODELS

σ_Y^2	$\sigma_{S_1}^2$	$\sigma_{S_2}^2$	$\rho_{r_1 r_2}$
1	1	4, 5	≥ 0.9
1	1	6, 7, 8	≥ 0.8
1	1	9, 10, 11, 12	≥ 0.7
1	1	13, ..., 20	≥ 0.6
1	2	6, 7, 8, 9	≥ 0.9
1	2	10, ..., 14	≥ 0.8
1	2	15, ..., 20	≥ 0.7
1	3	9, ..., 13	≥ 0.9
1	3	14, ..., 20	≥ 0.8
1	4	11, ..., 17	≥ 0.9
1	4	18, 19, 20	≥ 0.8
1	5	14, ..., 20	≥ 0.9
1	6	17, ..., 20	≥ 0.9
1	7	19, 20	≥ 0.9
2	2	7, 8, 9, 10	≥ 0.9
2	2	11, ..., 16	≥ 0.8
2	2	17, ..., 20	≥ 0.7
2	3	10, ..., 14	≥ 0.9
2	3	15, ..., 20	≥ 0.8
2	4	12, ..., 18	≥ 0.9
2	4	19, 20	≥ 0.8
2	5	15, ..., 20	≥ 0.9
2	6	17, ..., 20	≥ 0.9
2	7	20	≥ 0.9
3	3	10, ..., 16	≥ 0.9
		17, ..., 20	≥ 0.8
3	4	13, ..., 20	≥ 0.9
3	5	15, ..., 20	≥ 0.9
3	6	18, 19, 20	≥ 0.9
4	4	13, ..., 20	≥ 0.9
4	5	16, ..., 20	≥ 0.9
4	6	19, 20	≥ 0.9
5	5	17, ..., 20	≥ 0.9
5	6	19, 20	≥ 0.9
6	6	20	≥ 0.9

Using Eq.(4.1.2), and Table 9, we need to choose an augmented variable which has small variance, and hence we may have a short confidence interval for $\hat{\mu}_{S_i}$. Using Eq.(4.1.3), and Table 10, we also need an augmented variable which has small variance. However, an augmented variable which has a wide range of values will give more confidence to the respondent, particularly if a low value is highly sensitive, and so the concealing effect is high using a wide range augmented variable. To estimate unbiased correlations it is critical that the variance of the augmented variable be smaller than the variance for the sensitive variables, but the variances of the sensitive variables are unknown and hence it is difficult to choose a good augmented variable. Another problem is that extreme values cannot be protected by adding a random number. All these are disadvantages of the additive randomized response models.

4.2 Scrambled Randomized Response Models

Scrambled randomized response models can be used to obtain responses for sensitive questions when the answers are quantitative. The respondent is asked to multiply his(her) sensitive attribute (S) by a random value (Y). The random value is generated from a random device and is not known to the interviewer, but the distribution of the multiplier variable is completely known, and the multiplier variable is independent of the sensitive variables. We assumed that $S \geq 0$ and $Y > 0$, since the scrambled answer, r is SY . If $S = 0$ then $SY = 0$, and as long as $S = 0$ is a nonsensitive response, the fact that this answer is not protected by the randomization technique will not be a problem.

For the correlated two sensitive continuous variables case, the respondent is asked to generate a random number by a random device, and multiply his (her) own sensitive attribute for each question by that value, and then the response values for each question can be written

$$\begin{aligned} r_1 &= S_1 Y \\ r_2 &= S_2 Y \end{aligned} \tag{4.2.1}$$

where S_i and Y are independent.

And since $E r_i = E S_i E Y$, the unbiased estimate of μ_{S_i} is given by

$$\hat{\mu}_{S_i} = \frac{\hat{\mu}_{r_i}}{\mu_Y}, \tag{4.2.2}$$

where μ_Y is known and $\mu_Y \neq 0$.

The estimated variance of the response variable is

$$\begin{aligned} \text{Var}(r_i) &= E S_i^2 E Y^2 - (E S_i)^2 (E Y)^2, \text{ because } S_i \text{ and } Y \text{ are independent,} \\ &= \sigma_Y^2 (\sigma_{S_i}^2 + \mu_{S_i}^2) + \sigma_{S_i}^2 \mu_Y^2 = \sigma_{S_i}^2 (\sigma_Y^2 + \mu_Y^2) + \sigma_Y^2 \mu_{S_i}^2. \end{aligned} \tag{4.2.3}$$

From Eq.(4.2.3), the estimated variance of $\sigma_{S_i}^2$ is given by

$$\hat{\sigma}_{S_i}^2 = \frac{\hat{\sigma}_{r_1}^2 - \sigma_Y^2 \hat{\mu}_{S_i}^2}{\sigma_Y^2 + \mu_Y^2}, \quad (4.2.4)$$

where σ_Y^2 and μ_Y are known.

Now from Eq.(4.2.2) and Eq.(4.2.3), the estimated variance of $\hat{\mu}_{S_i}$ is

$$\text{Var}(\hat{\mu}_{S_i}) = \frac{1}{n \mu_Y^2} \text{Var}(r_i), \text{ because } \mu_Y \text{ is a known constant,}$$

$$= \frac{1}{n \mu_Y^2} [\sigma_Y^2 (\sigma_{S_i}^2 + \mu_{S_i}^2) + \sigma_{S_i}^2 \mu_Y^2]$$

$$= \frac{1}{n} [\frac{\sigma_Y^2}{\mu_Y^2} (\sigma_{S_i}^2 + \mu_{S_i}^2) + \sigma_{S_i}^2]$$

since $\sigma_{S_i}^2 + \mu_{S_i}^2 = E S_i^2$,

$$= \frac{1}{n} [\sigma_{S_i}^2 + \frac{\sigma_Y^2}{\mu_Y^2} E S_i^2]. \quad (4.2.5)$$

Since S_i and Y are independent, for fixed distributions of S_i , to reduce the variance of the estimated population mean, $\hat{\mu}_{S_i}$, we have to choose the multiplier variable

which makes the ratio, $\frac{\sigma_Y^2}{\mu_Y^2}$ as small as possible. Suppose the multiplier variable

follows a gamma distribution with parameters α and β , then the ratio, $\frac{\sigma_Y^2}{\mu_Y^2}$ is $\frac{1}{\alpha}$.

Therefore if we increase α , the variance of the estimated mean will be reduced.

Now from Eq.(4.2.1) and independent relationship between S_1 and Y , the covariance between two reported variables is

$$\text{Cov}(r_1, r_2) = \text{Cov}(S_1 Y, S_2 Y)$$

$$= E S_1 Y S_2 Y - E S_1 Y E S_2 Y.$$

since S_1 and Y are independent,

$$\sigma_{r_1 r_2} = E S_1 S_2 E Y^2 - E S_1 E S_2 (E Y)^2$$

$$= E S_1 S_2 E Y^2 - E S_1 E S_2 E Y^2 + E S_1 E S_2 E Y^2 - E S_1 E S_2 (E Y)^2$$

$$= \sigma_{S_1 S_2} E Y^2 + \mu_{S_1} \mu_{S_2} \sigma_Y^2$$

$$= \sigma_{S_1 S_2} (\sigma_Y^2 + \mu_Y^2) + \mu_{S_1} \mu_{S_2} \sigma_Y^2.$$

Therefore the estimated covariance between two sensitive variables is given by

$$\sigma_{S_1 S_2} = \frac{\sigma_{r_1 r_2} - \mu_{S_1} \mu_{S_2} \sigma_Y^2}{\sigma_Y^2 + \mu_Y^2}. \quad (4.2.6)$$

From Eq.(4.2.4) and (4.2.6), the correlation between two sensitive variables is given by

$$\rho_{S_1 S_2} = \frac{\sigma_{r_1 r_2} - \mu_{S_1} \mu_{S_2} \sigma_Y^2}{\sqrt{[\sigma_{r_1}^2 - \sigma_Y^2 \mu_{S_1}^2] [\sigma_{r_2}^2 - \sigma_Y^2 \mu_{S_2}^2]}}$$

Divide both the numerator and the denominator by $\sigma_{r_1} \sigma_{r_2}$

$$\rho_{S_1 S_2} = \frac{\rho_{r_1 r_2} + \frac{\mu_{S_1} \mu_{S_2} \sigma_Y^2}{\sigma_{r_1} \sigma_{r_2}}}{\sqrt{\left[1 - \frac{\sigma_Y^2 \mu_{S_1}^2}{\sigma_{r_1}^2}\right] \left[1 - \frac{\sigma_Y^2 \mu_{S_2}^2}{\sigma_{r_2}^2}\right]}} \quad (4.2.7)$$

By substituting Eq(4.2.3) and Eq(4.2.4) into Eq.(4.2.7), we may express $\rho_{S_1 S_2}$ as

a function of $\frac{\mu_Y^2}{\sigma_Y^2}$.

$$\rho_{S_1 S_2} = \frac{\rho_{r_1 r_2}}{\sqrt{\frac{\sigma_{S_1}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) \sigma_{S_2}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right)}{\left(\sigma_{S_1}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) + \mu_{S_1}^2\right) \left(\sigma_{S_2}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) + \mu_{S_2}^2\right)}} \frac{\mu_{S_1} \mu_{S_2}}{\sqrt{\left(\sigma_{S_1}^2 \left(1 + \frac{\sigma_Y^2}{\sigma_Y^2}\right) + \mu_{S_1}^2\right) \left(\sigma_{S_2}^2 \left(1 + \frac{\sigma_Y^2}{\sigma_Y^2}\right) + \mu_{S_2}^2\right)}} \sqrt{\frac{\sigma_{S_1}^2 \left(1 + \frac{\sigma_Y^2}{\sigma_Y^2}\right) \sigma_{S_2}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right)}{\left(\sigma_{S_1}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) + \mu_{S_1}^2\right) \left(\sigma_{S_2}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) + \mu_{S_2}^2\right)}} \quad (4.2.8)$$

Let X be the denominator of the first term,

$$X = \sqrt{\frac{\sigma_{S_1}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) \sigma_{S_2}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right)}{\left(\sigma_{S_1}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) + \mu_{S_1}^2\right) \left(\sigma_{S_2}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) + \mu_{S_2}^2\right)}}$$

Since $\frac{1}{X} = 1 + \frac{1 - X}{X}$

the first term of Eq.(4.2.8) can be written

$$\rho_{r_1 r_2} \left[1 + \frac{1 - X}{X} \right]$$

and the second term of Eq.(4.2.8) can be written

$$\frac{\mu_{S_1}^2 \mu_{S_2}^2}{\sqrt{\left(\sigma_{S_1}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) + \mu_{S_1}^2\right) \left(\sigma_{S_2}^2 \left(1 + \frac{\mu_Y^2}{\sigma_Y^2}\right) + \mu_{S_2}^2\right)}}$$

Now, let $\frac{\mu_{S_1}}{\sigma_{S_1}} = f_1$, $\frac{\mu_{S_2}}{\sigma_{S_2}} = f_2$, and the multiplier variable follows a gamma

distribution with parameters α and β , hence $\frac{\mu_Y^2}{\sigma_Y^2} = \alpha$.

Then by simple algebra

$$\rho_{S_1 S_2} = \rho_{r_1 r_2} + \rho_{r_1 r_2} \left[\sqrt{\left[1 + \frac{f_1^2}{(1 + \alpha)} \right] \left[1 + \frac{f_2^2}{(1 + \alpha)} \right]} - 1 \right] - \frac{f_1 f_2}{(1 + \alpha)}$$

where the last two terms are due to the random device. Hence if we estimate the correlation from the observed response data, we may have some bias which is given by

$$\text{Bias} = \rho_{r_1 r_2} \left[\sqrt{\left[1 + \frac{f_1^2}{(1 + \alpha)}\right] \left[1 + \frac{f_2^2}{(1 + \alpha)}\right]} - 1 \right] - \frac{f_1 f_2}{(1 + \alpha)}. \quad (4.2.9)$$

Since $\rho_{r_1 r_2}$, f_1 , and f_2 can be estimated from the observed data and α is known, we may estimate the bias and hence the estimated correlation between two sensitive variables, which is $\hat{\rho}_{r_1 r_2} + \hat{\text{Bias}}$. If we observe positive (negative) bias, the estimated correlation, $\rho_{r_1 r_2}$ from the observed data will over (under) estimate the correlation $\rho_{S_1 S_2}$ between the two sensitive variables. For fixed f_1 and f_2 , if we

increase α , the bias decreases, since for fixed f_1 and f_2 , $\frac{f_1 f_2}{(1 + \alpha)}$ decreases as α

increases and $\sqrt{\left[1 + \frac{f_1^2}{(1 + \alpha)}\right] \left[1 + \frac{f_2^2}{(1 + \alpha)}\right]} - 1$ also decreases as α increases. Therefore if the shape parameter, α , of the multiplier variable is large enough, the bias can be zero. This is shown in table 12 and 13.

To illustrate this procedure, we simulated a bivariate gamma distribution, the true correlation was set at 0.6, with the means (μ_{S_1}, μ_{S_2}) equal to (30, 37.037), and the variances $(\sigma_{S_1}^2, \sigma_{S_2}^2)$ equal to (300, 370.37). For the multiplier variable, we use a gamma distribution with various parameters.

The results of the simulations are presented in table 12, 13, 14, and 15.

Table 14 gives the expected means and standard deviations of the sampling distributions for population correlation value of 0.6 for sample size 100. The standard deviation values in table 14 decrease slowly as the α increases, as we expected from Eq.(4.2.4). Table 12 and 13 presents the results for estimated

TABLE 12

ESTIMATED CORRELATIONS & BIASES FOR THE
SCRAMBLED RANDOMIZED RESPONSE MODELS

α	$\rho_{r_1 r_2}$	Bias	$\rho_{S_1 S_2}$	Std($\rho_{S_1 S_2}$)
5	0.74195	-0.14500	0.59690	0.11630
6	0.72667	-0.13178	0.59488	0.11293
7	0.71722	-0.11936	0.59785	0.10998
8	0.70842	-0.10914	0.59927	0.10922
9	0.70051	-0.10097	0.59954	0.10243
10	0.69303	-0.09401	0.59901	0.10472
15	0.66736	-0.07027	0.59708	0.09897
20	0.65284	-0.05590	0.59694	0.09590
25	0.64179	-0.04660	0.59519	0.09313
30	0.63378	-0.04015	0.59362	0.09354
35	0.62991	-0.03491	0.59499	0.09128
40	0.62677	-0.03088	0.59589	0.09115
45	0.62301	-0.02781	0.59519	0.09159
50	0.61948	-0.02536	0.59412	0.09275
55	0.61800	-0.02318	0.59482	0.09104
60	0.61585	-0.02143	0.59441	0.09080
80	0.61739	-0.01597	0.60141	0.08682
100	0.61018	-0.01310	0.59707	0.08584
200	0.60351	-0.00669	0.59681	0.08504
300	0.60169	-0.00452	0.59717	0.08599
400	0.60028	-0.00340	0.59688	0.08282
500	0.60267	-0.00270	0.59997	0.08862

$n = 100, \rho_{S_1 S_2} = 0.6.$

TABLE 13

ESTIMATED CORRELATIONS & BIASES FOR THE
SCRAMBLED RANDOMIZED RESPONSE MODELS

α	$\rho_{r_1 r_2}$	Bias	$\rho_{S_1 S_2}$	Std($\rho_{S_1 S_2}$)
5	0.74971	-0.15405	0.59566	0.10048
6	0.73530	-0.13729	0.59801	0.09593
7	0.71917	-0.12127	0.59790	0.08667
8	0.70494	-0.11325	0.59168	0.08645
9	0.69695	-0.10496	0.59198	0.08392
10	0.69214	-0.09636	0.59578	0.08212
15	0.66813	-0.07031	0.59782	0.07440
20	0.65200	-0.05614	0.59585	0.07084
25	0.64598	-0.04587	0.60010	0.06832
30	0.63817	-0.03952	0.59864	0.06694
35	0.62932	-0.03496	0.59436	0.06723
40	0.62906	-0.03044	0.59886	0.06358
45	0.62484	-0.02748	0.59735	0.06659
50	0.62404	-0.02483	0.59720	0.06318
55	0.62027	-0.02277	0.59750	0.06346
60	0.61891	-0.02102	0.59789	0.06459
80	0.61221	-0.01615	0.59606	0.05941
100	0.61151	-0.01289	0.59862	0.05955
200	0.60656	-0.00660	0.59996	0.06175
300	0.60407	-0.00439	0.59967	0.05647
400	0.59966	-0.00336	0.59630	0.06132
500	0.60191	-0.00267	0.59923	0.05757

$n = 200, \rho_{S_1 S_2} = 0.6.$

TABLE 14

ESTIMATED MEANS & STANDARD DEVIATIONS FOR THE
SCRAMBLED RANDOMIZED RESPONSE MODELS

α	μ_{S_1}	$\text{Std}(\mu_{S_1})$	μ_{S_2}	$\text{Std}(\mu_{S_2})$
5	29.98168	2.32671	37.05844	2.66994
6	30.02349	2.19366	37.06843	2.58392
7	29.99817	2.17698	36.97602	2.48384
8	29.93463	2.15973	36.93093	2.44235
9	29.97757	2.01765	37.02887	2.31318
10	29.94107	2.10216	36.95828	2.39717
11	29.97544	2.02668	37.01020	2.25791
12	29.99186	2.02930	37.02292	2.25821
15	29.99598	2.02910	37.02791	2.20981
20	29.98787	1.87565	36.98916	2.12555
25	29.96865	1.90332	36.96799	2.08319
30	29.95429	1.90670	36.96342	2.06246
35	29.94810	1.87072	36.95879	2.05674
40	29.94579	1.86662	36.95485	2.05599
45	29.94225	1.82302	36.93484	2.04361
50	29.90725	1.80201	36.91562	2.01575
55	29.91269	1.77219	36.92020	1.97135
60	29.91406	1.76804	36.91732	1.98227
80	30.04758	1.76785	37.08747	2.02706
100	29.98567	1.74211	36.99157	1.94889
200	29.96098	1.76597	37.03351	1.97908
300	29.95310	1.74834	37.03675	1.91635

$$n = 100, \rho_{S_1 S_2} = 0.6, \mu_{S_1} = 30, \mu_{S_2} = 37.0373$$

TABLE 15
 CONDITIONS FOR THE POSITIVE BIAS OF THE
 SCRAMBLED RANDOMIZED RESPONSE MODELS

f_1	f_2	$f_1 f_2$	α	$\rho_{r_1 r_2}$
1	2	2	1,.....,20	≥ 0.9
1	3	3	2,3,4,5	≥ 0.8
			6,.....,20	≥ 0.7
1	4	4	1	≥ 0.8
			2,3,4,5,6	≥ 0.7
			7,.....,20	≥ 0.6
1	5	5	2,3	≥ 0.7
			4,.....,11	≥ 0.6
			12,.....,20	≥ 0.5
1	6	6	1,2	≥ 0.7
			3,4,5,6,7	≥ 0.6
			8,.....,20	≥ 0.5
1	7	7	1	≥ 0.7
			3,4,5	≥ 0.6
			6,.....,17	≥ 0.5
			18,19,20	≥ 0.4
1	8	8	1	≥ 0.7
			2,3,4	≥ 0.6
			5,.....,12	≥ 0.5
			13,.....,20	≥ 0.4
1	9	9	2,3,4	≥ 0.6
			5,.....,10	≥ 0.5
			11,.....,20	≥ 0.4
1	10	10	2,3	≥ 0.6
			4,5,6,7,8,9	≥ 0.5
			10,.....,20	≥ 0.4
1	11	11	2,3	≥ 0.6
			4,5,6,7,8	≥ 0.5
			9,.....,20	≥ 0.4
1	12	12	2	≥ 0.6
			4,5,6,7	≥ 0.5
			8,.....,20	≥ 0.4
1	13	13	2	≥ 0.6
			4,5,6,7	≥ 0.5
			8,.....,20	≥ 0.4
1	14	14	2	≥ 0.6
			4,5,6,7	≥ 0.5
			8,.....,18	≥ 0.4
			19,20	≥ 0.3
1	15	15	4,5,6	≥ 0.5
			7,.....,17	≥ 0.4
			18,19,20	≥ 0.3
1	16	16	3,4,5,6	≥ 0.5
			7,.....,16	≥ 0.4
			17,18,19,20	≥ 0.3

TABLE 15 (Continued)

f_1	f_2	$f_1 f_2$	α	$\rho_{r_1 r_2}$
1	17	17	3,4,5	≥ 0.5
			6,.....,15	≥ 0.4
			16,17,18,19,20	≥ 0.3
1	18	18	3,4	≥ 0.5
			5,.....,14	≥ 0.4
			15,.....,20	≥ 0.3
1	19	19	3,4	≥ 0.5
			7,.....,14	≥ 0.4
			15,.....,20	≥ 0.3
1	20	20	3,4	≥ 0.5
			6,.....,14	≥ 0.4
			15,.....,20	≥ 0.3
2	4	8	6,.....,20	≥ 0.9
2	5	10	3,.....,13	≥ 0.9
			14,.....,20	≥ 0.8
2	6	12	2	≥ 0.9
2	7	14	8,.....,20	≥ 0.8
			6,.....,14	≥ 0.8
2	8	16	15,.....,20	≥ 0.7
			1	≥ 0.9
2	9	18	5,.....,10	≥ 0.8
			11,.....,20	≥ 0.7
			4,5,6	≥ 0.8
2	10	20	10,.....,20	≥ 0.7
			4	≥ 0.8
2	11	22	8,.....,18	≥ 0.7
			19,20	≥ 0.6
			3,4	≥ 0.8
2	12	24	8,.....,15	≥ 0.7
			16,.....,20	≥ 0.6
			3	≥ 0.8
2	13	26	7,.....,14	≥ 0.7
			15,.....,20	≥ 0.6
			3	≥ 0.8
2	14	28	7,.....,12	≥ 0.7
			13,.....,20	≥ 0.6
			3	≥ 0.8
2	15	30	6,7,8,9	≥ 0.7
			13,.....,20	≥ 0.6
			6,7,8	≥ 0.7
2	16	32	12,.....,20	≥ 0.6
			6,7	≥ 0.7
2	17	34	11,.....,20	≥ 0.6
			6,7	≥ 0.7
2	18	36	11,.....,20	≥ 0.6
			5,6	≥ 0.7
2	18	36	11,.....,20	≥ 0.6

TABLE 15 (Continued)

f_1	f_2	$f_1 f_2$	α	$\rho_{r_1 r_2}$
2	19	38	5,6	≥ 0.7
			10,.....,19	≥ 0.6
			20	≥ 0.5
2	20	40	2	≥ 0.8
			5,6	≥ 0.7
			10,.....,18	≥ 0.6
			19,20	≥ 0.5
3	6	18	15,.....,20	≥ 0.9
3	7	21	9,.....,20	≥ 0.9
3	8	24	6,.....,11	≥ 0.9
3	9	27	5,6	≥ 0.9
			18,19,20	≥ 0.8
			4,5	≥ 0.9
3	10	30	14,.....,20	≥ 0.8
			4	≥ 0.9
3	11	33	12,.....,20	≥ 0.8
			11,.....,20	≥ 0.8
3	12	36	11,.....,20	≥ 0.8
3	13	39	3	≥ 0.9
			10,.....,17	≥ 0.8
			3	≥ 0.9
3	14	42	9,.....,14	≥ 0.8
			9,10,11,12	≥ 0.8
3	15	45	19,20	≥ 0.7
			8,9,10	≥ 0.8
3	16	48	18,19,20	≥ 0.7
			8,9,10	≥ 0.8
3	17	51	17,18,19,20	≥ 0.7
			8,9	≥ 0.8
3	18	54	16,17,18,19,20	≥ 0.7
			8	≥ 0.8
3	19	57	16,17,18,19,20	≥ 0.7
			7,8	≥ 0.8
3	20	60	15,.....,20	≥ 0.7
			17,18,19,20	≥ 0.9
4	9	36	13,.....,20	≥ 0.9
4	10	40	11,.....,18	≥ 0.9
4	11	44	10,11,12,13	≥ 0.9
4	12	48	9,10	≥ 0.9
4	13	52	8,9	≥ 0.9
4	14	56	7,8	≥ 0.9
4	15	60	7,20	≥ 0.9
4	16	64	7	≥ 0.9
4	17	68	19,20	≥ 0.8

TABLE 15 (Continued)

f_1	f_2	$f_1 f_2$	α	$\rho_{r_1 r_2}$
4	18	72	6	≥ 0.9
			18, 19, 20	≥ 0.8
4	19	76	6	≥ 0.9
			17, 18, 19, 20	≥ 0.8
4	20	80	16,, 20	≥ 0.8
5	13	65	19, 20	≥ 0.9
5	14	70	17, 18, 19, 20	≥ 0.9
5	15	75	15,, 20	≥ 0.9
5	16	80	14,, 18	≥ 0.9
5	17	85	13, 14, 15	≥ 0.9
5	18	90	12, 13, 14	≥ 0.9
5	19	95	12, 13	≥ 0.9
5	20	100	11, 12	≥ 0.9
6	20	120	19, 20	≥ 0.9

correlation, standard deviations, and biases. The amount of bias decreases as α increases. The standard deviations for the estimated correlation $\rho_{S_1 S_2}$ also decreases as α increases. It was expected that as values of α increase, the observed correlation $\rho_{r_1 r_2}$ will be close to the correlation between the two sensitive variables, and hence the bias will be close to zero. For $\alpha = 500$, $\rho_{r_1 r_2}$ was 0.60067, the bias was -0.0027 and hence $\hat{\rho}_{S_1 S_2}$ was 0.59997. It is close enough to the true correlation value $\rho_{S_1 S_2} = 0.6$. Table 15 gives the conditions for getting a positive bias in terms of f_1 , f_2 , and $\rho_{r_1 r_2}$. Values were set from 1 to 20 and $\rho_{r_1 r_2}$ was increased by 0.1. For fixed f_1 and f_2 , conditions which give positive bias depend on α and the observed $\rho_{r_1 r_2}$. When $f_1 = 1$ or 2 if we use large value of α for the low values of the observed $\rho_{r_1 r_2}$ we can get a positive bias. When $f_1 = 3$ and for various values of f_2 and α , to get a positive bias the minimum observed $\rho_{r_1 r_2}$ was 0.7. When $f_1 = 4$ and for various values of f_2 and α , to get a positive bias, the minimum observed $\rho_{r_1 r_2}$ was 0.8. When $f_1 = 4$ or 5 or 6 and for various values of f_2 and α , to get a positive bias, the minimum observed $\rho_{r_1 r_2}$ was 0.9. When $f_1 \geq 6$ for any combinations of f_2 and α , we never observed a positive bias.

A major field problem in conducting a survey using the scrambled randomized response technique is how does the interviewer furnish a random device which can generate the multiplier value. Eichhorn and Hayre (1983) discussed this problem. Here we may propose a simple and familiar method.

After generating multiplier values, we may write the values on a card and put those cards into an urn and ask the respondent to pick one card from the urn and multiply his (her) own S values and report product, SY to the interviewer.

LITERATURE CITED

- Abul-Ela, A. A., Greenberg, B. G., & Horvitz, D. G. (1967).
A Multiproportions Randomized Response Model. JASA, 62, 990–1008.
- Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis. 2nd ed., John Wiley & Sons, New York.
- Bishop, Y. M., Fienberg, S. E. & Holland, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice. The MIT Press, Cambridge.
- Cochran, W. G. (1977). Sampling Techniques. 3rd ed. John Wiley & Sons, New York
- Deming, W. E. (1960). Sample Design in Business Research. John Wiley & Sons, New York.
- Edgell, S. E., Himmelfarb, S. & Cira, D. J. (1986). Techniques to Estimate Correlation. Psychological Bulletin. 100(2), 251–256.
- Eichhorn, B. H., & Hayre, L. S. (1983). Scrambled Randomized Methods for Obtaining Sensitive Quantitative Data. Journal of Statistical Planning & Inference, 7, 307–316.
- Fishman, G. S. (1978). Principles of Discrete Event Simulation. John Wiley & Sons, New York.
- Fox, J. A. & Tracy, P. E. (1984), Measuring Associations with Randomized Response. Social Science Research, 13, 188–197.

- Gould, A. L., Shah, B. V. & Abernathy, J. R. (1969). Unrelated Question Randomized Response Techniques with two Trials per Respondent. Proceedings of Social Statistics Section, American Statistical Association, 351–359.
- Greenberg, B. G., Abul-Ela, A. A., Simmons, W. R., & Horvitz, D. G. (1969). The Unrelated Question Randomized Response Model Theoretical Framework. JASA, 64, 520–539.
- Hamdan, M. A. & Martinson, E. O. (1971). Maximum Likelihood Estimation in The Bivariate Binomial (0,1) Distribution: Application to 2 x 2 tables. Austral. J. Statist., 13 154–158.
- Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1951). Sample Survey Methods and Theory. Vol 1. John Wiley & Sons, New York.
- Hatchett, S. & Schuman, H. (1975). White Respondents and Race-of-Interviewer Effect. Public Opinion Quarterly, 39 523–528.
- Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The Unrelated Question Randomized Response Model. Proceedings of Social Statistics Section, American Statistical Association, 65–72.
- Johnson, R. A. & Wichern, D. W. (1982). Applied Multivariate Statistical Analysis. Prentice Hall, Englewood Cliffs.
- Kraemer, H. C. (1980). Estimation and Testing of Bivariate Association Using Data Generated by the Randomized Response Technique. Psychological Bulletin, 87, 304–308.
- Kim, Jong Ik. & Flueck, J. A. (1978). An Additive Randomized Response Model. Proceedings of Survey Research Methods Section, American Statistical Association, 351–355.

- Kocherlakota, S. & Kocherlakota, K. (1992). Bivariate Discrete Distributions. Marcel Dekker, Inc., New York.
- Liu, P. T. & Chow, L. P. (1976 a). A New Discrete Quantitative Randomized Response Model. JASA, 71, 72–73.
- Liu, P. T. & Chow, L. P. (1976 b). The Efficiency of The Multiple Trial Randomized Response Technique. Biometrics, 32, 607–618.
- Mardia, K. V. (1970). Families of Bivariate Distributions. Hafner Publishing Company, Darien.
- Mood, A. M., Graybill, F. A. & Boes, D, C (1974). Introduction to The Theory of Statistics. 3rd ed., McGraw–Hill, London.
- Moors, J. J. (1971) Optimization of The Unrelated Question Randomized Response Model. JASA, 66, 627–629.
- Muirhead, R. J. (1982). Aspects of Multivariate Statistical Theory. John Wiley & Sons, New York.
- Ong, S. H. (1992). The Computer Generation of Bivariate Gamma Variates. Commun. Stat. Simul., 21 285–299.
- Pollock, K. H. & Bek, Y.(1976). A Comparison of Three Randomized Response Model for Quantitative Data. JASA, 71, 884–886.
- Rao, C. R. (1973). Linear Statistical Inference and Its Applications. 2nd ed. John Wiley & Sons, New York.
- SAS Institute Inc. (1985). SAS User's Guide : Basics, Version 5 ed., Cary.
- Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics, John Wiley & Sons, New York.

- Sudman, S. & Bradburn, N. (1983). Asking Questions: A practical Guide to Questionnaire Design. Jossey-Bass Publishers, San Francisco.
- Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. & Asok, C. (1984). Sampling Theory of Surveys with Applications. 3rd ed., Iowa State University Press, Ames.
- Warde, W. D. (1991) Sampling Method Class Lecture Note. Oklahoma State University
- Warner, S. L. (1965). Randomized Response : A Survey Technique for Eliminating Evasive Answer Bias. JASA, 60, 63-69.

VITA ²

Geun-Shik Han

Candidate for the Degree of

Doctor of Philosophy

Thesis: CORRELATION ANALYSIS FOR THE RANDOMIZED
RESPONSE MODELS

Major Field: Statistics

Biographical:

Personal Data: Born in Chung-Ju, Korea, Oct 22, 1957.

Education: Graduated from Chung-Ju High School, Chung-Ju, Korea, in February, 1976; Received a Bachelor of Science Degree with a Major in Statistics from Korea University, Seoul, Korea, in August 1984; Received the Master of Science Degree with a Major in Statistics from Iowa State University in May 1988. Completed Requirements for a Doctor of Philosophy Degree in Statistics of Oklahoma State University in May, 1993.

Professional Experience: Working for the Korean Reinsurance Company, Seoul, Korea, August, 1984, to July, 1985. Teaching Assistant (Business Statistics, Engineering Statistics), Department of Statistics, Oklahoma State University, August, 1991, 1992.

Scholarly Organization: Member of Mu Sigma Rho.