1. Introduction

Consider a population of size N which can be divided into three subpopulations of size N₁, N₂, and N₃, where N₁ + N₂ + N₃ = N. Suppose that we are interested in estimating N₁/(N₁ + N₂), as we have no direct interest in the third group and only secondary interest in N₁/N and (N₁ + N₂)/N. Procedures are well defined in Cochran [1963] for point and interval estimation of ratios. What is proposed here is methods of accomplishing these goals in situations requiring use of a randomized response technique to eliminate untruthful responses to questions which might cause embarrassment or attach stigma.

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

Set A

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

Set B

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

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

We shall define the following notation:

\[ \Pi_1 = \text{probability of answering "yes" to question 1 set A.} \]
\[ \Pi_2 = \text{probability of answering "yes" to question 2 set A given that question 1 set A was answered "yes".} \]
\[ \theta_1 = \text{probability of answering "yes" to question 1 set B.} \]
\[ \theta_2 = \text{probability of answering "yes" to question 2 set B given that question 1 set B was answered "yes".} \]
\[ n_1 = \text{number of people in a sample of size n who answer "yes, yes".} \]
\[ n_2 = \text{number of people in a sample of size n who answer "yes, no".} \]
\[ \Lambda_1 = \text{probability of answering "yes" to question 1 (either set).} \]
\[ \Lambda_2 = \text{probability of answering "yes" to question 2 (either set).} \]
\[ \lambda_1 = (n_1 + n_2)/n. \]
\[ \lambda_2 = n_1/n. \]

We thus have the following equalities:

\[ \Lambda_1 = p\Pi_1 + (1-p)\theta_1, \]
\[ \Lambda_2 = p\Pi_2 + (1-p)\theta_2. \]

From (1) we have

\[ \Pi_1 = \frac{\Lambda_1 - (1-p)\theta_1}{p}, \]

and from (2) we have

\[ \Pi_2 = \frac{\Lambda_2 - (1-p)\theta_1\theta_2}{p\Pi_1}. \]

Thus,

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

2. Estimation

Since n₁ + n₂ is distributed binomially \((n,A₁)\), similar arguments are used to show that n₁ given n₁ + n₂ is distributed binomially \((n₁ + n₂, A₁/A₁)\). Also, since the joint probability density of n₁ and n₁ + n₂ is the product of the density of n₁ + n₂ and of n₁ given n₁ + n₂, we have the joint likelihood function:

\[ L = \lambda_1^{n_1} \lambda_2^{n_2} (1-\lambda_1)^{n_1-n_2} \lambda_1^{n_1} (1-\lambda_2)^{n_2-n_2} \]

Solving for \(\Lambda_1\) and \(\Lambda_2\) we have:

\[ \tilde{\Lambda}_1 = \frac{n_1}{n_1 + n_2} \]

and

\[ \tilde{\Lambda}_2 = \frac{n_1}{n} \]

as our maximum likelihood estimators. But since

\[ \Lambda_1 = p\Pi_1 + (1-p)\theta_1 \]

and

\[ \Lambda_2 = p\Pi_2 + (1-p)\theta_1\theta_2, \]

then

\[ \tilde{\Pi}_1 = \left( \frac{n_1 + n_2}{n} \right) \theta_1 \]

and

\[ \tilde{\Pi}_2 = \left( \frac{n_1}{n} \right) \theta_2 \]

are maximum likelihood estimators for \(\Pi_1\) and \(\Pi_2\), respectively, by the invariance property.

It may be easily shown that \(\tilde{\Pi}_2\) is not unbiased.

3. Accuracy of \(\widehat{\Pi}_2\)

Recalling the literature of ratio estimation we recognize \(\widehat{\Pi}_2\) as a ratio estimate. The approximate variance of \(\widehat{\Pi}_2\), then, can be arrived at by
direct application of procedures well defined in the literature. There are various equivalent expressions for Var \( \Pi_2 / n \). Among the simplest is

\[
\text{Var}(\Pi_2 / n) = \frac{[\Pi_2 - (1-p)\theta_1 \theta_2]^2}{[\Pi_2 - (1-p)\theta_1 \theta_2]^2} \cdot \left[ \frac{\Pi_2 - (1-p)\theta_1 \theta_2}{n[\Pi_2 - (1-p)\theta_1 \theta_2]^2} \right]^2 + \frac{2\Pi_2 - (1-p)\theta_1 \theta_2}{n[\Pi_2 - (1-p)\theta_1 \theta_2]^2}.
\]

This is equivalent to

\[
\Pi_2^2(C_{\theta_1} \lambda_2 + C_{\theta_1} \lambda_1 - 2C_{\theta_1} \lambda_2),
\]

where

\[
C_{\theta_1} \lambda_2 = \frac{\Pi_2^2}{n[\Pi_2 - (1-p)\theta_1 \theta_2]^2}
\]

and

\[
C_{\theta_1} \lambda_1 = \frac{\Pi_2^2}{n[\Pi_2 - (1-p)\theta_1 \theta_2]^2}
\]

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

\[
\Pi_2^2(1-\lambda_1) = \frac{\Pi_2^2}{n[\Pi_2 - (1-p)\theta_1 \theta_2]^2}[\Pi_2 - (1-p)\theta_1 \theta_2]
\]

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

\[
E(\hat{\Pi}_2 - \Pi_2)^2 = (\Pi_2^2)(C_{\theta_1} \lambda_1 + C_{\theta_1} \lambda_2 - 2C_{\theta_1} \lambda_2)(1 + 3C_{\lambda_1} \lambda_1)
\]

\[+ \frac{6C_{\lambda_1} \lambda_1^2 (C_{\theta_1} \lambda_2^2 + C_{\theta_1} \lambda_1 - 2C_{\theta_1} \lambda_2)}{n},\]

\[= (\Pi_2^2)(C_{\theta_1} \lambda_1 + C_{\theta_1} \lambda_2 - 2C_{\theta_1} \lambda_2)(1 + 3C_{\lambda_1} \lambda_1)
\]

\[+ 6C_{\lambda_1} \lambda_1^2 (C_{\theta_1} \lambda_2^2 + C_{\theta_1} \lambda_1 - 2C_{\theta_1} \lambda_2).
\]

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

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

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

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

Since

\[
C_{\lambda_1} \lambda_1 = \frac{(1-\lambda_1)\lambda_1}{n[\Pi_2 - (1-p)\theta_1 \theta_2]^2},
\]

this is equivalent to

\[
\frac{100\Pi_2 - (1-p)\theta_1 \theta_2}{(\Pi_2^2)^2}.
\]

From this expression it is easy to see that we need to keep \( \Pi_2 / n \) as large as possible to keep the sample size small. In order to get a better feeling for what we mean by "large" and "small", the following tables of \( n \) for selected values of \( \Pi_1 \) are presented:

<table>
<thead>
<tr>
<th>( \Pi_1 = .5 )</th>
<th>.9</th>
<th>.7</th>
<th>.5</th>
<th>.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_1 )</td>
<td>708</td>
<td>811</td>
<td>1012</td>
<td>1212</td>
</tr>
<tr>
<td>( \Pi_1 = .3 )</td>
<td>.9</td>
<td>.7</td>
<td>.5</td>
<td>.3</td>
</tr>
<tr>
<td>( \theta_1 )</td>
<td>588</td>
<td>682</td>
<td>782</td>
<td>882</td>
</tr>
<tr>
<td>( \Pi_1 = .1 )</td>
<td>.9</td>
<td>.7</td>
<td>.5</td>
<td>.3</td>
</tr>
<tr>
<td>( \theta_1 )</td>
<td>482</td>
<td>582</td>
<td>682</td>
<td>782</td>
</tr>
<tr>
<td>( \Pi_1 = .05 )</td>
<td>.9</td>
<td>.7</td>
<td>.5</td>
<td>.3</td>
</tr>
<tr>
<td>( \theta_1 )</td>
<td>482</td>
<td>582</td>
<td>682</td>
<td>782</td>
</tr>
</tbody>
</table>

To see exactly in terms of variances how these values perform, we selectively choose values for \( \theta_2 \) and \( \theta_3 \) also. To reduce the effect of these parameters we choose \( \Pi_2 = \theta_3 \) and arbitrarily...
choose \( p_0 = .5 \). The variances for the \( n \)-values in boxes above are:

\[
\begin{array}{ccccccc}
\Pi = .5 & p = .7 & \Pi = .5 & Var \tilde{\theta}_2 = .0030 & Var \tilde{\theta}_2 = .0030 \\
\Pi = .1 & .3 & .1 & .0032 & .0036 \\
\Pi = .05 & .0031 & .0034 \\
\end{array}
\]

From these tables it can be concluded that to keep the MSE low, we need to have \( p \) as large as possible and \( \Pi \) as small as possible without compromising the anonymity of the respondents. It is also clear that for fixed \( \Pi \), \( \theta_i \), \( p \), if \( \theta_0 > \Pi > \frac{1}{2} \) or if \( \theta_0 < \Pi < \frac{1}{2} \), the MSE will be lower than for the case \( \Pi = \Pi_0 \). For \( \Pi = \frac{1}{2} \), however, all \( \Pi_0 \neq \Pi_0 \) result in smaller MSE, for the case \( \Pi = .5 \) and larger MSE for \( \Pi = .1 \).

\[
\begin{array}{ccccccc}
\Pi = .5 & .1 & .3 & .5 & .7 & .9 \\
p = .7 & .00180 & .00273 & .00358 & .00436 & .00507 \\
\Pi = .1 & .3 & .5 & .7 & .9 \\
\Pi = .05 & .00369 & .00420 & .00465 & .00502 & .00532 \\
\Pi = .5 & .00486 & .00497 & .00500 & .00497 & .00486 \\
\Pi = .7 & .00532 & .00502 & .00465 & .00420 & .00369 \\
\Pi = .9 & .00507 & .00436 & .00358 & .00273 & .00180 \\
\end{array}
\]

\[
\theta_2
\]

\[
\begin{array}{ccccccc}
\Pi = .5 & .1 & .3 & .5 & .7 & .9 \\
p = .7 & .00115 & .00250 & .00374 & .00488 & .00694 \\
\Pi = .1 & .3 & .5 & .7 & .9 \\
\Pi = .05 & .00299 & .00315 & .00320 & .00315 & .00299 \\
\Pi = .5 & .00432 & .00388 & .00334 & .00269 & .00194 \\
\Pi = .7 & .00694 & .00436 & .00374 & .00273 & .00115 \\
\Pi = .9 & .00694 & .00436 & .00374 & .00273 & .00115 \\
\end{array}
\]

From these tables it can be concluded that to keep the MSE low, we need to have \( p \) as large as possible and \( \Pi \) as small as possible without compromising the anonymity of the respondents. It is also clear that for fixed \( \Pi \), \( \theta_i \), \( p \), if \( \theta_0 > \Pi > \frac{1}{2} \) or if \( \theta_0 < \Pi < \frac{1}{2} \), the MSE will be lower than for the case \( \Pi = \Pi_0 \). For \( \Pi = \frac{1}{2} \), however, all \( \Pi_0 \neq \Pi_0 \) result in smaller MSE, for the case \( \Pi = .5 \) and larger MSE for \( \Pi = .1 \).

However, \( \Pi_0 \) is never known, and if we assume that it can achieve any value between 0 and 1, then we must look at each column of these tables for the maximal MSE. A quick inspection leads to the conclusion that for every value of \( \theta_0 \neq \frac{1}{2} \), the maximum MSE is greater than the maximum MSE for \( \theta_0 = \frac{1}{2} \). We have numerically derived then, a minimax rule for choice of \( \theta_0 \): always choose \( \theta_0 = \frac{1}{2} \). If we use different assumptions about the range of \( \Pi_0 \), for example \( \Pi_0 > \frac{1}{2} \) or \( \Pi_0 < \frac{1}{2} \), we would use \( \theta_0 > \frac{1}{2} \) and \( \theta_0 < \frac{1}{2} \), respectively, as minimax rules.

\[\text{4. Extensions}\]

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

\[
\hat{\theta}_k = \frac{\lambda_k - (1-p)\theta_1 \cdots \theta_k}{\lambda_{k-1} - (1-p)\theta_1 \cdots \theta_{k-1}}
\]

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

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

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

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

\[\text{5. Summary}\]

Randomized response techniques can be used in a census of human populations, for obtaining information on a sensitive characteristic. In surveys of human populations, it might be of interest to measure the proportion of individuals belonging to group \( A \), the members of which are associated with a characteristic that is stigmatized in the opinion of the population in general. Hence a member of such a group might suffer embarrassment in conceding explicitly his association with the group. The randomized response technique is devised to mask the respondent's answer so that he can feel assured that his anonymity as to the response is preserved.

In certain surveys it might be of interest to obtain an estimate of the membership of a subgroup of \( A \). In this case, the following procedure, which is called randomized conditional response model, can be applied.

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

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

\[\text{6. Bibliography}\]


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