Ratio estimation in randomized response designs
by Reider Sverre Peterson

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY in Mathematics
Montana State University

Abstract:
In this work, estimation of a ratio of sensitive characteristics using Warner's randomized response type of design is investigated. Estimators for the mean of the ratios and its mean squared error is obtained. An unbiased Hartley-Boss type of ratio estimate is also found along with an unbiased estimate of the variance of this estimator. The asymptotic distribution of the estimator for the ratio of means is also obtained. A method of setting confidence intervals for the ratio of means for the normal case, which is an application of Fieller's Theorem, is obtained. The usually quite-robust method of setting confidence intervals using the Jackknife procedure is also given. A Monte-Carlo study was done to investigate the properties of the various estimators for normal populations and for Chi-Squared populations.

RATIO ESTIMATION IN RANDOMIZED RESPONSE DESIGNS

by

REIDER SVERRE PETERSON

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematics

Approved:

Robert D. Engle
Head, Major Department

Kenneth J. Siebert
Chairman, Examining Committee

Henry L. Parsons
Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

June, 1974

ACKNOWLEDGEMENT

The author wishes to express his gratitude to his thesis advisor, Dr. Kenneth J. Tiahrt, for the guidance and the many helpful suggestions made during the preparation of this thesis.

The author is also very grateful to Dr. Martin Hamilton, who gave willingly of his time to aid in many areas.

Appreciation is also extended to Professors Dennis O. Blackketter, Rodney T. Hansen, Richard E. Lund, Franklin S. McFeely and Eldon J. Whitesitt for serving on his graduate committee.

TABLE OF CONTENTS

ABSTRACT

In this work, estimation of a ratio of sensitive characteristics using Warner's randomized response type of design is investigated. Estimators for the mean of the ratios and its mean squared error is obtained. An unbiased Hartley-Ross type of ratio estimate is also found along with an unbiased estimate of the variance of this estimator. The asymptotic distribution of the estimator for the ratio of means is also obtained. A method of setting confidence intervals for the ratio of means for the normal case, which is an application of Fieller's Theorem, is obtained. The usually quite-robust method of setting confidence intervals using the Jackknife procedure is also given. A Monte-Carlo study was done to investigate the properties of the various estimators for normal populations and for Chi-Squared populations.

CHAPTER I

INTRODUCTION

Obtaining information about sensitive characteristics of a population can be of great importance to such people as social scientists and to policy makers and administrators of welfare programs. Obtaining unbiased information under these conditions is extremely difficult because of the propensity for a person to lie, especially to an interviewer who is probably a complete stranger, when asked to reveal information about himself which he may consider personal. One method of combating this reluctance to cooperate with an interviewer has been termed "Randomized Response" designs. Originally proposed by Samuel Warner (1965) [13], his design and several of its modifications appear to be quite successful in obtaining information on sensitive characteristics.

Warner's original design gives an unbiased estimator for the proportion of people who are members of a group possessing a sensitive characteristic, for example, the proportion of women who have had abortions [1], or the proportion who have driven an automobile while intoxicated, et cetera. Warner's design uses a randomizing device to determine if the person being interviewed should respond to

the question: "Are you a member of group A?", or to the question: "Are you not a member of group A?" The first question is asked with a probability of p (not equal to .5) and the second with a probability of 1-p. Obviously, the value of p is chosen as large as possible, but not so large as to lose the confidence of the respondent. Any easy to use randomization device may be used such as a spinner (marked off into two regions), a die or a pair of dice, et cetera. It should be noted that the respondent uses the randomization device in complete privacy. In Warner's design the response is either a yes or no, and the interviewer does not know to which question the person has responded.

If the proportion of people who actually lie is quite small, then the randomized response design is fairly inefficient when compared to asking the sensitive question directly [13]. Therefore, a number of modifications of Warner's original design have been made to improve the efficiency of the randomized response design.

One attempt at improving efficiency is to incorporate an "unrelated question" [10]. In this design, the respondent is asked either the sensitive question (with

probability p) or a question which is unrelated to the sensitive question. For instance, the two questions might be: "Have you ever driven while intoxicated?" (sensitive), or "Do you own two automobiles?" (nonsensitive). If the proportion of the population that is in the nonsensitive group is unknown, two independent samples are needed in order to estimate the proportion in the sensitive group. An obvious improvement would be to use a question whose proportion of yes (or no) responses is known. One such possibility would be, in the event the randomization device chooses the nonsensitive question, to have the respondent roll a die and answer the question: "Does the die show a number less than or equal to four?" Using this type of randomization design, only one sample would have to be taken since the moments of the nonsensitive distribution would be known.

Other modifications that have been proposed include: i) two alternate (nonsensitive) questions used in conjunction with a sensitive question [7]. In this design, one of the nonsensitive questions is asked directly (with probability one), and in addition, either the sensitive question or the other nonsensitive question is asked, depending upon the outcome of the randomization device.

ii) always asking the sensitive question directly, but then instructing the respondent to either lie or tell the truth depending upon the outcome of a randomization device. This type of design is called a contamination design [2].

iii) multiple responses from each respondent.

Greenberg et.al. [11] showed that the randomized response design can be used for obtaining information about quantitative as well as qualitative data. They used the unrelated, innocuous question type of design, which means that two independent samples must be taken in order to estimate the parameters of both the "sensitive distribution" and the "nonsensitive distribution."

Let p_1 and p_2 be the probability of selecting the sensitive question in the first and second samples respectively. If $\bar{z}_1$ and $\bar{z}_2$ are the mean responses from the first and second samples respectively, then unbiased estimators of the sensitive and nonsensitive means are respectively:

$$\hat{\mu}_A = \frac{(1-p_2)\bar{z}_1 - (1-p_1)\bar{z}_2}{p_1 - p_2},$$

$$\hat{\mu}_Y = \frac{p_2\bar{z}_1 - p_1\bar{z}_2}{p_2 - p_1}.$$

In order to maintain the confidence of the respondent, the plausible responses to the nonsensitive question should be plausible responses to the sensitive question and vice versa.

An improvement in Greenberg's design would be to incorporate a simple game (randomizing device) as the nonsensitive question, whose moments are known, and whose outcomes could be plausible responses to the sensitive question. Again, the advantage is that only one sample would have to be taken. This could either decrease the cost of running the survey or increase information if both samples of the original design were combined into the one required for this "improved" design.

In this paper, estimation of a ratio will be considered. Suppose that both the numerator and denominator questions, that are of interest, are sensitive. For instance, we might be interested in estimating the ratio of the amount spent on gambling to the amount spent on liquor, or the ratio of the amount given to charity to the amount spent on liquor, et cetera.

The interviewing procedure is to have the respondent use a randomization device (in private) to determine to which question, the sensitive or the innocuous, nonsensitive

question, he should respond for the numerator, and then give the response. The same respondent then uses the randomization device again to determine which question, sensitive or nonsensitive, he should respond to for the denominator. Therefore, each respondent will give two responses, one for the numerator and one for the denominator, and these will be recorded by the interviewer.

In this paper the technique discussed previously will be used. That is, distributions whose moments are known will be used for the nonsensitive questions in the numerator and denominator.

As an example of this technique, suppose we want to estimate the ratio of gambling expenditure to liquor expenditure per household per year. The randomization device that will be considered here is a simple child's spinner. This type of device has two advantages. First, it is easy for the respondent to operate, and secondly, the areas are easy to mark so that the probability of asking the sensitive question can be made to have virtually any desired value. The circle under the spinner is then marked off into two regions, say A_1 and $\tilde{A}_1$, for the numerator and A_2 and $\tilde{A}_2$ for the denominator. (It might simplify the procedure if a

second one were used for the denominator.) If the spinner then stops in region A_1, the respondent is supposed to answer the sensitive question: "How much did you spend on gambling last year?" And if the spinner stops in $\tilde{A}_1$, the respondent is supposed to answer the nonsensitive question. One simple possibility for the nonsensitive part would be to have another spinner marked so that the values obtained from it would be plausible responses to the sensitive question. Continuing with the example, suppose it is estimated that the range of the amount spent on gambling is from \$0 to \$1000. Then the spinner to be used for the nonsensitive question could be constructed so that the numbers from 0 to 1000 were laid out uniformly around the circle. Then the mean of this (uniform) distribution is 500 and its standard deviation is $\sqrt{(1000)^2/12} = 288.67$. A similar device can be constructed for the nonsensitive question in the denominator.

CHAPTER II

RATIO ESTIMATION

Case I: Numerator is a sensitive characteristic, the denominator is nonsensitive. As an example, we might be interested in determining the ratio of the amount spent on gambling to the amount spent on rent per time interval. Or possibly, the amount spent on gambling is our primary concern and we are using the amount spent on rent as a concommitant variable. The interviewing procedure is to have each respondent use the randomization device (in private) to determine which question to respond to in the numerator. The question in the denominator is asked of each respondent directly.

The notation which is required in the development of this ratio estimation procedure follows. Let

n = sample size;

p = probability that the sensitive question, x_1, is selected by the randomization device to be answered by the respondent in the numerator;

x_{1i} = real value of the sensitive characteristic for respondent i;

Z_i = response from individual i for the numerator;

X_{2i} = response from individual i for the denominator;

$f_1(X_1)$ = probability density function associated with the sensitive question (numerator);

$E_{f_1}(X_1) = \mu_1$ = population mean for the sensitive question;

$g_1(Y)$ = probability density function associated with the unrelated question (distribution);

$E_{g_1}(Y) = \mu_Y$, chosen to be approximately equal to μ_1;

$f_2(X_2)$ = probability density function associated with the nonsensitive question, (this question is thus asked directly of a respondent);

$E_{f_2}(X_2) = \mu_2$.

Using this notation, the probability density function for each response, Z, in the numerator of the sample is obtained from the randomized selection procedure:

$$\psi(Z) = p\ f_1(X_1) + (1-p)g_1(Y).$$

Then

$$\begin{aligned}\mu_Z &= E[\psi(Z)] \\ &= p\ E_{f_1}(X_1) + (1-p)E_{g_1}(Y) \\ &= p\ \mu_1 + (1-p)\mu_Y.\end{aligned}$$

Hence,

$$\mu_1 = [\mu_Z - (1-p)\mu_Y]/p.$$

And since $\bar{Z}_1$, the numerator sample response mean, is an unbiased estimate of μ_{Z_1}, $\hat{\mu}_1 = [\bar{Z} - (1-p)\mu_Y]/p$ is an unbiased estimate of μ_1.

For the nonsensitive question, we have that $\hat{\mu}_2 = \bar{X}_2$, the sample mean of the response from the nonsensitive question, is an unbiased estimate of μ_2.

Hence, a ratio estimate of $R = \mu_1/\mu_2$ is given by $\hat{R} = \hat{\mu}_1/\hat{\mu}_2 = [\bar{Z} - (1-p)\mu_Y]/p\ \bar{X}_2$.

To investigate the bias of this estimator, consider:

$$\hat{R} - R = (\hat{\mu}_1/\hat{\mu}_2) - (\mu_1/\mu_2)$$

$$= \frac{\bar{Z} + (p-1)\mu_Y}{p\ \bar{X}_2} - \frac{\mu_Z + (p-1)\mu_Y}{p\ \mu_2}$$

$$= \frac{1}{p}\left(\frac{\bar{Z}}{\bar{X}_2} - \frac{\mu_Z}{\mu_2}\right) + \frac{p-1}{p}\left(\frac{\mu_Y}{\bar{X}_2} - \frac{\mu_Y}{\mu_2}\right)$$

If the sample size is large, $\bar{X}_2$ should be close to μ_2, and this would imply $\hat{R} - R \doteq (\bar{Z} - \mu_Z)/p\ \mu_2$ and $E(\hat{R} - R) \doteq 0$. Thus $\hat{R}$ is unbiased for R when it is assumed that $\bar{X}_2 \doteq \mu_2$.

The variance of the estimator R is:

$$\mathrm{Var}(R) = E(R - R)^2$$

$$= E\left[\frac{1}{p}\left(\frac{\bar{Z}}{\bar{X}_2} - \frac{\mu_Z}{\mu_2}\right) + \frac{p-1}{p}\left(\frac{\mu_Y}{\bar{X}_2} - \frac{\mu_Y}{\mu_2}\right)\right]^2$$

Again, assuming that $\bar{X}_2$ is close to μ_2, it follows that

$$\mathrm{Var}(R) \doteq E\left[\frac{1}{p}\left(\frac{\bar{Z}}{\mu_2} - \frac{\mu_Z}{\mu_2}\right)\right]^2 = E(\bar{Z} - \mu_Z)^2/p^2\mu_2^2$$

$$= \frac{1}{p^2\mu_2^2}\ \mathrm{Var}(\bar{Z})$$

which could be estimated by $\widehat{\mathrm{Var}(R)} = s_Z^2/n\ p^2\bar{X}_2^2$, where s_Z^2 is the usual sample variance, i.e., $s_Z^2 = \Sigma_{i=1}^{n}(Z_i - Z)^2/n-1$, for an infinite population. This estimate would be unbiased if μ_2 is known.

In ratio estimation where the numerator is a sensitive characteristic and the denominator is nonsensitive, a biased estimator is obtained. This estimate is unbiased if the sample in the denominator is close to the true population mean. In this case, the variance of the ratio estimator may also be estimated without bias.

Case II: Both numerator and denominator are sensitive. Each respondent is asked to use the randomization device to determine if he will answer the sensitive question in the numerator or pick a number from a known distribution (which closely approximates the distribution of the sensitive question), and he will complete a similar procedure for the denominator response.

The notation used for this case is similar to that for Case I:

n = sample size;

p_1 = the probability that the sensitive question will be chosen by the randomization device for the numerator response;

p_2 = the probability that the sensitive question will be chosen by the randomization device for the denominator response.

Z_{1i} = response from individual i for the numerator;

Z_{2i} = response from individual i for the denominator;

X_{1i} = actual value for the sensitive question in the numerator for individual i;

X_{2i} = actual value for the sensitive question in the denominator for individual i;

$f_1(X_1)$ = probability density function associated with the sensitive question in the numerator;

$f_2(X_2)$ = probability density function associated with the sensitive question in the denominator;

$g_1(Y_1)$ = probability density function associated with the "unrelated" question in the numerator;

$g_2(Y_2)$ = probability density function associated with the "unrelated" question in the denominator;

$E_{f_1}(X_1) = \mu_1$;

$E_{f_2}(X_2) = \mu_2$;

$E_{g_1}(Y_1) = \mu_{Y_1}$;

$E_{g_2}(Y_2) = \mu_{Y_2}$.

The probability density function for the response in the numerator is:

$$\psi_1(Z_1) = p_1 f_1(X_1) + (1-p_1)g_1(Y_1),$$

and for the denominator,

$$\psi_2(Z_2) = p_2 f_2(X_2) + (1-p_2)g_2(Y_2).$$

Then,

$$\mu_{Z_1} = E[\psi_1(Z_1)] = p_1\mu_1 + (1-p_1)\mu_{Y_1},$$

$$\mu_{Z_2} = E[\psi_2(Z_2)] = p_2\mu_2 + (1-p_2)\mu_{Y_2},$$

So,

$$\mu_1 = [\mu_{Z_1} - (1-p_1)\mu_{Y_1}]/p_1,$$

$$\mu_2 = [\mu_{Z_2} - (1-p_2)\mu_{Y_2}]/p_2.$$

Therefore, the ratio, $R = \mu_1/\mu_2$, of the means of the two sensitive characteristics is

$$R = \frac{\mu_1}{\mu_2} = \frac{(\mu_{Z_1} - (1-p_1)\mu_{Y_1})p_2}{(\mu_{Z_2} - (1-p_2)\mu_{Y_2})p_1}.$$

Using the unbiased estimators $\bar{Z}_1$ and $\bar{Z}_2$ for μ_{Z_1} and μ_{Z_2} respectively, unbiased estimators for μ_1 and μ_2 are:

$$\hat{\mu}_1 = [\bar{Z}_1 - (1-p_1)\mu_{Y_1}]/p_1,$$

$$\hat{\mu}_2 = [\bar{Z}_2 - (1-p_2)\mu_{Y_2}]/p_2.$$

And the estimator of the ratio, R, is:

$$\hat{R} = \frac{\hat{\mu}_1}{\hat{\mu}_2} = \frac{p_2(\bar{Z}_1 - (1-p_1)\mu_{Y_1}}{p_1(\bar{Z}_2 - (1-p_2)\mu_{Y_2}}.$$

To obtain approximate values for the expected value of the estimator $\hat{R}$ and also to find $MSE(\hat{R})$, the mean squared error of $\hat{R}$, it will be useful to introduce some notation that will make the derivations less complicated. Let

$Z_{1i} = \mu_{Z_1} + \epsilon_{1i}$ so that $\bar{Z}_1 = \frac{1}{n} \Sigma_{i=1}^{n} Z_{1i} = \mu_{Z_1} + \bar{\epsilon}_1$,

$Z_{2i} = \mu_{Z_2} + \epsilon_{2i}$ so that $\bar{Z}_2 = \frac{1}{n} \Sigma_{i=1}^{n} Z_{2i} = \mu_{Z_2} + \bar{\epsilon}_2$.

Then,

$E(\bar{\epsilon}_1) = E(\bar{\epsilon}_2) = 0$

and

$$E(\bar{\epsilon}_1^2) = Var(\bar{Z}_1) = \sigma^2_{\bar{Z}_1} = \sigma^2_{Z_1}/n,$$

$$E(\bar{\epsilon}_2^2) = Var(\bar{Z}_2) = \sigma^2_{\bar{Z}_2} = \sigma^2_{Z_2}/n,$$

$$E(\bar{\epsilon}_1\bar{\epsilon}_2) = E[(\bar{Z}_1 - \mu_{Z_1})(\bar{Z}_2 - \mu_{Z_2})] = Cov(\bar{Z}_1,\bar{Z}_2)$$
$$= Cov(Z_1,Z_2)/n = \sigma_{Z_1Z_2}/n$$

Also, let $k_1 = 1/p_1\mu_1$ and $k_2 = 1/p_2\mu_2$.

Now, to find the bias of the estimator $\hat{R}$, consider,

$$E(\hat{R}) = E\left[\frac{(\bar{Z}_1 - (1-p_1)\mu_{Y_1})/p_1}{(\bar{Z}_2 - (1-p_2)\mu_{Y_2})/p_2}\right]$$

$$= \frac{p_2}{p_1} E\left[\frac{\mu_{Z_1} - (1-p_1)\mu_{Y_1} + \bar{\epsilon}_1}{\mu_{Z_2} - (1-p_2)\mu_{Y_2} + \bar{\epsilon}_2}\right]$$

Since $\mu_{Z_i} - (1-p_i)\mu_{Y_i} = p_i\mu_i + (1-p_i)\mu_{Y_i} - (1-p_i)\mu_{Y_i}$

$= p_i\mu_i$ for $i = 1$ or 2,

$$E(\hat{R}) = \frac{p_2}{p_1} E\left[\frac{p_1\mu_1 + \bar{\epsilon}_1}{p_2\mu_2 + \bar{\epsilon}_2}\right]$$

$$= \frac{p_2}{p_1}\frac{p_1\mu_1}{p_2\mu_2} E\left[\frac{1 + \bar{\epsilon}_1/p_1\mu_1}{1 + \bar{\epsilon}_2/p_2\mu_2}\right]$$

$$= R\, E\left[(1 + \bar{\epsilon}_1/p_1\mu_1)(1 + \bar{\epsilon}_2/p_2\mu_2)^{-1}\right] .$$

Since $(1 + \bar{\epsilon}_2/p_2\mu_2)^{-1} = (1 + k_2\bar{\epsilon}_2)^{-1}$ is to be expanded in a power series, $(\bar{\epsilon}_2/p_2\mu_2)^2$ must be less than one or $\bar{\epsilon}_2^2 < p_2^2\mu_2^2$. $\bar{\epsilon}_2$ is the quantity $\bar{Z}_2 - \mu_{Z_2}$ which should be relatively small. μ_2 is the population mean of the sensitive question and generally will not be close to zero. Therefore, it is a reasonable assumption that $\bar{\epsilon}_2^2 < p_2^2\mu_2^2$. Hence, expanding $(1 + k_2\bar{\epsilon}_2)^{-1}$ in a power series,

$$E(\hat{R}) = R\, E\ (1 + k_1\bar{\epsilon}_1)(1 - k_2\bar{\epsilon}_2 + k_2^2\bar{\epsilon}_2^2 - k_2^3\bar{\epsilon}_2^3 + k_2^4\bar{\epsilon}_2^4 \ldots)$$

$$= R\left[1 + k_1\, E(\bar{\epsilon}_1) - k_2\, E(\bar{\epsilon}_2) - k_1 k_2\, E(\bar{\epsilon}_1\bar{\epsilon}_2) + k_2^2\, E(\bar{\epsilon}_2^2) + k_1 k_2^2\, E(\bar{\epsilon}_1\bar{\epsilon}_2^2) + \ldots\right].$$

But since $E(\bar{\epsilon}_1) = E(\bar{\epsilon}_2) = 0$, this can be written as:

$$E(\hat{R}) = R\left[1 + k_2^2\, E(\bar{\epsilon}_2^2) - k_1 k_2\, E(\bar{\epsilon}_1\bar{\epsilon}_2) + k_1 k_2^2\, E(\bar{\epsilon}_1\bar{\epsilon}_2^2) + \ldots\right]$$

If the contributions to $E(\hat{R})$ of the terms involving $\bar{\epsilon}_1\bar{\epsilon}_2^2$ and higher powers of $\bar{\epsilon}_2$ are negligible, then $E(\hat{R})$ is approximately:

$$E(\hat{R}) \doteq R\left[1 + k_2^2\, E(\bar{\epsilon}_2^2) - k_1 k_2\, E(\bar{\epsilon}_1\bar{\epsilon}_2)\right]$$

$$= R\left[1 + k_2^2\, \sigma^2_{\bar{Z}_2} - k_1 k_2\, \sigma_{\bar{Z}_1\bar{Z}_2}\right]$$

$$= R\left[1 + k_2^2\, \sigma^2_{Z_2}/n - k_1 k_2\, \sigma_{Z_1 Z_2}/n\right]$$

So that an approximation of $E(\hat{R})$ is $E_1(\hat{R})$ given by

$$E_1(\hat{R}) = R\left[1 + k_2^2\, \sigma^2_{Z_2}/n - k_1 k_2\, \sigma_{Z_1 Z_2}/n\right].$$

Therefore, the bias of $\hat{R}$ is approximately:

$$\text{bias}(\hat{R}) \doteq E(\hat{R}_1) - R$$

$$= R\left[k_2^2\, \sigma^2_{Z_2}/n - k_1 k_2\, \sigma_{Z_1 Z_2}/n\right].$$

Since $\hat{R}$ is a biased estimate of R, the mean squared error of $\hat{R}$ will be obtained as follows:

$$MSE(\hat{R}) = E\,(\hat{R} - R)^2 = E\left[\hat{R}^2 - 2R\hat{R} + R^2\right]$$

$$= E(\hat{R}^2) - 2R\,E(\hat{R}) + R^2.$$

By using $E_1(\hat{R})$ as an approximation of $E(\hat{R})$ and substituting for $\hat{R}$ in the first term, $MSE(\hat{R})$ can be written as:

$$MSE(\hat{R}) \doteq E\left[\frac{p_2(\bar{Z}_1 - (1-p_1)\mu_{Y_1})}{p_1(\bar{Z}_2 - (1-p_2)\mu_{Y_2})}\right]^2 - 2\,R\,E_1(\hat{R}) + R^2$$

$$= E\left[\frac{p_2(\bar{Z}_1 - (1-p_1)\mu_{Y_1})}{p_1(\bar{Z}_2 - (1-p_2)\mu_{Y_2})}\right]^2 - 2\,R^2\left[1 + k_2^2\sigma_{Z_2}^2/n - k_1k_2\,\sigma_{Z_1Z_2}/n\right] + R^2$$

$$= E\left[\frac{p_2(\bar{Z}_1 - (1-p_1)\mu_{Y_1})}{p_1(\bar{Z}_2 - (1-p_2)\mu_{Y_2})}\right]^2 - R^2\left[1 + 2k_2^2\sigma_{Z_2}^2/n - 2k_1k_2\,\sigma_{Z_1Z_2}/n\right] \quad (1)$$

Again, letting $\bar{Z}_i = \mu_{Z_i} + \bar{\epsilon}_i$, $i = 1, 2$, the first term in the above expression can be expanded by essentially duplicating the steps in the derivation of $E(\hat{R})$ and is

$$E\left[\frac{p_2(\mu_{Z_1} - (1-p_1)\mu_{Y_1} + \bar{\epsilon}_1)}{p_1(\mu_{Z_2} - (1-p_2)\mu_{Y_2} + \bar{\epsilon}_2)}\right]^2$$

$$= R^2 \quad E\left[\frac{1 + \frac{\bar{\epsilon}_1}{p_1\mu_1}}{1 + \frac{\bar{\epsilon}_2}{p_2\mu_2}}\right]^2$$

$$= R^2 \quad E\left[(1 + k_1\bar{\epsilon}_1)^2(1 + k_2\bar{\epsilon}_2)^{-2}\right]$$

$$= R^2 \quad E(1 + 2k_1\bar{\epsilon}_1 + k_1^2\bar{\epsilon}_1^2)(1 - 2k_2\bar{\epsilon}_2 + 3k_2^2\bar{\epsilon}_2^2 - 4k_2^3\bar{\epsilon}_2^3 + 5k_2^4\bar{\epsilon}_2^4 \ldots),$$

by again making the assumption that $\bar{\epsilon}_2^2 < p_2^2\mu_2^2$.

By expanding this result and assuming that terms of order $\bar{\epsilon}_1^i\bar{\epsilon}_2^j$ for $i + j \geq 3$ are negligible, and hence retaining the first four terms, the first term in (1) is approximately

$$R^2 \quad E\left[1 + 2(k_1\bar{\epsilon}_1 - k_2\bar{\epsilon}_2) + k_2^2\bar{\epsilon}_1^2 + 3k_2^2\bar{\epsilon}_2^2 - 4k_1k_2\bar{\epsilon}_1\bar{\epsilon}_2\right]$$

$$= R^2\left[1 + k_1^2\sigma_{Z_1}^2/n + 3k_2^2\sigma_{Z_2}^2/n - 4k_1k_2\sigma_{Z_1Z_2}/n\right]$$

recalling that $E(\bar{\epsilon}_1) = E(\bar{\epsilon}_2) = 0$.

Upon combining the two terms in (1), the $MSE(\hat{R})$ can be written as

$$MSE(\hat{R}) \doteq R^2 \left[1 + k_1^2\sigma_{Z_1}^2/n + 3k_2^2\sigma_{Z_2}^2/n - 4k_1k_2\sigma_{Z_1Z_2}/n\right]$$

$$- R^2 \left[1 + 2k_2^2\sigma_{Z_2}^2/n - 2k_1k_2\sigma_{Z_1Z_2}/n\right]$$

$$= R^2 \left[1 + k_1^2\sigma_{Z_1}^2/n + 3k_2^2\sigma_{Z_2}^2/n - 4k_1k_2\sigma_{Z_1Z_2}/n\right.$$

$$\left. - 1 - 2k_2^2\sigma_{Z_2}^2/n + 2k_1k_2\sigma_{Z_1Z_2}/n\right]$$

$$= \frac{R^2}{n}\left[k_1^2\sigma_{Z_1}^2 + k_2^2\sigma_{Z_2}^2 - 2k_1k_2\sigma_{Z_1Z_2}\right]$$

$$= \frac{R^2}{n}\left[\frac{\sigma_{Z_1}^2}{p_1^2\mu_1^2} + \frac{\sigma_{Z_2}^2}{p_2^2\mu_2^2} - \frac{2\,\sigma_{Z_1Z_2}}{p_1p_2\mu_1\mu_2}\right]$$

An estimator of $MSE(\hat{R})$ would be to use the same expression as above, but replacing all parameters with estimators. Hence

$$MSE(\hat{R}) = \hat{R}^2/n\left[\hat{k}_1^2s_{Z_1}^2 + \hat{k}_2^2s_{Z_2}^2 - 2\hat{k}_1\hat{k}_2\,s_{Z_1Z_2}\right]$$

where $s^2_{Z_1} = \frac{1}{n-1} \Sigma_{i=1}^{n} (Z_{1i} - \bar{Z}_1)^2 = \left[n\Sigma Z^2_{1i} - (\Sigma Z_{1i})^2 \right] / n(n-1)$

$$s^2_{Z_2} = \frac{1}{n-1} \Sigma_{i=1}^{n} (Z_{2i} - \bar{Z}_2)^2 = \left[n\Sigma Z^2_{2i} - (\Sigma Z_{2i})^2 \right] / n(n-1)$$

$$\hat{k}_1 = 1/p_1\hat{\mu}_1 = 1/(\bar{Z}_1 - (1-p_1)\mu_{Y_1})$$

$$\hat{k}_2 = 1/p_2\hat{\mu}_2 = 1/(\bar{Z}_2 - (1-p_2)\mu_{Y_2})$$

$$s_{Z_1Z_2} = \frac{1}{n-1} \Sigma_{i=1}^{n} (Z_{1i} - \bar{Z}_1)(Z_{2i} - \bar{Z}_2)$$

$$= \left[n \Sigma Z_{1i}Z_{2i} - (\Sigma Z_{1i})(\Sigma Z_{2i}) \right] / n(n-1).$$

In ratio estimation using the randomized response technique where both numerator and denominator characteristics of interest are of a sensitive nature, the estimator that is obtained is biased. Also, the mean squared error of the estimator is an approximation to the true mean squared error. Normally this discrepancy is quite small. The approximate mean squared error cannot be estimated without bias.

Case III: In cases I and II, an estimator of $R = \mu_1/\mu_2$, the ratio of means, and its mean square error were found. In this section, estimation of a different parameter will be considered. Let r_{X_i} be the true ratio of two sensitive characteristics for individual i. Then the parameter of interest is the mean of all such ratios. As indicated above, both the numerator and denominator are sensitive characteristics.

Since the estimator is going to be the mean of the ratios, the procedure of randomizing the responses will be altered somewhat. In this case, each respondent is asked to respond to either: a) both sensitive questions, one for the numerator and one for the denominator or b) both nonsensitive questions, one for the numerator and one for the denominator. The randomization device is then used to determine to which set of questions, a or b, he should respond.

Assuming that the response in the denominator of the ratios will not be zero, the ratio of the two sensitive or the ratio of the two nonsensitive responses can be considered as being one observation and hence not really a ratio estimate at all. For the sake of completeness, however,

the derivation of the estimator and its variance will be included.

The notation required is as follows:

n = sample size;

p = the probability that the sensitive questions will be chosen by the randomization device;

r_{Z_i} = response from individual i;

r_{X_i} = actual value of the ratio of the sensitive questions for the ith respondent;

r_{Y_i} = value of the ratio of the non-sensitive questions for the ith respondent;

$f(r_X)$ = probability density function associated with the sensitive ratio;

$g(r_Y)$ = probability density function associated with the non-sensitive ratio;

then $E_f(r_X) = R_X$ and $E_g(r_Y) = R_Y$. The probability density function for each response is:

$$(r_Z) = pf(r_X) + (1-p)g(r_Y)$$

giving

$$R_Z = E\left[\psi(r_Z)\right] = pE_f(r_X) + (1-p)E_g(r_Y)$$

$$= pR_X + (1-p)R_Y.$$

Hence, $R_X = \left[R_Z - (1-p)R_Y\right]/p$. If we again assume that the mean (and variance) of the non-sensitive distribution is known, an estimate of R_X is:

$$\hat{R}_X = \left[\bar{r}_Z - (1-p)R_Y\right]/p$$

where $\bar{r}_Z$ is the mean of the ratio of responses from the sample.

The estimator $\hat{R}_X$ of R_X is unbiased, since

$$E(\hat{R}_X) = E\left[\bar{r}_Z - (1-p)R_Y/p\right] = \left[E(\bar{r}_Z) - (1-p)R_Y\right]/p$$

$$= p^{-1}\left[E\left\{p\,\bar{r}_X + (1-p)\bar{r}_Y\right\} - (1-p)R_Y\right]$$

$$= p^{-1}\left[E\left\{p\,\frac{\Sigma_{i=1}^{n} r_{X_i}}{n} + (1-p)\,\frac{\Sigma_{i=1}^{n} r_{Y_i}}{n}\right\} - (1-p)R_Y\right]$$

$$= p^{-1}\left[\frac{p}{n}\,\Sigma_{i=1}^{n} E_f(r_{X_i}) + \frac{1-p}{n}\,\Sigma_{i=1}^{n} E_g(r_{Y_i}) - (1-p)R_Y\right]$$

$$= p^{-1}\left[\frac{p}{n}\,\Sigma_{i=1}^{n} R_X + \frac{1-p}{n}\,\Sigma_{i=1}^{n} R_Y - (1-p)R_Y\right]$$

$$= p^{-1}\left[p \cdot R_X + (1-p)R_Y - (1-p)R_Y\right]$$

$$= R_X$$

which completes the proof.

The variance of this estimate is given by:

$$\mathrm{Var}(\hat{R}_X) = E\left[(\hat{R}_X - R_X)^2\right] = E\left[\frac{\bar{r}_Z - (1-p)R_Y}{p} - \frac{R_Z - (1-p)R_Y}{p}\right]^2$$

$$= E\left[\bar{r}_Z - (1-p)R_Y - R_Z + (1-p)R_Y\right]^2/p^2$$

$$= \frac{1}{p^2} E(\bar{r}_Z - R_Z)^2 = \frac{1}{p^2} \mathrm{Var}(\bar{r}_Z).$$

The unbiased estimator of $\mathrm{Var}(\hat{R}_X)$ is

$$\widehat{\mathrm{Var}(\hat{R}_X)} = \frac{1}{np^2} s^2_{r_Z},$$

where $s^2_{r_Z}$ is the sample variance of responses, i.e.

$$s^2_{r_Z} = \Sigma_{i=1}^{n} (r_{Z_i} - \bar{r}_Z)^2/n-1.$$

CHAPTER III

UNBIASED RATIO TYPE ESTIMATORS

If the mean of the population of the sensitive question, μ_2, is known in either Case I or II, then an unbiased ratio-type (Hartley-Ross) estimator of μ, can be found [8]. Since this estimator uses the same type of variables as defined in Case III, Chapter II, the same notation will be observed here. Namely,

$$r_Z = Z_1/Z_2$$

and

$$\bar{r}_Z = \frac{1}{n} \Sigma_{i=1}^{n} (Z_{1i}/Z_{2i}) = \frac{1}{n} \Sigma_{i=1}^{n} r_{Z_i}.$$

In order to obtain the unbiased ratio estimate, consider the following quantity:

$$E\left[r_Z(Z_2 - \mu_{Z_2})\right] = E\left[(Z_1/Z_2)(Z_2 - \mu_{Z_2})\right]$$

$$= \mu_{Z_1} - \mu_{Z_2} E(Z_1/Z_2)$$

$$= \mu_{Z_1} - \mu_{Z_2} E(r_Z)$$

But $E(r_Z) = E(\bar{r}_Z)$, so the above can be written as:

$$E\left[r_Z(Z_2 - \mu_{Z_2})\right] = \mu_{Z_1} - \mu_{Z_2} E(\bar{r}_Z)$$

$$= \mu_{Z_2}\left[(\mu_{Z_1}/\mu_{Z_2}) - E(\bar{r}_Z)\right]$$

$$= \mu_{Z_2}\left[R_Z - E(\bar{r}_Z)\right]$$

$$= -\mu_{Z_2}\left[E(\bar{r}_Z) - R_Z\right]$$

Now the quantity in the brackets in the expression above is the bias of the estimator $\bar{r}_Z$, say $B(\bar{r}_Z)$.

Therefore,

$$E\left[r_Z(Z_2 - \mu_{Z_2})\right] = -\mu_{Z_2}\left[B(\bar{r}_Z)\right].$$

Or upon solving for the bias, $B(\bar{r}_Z)$:

$$B(\bar{r}_Z) = -\mu_{Z_2}^{-1} E\left[r_2(Z_2 - \mu_{Z_2})\right] \qquad (2)$$

Before proceding further, the following should be noted:

$$\mathrm{Cov}(r_Z, Z_2) = \frac{1}{n-1}\Sigma r_{Z_i}(Z_{2i} - \bar{Z}_2)$$

$$= \frac{n}{n-1}(\bar{Z}_1 - \bar{r}_Z\bar{Z}_2)$$

which will now be shown. In the derivation that follows, the range on all summations is from one to n.

$$\mathrm{Cov}(r_Z, Z_2) = \frac{1}{n-1}\Sigma(r_{Z_i} - \bar{r}_Z)(Z_{2i} - \bar{Z}_2)$$

$$= \frac{1}{n-1}\Sigma(r_{Z_i}Z_{2i} - r_{Z_i}\bar{Z}_2 - \bar{r}_Z Z_{2i} + \bar{r}_Z\bar{Z}_2)$$

$$= \frac{1}{n-1}\left[\Sigma Z_{1i} - \bar{Z}_2 \Sigma r_{Z_i} - \bar{r} \Sigma Z_{2i} + n \bar{r}_Z \bar{Z}_2\right]$$

$$= \frac{1}{n-1}\left[n \bar{Z}_1 - n \bar{r}_Z \bar{Z}_2 - n \bar{r}_Z \bar{Z}_2 + n \bar{r}_Z \bar{Z}_2\right]$$

$$= \frac{n}{n-1}\left[\bar{Z}_1 - \bar{r}_Z \bar{Z}_2\right]$$

$$= \frac{1}{n-1}\left[\Sigma Z_{1i} - \bar{Z}_2 \Sigma r_{Z_i}\right]$$

$$= \frac{1}{n-1} \Sigma(Z_{1i} - r_{Z_i} \bar{Z}_2)$$

$$= \frac{1}{n-1} \Sigma(r_{Z_i} \cdot Z_{2i} - r_{Z_i} \bar{Z}_2)$$

$$= \frac{1}{n-1} \Sigma r_{Z_i}(Z_{2i} - \bar{Z}).$$

Another result that is needed is that an unbiased estimator of $E\left[r_Z(Z_2 - \mu_{Z_2})\right] = \mu_{Z_1} - \mu_{Z_2} E(r_Z)$ is $\widehat{Cov(r_Z, Z_2)}$. To show this, consider the expected value of the following form of $\widehat{Cov(r_Z, Z_2)}$:

$$E\left[\frac{1}{n-1} \Sigma_{i=1}^{n} r_{Z_i}(Z_{2i} - \bar{Z}_2)\right]$$

$$= E\left[\frac{1}{n-1}\left\{\Sigma_{i=1}^{n} \frac{Z_{1i}}{Z_{2i}} Z_{2i} - \bar{Z}_2 \Sigma_{i=1}^{n} r_{Z_i}\right\}\right]$$

$$= \frac{1}{n-1}\left[\Sigma_{i=1}^{n} E Z_{1i} - \frac{1}{n} E \Sigma_{i=1}^{n} Z_{2i} \Sigma_{i=1}^{n} r_{Z_i}\right]$$

$$= \frac{1}{n-1}\left\{n\,\mu_{Z_1} - \frac{1}{n}\,E\;(Z_{21} + Z_{22} + \cdots + Z_{2n})\right.$$

$$\left.\left(\frac{Z_{11}}{Z_{21}} + \frac{Z_{12}}{Z_{22}} + \frac{Z_{13}}{Z_{23}} + \cdots + \frac{Z_{1n}}{Z_{2n}}\right)\right\}$$

$$= \frac{1}{n-1}\left\{n\,\mu_{Z_1} - \frac{1}{n}\,E\left[\left(Z_{11} + \frac{Z_{12}}{Z_{22}}\,Z_{21} + \frac{Z_{13}}{Z_{23}}\,Z_{21} + \cdots + \frac{Z_{1n}}{Z_{2n}}\,Z_{21}\right)\right.\right.$$

$$+ \left(Z_{12} + \frac{Z_{11}}{Z_{21}}\,Z_{22} + \frac{Z_{13}}{Z_{23}}\,Z_{22} + \cdots + \frac{Z_{1n}}{Z_{2n}}\,Z_{22}\right) + \cdots$$

$$\left.\left. + \left(Z_{1n} + \frac{Z_{11}}{Z_{21}}\,Z_{2n} + \frac{Z_{12}}{Z_{22}}\,Z_{2n} + \cdots + \frac{Z_{1,n-1}}{Z_{2,n-1}}\right)\right]\right\}$$

$$= \frac{1}{n-1}\left\{n\,\mu_{Z_1} - \frac{1}{n}\left[E\;\Sigma_{i=1}^{n} Z_{1i} + E\left(\Sigma_{i=2}^{n} r_{Z_i} Z_{21}\right.\right.\right.$$

$$\left.\left.\left. + \Sigma_{\substack{i=1,\\ i\neq 2}}^{n} r_{Z_i} Z_{22} + \cdots + \Sigma_{i=1}^{n-1} r_{Z_i} Z_{2n}\right)\right]\right\}$$

$$= \frac{1}{n-1}\left\{n\,\mu_{Z_1} - \frac{1}{n}\left[n\,\mu_{Z_1} + \Sigma_{i=2}^{n}\,E\,r_{Z_i}\,Z_{21} + \cdots\right.\right.$$

$$\left.\left. + \Sigma_{i=1}^{n-1}\;E\,r_{Z_i}\,Z_{2n}\right]\right\}$$

and since r_{Z_i} is independent of Z_{2j}, $i \neq j$,

$$= \mu_{Z_1} - \frac{1}{n(n-1)} \left[\Sigma_{i=2}^{n} (E\, r_{Z_i}\, E\, Z_{21}) + \right.$$

$$\left. + \Sigma_{\substack{i=1,\\ i\neq 2}}^{n} (E\, r_{Z_i}\, E\, Z_{22}) + \ldots + \Sigma_{i=1}^{n-1} E\, r_{Z_i}\, E\, Z_{2n} \right]$$

$$= \mu_{Z_1} - \frac{1}{n(n-1)} \left[(n-1)\, E(\bar{r}_Z)\, \mu_{Z_2} + (n-1)\, E(\bar{r}_Z)\, \mu_{Z_2} \right.$$

$$\left. + \ldots + (n-1)\, E(\bar{r}_Z)\, \mu_{Z_2} \right]$$

$$= \mu_{Z_1} - \frac{1}{n(n-1)} \left[n(n-1)\, \mu_{Z_2}\, E(\bar{r}_Z) \right] = \mu_{Z_1} - \mu_{Z_2}\, E(\bar{r}_Z).$$

which completes the proof.

Using the estimator $\frac{1}{n-1} \Sigma_{i=1}^{n} r_{Z_i} (Z_{2i} - \bar{Z}_2)$ in (2), the unbiased estimator of

$R_Z = \frac{\mu_{Z_1}}{\mu_{Z_2}}$ is:

$$\bar{r}'_Z = \bar{r}_Z + \frac{1}{(n-1)\mu_{Z_2}} \Sigma_{i=1}^{n} r_{Z_i} (Z_{2i} - \bar{Z}_2)$$

$$= \bar{r}_Z + \frac{n(\bar{Z}_1 - \bar{r}_Z\, \bar{Z}_2)}{(n-1)\mu_{Z_2}}.$$

Since $\mu_{Z_1} = R_Z \mu_{Z_2}$, $R = \mu_1/\mu_2$ can be written in the following form:

$$R = \frac{p_2[\mu_{Z_1} - (1-p_1)\mu_{Y_1}]}{p_1[\mu_{Z_2} - (1-p_2)\mu_{Y_2}]}$$

$$= \frac{p_2[R_Z\mu_{Z_2} - (1-p_1)\mu_{Y_1}]}{p_1[\mu_{Z_2} - (1-p_2)\mu_{Y_2}]}$$

$$= \frac{R_Z\mu_{Z_2} - (1-p_1)\mu_{Y_1}}{p_1\mu_2}.$$

Therefore, the unbiased estimator of $R = \mu_1/\mu_2$ is:

$$\bar{r}' = \frac{\bar{r}_Z' \mu_{Z_2} - (1-p_1)\mu_{Y_1}}{p_1\mu_2}$$

$$= \frac{1}{p_1\mu_2}\left[\bar{r}_Z \mu_{Z_2} + \frac{n}{n-1}(\bar{Z}_1 - \bar{r}_Z\bar{Z}_2) - (1-p_1)\mu_{Y_1}\right].$$

$$= \frac{1}{p_1\mu_2}\left[\bar{r}_Z \mu_{Z_2} + \overbrace{\mathrm{Cov}(r_Z, Z_2)} - (1-p_1)\mu_{Y_1}\right]$$

then, since μ_2 is assumed to be known, as unbiased estimate of μ_1 is:

$$\tilde{\mu}_1 = \bar{r}'\mu_2 = \frac{1}{p_1}\left[\bar{r}_Z\mu_{Z_2} + \frac{n}{n-1}(\bar{Z}_1 - \bar{r}_Z\bar{Z}_2) - (1-p_1)\mu_{Y_1}\right]$$

$$= \frac{1}{p_1}\left[\bar{r}_Z\,\mu_{Z_2} + \widehat{Cov(r_Z,Z_2)} - (1-p_1)\mu_{Y_1}\right]$$

Note that μ_{Z_2} is known since $\mu_{Z_2} = p_2\mu_2 + (1-p_2)\mu_{Y_2}$ and both μ_2 and μ_{Y_1} are known.

The unbiased estimator $\bar{r}'$ of R is a function of the mean of the sample responses Z_{1i}/Z_{2i}, and of the sample covariance. So this is quite a different type of estimator than $\hat{R}$, which is a function of the sample means of the numerator and denominator responses. But since μ_2 must be known for this estimator, its value lies not with estimating the ratio R but with estimating the mean μ_1.

The exact variance of the estimator $\tilde{\mu}_1$ is

$$Var(\tilde{\mu}_1) = \left[\mu_{Z_2}^2\, Var(\bar{r}_Z) + 2\,\mu_{Z_2}\, Cov(\bar{r}_Z,C) + Var(C)\right]/p_1^2 \qquad (3)$$

where

$$C = \widehat{Cov(r_Z, Z_2)} = \left[\Sigma_{i=1}^{n} (r_{Z_i} - \bar{r}_Z)(Z_{2i} - \bar{Z}_2)\right]/n\text{-}1.$$

In order to estimate $Var(\tilde{\mu}_1)$, each of the three terms in the expression above will be given in a form which will allow for estimation.

The first term follows readily since $Var(\bar{r}_Z) = \frac{1}{n} Var(r_Z)$ which has an unbiased estimate:

$$s^2_{r_Z}/n = \frac{1}{n(n-1)} \Sigma(r_{Z_i} - \bar{r}_Z) = \frac{1}{n(n-1)} \left[n\Sigma r^2_{Z_i} - (\Sigma r_{Z_i})^2\right]$$

The second term can be estimated by rewriting it in quite a different form as follows:

$$\begin{aligned}
&Cov(\bar{r}_Z, C) \\
&= E\left[(\bar{r}_Z - E(\bar{r}_Z))(C - E(C))\right] \\
&= E\left[(\bar{r}_Z - R_Z)\left\{\frac{1}{n-1} \Sigma(r_{Z_i} - \bar{r}_Z)(Z_{2i} - \bar{Z}_2) - Cov(r_Z, Z_2)\right\}\right] \\
&= E\left[\frac{\bar{r}_Z - R_Z}{n-1} \Sigma(r_{Z_i} - \bar{r}_Z)(Z_{2i} - \bar{Z}_2)\right] - Cov(r_Z, Z_2)E(\bar{r}_Z - R_Z)
\end{aligned}$$

The last term is zero because $E(\bar{r}_Z - R_Z) = 0$. Therefore,

$$\mathrm{Cov}(\bar{r}_Z, C)$$

$$= \frac{1}{n-1} \; E\left[\frac{1}{n} \; \Sigma_{j=1}^{n} \; (r_{Z_j} - R_Z) \; \Sigma_{i=1}^{n} \; (r_{Z_i} - \bar{r}_Z)(Z_{2i} - \bar{Z}_2)\right]$$

$$= \frac{1}{n(n-1)} \; E\left[\Sigma_{i=1}^{n} \; (r_{Z_i} - R_Z)(r_{Z_i} - \bar{r}_Z)(Z_{2i} - \bar{Z}_2)\right.$$

$$\left. + \; \Sigma_{\substack{i=1 \\ i \neq j}}^{n} \; \Sigma_{j=1}^{n} \; (r_{Z_i} - R_Z)(r_{Z_j} - \bar{r}_Z)(Z_{2j} - \bar{Z}_2)\right]$$

By adding and subtracting the parameters R_Z in the middle term and μ_{Z_2} in the last term in each summation, letting $\Delta r_{Z_i} = r_{Z_i} - R_Z$, $\Delta Z_{2i} = Z_{2i} - \mu_{Z_2}$, $\Delta \bar{r}_Z = \bar{r}_Z - R_Z$ and $\Delta \bar{Z}_2 = \bar{Z}_2 - \mu_{Z_2}$, and expanding the factors in each sum, the above can be written as:

$$\mathrm{Cov}(\bar{r}_Z, C)$$

$$= \frac{1}{n(n-1)} \; E\left\{\Sigma(r_{Z_i} - R_Z) \; \left[(r_{Z_i} - R_Z) - (\bar{r}_Z - R_Z)\right]\right.$$

$$\left. \left[(Z_{2i} - \mu_{Z_2}) - (\bar{Z}_2 - \mu_{Z_2})\right]\right.$$

$$+ \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \left[(r_{Z_i} - R_Z)\ (r_{Z_j} - R_Z) - (\bar{r}_Z - R_Z) \right] \left[(Z_{2j} - \mu_{Z_2}) - (\bar{Z}_2 - \mu_{Z_2}) \right] \Bigg\}$$

$$= \frac{1}{n(n-1)} E \Bigg\{ \sum_{i=1}^{n} \Big[(\Delta r_{Z_i})^2 \Delta Z_{2i} - (\Delta r_{Z_i})^2 \Delta \bar{Z}_2 - \Delta r_{Z_i} \Delta \bar{r}_Z \Delta Z_{2i} + \Delta r_{Z_i} \Delta \bar{r}_Z \Delta \bar{Z}_2 \Big]$$

$$+ \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \Big[\Delta r_{Z_i} \Delta r_{Z_j} \Delta Z_{2j} - \Delta r_{Z_i} \Delta r_{Z_j} \Delta \bar{Z}_2 - \Delta r_{Z_i} \Delta \bar{r}_Z \Delta Z_{2j} + \Delta r_{Z_i} \Delta \bar{r}_Z \Delta \bar{Z}_2 \Big] \Bigg\}$$

$$= \frac{1}{n(n-1)} \Bigg\{ \sum_{i=1}^{n} \Big(E \left[(\Delta r_Z)^2 \Delta Z_{2i} \right] - \frac{1}{n} E \left[(\Delta r_{Z_i})^2 \sum_{j=1}^{n} \Delta Z_{2j} \right]$$

$$- \frac{1}{n} E \left[\Delta r_{Z_i} \Delta Z_{2i} \sum_{j=1}^{n} \Delta r_{Z_j} \right] + \frac{1}{n^2} E \left[\Delta r_{Z_i} \sum_{j=1}^{n} \Delta r_{Z_j} \sum_{k=1}^{n} \Delta Z_{2k} \right] \Big)$$

$$+ \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \Big(E(\Delta r_{Z_i})\ E \left[\Delta r_{Z_j} \Delta Z_{2j} \right] - \frac{1}{n} E \left[\Delta r_{Z_i} \Delta r_{Z_j} \sum_{k=1}^{n} \Delta Z_{2k} \right]$$

$$- \frac{1}{n} E \left[\Delta r_{Z_i} \Delta Z_{2j} \sum_{k=1}^{n} \Delta r_{Z_k} \right] + \frac{1}{n^2} E \left[\Delta r_{Z_i} \sum_{j=1}^{n} \Delta r_{Z_j} \sum_{k=1}^{n} \Delta Z_{2k} \right] \Big) \Bigg\}$$

The first three terms in the second summation are all zero. This is true because at least two of the subscripts i, j or k are different from each other, so that when these terms are expanded each one is of the form $E(\Delta r_{Z_i} \Delta r_{Z_j} \Delta Z_{2j})$ $= E(\Delta r_{Z_i}) E(\Delta r_{Z_j} \Delta Z_{2j})$ for $i \neq j$ and the first expectation has the value zero. Thus,

$$\mathrm{Cov}(\bar{r}_Z, C)$$

$$= \frac{1}{n-1} \left\{ E\left[(\Delta r_Z)^2 \Delta Z_2\right] - \frac{1}{n} E\left[(\Delta r_Z)^2 \Delta Z_2\right] - \frac{1}{n} E\left[(\Delta r_Z)^2 \Delta Z_2\right] + \frac{1}{n^2} E\left[(\Delta r_Z)^2 \Delta Z_2\right] \right\} + \frac{1}{n^2} E\left[(\Delta r_Z)^2 \Delta Z_2\right]$$

$$= E\left[(\Delta r_Z)^2 \Delta Z_2\right] \left[\frac{1}{n-1}\left(1 - \frac{2}{n} + \frac{1}{n^2}\right) + \frac{1}{n^2}\right]$$

$$= \frac{1}{n} E\left[(\Delta r_Z)^2 \Delta Z_2\right].$$

Now by using the method of moments of bivariate cumulants [4], this expectation can be written as:

$$E\left[(\Delta r_Z)^2\,\Delta Z_2\right] = \mu'_{21} - \mu'_{20}\mu'_{01} - 2\mu'_{11}\mu'_{10} + 2\mu'_{10}\mu'_{01}$$

$$= E(r_Z^2 Z_2) - E(r_Z^2)E(Z_2) - 2E(r_Z Z_2)E(r_Z) + 2(E\,r_Z)^2 E(Z_2)$$

where $\mu'_{st} = E(r_Z^s Z_2^t)$.

Similarly, the third term in (3) can be written as:

$$\mathrm{Var}(C) = \frac{1}{n}\left[E(\Delta r_Z)^2(\Delta Z_2)^2 + \frac{\mathrm{Var}(r_Z)\mathrm{Var}(Z_2)}{n-1} - \frac{n-2}{n-1}\,\mathrm{Cov}^2(r_Z, Z_2)\right].$$

Again using bivariate cumulants, this can be written as:

$$\mathrm{Var}(C) = \frac{1}{n}\Big[\mu'_{22} - 2\mu'_{21}\mu'_{01} + 2\mu'_{20}\mu'^2_{01} - \mu'_{20}\mu'_{02} - 2\mu'_{12}\mu'_{10} - 2\mu'^2_{11} + 8\mu'_{11}\mu'_{10}\mu'_{01} - 6\mu'^2_{10}\mu'^2_{01} + 2\mu'^2_{10}\mu'_{02} + \frac{1}{n-1}(\mu'_{20} - \mu'^2_{10})(\mu'_{02} - \mu'^2_{01}) - \frac{n-2}{n-1}(\mu'_{11} - \mu'_{10}\mu'_{01})^2\Big]$$

The above could be expressed in terms of expectations by replacing μ'_{st} with $E(r_Z^s Z_2^t)$.

Upon substituting these terms into (3), $\mathrm{Var}(\tilde{\mu}_1)$ becomes:

$$\mathrm{Var}(\tilde{\mu}_1) = \frac{1}{np_1^2}\left\{\mu_{Z_2}^2\ \mathrm{Var}(r_Z) + 2\,\mu_{Z_2}\,E\left[(\Delta r_Z)^2(\Delta Z_2)\right]\right.$$

$$\left. + E\left[(\Delta r_Z)^2(\Delta Z_2)^2\right] + \frac{\mathrm{Var}(r_Z)\mathrm{Var}(Z_2)}{n-1} - \frac{n-2}{n-1}\ \mathrm{Cov}^2(r_Z, Z_2)\right\}$$

An unbiased estimator for $\mathrm{Var}(\tilde{\mu}_1)$ can be found using bivariate k-statistics [4] which are unbiased estimates of the corresponding population cumulants. Using the results of Goodman and Hartley [8] (after correcting for typographical errors), an unbiased estimate of $\mathrm{Var}(\tilde{\mu}_1)$ is:

$$\widehat{\mathrm{Var}(\tilde{\mu}_1)} = \frac{1}{p_1^2}\left[\frac{1}{n}\mu_{Z_2}^2 s_{r_Z}^2 + \frac{2}{n-2}\mu_{Z_2}\,C'\right.$$

$$\left. + \frac{(n-1)s_{r_Z}^2 s_{Z_2}^2 + (n-3)C^2 + (1-\frac{2}{n})(n-1)k_{22}}{n^2 - n - 2}\right]$$

The symbols used in this expression are defined and their computational forms are given by:

$$s_{r_Z}^2 = \frac{1}{n-1}\ \Sigma_{i=1}^n\ (r_{Z_i} - \bar{r}_Z)^2 = \left[n\ \Sigma\left(\frac{Z_{1j}}{Z_{2i}}\right)^2 - \left(\Sigma\ \frac{Z_{ij}}{Z_{2i}}\right)^2\right]/n(n-1),$$

$$s_{Z_2}^2 = \frac{1}{n-1}\ \Sigma_{i=1}^n\ (Z_{2i} - \bar{Z}_2)^2 = \left[n\ \Sigma_{i=1}^n\ Z_{2i}^2 - (\Sigma Z_{2i})^2\right]/n(n-1),$$

$$c' = \frac{1}{n-1} \sum_{i=1}^{n} (r_{Z_i} - \bar{r}_Z)^2 (Z_{2i} - \bar{Z}_2)$$

$$= \frac{1}{n-1} \left[\Sigma Z_{1j} r_{Z_i} - 2\bar{r}_Z \Sigma Z_{1j} + \bar{r}_Z^2 \Sigma Z_{2i} - (n-1)\bar{Z}_2 s^2_{r_Z} \right],$$

$$c = \frac{1}{n-1} \Sigma (r_{Z_i} - \bar{r}_Z)(Z_{2i} - \bar{Z}_2) = \frac{1}{n(n-1)} \left[n \Sigma Z_{ij} - (\Sigma Z_{2i})(\Sigma r_{Z_i}) \right],$$

and

$$k_{22} = \frac{1}{(n-1)(n-2)(n-3)} \left[n(n+1) S_{22} - 2(n+1) \left(S_{21} S_{01} + S_{12} S_{10} \right) \right.$$

$$- (n-1) \left(S_{20} S_{02} + 2 S_{11}^2 \right) + 8 S_{11} S_{10} S_{01} + 2 S_{20} S_{01}^2$$

$$\left. + 2 S_{02} S_{10}^2 - \frac{6}{n} S_{10}^2 S_{01}^2 \right]$$

where $S_{rt} = \Sigma Z_{2i}^r r_{Z_i}^t$.

The computational form of k_{22} is:

$$k_{22} = \frac{1}{(n-1)(n-2)(n-3)} \left\{ n(n+1) \Sigma Z_{1j}^2 \right.$$

$$- 2(n+1) \left[(\Sigma Z_{2i} Z_{1i})(\Sigma r_{Z_i}) + (\Sigma Z_{1i} r_{Z_i})(\Sigma Z_{2i}) \right]$$

$$- (n-1)\left[\Sigma Z_{2i}^2 \Sigma r_{Z_i}^2 + 2(\Sigma Z_{1i})^2\right]$$

$$+ 8(\Sigma Z_{1i})(\Sigma Z_{2i})(\Sigma r_{Z_i}) + 2(\Sigma Z_{2i}^2)(\Sigma r_{Z_i})^2 + 2(\Sigma r_{Z_i}^2)(\Sigma Z_{2i})^2$$

$$- \frac{6}{n}(\Sigma Z_{2i})^2(\Sigma r_{Z_i})^2\Big\},$$

since $S_{11} = \Sigma\left(Z_{2i}\ r_{Z_i}\right) = \Sigma\left(Z_{2i}\ \frac{Z_{1i}}{Z_{2i}}\right) = \Sigma\ Z_{1i}$,

$$S_{22} = \Sigma\left(Z_{2i}^2\ r_{Z_i}^2\right) = \Sigma\left(Z_{2i}^2\ \frac{Z_{1i}^2}{Z_{2i}^2}\right) = \Sigma\ Z_{1i}^2,$$

$$S_{12} = \Sigma\left(Z_{2i}\ r_{Z_i}^2\right) = \Sigma\left(Z_{2i}\ \frac{Z_{1i}^2}{Z_{2i}^2}\right) = \Sigma\left(Z_{1i}\ \frac{Z_{ij}}{Z_{2i}}\right) = \Sigma\left(Z_{1i}\ r_{Z_i}\right),$$

and $S_{21} = \left(\Sigma\ Z_{2i}^2\ r_{Z_i}\right) = \Sigma\left(Z_{2i}^2\ \frac{Z_{1i}}{Z_{2i}}\right) = \Sigma\left(Z_{2i}\ Z_{1i}\right)$.

In this chapter, an unbiased ratio-type estimator was found for μ_1. So that in the event that μ_2 is known, this extra information can be used to obtain a better estimate of μ_1 than is possible with just using information about the sensitive characteristic X_1 alone. The exact

variance of the estimator μ_1 was also found. It should be noted that since $\bar{r}' = \tilde{\mu}_1/\mu_2$, that the exact variance of $\bar{r}'$ was essentially also obtained. Also by using bivariate cumulants and k-statistics, unbiased estimators of these variances were also obtained.

CHAPTER IV

FINITE POPULATIONS

In the previous chapters, an infinite population size has been assumed. In this chapter, the effects on the estimators of sampling without replacement from a finite population of N elements will be investigated. The interviewing scheme and notation of Case II will be used here, i.e., both numerator and denominator characteristics of interest are sensitive.

The only change in notation required is that X_1 and X_2 now have discrete probability distributions and therefore will be labeled $P_1(X_1)$ and $P_2(X_2)$ respectively. The probability density function for the numerator is

$$\psi_1(Z_1) = p_1 P_1(X_1) + (1-p_1) g_1(Y_1)$$

and for the denominator

$$\psi_2(Z_2) = p_2 P_2(X_2) + (1-p_2) g_2(Y_2)$$

where $g_1(Y_1)$ and $g_2(Y_2)$ are again the probability density functions of the nonsensitive characteristics in the numerator and denominator respectively.