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
© Copyright by Reider Sverre Peterson (1974)

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.

Consider the random variable

$$Z_1 = p_1 I_{X_1} X_1 + (1-p_1) I_{Y_1} Y_1$$

where I_{X_1} is an indicator variable equal to one if the sensitive question is selected, zero otherwise, and I_{Y_1} is one if the nonsensitive question is selected, zero otherwise. Using these assumptions, Z_1 is a random variable which is a mixture of a discrete and a continuous random variable. Hence,

$$\mu_{Z_1} = E_1(Z_1) = p_1 E_{p(X_1)} X_1 + (1-p_1) E_{g(Y_1)} Y_1,$$

where the first expectation is over the discrete probability distribution $P(X_1)$ and the second over the continuous probability density function $g(Y_1)$. Therefore,

$$\mu_{Z_1} = p_1 \sum_{i=1}^{N} X_{1i} P_1(X_{1i}) + (1-p_1) \int Y_1 g_1(Y_1) dY_1$$

$$= p_1 \mu_1 + (1-p_1) \mu_{Y_1},$$

which is the same result as when both numerator populations were considered to be infinite. Similarly, for the denominator, it may be shown that:

$$\mu_{Z_2} = p_2\mu_2 + (1-p_2)\mu_{Y_2}.$$

Hence, as before,

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

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

and

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

Suppose a simple random sample of n observations is drawn without replacement from the finite population. Then

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

and

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

where $\bar{Z}_1$ and $\bar{Z}_2$ are again sample means for the numerator and denominator respectively, are unbiased estimates of μ_1 and μ_2. Thus,

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

is a biased estimate of the ratio, R.

The derivation of $E(\hat{R})$ exactly parallels that for the infinite population case and the approximation is,

$$E(\hat{R}) \doteq R\left[1 + k_2^2\,\sigma^2_{\bar{Z}_2} - k_1 k_2\,\mathrm{Cov}(\bar{Z}_1,\bar{Z}_2)\right] .$$

From standard well known results for finite populations with simple random sampling,

$$\mathrm{Cov}(\bar{Z}_1,\bar{Z}_2) = \frac{N-n}{n(N-1)} \quad \mathrm{Cov}(Z_1,Z_2) = \sigma_{Z_1Z_2}$$

and

$$\sigma^2_{\bar{Z}_2} = \frac{N-n}{n(N-1)} \quad \sigma^2_{Z_2} .$$

Hence,

$$E(\hat{R}) \doteq R\left\{1 + \frac{N-n}{n(N-1)}\left[\frac{\sigma^2_{Z_2}}{p_2^2\mu_2^2} - \frac{\sigma_{Z_1Z_2}}{p_1p_2\mu_1\mu_2}\right]\right\}$$

and the $MSE(\hat{R})$ would be approximately:

$$MSE(\hat{R}) \doteq \frac{R^2}{n}\,\frac{N-n}{N-1}\left\{\frac{\sigma^2_{Z_1}}{p_1^2\mu_1^2} + \frac{\sigma^2_{Z_2}}{p_2^2\mu_2^2} - \frac{2\,\sigma_{Z_1Z_2}}{p_1p_2\mu_1\mu_2}\right\}.$$

To obtain the unbiased ratio-type estimator for simple random sampling in a finite population,

let $r_{Z_i} = \frac{Z_{1i}}{Z_{2i}}$ and $\bar{r}_Z = \frac{1}{n} \Sigma_{i=1}^{n} \frac{Z_{1i}}{Z_{2i}}$.

Now consider:

$$\frac{1}{N} \Sigma_{i=1}^{N} r_{Z_i}(Z_{2i} - \mu_{Z_2}) = \frac{1}{N} \Sigma_{i=1}^{N} \left(\frac{Z_{1i}}{Z_{2i}} Z_{2i} \right) - \frac{1}{N} \Sigma_{i=1}^{N} \left(\frac{Z_{1i}}{Z_{2i}} \mu_{Z_2} \right)$$

$$= \frac{1}{N} \Sigma_{i=1}^{N} Z_{1i} - \mu_{Z_2} \left(\frac{1}{N} \Sigma_{i=1}^{N} \frac{Z_{1i}}{Z_{2i}} \right)$$

$$\doteq \mu_{Z_1} - \mu_{Z_2} E(r_{Z_i}) \text{ if } N \text{ is large.}$$

But $E(r_{Z_i}) = E(\bar{r}_Z)$ in simple random sampling, hence

$$\mu_{Z_1} - \mu_{Z_2} E(r_{Z_i}) = \mu_{Z_1} - \mu_{Z_2} E(\bar{r}_Z) = \mu_{Z_2}(R_Z - E(\bar{r}_Z)),$$

where $R_Z = \mu_{Z_1}/\mu_{Z_2}$.

Thus the bias in $\bar{r}_Z = E(\bar{r}_Z) - R_Z$

$$= - \frac{1}{N\mu_{Z_2}} \Sigma_{i=1}^{N} r_{Z_i}(Z_{2i} - \mu_{Z_2}) \qquad (4)$$

For simple random sampling, an unbiased estimate of

$$\frac{1}{N-1} \Sigma_{i=1}^{N} r_{Z_i}(Z_{2i} - \mu_{Z_2})$$

is

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

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

Substituting this into (4), the unbiased estimate of R_Z is: $\bar{r}'_Z = \bar{r}_Z + \frac{n(N-1)}{N(n-1)\mu_{Z_2}}(\bar{Z}_1 - \bar{r}_Z\bar{Z}_2)$. The corresponding unbiased estimate of the population total of Z_1 (numerator total for the population) is:

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

Now since $\mu_{Z_1} = R_Z\, \mu_{Z_2}$, R can be written as:

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

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

Hence the unbiased estimator of R is:

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

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

where $\mu_{Z_2} = p_2 \mu_2 + (1-p_2)\mu_{Y_2}$.

Thus the unbiased estimate of the total for the sensitive question in the numerator is:

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

Most of the results obtained in the infinite population case carry over to the finite population case by supplying the finite population correction factor in the appropriate places in the estimators.

CHAPTER V

ASYMPTOTIC DISTRIBUTION OF $\hat{R}$ AND CONFIDENCE INTERVALS

To obtain the asymptotic distribution of the estimator R, we need the following.

Theorem ii), sec. 6a.2, page 387 in Rao [13]:

"Let T_n be a k-dimensional statistic $(t_{1n}, t_{2n}, \ldots, t_{kn})$ such that the asymptotic distribution of $\sqrt{n}(t_{1n} - \theta_1)$, $\sqrt{n}(t_{2n} - \theta_2)$, ..., $\sqrt{n}(t_{kn} - \theta_k)$ is a k-variate normal with mean $\underline{0}$ and dispersion matrix $\Sigma = (\sigma_{ij})$. Let g be a function of k variables which is totally differentiable. Then the asymptotic distribution of $u = \sqrt{n}\left[g(t_{1n}, t_{2n}, \ldots, t_{kn}) - g(\theta_1, \theta_2, \ldots, \theta_k)\right]$ is normal with mean 0 and $v(\theta) = \Sigma_i \Sigma_j \sigma_{ij} \frac{\partial g}{\partial \theta_i} \frac{\partial g}{\partial \theta_j}$."

To apply this theorem to the estimator R, associate

t_{1n} with $\bar{Z}_1 - (1-p_1)\mu_{Y_1}$, θ_1 with $\mu_{Z_1} - (1-p_1)\mu_{Y_1}$,

t_{2n} with $\bar{Z}_2 - (1-p_2)\mu_{Y_2}$, θ_2 with $\mu_{Z_2} - (1-p_2)\mu_{Y_2}$.

Then,

$$\sqrt{n}(t_{1n} - \theta_1) = \sqrt{n}\left\{\bar{Z}_1 - (1-p_1)\mu_{Y_1} - \left[\mu_{Z_1} - (1-p_1)\mu_{Y_1}\right]\right\}$$

$$= \sqrt{n}(\bar{Z}_1 - \mu_{Z_1}),$$

$$\sqrt{n}(t_{2n} - \theta_2) = \sqrt{n}\left\{\bar{Z}_2 - (1-p_2)\mu_{Y_2} - \left[\mu_{Z_2} - (1-p_2)\mu_{Y_2}\right]\right\},$$

$$= \sqrt{n}(\bar{Z}_2 - \mu_{Z_2}).$$

By the multivariate Central Limit Theorem, these variates have a limiting distribution which is a bivariate normal with 0 means and dispersion matrix given by

$$\sum = \begin{pmatrix} \sigma^2_{Z_1} & \sigma_{Z_1Z_2} \\ \sigma_{Z_1Z_2} & \sigma^2_{Z_2} \end{pmatrix}$$

The function $g(t_{1n}, t_{2n})$ is the estimator $\hat{R}$, so that

$$g(t_{1n}, t_{2n}) = \frac{p_2\left[\bar{Z}_1 - (1-p_1)\mu_{Y_1}\right]}{p_1\left[\bar{Z}_2 - (1-p_2)\mu_{Y_2}\right]}$$

and for the parameters,

$$g(\theta_1, \theta_2) = \frac{p_2\left[\mu_{Z_1} - (1-p_1)\mu_{Y_1}\right]}{p_1\left[\mu_{Z_2} - (1-p_2)\mu_{Y_2}\right]}.$$

Hence,

$$u = \sqrt{n}\left[g(t_{1n}, t_{2n}) - g(\theta_1, \theta_2)\right]$$

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

$$= \sqrt{n} \, (\hat{R} - R)$$

Differentiating $g(\theta_1, \theta_2)$ with respect to the parameters gives

$$\frac{\partial g}{\partial \theta_1} = \frac{\partial g}{\partial \mu_{Z_1}} = \frac{p_2}{p_1} \cdot \frac{1}{\mu_{Z_2} - (1-p_2)\mu_{Y_2}} = \frac{1}{p_1} \frac{1}{\dfrac{\mu_{Z_2} - (1-p_2)\mu_{Y_2}}{p_2}}$$

$$= \frac{1}{p_1 \mu_2}$$

and

$$\frac{\partial g}{\partial \theta_2} = \frac{\partial g}{\partial \mu_{Z_2}} = \frac{p_2}{p_1} \left[- \frac{\mu_{Z_1} - (1-p_1)\mu_{Y_1}}{(\mu_{Z_2} - (1-p_1)\mu_{Y_1})^2} \right] = - \frac{\mu_1}{p_2 \mu_2^2} \, .$$

Therefore, the asymptotic variance is

$$v(\theta) = \sigma_{Z_1}^2 \left(\frac{\partial g}{\partial \mu_{Z_1}} \right)^2 + \sigma_{Z_2}^2 \left(\frac{\partial g}{\partial \mu_{Z_2}} \right)^2 + 2 \, \sigma_{Z_1 Z_2} \frac{\partial g}{\partial \mu_{Z_1}} \frac{\partial g}{\partial \mu_{Z_2}}$$

$$= \frac{\sigma_{Z_1}^2}{p_1^2 \mu_2^2} + \frac{\mu_1^2 \sigma_{Z_2}^2}{p_2^2 \mu_2^4} - \frac{2 \, \mu_1 \, \sigma_{Z_1 Z_2}}{p_1 p_2 \, \mu_2^3} \, . \qquad (5)$$

Thus the asymptotic distribution of $u = \sqrt{n}\,(\hat{R} - R)$ is normal with mean 0 and variance $v(\theta)$, i.e., $\sqrt{n}(R - R) \xrightarrow[n\to\infty]{} N(0, v(\theta))$, where $v(\theta)$ is given by (5).

Confidence Intervals For $R = \mu_1/\mu_2$

Two methods of setting confidence intervals for the ratio of means will be considered:

i) use of Fieller's Theorem [6]

and

ii) use of the Jackknife Technique [12].

Method I (Normal Case). The following is Fieller's Theorem.

Let $T \sim MVN(\tau, \Gamma)$ where $T = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$, $\tau = \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix}$ and $\Gamma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$. Let $S \sim$ Wishart (k, Γ), independent of T, where $S = \begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}$. Then $100(1 - \alpha)\%$ confidence limits for $\theta = \mu_1/\mu_2$ are given by:

i) $\theta \in (\theta_L, \theta_U)$ if $1 - g > 0$;

ii) $\theta < \theta_L$ or $\theta > \theta_U$ if $1 - g < 0$, θ_L, θ_U are real; and

iii) $\theta \in (-\infty, \infty)$ if the roots θ_L, θ_U are imaginary, and $1 - g < 0$, where θ_L, θ_U, and g are given by

$$\theta_L = \left[Q_{12} - \left(Q_{12}^2 - Q_1 Q_2\right)^{1/2}\right]/Q_2,$$

$$\theta_U = \left[Q_{12} + \left(Q_{12}^2 - Q_1 Q_2\right)^{1/2}\right]/Q_2,$$

$$g = t^2 s_2^2/n \bar{T}_2^2$$

and in θ_L and θ_U,

$$Q_1 = \bar{T}_1^2 - t^2 s_1^2/n,$$

$$Q_2 = \bar{T}_2^2 - t^2 s_2^2/n,$$

$$Q_{12} = \bar{T}_1 \bar{T}_2 - t^2 s_{12}/n.$$

Fieller's Theorem is based on normal theory and the robustness of the confidence interval developed below is investigated by a Monte Carlo study in Chapter VI.

To apply these results to the estimators in R, assume that Z'_{1i} and Z'_{2i} are normally distributed, i.e., that

$$\begin{pmatrix} Z'_{1i} \\ Z'_{2i} \end{pmatrix} = \begin{pmatrix} \left[Z_{1i} - (1-p_1)\mu_{Y_1}\right]/p_1 \\ \left[Z_{2i} - (1-p_2)\mu_{Y_2}\right]/p_2 \end{pmatrix}$$

$$\sim \text{MVN}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2_{Z_1}/p_1^2 & \sigma_{Z_1Z_2}/p_1p_2 \\ \sigma_{Z_1Z_2}/p_1p_2 & \sigma^2_{Z_2}/p_2^2 \end{pmatrix}\right] .$$

The means and variances can be justified by noting the following:

$$E(Z'_{ji}) = E\left[(Z_{ji} - (1-p_j)\mu_{Y_j})/p_j\right] = (\mu_{Z_j} - (1-p_j)\mu_{Y_j})/p_j$$

$$= \mu_j, \text{ for } i = 1 \text{ or } 2,$$

$$\text{Var}(Z'_{ji}) = E\left[\frac{1}{p_j}(Z_j - (1-p_j)\mu_{Y_j}) - \frac{1}{p_j}(\mu_{Z_j} - (1-p_j)\mu_{Y_j})\right]^2$$

$$= \frac{1}{p_j^2} E\left[(Z_j - \mu_{Z_j})^2\right]$$

$$= \sigma^2_{Z_j}/p_j^2, \text{ for } j = 1 \text{ or } 2,$$

and

$$\text{Cov}(Z_{1j}, Z_{2i}) = E\left[\frac{1}{p_1}(Z_1 - \mu_{Z_1}) \cdot \frac{1}{p_2}(Z_2 - \mu_{Z_2})\right]$$

$$= \frac{1}{p_1 p_2} \operatorname{Cov}(Z_1, Z_2)$$

$$= \sigma_{Z_1 Z_2}/p_1 p_2 \cdot$$

Hence $100(1 - \alpha)\%$ confidence limits for $R = \mu_1/\mu_2$ are given by:

$$\theta_L = \left[Q_{12} - (Q_{12}^2 - Q_1 Q_2)^{1/2}\right]/Q_2$$

$$\theta_U = \left[Q_{12} + (Q_{12} - Q_1 Q_2)^{1/2}\right]/Q_2$$

where, for the estimators given here,

$$Q_1 = \bar{Z}'_{1i} - t^2 s^2_{Z_1}/p_1^2 n = \left[(\bar{Z}_1 - (1-p_1)\mu_{Y_1})^2 - t^2 s^2_{Z_1}/n\right]/p_1^2$$

$$Q_2 = \bar{Z}'_{2i} - t^2 s^2_{Z_2}/p_2^2 n = \left[(\bar{Z}_2 - (1-p_2)\mu_{Y_2})^2 - t^2 s^2_{Z_2}/n\right]/p_2^2$$

$$Q_{12} = \bar{Z}'_{1i}\bar{Z}'_{2i} - t^2 \widehat{\operatorname{Cov}(Z_1, Z_2)}/n\, p_1 p_2$$

$$= (\bar{Z}_1 - (1-p_1)\mu_{Y_1})(\bar{Z}_2 - (1-p_2)\mu_{Y_2})$$

$$- t^2 \widehat{\operatorname{Cov}(Z_1 Z_2)}/n\, p_1 p_2$$

and t is the upper $\alpha/2$ value of the student's t distribution with $n - 1$ degrees of freedom. It should be noted that θ_L and θ_U give the $100(1 - \alpha)\%$ confidence limits only if the

quantity $1 - t^2 s^2_{Z_2}/n \bar{Z}'_2 = 1 - g$ is greater than zero.

The confidence interval considered above has been developed for the case where the sampling is done from normal populations. The Jackknife Method is another procedure of obtaining confidence intervals which has been shown to be quite robust; that is, it provides the expected confidence levels even when the populations are not normal. A short outline of the general Jackknife procedure will be given first.

Let $\hat{\theta}$ be the estimator (biased or unbiased) of the unknown paramenter θ. The entire sample of size n is divided into r groups each of size k, i.e., $n = rk$. $\hat{\theta}$ is then the estimate of θ computed from all n observations. Now let $\hat{\theta}_{-i}$, $i = 1, 2, \ldots, r$ denote the same estimate of θ, but computed from all the observations except for the ith group, i.e., delete the ith group and compute the estimate of θ on the remaining n-k observations. Then find the "psuedo-values" $\tilde{\theta}_i = r\hat{\theta} - (r-1)\hat{\theta}_{-i}$, $i = 1, 2, \ldots, r$. The Jackknife estimate is the mean of the $\tilde{\theta}_i$'s, i.e.,

$$\hat{\theta}_J = \frac{1}{r} \Sigma_{i=1}^{r} \tilde{\theta}_i = \frac{1}{r} \Sigma_{i=1}^{r} (r\hat{\theta} - (r-1)\hat{\theta}_{-i})$$

$$= r\hat{\theta} - \frac{r-1}{r} \Sigma_{i=1}^{r} \hat{\theta}_{-i} .$$

The Jackknife estimate is useful for biased estimates, which ratio estimates invariably are, since it eliminates the 1st order bias term, when expanded as a function of the sample size, i.e., if $E\hat{\theta} = \theta + a/n + O(1/n^2)$, then $E(\hat{\theta}_J) = \theta + O(1/n^2)$.

Method II. To obtain confidence limits using the Jackknife procedure, the variance of the pseudo-values are used as an estimate of the variance of the Jackknife estimate of $\hat{R}$. That is, find

$$s_J^2 = \Sigma_{i=1}^{r} (\tilde{\theta}_i - \bar{\tilde{\theta}}.)^2/r(r-1)$$

$$= \left[r \Sigma_{i=1}^{r} \tilde{\theta}_i^2 - (\Sigma \tilde{\theta}_i)^2 \right] /r(r-1),$$

where $\tilde{\theta}_i$ are the psuedo-values discussed previously. In terms of the parameter R and its estimator $\hat{R}$, $\tilde{\theta}_{-i} = \tilde{\hat{R}}_{-i}$ is the estimator $\hat{R}$ but computed from all observations except those in the ith group. $\tilde{\theta}_i = \tilde{\hat{R}}_i$ are the pseudo-values, i.e. $\tilde{\hat{R}}_i = r\hat{R} - (r-1)\tilde{\hat{R}}_{-i}$, and $\tilde{\theta}_J = \hat{R}_J$ where $\hat{R}_J = \frac{1}{r} \Sigma_{i=1}^{r} \tilde{\hat{R}}_i$ is

the Jackknife estimate of R. Then the $100(1 - \alpha)\%$ confidence limits for the parameter R are given by:

$$R_L^J = \hat{R}_J - t_{\alpha/2,r-1}\, s_J$$

$$R_U^J = \hat{R}_J + t_{\alpha/2,r-1}\, s_J$$

where $t_{\alpha/2,r-1}$ is the upper value of the student's t with r-1 degrees of freedom.

This procedure for obtaining confidence intervals using the Jackknife technique is straightforward, but computationally lengthy. The robustness of the Jackknife technique is also investigated in the Monte Carlo study discussed in Chapter 6.

CHAPTER VI

MONTE-CARLO STUDY

The Monte-Carlo study was done using a uniform [0,1] pseudo-random number generator to indicate which distribution (sensitive or nonsensitive) to sample for each observation. This procedure was used to simulate as closely as possible the real-world situations. The uniform random number generator was also used to obtain the observations by first generating a uniform number and then using the Box-Mueller transformation to obtain a normal deviate. These numbers were truncated at four standard deviations from the mean so that the numbers obtained had no real outliers.

There are essentially thirteen types of variables (not counting different populations) that could be altered, making the number of different runs that were possible very large. An attempt was made to discover what change in the parameters, or combination of changes, would produce the most striking results, whether desirable or undesirable. Since computer time seemed to be mostly a function of r = NJCK, which is the number of times the same estimator is computed leaving out k = KJCK observations at a time,

most of the runs of the first set (Normal distributions) were run at a relatively small value. Note that the total sample size is equal to r · k = NJCK · KJCK.

The runs were separated into three groups, the order in this paper being the order in which they were run. Thus by observing the results of each set of runs, hopefully a more intelligent design was obtained for the next set of runs. The three sets are: Normal distributions ("large" samples), Normal distributions (small samples), Chi-Squared distributions. A table of the parameters used for each run is given. Only when the parameter was changed was an entry made in the table. Hopefully, this will facilitate discerning which parameter(s) were altered for each run. Also, a table of results is given for each set. Each run consisted of N = 100 samples, each sample being of size r · k = NJCK · KJCK.

The column headings of the results are as follows. The second column is the true ratio of means, $R = \mu_1/\mu_2$, $\hat{R}$ and $\hat{R}_J$, columns three and four, being the estimates of R from the entire sample and the Jackknife technique respectively.

The fifth column is the approximate theoretical mean-squared error of the estimator $\hat{R}$; approximate because of the truncation of the terms in its derivation. The sixth column is the mean over the N = 100 samples of the squared deviations of $\hat{R}$ from the actual value of R; column seven, the mean of the squared deviations of $\hat{R}$ from the mean of $\hat{R}$ for the N = 100 samples. Column eight is the mean of the estimated mean-squared error of the estimator $\hat{R}$. If one were to actually use $\hat{R}$ to estimate R, one of these (column eight) would be used as an estimate of $MSE(\hat{R})$, so this is one of the columns that should be studied quite closely. Column nine is the same type of quantity as is found in column seven, only for the Jackknife estimate, and column ten is the counterpart of column eight, again for the Jackknife estimate.

Columns eleven, twelve and thirteen are respectively, the confidence coefficient; the fraction (out of the N = 100) of confidence intervals that bracket R using the Jackknife (column twelve) and Fieller's Theorem (column thirteen). The explanation under Fieller's Theorem for the Chi-Squared distributions will be given in that section.

Finally, column fourteen is the real value of μ_1, the mean of the sensitive question in the numerator and the mean of the estimates of μ_1 in column fifteen. Column sixteen is the mean of the squared deviations of the estimate $\hat{\mu}_1$ from the mean of the estimates, and in column seventeen, the mean of the variance estimates.

The same starting number (14689) for the random number generator was used throughout except in run number eight of the first set. For large sample sizes, different starting numbers should not make much difference, but for small sample sizes, the effect could be quite large.

PARAMETERS FOR MONTE-CARLO STUDY
NORMAL DISTRIBUTIONS

Run No.	XPN = p_1	XPD = p_2	XMN1 = μ_{Y_1}	XMN2 = μ_1	XMN3 = μ_{Y_2}	XMN4 = μ_2	VN1 = $\sigma^2_{Y_1}$	VN2 = σ^2_1	VN3 = $\sigma^2_{Y_2}$	VN4 = σ^2_2	KJCK = k	NJCK = r	TDF = t_{n-1}	1 - α
1	.6	.6	50	50	10	10	4	4	4	4	1	100	1.96	.95
2											1	100	1.645	.90
3											1	200	1.96	.95
4											4	100		
5											4	25		
6											5	20		
7											5	40		
8											4	25		
9				40										
10	.8													
11		.6		25										
12	.8													
13	.9													
14				50	20									
15		.8												
16		.6	60	60		30								
17		.8												
18		.6			30		16							
19							4	16						
20		.8												
21		.6						4		16				
22		.8												
23		.6					16		16	4				
24							4	16	4	16				
25		.8												
26		.6	40		50									
27	.8	.8												
28	.6	.6					16	36	36	16				

MONTE CARLO STUDY - NORMAL DISTRIBUTIONS

Run No.	R	$\bar{\hat{R}}$	$\bar{\hat{R}}_J$	Approx. Theor. $MSE(\hat{R})$	$\frac{\Sigma(\hat{R}-R)^2}{N}$	$\frac{\Sigma(\hat{R}-\bar{\hat{R}})^2}{N}$	$\widehat{MSE\ R}$	$\frac{\Sigma(\hat{R}_J-\bar{\hat{R}}_J)^2}{N}$	$\widehat{MSE(R_J)}$	$1-\alpha$ Nominal Value	$\widehat{1-\alpha}$ Using Jackknife	$\widehat{1-\alpha}$ Using Fieller's Theorem	μ_1	$\bar{\hat{\mu}}_1$	$\frac{\Sigma(\hat{\mu}_1-\bar{\hat{\mu}}_1)^2}{N}$	$\widehat{Var(\hat{\mu}_1)}$
1	5	5.0133	5.0073	.02889	.03172	.03242	.03035	.03207	.03035	.95	.95	.96	50	50.078	3.4773	3.2854
2	5	5.0133	5.0073	.02889	.03172	.03242	.03035	.03207	.03035	.90	.92	.92	50	50.078	3.4773	3.2854
3	5	5.0165	5.0135	.01444	.01700	.01744	.01497	.01734	.01496	.95	.95	.93	50	50.140	1.8453	1.6240
4	5	5.0054	5.0040	.00722	.00921	.00976	.00737	.00977	.00746	.95	.91	.95	50	50.041	1.0442	.8025
5	5	5.0133	5.0072	.02889	.03172	.03242	.03035	.03220	.03049	.95	.95	.96	50	50.078	3.4773	3.2854
6	5	5.0133	5.0073	.02889	.03172	.03242	.03035	.03215	.02988	.95	.92	.96	50	50.078	3.4773	3.2854
7	5	5.0165	5.0135	.01444	.01700	.01742	.01497	.01736	.01497	.95	.93	.93	50	50.140	1.8491	1.6236
8	5	4.9969	4.9912	.02889	.03130	.03210	.02903	.03192	.02860	.95	.91	.92	50	49.908	3.4590	3.1431
9	4	4.0075	4.0047	.01889	.02915	.02058	.01813	.02058	.01764	.95	.93	.95	40	40.069	3.5650	3.2211
10	4	4.0032	3.9984	.01840	.02476	.02531	.02178	.02513	.02182	.95	.92	.94	40	39.975	1.8277	1.5672
11	2.5	2.5074	2.5047	.00806	.05479	.05525	.05005	.05512	.04877	.95	.93	.95	25	25.055	6.5178	5.8031
12	2.5	2.4884	2.4854	.00757	.03005	.03018	.02323	.03014	.02269	.95	.90	.91	25	24.850	2.9169	2.2298
13	2.5	2.4957	2.4927	.00744	.01638	.01649	.01449	.01643	.01439	.95	.92	.92	25	24.919	1.3098	1.1422
14	5	5.0503	5.0108	.02889	.18859	.18809	.21070	.18032	.22033	.95	.96	.97	50	50.132	12.562	12.630
15	5	5.0690	5.0529	.01674	.07769	.07420	.08182	.07374	.08233	.95	.96	.98	50	50.670	11.083	12.471
16	2	2.0017	1.9998	.00062	.00375	.00375	.00364	.00373	.00383	.95	.92	.96	60	59.994	4.9508	4.6943
17	2	1.9951	1.9943	.00040	.00129	.00126	.00145	.00125	.00154	.95	.97	.97	60	59.746	2.4293	2.7619
18	2	2.0008	2.0005	.00076	.00091	.00090	.00078	.00090	.00081	.95	.88	.93	60	60.017	.8876	.7097
19	2	2.0025	2.0022	.00083	.00090	.00088	.00086	.00089	.00087	.95	.93	.95	60	60.067	.8763	.7753
20	2	2.0022	2.0020	.00062	.00068	.00066	.00063	.00065	.00064	.95	.93	.96	60	60.067	.8763	.7753
21	2	2.0018	2.0010	.00151	.00171	.00171	.00153	.00170	.00154	.95	.95	.95	60	60.033	1.6465	1.4081
22	2	2.0019	2.0014	.00107	.00113	.00111	.00111	.00111	.00111	.95	.95	.96	60	60.055	1.8239	1.7312
23	2	2.0021	2.0015	.00136	.00150	.00149	.00143	.00148	.00146	.95	.90	.93	60	60.046	1.4287	1.3078
24	2	2.0029	2.0021	.00173	.00194	.00193	.00173	.00193	.00177	.95	.92	.94	60	60.064	1.8497	1.6123
25	2	2.0030	2.0024	.00129	.00133	.00131	.00133	.00131	.00135	.95	.96	.96	60	60.086	1.9754	1.9355
26	2	2.0108	2.0037	.00173	.01536	.01538	.01696	.01514	.01757	.95	.95	.94	60	60.110	8.8359	9.4609
27	2	2.0224	2.0917	.00118	.00662	.00616	.00682	.00616	.00687	.95	.94	.96	60	60.543	4.5827	4.9126
28	2	2.0138	2.0059	.00383	.01743	.01740	.01924	.01713	.01991	.95	.96	.95	60	60.164	10.195	10.838

Normal Distributions

In the first set, twenty-eight runs were made with a relatively large sample size: greater than or equal to 100. The first eight runs were made with equal numerator means (50), equal denominator means (10) and all variances equal (4). The purpose of these was to determine if the sample size and/or Jackknifing numbers seemed to make a significant difference. A confidence coefficient of .95 was used almost exclusively (in all three sets of runs), with an occasional .90 level being tried.

A different starting number for the random number generator was used in run number eight (86771), so that comparing this run with run number five indicates the effect that the starting number might have. For these two, it made a considerable difference: negative rather than positive bias in estimating R, and a reduction in the fraction of confidence intervals that actually bracketed R using both the Jackknife and Fieller's Theorem.

Different Jackknifing numbers did not seem to have much effect on reducing the (average) bias, and as a general trend, did not seem to reduce the bias very much over the regular estimate. Notice also that when the bias was

negative, in run numbers eight and twelve, the Jackknife actually increased the bias.

In run number eighteen, where the variance of the non-sensitive distribution in the numerator was quadrupled, the Jackknife estimate fared badly (at least for the particular starting number and Jackknifing numbers used) for the confidence coefficient.

The means in the denominator were initially so small that the variances could not be increased without generating observations close to zero. This resulted in unrealistically large ratios. To overcome this problem, the mean in the denominator was increased and the mean in the numerator was also increased in order to provide an integer value for the theoretical ratio. This change in means following the first thirteen runs allowed investigation of unequal variances while preserving the invested computer time used in obtaining the preliminary results. It is reasonable that the relative changes in both the numerator and denominator means did not affect the relevant comparisons.

In runs eleven through seventeen and again in twenty-six, twenty-seven and twenty-eight, there is a big discrep-

ancy between the theoretical $MSE(\hat{R})$ and all of the other indicators of what the $MSE(\hat{R})$ really is. This discrepancy ranges from a factor of approximately four to seven. If the actual value of $MSE(\hat{R})$ is really as low as the theoretical $MSE(\hat{R})$ indicates, then one would expect the confidence coefficients to be almost always one, since these confidence intervals are set using $\widehat{MSE(\hat{R})}$ and $\widehat{MSE(\hat{R}_J)}$. Since this is not the case, the conclusion must be that the terms not included in the approximation will contribute substantially to the theoretical $MSE(\hat{R})$ for these parameter values. This large discrepancy occurs when the denominator and/or numerator means are very different. If the numerator means were not too different (50 versus 40 in runs nine and ten), then the large discrepancy did not appear. But when the actual difference is larger (runs eleven, twelve and thirteen), then the large discrepancy between the theoretical $MSE(\hat{R})$ and its estimates appears. The denominator means differed by ten units in runs fourteen and fifteen which is the same difference as the numerator means in runs nine and ten. However, the ten unit difference in the denominator is a much larger relative (to the mean values) difference. Increasing the

probability of sampling from the distribution(s) of interest (the sensitive distribution) did reduce the discrepancy somewhat, but they were still quite large. Increasing the variances so that the observations in the numerator and/or denominator could overlap, did decrease the discrepancy. For instance, in run twenty-eight, the factor was only two.

It is also rather interesting that increasing the value of p_1 and/or p_2 does not seem to increase the precision of estimation appreciably. As a matter of fact, increasing both p_1 and p_2 from run twenty-six to twenty-seven, decreased the precision from 2.0108 to 2.0224. But overall, increasing p_2 did seem to increase the confidence levels (runs sixteen to seventeen, nineteen to twenty, twenty-one to twenty-two and twenty-four to twenty-five). In runs eleven, twelve and thirteen, where p_1 was increased from .6 to .8 to .9, the confidence levels decreased from runs eleven to twelve and increased a little from twelve to thirteen. So it appears that even with a relatively small p_1 and p_2, say in the neighborhood of .6, that good results for confidence intervals are still obtained, even when the means and/or variances of the sensitive versus nonsensitive distributions are quite different. Thus it would appear

that matching the means is a big factor only in terms of the respondents assurance that the interviewer cannot discern the question by looking at the response, and that the value of p_1 and p_2 would be most important only in terms of the sample size that would be required.

It also would appear that the Jackknife is of little value if the populations are normal and the sample size is large. Generally, Fieller's Theorem gave better results, which should not be surprising, and since the Jackknife is rather lengthy computationally when compared with Fieller's Theorem, should probably not be used if the assumption of normal populations can be made.

PARAMETERS FOR MONTE-CARLO STUDY
NORMAL DISTRIBUTIONS
(Small Sample Sizes)

Run No.	XPN = p_1	XPD = p_2	XMN1 = μ_{Y_1}	XMN2 = μ_1	XMN3 = μ_{Y_2}	XMN4 = μ_2	VN1 = $\sigma^2_{Y_1}$	VN2 = σ^2_1	VN3 = $\sigma^2_{Y_2}$	VN4 = σ^2_2	KJCK = k	NJCK = r	TDF = t_{n-1}	1 - α
1	.6	.6	60	60	20	20	4	4	4	4	1	20	2.093	.95
2												10	2.262	
3							8	8	8	8		20	2.093	
4				50			4	4	4	4		10	1.833	.90
5												20	2.093	.95
6												10	2.262	
7						30								
8				60										
9			50											
10	.8	.8												
11												20	2.093	.95

MONTE CARLO STUDY - NORMAL DISTRIBUTIONS

(Small Sample Sizes)

Run No.	R	$\bar{\hat{R}}$	$\bar{\hat{R}}_J$	Approx. Theor. MSE($\hat{R}$)	$\frac{\Sigma(\hat{R}-R)^2}{N}$	$\frac{\Sigma(\hat{R}-\bar{\hat{R}})^2}{N}$	$\widehat{MSE\ R}$	$\frac{\Sigma(\hat{R}_J-\bar{\hat{R}}_J)^2}{N}$	$\widehat{MSE(\hat{R}_J)}$	a	b	c	μ_1	$\bar{\mu}_1$	$\frac{\Sigma(\hat{\mu}_1-\bar{\mu}_1)^2}{N}$	$\widehat{Var(\mu_1)}$
1	3	3.0122	3.0077	.01389	.01191	.01194	.01508	.01182	.01508	.95	.96	.97	60	60.164	4.86	6.06
2	3	3.0165	3.0076	.02778	.03107	.03118	.02987	.03052	.02996	.95	.95	.96	60	60.157	12.44	11.81
3	3	3.0193	3.0102	.02778	.02409	.02403	.03065	.02351	.03068	.95	.96	.97	60	60.226	9.84	12.42
4	3	3.0287	3.0103	.05556	.06391	.06377	.06202	.06108	.06240	.90	.91	.86	60	60.227	25.44	24.32
5	2.5	2.5158	2.5121	.01007	.01214	.01197	.01854	.01191	.01853	.95	.99	.99	50	50.259	5.19	8.03
6	2.5	2.5074	2.5003	.02014	.02984	.03005	.03613	.02971	.03615	.95	.98	.90	50	50.029	12.90	15.47
7	1.67	1.6679	1.6529	.00466	.03049	.03081	.03405	.02903	.03406	.95	.96	.97	50	49.507	40.11	42.88
8	2	2.0066	1.9882	.00617	.03852	.03885	.03955	.03627	.03940	.95	.94	.97	60	59.571	46.62	46.15
9	2	2.0110	1.9922	.00617	.05033	.05070	.04866	.04784	.04841	.95	.93	.93	60	59.713	51.73	48.33
10	2	1.9967	1.9899	.00347	.01691	.01705	.01750	.01637	.01700	.95	.94	.95	60	59.652	16.31	15.89
11	2	2.0000	1.9965	.00174	.00896	.00903	.00875	.00875	.00890	.95	.95	.95	60	59.873	8.84	8.66

.06*

Column a: 1 - α Nominal Value; Column b: 1 - α Using Jackknife; Column c: 1 - α Using Fieller's Theorem;

* Fraction of confidence intervals that cannot be constructed.

Normal Distributions (small sample sizes)

Since the values of p_1 and p_2 did not seem to have a large effect on the results on the first set of runs, values of .6 were used for both p_1 and p_2 in all runs except the last two. True to form, the increase from .6 to .8 does not seem to improve the estimation. The Jackknife method actually decreased in precision with a sample size of ten, while the regular estimate increased (runs nine and ten).

Again, the method using Fieller's Theorem did as well or better than the Jackknife in almost every case (notable exception in run four with $1 - \alpha = .90$), so that again it would be recommended that this method be used rather than the Jackknife.

Using greatly different variances did not affect the results in the first run set, so the same variances were used throughout. With a small sample size this could possibly have a more profound effect than is now suspected.

The best results using the Jackknife for other types of estimation is when $k = 1$. So for the small sample situation, this was the only value tried. Another reason for

using $k = 1$ for small sample sizes would be that the maximum number of degrees of freedom for making inferences could be used.

Again notice the large discrepancy between the theoretical $MSE(\hat{R})$ and its estimators in runs seven through eleven. In run number seven, the denominator means were made different and were so for the remainder of the runs.

The estimate of μ_1 suffers under a small sample size, since the $Var(\tilde{\mu}_1)$ increases drastically as the sample size decreases. This is not true in general of the estimator $\hat{R}$.

PARAMETERS FOR MONTE-CARLO STUDY
CHI-SQUARED DISTRIBUTIONS

Run No.	XPN = p_1	XPD = p_2	XMN1 = μ_{Y_1}	XMN2 = μ_1	XMN3 = μ_{Y_2}	XMN4 = μ_2	VN1 = $\sigma^2_{Y_1}$	VN2 = σ^2_1	VN3 = $\sigma^2_{Y_2}$	VN4 = σ^2_2	KJCK = k	NJCK = r	TDF = t_{n-1}	1 - α
1	.6	.6	82	82	50	50	326	326	198	198	1	20	2.093	.95
2											1	50	1.96	
3											2	25	1.96	
4			50	82	50	82	198	326	198	326	4	25	1.96	
5											1	20	2.093	
6	.8	.8									1	20	2.093	
7	.6	.6									2	50	1.96	
8											1	100	1.96	
9	.8	.8									2	50	1.96	
10	.6	.6	83	65	82	50	656	258	326	198	1	20	2.093	
11											1	50	1.96	
12											2	50	1.96	
13											2	100	1.96	
14	.8	.8									1	20	2.093	
15											1	50	1.96	
16											2	50	1.96	
17	.6	.6									2	50	1.645	.90
18	.6	.6									1	100	1.645	.90

MONTE CARLO STUDY - CHI-SQUARED DISTRIBUTIONS

Run No.	R	$\overline{\hat{R}}$	$\overline{\hat{R}_J}$	Approx. Theor. $MSE(\hat{R})$	$\frac{\Sigma(\hat{R}-R)^2}{N}$	$\frac{\Sigma(\hat{R}-\bar{\hat{R}})^2}{N}$	$\overline{MSE\ R}$	$\frac{\Sigma(\hat{R}_J-\bar{\hat{R}}_J)^2}{N}$	$\overline{MSE(\hat{R}_J)}$	a	b	Fieller's Theorem i	Fieller's Theorem ii	Fieller's Theorem iii	μ_1	$\bar{\hat{\mu}}_1$	$\frac{\Sigma(\hat{\mu}_1-\bar{\hat{\mu}}_1)^2}{N}$	$\overline{Var(\hat{\mu}_1)}$
1	1.64	1.6784	1.6580	.04770	.04533	.04308	.05358	.04239	.05454	.95	.97	.86	.14	1.00	82	83.056	118.53	143.41
2	1.64	1.6580	1.6503	.01908	.01754	.01740	.02010	.01706	.02022	.95	.98	0	.97	.97	82	82.594	48.18	55.11
3	1.64	1.6580	1.6503	.01908	.01754	.01740	.02010	.01712	.02041	.95	.98	0	.97	.97	82	82.594	48.18	55.11
4	1.00	1.0037	1.0014	.00227	.00489	.00493	.00443	.00488	.00464	.95	.96	0	.93	.93	82	82.115	37.77	33.92
5	1.00	1.0175	1.0051	.01135	.02301	.02294	.02465	.02194	.01158	.95	.96	.88	.11	.92	82	82.456	173.59	182.74
6	1.00	1.0136	1.0078	.00698	.01223	.01217	.01151	.01194	.01158	.95	.94	.70	.21	.93	82	82.585	88.64	84.46
7	1.00	1.0037	1.0015	.00227	.00489	.00493	.00443	.00489	.00451	.95	.93	0	.93	.93	82	82.115	37.77	33.92
8	1.30	1.0037	1.0015	.00227	.00489	.00493	.00443	.00489	.00443	.95	.94	0	.93	.93	82	82.115	37.77	33.92
9	1.00	1.0017	1.0006	.00140	.00217	.00220	.00217	.00219	.00223	.95	.95	0	.95	.95	82	82.025	16.34	16.18
10	1.30	1.3622	1.3203	.04657	.07481	.07168	.08845	.06390	.09295	.95	.98	.96	.04	1.00	65	66.366	156.53	191.34
11	1.30	1.3217	1.3061	.01863	.02838	.02820	.03231	.02686	.03280	.95	.94	.12	.84	.75	65	65.389	69.49	76.82
12	1.30	1.3104	1.3027	.00932	.01672	.01679	.01552	.16442	.01588	.95	.95	0	.93	.93	65	65.158	41.76	38.10
13	1.30	1.3070	1.3033	.00466	.00726	.00730	.00754	.00721	.00765	.95	.95	0	.95	.95	65	65.183	18.24	18.74
14	1.30	1.3394	1.3225	.02236	.02877	.02750	.03523	.02679	.03628	.95	.93	.02	.90	.92	65	65.393	37.68	35.36
15	1.30	1.3139	1.3074	.00894	.01431	.01427	.01337	.01411	.01351	.95	.93	.02	.90	.92	65	65.393	37.68	35.36
16	1.30	1.3100	1.3069	.00472	.00792	.00790	.00661	.00786	.00659	.95	.93	0	.94	.94	65	65.354	20.94	17.59
17	1.30	1.3104	1.3027	.00932	.01672	.01679	.01552	.01644	.01588	.90	.89	0	.89	.89	65	65.158	41.76	38.10
18	1.30	1.3104	1.3029	.00932	.01672	.01679	.01552	.01645	.01562	.90	.87	0	.89	.89	65	65.158	41.76	38.10

Column a: $1 - \alpha$ Nominal Value; Column b: $\widehat{1 - \alpha}$ Using Jackknife; Column i: fraction of confidence intervals that cannot be constructed;
Column ii: actual fraction of confidence intervals that bracket the mean; Column iii: fraction of confidence intervals that can be constructed which do bracket the mean.

Chi-Squared Distributions

In order to study the effects of having non-symmetric distributions, normal deviates were obtained and then transformed to a Chi-Squared distribution by adding three to each normal random number and then squaring the result [15].

The three columns under Fieller's Theorem in the results are: i) The fraction (out of the 100) samples that could not be constructed using Fieller's Theorem because of the quantity 1 - g being less than zero. It would be expected that this fraction is quite large, especially for small samples because of the non-normality of the distributions; however, this was not the case in run fourteen. ii) The fraction of confidence intervals out of the total of 100 that actually did bracket the true ratio. iii) The proportion of the confidence intervals that could be constructed that actually did bracket the ratio.

Both large and small sample sizes were tried with various values of the Jackknifing constants, r and k. In runs two and three, where the sample size was 100, letting k = 1 or 2 did not make any difference on the indicated confidence level, which was .98 for both. It appears that the use of the Jackknife method is a must for small sample

sizes, at least for these types of distributions. The confidence intervals based on Fieller's Theorem became better as the sample size increased, but had a larger indicated confidence level than the Jackknife only twice out of the eighteen runs.

The discrepancy between the approximate theoretical $MSE(\hat{R})$ and its various estimates again became greater when the means were different. However, this discrepancy never did become nearly as large as for the first two sets of runs, possibly because the variances were always kept at quite large values.

Again the estimation of μ_1 suffered under small sample sizes by having a large variance. The exception was run fourteen where the probabilities of sampling from the sensitive distributions were increased to .8 in both the numerator and denominator.

Overall, the most striking characteristic of this set of runs was the good performance of the Jackknife technique in constructing confidence intervals.

CHAPTER VII

SUMMARY

Since Warner's original paper in 1965 on the idea of randomizing responses from individuals so as to obtain a more unbiased estimate of sensitive characteristics, many improvements and variations have been proposed. The results given in this paper are an application of the randomized response technique when it is desired to either estimate the ratio of two sensitive characteristics or to use a concomitant variable to aid in the estimation of one sensitive characteristic.

Like most other estimators of a ratio, the estimator developed here is biased, and the estimate of its mean squared error is also biased. But if the denominator means are known, an unbiased ratio-type estimator of the mean of the numerator sensitive characteristic can be found. This unbiased ratio-type estimator has an exact variance which also has an unbiased estimator.

Two different methods of setting confidence intervals for the ratio of the population means were discussed. Based on the Monte-Carlo study, the method of setting confidence intervals which is based on Fieller's Theorem works very well for normal populations, as was expected, since Fieller's

results were derived using normal theory. The method using the Jackknife procedure also worked quite well in the normal case, but the computations involved are more lengthy. Utilization of high-speed electronic computers, however, can overcome this factor. When the populations were non-normal, Chi-Squared distributions and the sample size was relatively small, the method of setting confidence intervals using the Jackknife techniques was far superior. However, if the sample size is increased, the method based on Fieller's Theorem appears to approach that of the Jackknife.

Also, having large values for p_1 and p_2 did not make as much noticeable difference as would be suspected. Therefore, the randomized response type of design is worthwhile because the probabilities of choosing the sensitive question can be in the neighborhood of .6, which should be small enough to ensure the confidentiality of the response and hence to maintain the truthfulness of the respondent.

As with most Monte-Carlo studies, there is almost an unlimited number of combinations of parameters and distributions that could be tried, but since computer and/or researchers time is limited, the study must be terminated at some point. If the Monte-Carlo study could be continued,

one of the more interesting possibilities for further investigation would be the mixture of distributions. Such possibilities would include a normal and a uniform distribution in both the numerator and denominator, a uniform and a Chi-Squared distribution in both the numerator and denominator, a uniform and a normal in the numerator and a uniform and a Chi-Squared distribution in the denominator, as well as other more exotic combinations of distributions. The uniform distribution or a simple binomial distribution are possibilities worthwhile considering since these are two distributions that might be easily incorporated into the sampling procedure as the nonsensitive distributions.

BIBLIOGRAPHY

1. Abernathy, James R., Greenberg, Bernard G., Horvitz, Daniel G., (1970) "Estimates of Induced Abortion in Urban North Carolina," Demography, 7:19-29.

2. Boruch, Robert F., (1972) "Relations Among Statistical Methods for Assuring Confidentiality of Social Research Data," Social Science Research 1, 403-414.

3. Cochran, William G., (1963) Sampling Techniques, 2nd Ed., New York: John Wiley and Sons, Inc.

4. Cook, M. B., (1951) "Bi-variate k-statistics and Cumulants of their Joint Sampling Distribution," Biometrika 38, 179-195.

5. Creasy, M. A., (1954) "Limits for the Ratio of Means," Journal of the Royal Statistical Society, 16:186-194.

6. Fieller, E. C., (1954) "Some Problems in Interval Estimation," Journal of the Royal Statistical Society, 16:175-185.

7. Folsom, Ralph E., Greenberg, Bernard G., Horvitz, Daniel G., Abernathy, James R., (1972) "The Two Alternate Questions Randomized Response Model for Human Surveys," Journal of the American Statistical Association 68, 525-530.

8. Goodman, Leo A. and Hartley, H. O., (1958) "The Precision of Unbiased Ratio-Type Estimators," Journal of the American Statistical Association 53, 491-508.

9. Gould, A. L., Shah, B. V., Abernathy, J. R., (1969) "Unrelated Question Randomized Response Techniques With Two Trials Per Respondent," Proceedings of the Social Statistics Section, American Statistical Association.

10. Greenberg, Bernard G., Abul-Ela, Abdel-Latif A., Simmons, Walt R., Horvitz, Daniel G., (1969) "The Unrelated Question Randomized Response Model: Theoretical Framework," J. Amer. Stat. Assn., 64:520-539.

11. Greenberg, Bernard G., Kuebler, Roy R. Jr., Abernathy James R., Horvitz, Daniel G., (1971) "Application of the Randomized Response Technique in Obtaining Quantitative Data," J. Amer. Stat. Assn., 66:243-250.

12. Miller, R. G. Jr., (1964) "A Trustworthy Jackknife," Annals of Mathematical Statistics 35, 1594-1605.

13. Rao, C. R., (1973) Linear Statistical Inference and It's Applications, 2nd Ed., New York: John Wiley and Sons, Inc.

14. Warner, S. L., (1965) "Randomized Response: A Survey Technique for Eliminating Evasive Anser Bias," J. Amer. Stat. Assn., 60:68-69.

15. Yates, Frank, (1972) "A Monte-Carlo Trial on the Behavior of the Non-Additivity Test With Non-Normal Data," Biometrika 59, 2:253-261.

APPENDIX

```
C      MONTE CARLO STUDY FOR RATIO ESTIMATION
C          IN THE RANDOMIZED RESPONSE DESIGN
C
C
C XPN = PROB OF SELECTING THE SENSITIVE QUESTION
C IN THE NUMERATOR
C XPD = PROB OF SELECTING THE SENSITIVE QUESTION
C IN THE DENOMINATOR
C XMN1 = MEAN OF NORMAL FOR THE NONSENSITIVE IN THE NUM
C XMN2 = MEAN OF NORMAL FOR SENSITIVE IN THE NUM
C XMN3 = MEAN OF NORMAL FOR THE NONSENSITIVE IN THE DEN
C XMN4 = MEAN OF NORMAL FOR THE SENSITIVE IN THE DEN
C VN1 = VAR OF THE NORMAL FOR THE NONSEN IN THE NUM
C VN2 = VAR OF THE NORMAL FOR THE SEN    IN THE NUM
C VN3 = VAR OF THE NORMAL FOR THE NONSEN IN THE DEN
C VN4 = VAR OF THE NORMAL FOR THE SEN    IN THE DEN
C KPS = SIZE OF POPULATION OF SAMPLES
C KJCK = NUMBER OF ORSER PER GROUP FOR THE JACKNIFE
C NJCK = NUMBER OF GROUPS FOR THE JACKNIFE
C TDF = UPPER VALUE OF THE T-DIST WITH N-1 D.F.
C
C
      DIMENSION AN(200,4),AD(200,4),RHAT(100),XMSE
     &(100),RHJCK(100),VRHJCK(100),
     &XUBI(100),VP(100)
C
   10 READ(105,12) XPN,XPD,NSTART,XMN1,XMN2,XMN3,XMN4,
     &VN1,VN2,VN3,VN4,KPS,KJCK,NJCK,TDF,NCELL1,NCELL2
   12 FORMAT(2F6.4,I7,8F4.0,3I4,F6.4,2I3)
      OUTPUT NSTART
      KSS=KJCK*NJCK
      KSSJ=KSS-KJCK
      XMNS = 0.0
      NCI=0.0
      NG=0.0
      NCIJ = 0.0
      DO 60 L=1,KPS
      XNUMS=0.0
      XDENS=.0
      XSCP=.0
      XDEN=.0
      XNUM=.0
      ZRAT=.0
```

```
      ZRATS=.0
      ZR=0.0
      DO 62 I=1,NJCK
      DO 62 J=1,KJCK
      CALL MYRAN(NSTART,YN)
      IF(XPN-YN)15,14,14
   14 CALL RNORM(XMN2,VN2,DEVN,NSTART)
      GO TO 16
   15 CALL RNORM(XMN1,VN1,DEVN,NSTART)
   16 AN(I,J)=DEVN
      XNUM=XNUM+DEVN
      CALL MYRAN(NSTART,YD)
      IF(XPD-YD)18,17,17
   17 CALL RNORM(XMN4,VN4,DEVD,NSTART)
      GO TO 19
   18 CALL RNORM(XMN3,VN3,DEVD,NSTART)
   19 AD(I,J)=DEVD
      XDEN=XDEN+DEVD
      ZRAT=ZRAT+(DEVN/DEVD)
      ZR=ZR+((DEVN*DEVN/DEVD))
      ZRATS=ZRATS+ ((DEVN*DEVN)/(DEVD*DEVD))
      XNUMS=XNUMS+(DEVN*DEVN)
      XDENS=XDENS+(DEVD*DEVD)
      XSCP=XSCP+(DEVN*DEVD)
C9991 OUTPUT YN,DEVN,YD,DEVD
   62 CONTINUE
      ZBAR1=XNUM/KSS
      ZBAR2=XDEN/KSS
      XK1 = ZBAR1-(1.-XPN)*XMN1
      XK2 = ZBAR2-(1.-XPD)*XMN3
      RHAT(L) = (XK1*XPD)/(XK2*XPN)
      SZ1=(XNUMS-((XNUM*XNUM)/KSS))/(KSS*(KSS-1.))
      SZ2=(XDENS-((XDEN*XDEN)/KSS))/(KSS*(KSS-1.))
      XCOV = (XSCP-(XNUM*XDEN/KSS))/(KSS*(KSS-1.))
C9981 GO TO 42
      T1=SZ1/(XK1*XK1)
      T2=SZ2/(XK2*XK2)
      T3=XCOV/(XK1*XK2)
      XMSE(L)=(RHAT(L)*RHAT(L))*(T1+T2-(2.*T3))
C9993 OUTPUT  RHAT(L),SZ1,SZ2,XCOV,T1,T2,T3,XMSE(L)
C9994 OUTPUT XSCP
C9992 GO TO 60
      RBAR =ZRAT/KSS
      XC = (KSS*(ZBAR1-RBAR*ZBAR2))/(KSS-1.)
```

```
      XBIAS=RHAT(L)-(XMN2/XMN4)
      XMNS = XMNS+(XBIAS*XBIAS)
      XMZ2=XPD*XMN4+(1.-XPD)*XMN3
      XUB1(L) = (RBAR*XMZ2+XC-(1.-XPN)*XMN1)/XPN
      T1 = ((ZRATS)-(ZRAT*ZRAT/KSS))/(KSS-1.)
      T2=KSS*SZ2
      T3=(1./(KSS-1.))*(ZR-(2.*XNUM*RBAR)+(RBAR*RBAR*XDEN)
     C-(KSS-1.)*ZBAR2*T1)
      TEMP1 = KSS*(KSS+1.)*XNUMS
      TEMP2 = 2.*(KSS+1.)*(XSCP*ZRAT+ZR*XDEN)
      TEMP3 = (KSS-1.)*(XDENS*ZRATS+2.*XNUM*XNUM)
      TEMP4 = 8.*XNUM*XDEN*ZRAT
      TEMP5 = 2.*XDENS*ZRAT*ZRAT
      TEMP6 = 2.*ZRATS*XDEN*XDEN
      TEMP7 = (6.*XDEN*XDEN*ZRAT*ZRAT)/KSS
      T8 = (KSS-1.)*(KSS-2.)*(KSS-3.)
      T4 = (TEMP1-TEMP2-TEMP3+TEMP4+TEMP5+TEMP6-TEMP7)/T8
      TEMP1 = (XMZ2*XMZ2*T1)/KSS
      TEMP2 = (2.*XMZ2*T3)/(KSS-2.)
      TEMP3 = (KSS-1.)*T1*T2
      TEMP4 = (KSS-3.)*XC*XC
      TEMP5 = (KSS-1.)*(1.-2./KSS)*T4
      TK = KSS*KSS-KSS-2.
      VP(L) = (TEMP1+TEMP2+((TEMP3+TEMP4+TEMP5)/TK))/(XPN*XPN)
   42 Y=1.-((TDF*TDF*SZ1*XPD)/(ZBAR2-(1.-XPD)*XMN3))
      IF(Y)30,30,31
   31 TEMP = XK1*XK1
      XQ1 = (TEMP-TDF*TDF*SZ1)/(XPN*XPN)
      TEMP = XK2*XK2
      XQ2 = (TEMP-TDF*TDF*SZ2)/(XPD*XPD)
      TEMP = XK1*XK2
      XQ12 = (TEMP-TDF*TDF*XCOV)/(XPN*XPD)
      TEMP=SQRT(XQ12*XQ12-XQ1*XQ2)
      XL=(XQ12-TEMP)/XQ2
      XU=(XQ12+TEMP)/XQ2
C9995 OUTPUT XL,XU
      IF(XL-XMN2/XMN4)34,34,35
   34 IF(XU-XMN2/XMN4)35,36,36
   36 NCI=NCI+1
      GO TO 35
   30 NG=NG+1
   35 CONTINUE
C9982 GO TO 60
```

```
C
C  USING THE JACKNIFE METHOD
C
      PSS = 0.0
      PSR=0.0
      PSV=0.0
      DO 20 I=1,NJCK
      IF (I.EQ.1)GO TO 23
      DO 22 IX=1,KJCK
      TEMP=AN(1,IX)
      AN(1,IX)=AN(I,IX)
      AN(I,IX)=TEMP
      TEMP=AD(1,IX)
      AD(1,IX)=AD(I,IX)
      AD(I,IX)=TEMP
   22 CONTINUE
   23 XNUM=0.0
      XDEN=0.0
      XNUMS=0.0
      XDENS=0.0
      XSCP=0.0
      DO 24 J=2,NJCK
      DO 24 K=1,KJCK
      XNUM=XNUM+AN(J,K)
      XDEN=XDEN+AD(J,K)
      XNUMS=XNUMS+(AN(J,K)*AN(J,K))
      XDENS=XDENS+(AD(J,K)*AD(J,K))
      XSCP=XSCP+(AN(J,K)*AD(J,K))
   24 CONTINUE
      ZBAR1=XNUM/KSSJ
      ZBAR2=XDEN/KSSJ
      XK1 = ZBAR1-(1.-XPN)*XMN1
      XK2 = ZBAR2-(1.-XPD)*XMN3
      RHATJP  = (XK1*XPD)/(XK2*XPN)
      SZ1=(XNUMS-((XNUM*XNUM)/KSSJ))/(KSSJ*(KSSJ-1.))
      SZ2=(XDENS-((XDEN*XDEN)/KSSJ))/(KSSJ*(KSSJ-1.))
      XCOV = (XSCP-(XNUM*XDEN/KSSJ))/(KSSJ*(KSSJ-1.))
      T1=SZ1/(XK1*XK1)
      T2=SZ2/(XK2*XK2)
      T3=XCOV/(XK1*XK2)
      TMSEJP=(RHATJP*RHATJP)*(T1+T2-(2.*T3))
      PSEUDR=NJCK*RHAT(L)-((NJCK-1.)*RHATJP)
      PSR=PSR+PSEUDR
      PSS = PSS+PSEUDR*PSEUDR
```

```
C9996 OUTPUT TMSEJP
   20 CONTINUE
      R = XMN2/XMN4
      TEMP=NJCK
      RHJCK(L)=PSR/TEMP
      T1 = NJCK*(NJCK-1.)
      VRHJCK(L) = (PSS-NJCK*RHJCK(L)*RHJCK(L))/T1
      SDJ = SQRT(VRHJCK(L))
      T1 = RHJCK(L)-TDF*SDJ
      T2 = RHJCK(L)+TDF*SDJ
C9987 OUTPUT T1,T2
      IF(R.LE.T1) GO TO 26
      IF(T2.LE.R) GO TO 26
      NCIJ = NCIJ+1.
   26 CONTINUE
   60 CONTINUE
      TEMP=KPS
      PNG=NG/TEMP
      XNCIJ = NCIJ/TEMP
      PNCI=NCI/TEMP
      X1 = XPN*XPN*XMN2*XMN2
      X2 = XPD*XPD*XMN4*XMN4
      T1 = (XPN*VN2+(1.-XPN)*VN1)/X1
      T2 = (XPD*VN4+(1.-XPD)*VN3)/X2
      TMSE =(R*R*(T1+T2))/KSS
      XMNS = XMNS/TEMP
      OUTPUT XPN,XPD,XMN1,XMN2,XMN3,XMN4,VN1,VN2,VN3,VN4,KPS
      OUTPUT KJCK,NJCK,TDF
      OUTPUT ' '
   56 FORMAT (28HTHE THEORETICAL MSE(RHAT) =   ,G11.4)
      WRITE(108,56) TMSE
      OUTPUT ' '
      OUTPUT ' '
      OUTPUT 'E(RHAT-R)**2 = '
      OUTPUT XMNS
      OUTPUT ' '
      OUTPUT 'THE FRACTION OF CON INT THAT BRACKET THE MEAN'
      OUTPUT 'USING THE JACKKNIFE IS  '
      OUTPUT XNCIJ
      OUTPUT ' '
   64 WRITE (108,91)
      WRITE (108,92) PNG
      WRITE (108,93)
      WRITE (108,94) PNCI
```

```
   82 FORMAT ('1')
      WRITE(108,82)
      OUTPUT ' FREQ DIST OF RHAT'
      OUTPUT ' '
      CALL DSUMRY(RHAT,KPS,NCELL1,0,-10000000.)
      OUTPUT ' FREQ DIST OF MSE OF RHAT'
      OUTPUT ' '
      CALL DSUMRY(XMSE,KPS,NCELL2,0,-10000000.)
      OUTPUT ' FREQ DIST OF RHAT USING JACKNIFE'
      OUTPUT ' '
      CALL DSUMRY(RHJCK,KPS,NCELL1,0,-10000000.)
      OUTPUT 'FREQ DIST OF MSE OF JACKKNIFE EST OF RHAT'
      OUTPUT ' '
      CALL DSUMRY(VRHJCK,KPS,NCELL2,0,-10000000.)
      OUTPUT ' FREQ DIST OF UNBIASED EST OF MU 1'
      OUTPUT ' '
      CALL DSUMRY(XUB1,KPS,NCELL1,0,-10000000.)
      OUTPUT ' FREQ DIST OF THE VARIANCE OF EST OF MU 1'
      OUTPUT ' '
      CALL DSUMRY(VP,KPS,NCELL2,0,-10000000.)
   91 FORMAT(/,/,44HTHE FRAC OF SAMPLES FOR WHICH NO
     &CONFIDENCE)
   92 FORMAT(32HINTERVAL CAN BE CONSTRUCTED IS  ,F6.4,/)
   93 FORMAT(38HTHE FRAC OF SAMPLES FOR WHICH THE C.I.)
   94 FORMAT(22HBRACKETS THE MEAN IS  ,F6.4)
      GO TO 10
  999 STOP
      END
C
C
C  SUBROUTINE MYRAN GENERATES A UNIFORM RANDOM DIGIT ON [0,1]
C
      SUBROUTINE MYRAN(K,Y)
      K=K*65539
      IF(K.LE.0)K=K+1+2147483647
      Y=K*.4656613E-09
      RETURN
      END
C
C
C
C  SUBROUTINE RNORM GENERATES A RANDOM NORMAL DEVIATE
```

```
C   SUBROUTINE RNORM GENERATES A RANDOM NORMAL DEVIATE
C   WHICH IS TRUNCATED AT FOUR STANDARD DEV.
C
      SUBROUTINE RNORM(XM,XV,DEV,NSTART)
      CALL MYRAN(NSTART,RA)
      CALL MYRAN(NSTART,RB)
      V=(-2.0*ALOG(RA))**0.5*COS(6.283*RB)
      IF(V.LE.-4.) V=-4.; GO TO 12
      IF(V.GE.4.) V=4.; GO TO 12
   12 DEV=V*SQRT(XV)+XM
      RETURN
      END
C
C
C
C...SUBROUTINE DSUMRY(X,N,NCELL,IZERO, UPPER)
C
C...X IS DATA VECTOR
C...N IS LENGTH OF DATA VECTOR
C...NCELL IS NO. OF CELLS TO FORM HISTOGRAM
C...IZERO = 1 IF LOWER CELL BDRY IS ZERO; = BLANK OTHERWISE
C...UPPER = UPPER CELL BOUNDRY; UPPER = X(N) IF ASSIGNED
     &.LT. -1.0E-10
      SUBROUTINE DSUMRY(X,N,NCELL,IZERO,UPPER)
      DIMENSION X(1000)
      DIMENSION NFREQ(100),GROUP(100),POINT(100)
      DATA IIA/1HX/
   10 FORMAT (3I4)
   20 FORMAT (F10.4)
   30 FORMAT (1H1)
   40 FORMAT (1H ,1HN,6X,I4,11X,5HRANGE,4X,G11.5,5X,
     &10HCOEF. VAR.,1X,F10.5)
   50 FORMAT(1H ,4HMEAN,2X,G11.5,5X,8HVARIANCE,1X,G11.5,
     &5X,8HSKEWNESS,3X,F10.5)
   60 FORMAT(1H ,6HMEDIAN,1X,G11.5,4X,9HSTD. DEV.,1X,G11.5,
     &4X,12HNO. OF CELLS,1X,I4)
   70 FORMAT (1H0)
   80 FORMAT (1H0,10HCELL. MID.,5X,5HFREQ.,5X,' CELL
     &WIDTH = ',G11.5)
   90 FORMAT (1H ,G11.5,6X,I4,2X,60A1)
  100 FORMAT (1H )
C...SORT
      L=N-1
```

```
      DO 120 J=1,L
      LL=L-J+1
      DO 110 I=1,LL
      LG=I+1
      IF (X(I) .LT. X(LG)) GO TO 110
      A=X(I)
      X(I)=X(LG)
      X(LG)=A
  110 CONTINUE
  120 CONTINUE
      RANGE = X(N) - X(1)
C...TO CALCULATE CELL BOUNDARIES, FREQUENCIES, MIDPOINTS
      IF (NCELL .EQ. 0) NCELL = 15
      DO 130 I=1,NCELL
  130 NFREQ(I)=0
      IF (IZERO .EQ. 1) GO TO 140
      WIDTH = (UPPER - X(1))/(NCELL-1.)
      IF (UPPER .LT. -1.0E5)  WIDTH =(X(N)-X(1))/(NCELL-1.)
      RIDPT = WIDTH/2.
      GROUP(1)=X(1)+RIDPT
      POINT(1)=X(1)
      GO TO 150
  140 WIDTH=(X(N)-0.0)/(NCELL-.5)
      IF (UPPER .GT. -1.0E5) WIDTH=(UPPER)/NCELL
      GROUP(1)=WIDTH
      POINT(1)=GROUP(1)/2.0
  150 DO 160 I=2,NCELL
  160 GROUP(I)=GROUP(I-1)+WIDTH
      DO 190 I=1,N
      DO 170 M=1,NCELL
      IF (X(I) .LE. GROUP(M)) GO TO 180
  170 CONTINUE
  180 NFREQ(M)=NFREQ(M)+1
  190 CONTINUE
      DO 200 I=2,NCELL
  200 POINT(I)=POINT(I-1)+WIDTH
C...CALCULATE THE MEAN
  210 XSUM=0.0
      DO 220 I=1,N
  220 XSUM=XSUM+X(I)
      AVE=XSUM/N
```

```
C...CALCULATE THE VARIANCE AND STD. DEVIATION
      TEXS = 0.0
      DO 230 I=1,N
  230 TEXS = TEXS+(X(I)*X(I))
      PART1 = TEXS-(XSUM*XSUM/N)
      VAR=PART1/N
      VR=PART1/(N-1)
      SD=SQRT(VAR)
      SD1=SQRT(VR)
C...CALCULATE SKEWNESS
      SKW1=0.0
      DO 240 I=1,N
  240 SKW1=SKW1+(X(I)-AVE)**3.
      SKW2=N*(SD**3.)
      SKEW=SKW1/SKW2
C...CALCULATE THE MEDIAN
      J=(N+1)/2
      K=(N+2)/2
      XMED=(X(J)+X(K))/2.
C...CALCULATE THE COEFFICIENT OF VARIATION
      COEFV=SD1/AVE
C...PRINT OUT
      WRITE(108,40)N,RANGE,COEFV
      WRITE (108,50) AVE,VR,SKEW
      WRITE (108,60) XMED,SD1,NCELL
      WRITE (108,70)
      WRITE (108,80) WIDTH
      WRITE (108,100)
      MAX=NFREQ(1)
      DO 250 I=2,NCELL
      IF (MAX .LT. NFREQ(I)) MAX=NFREQ(I)
  250 CONTINUE
      DO 290 I=1,NCELL
      IF (MAX .LE. 45) K=NFREQ(I); GO TO 270
      IF (NCELL .LE. 15) GO TO 260
      K=NFREQ(I)*45./MAX+.5
      GO TO 270
  260 K=NFREQ(I)*30./MAX+.5
  270 IF (NFREQ(I) .EQ. 0) GO TO 280
      WRITE (108,90) POINT(I),NFREQ(I),(IIA,L=1,K)
      GO TO 290
  280 WRITE (108,90) POINT(I),NFREQ(I)
  290 CONTINUE
```

```
290 CONTINUE
 13 FORMAT('1')
    WRITE (108,13)
    RETURN
    END
```

MONTANA STATE UNIVERSITY LIBRARIES
3 1762 10184014 6

D378
P445
cop 2