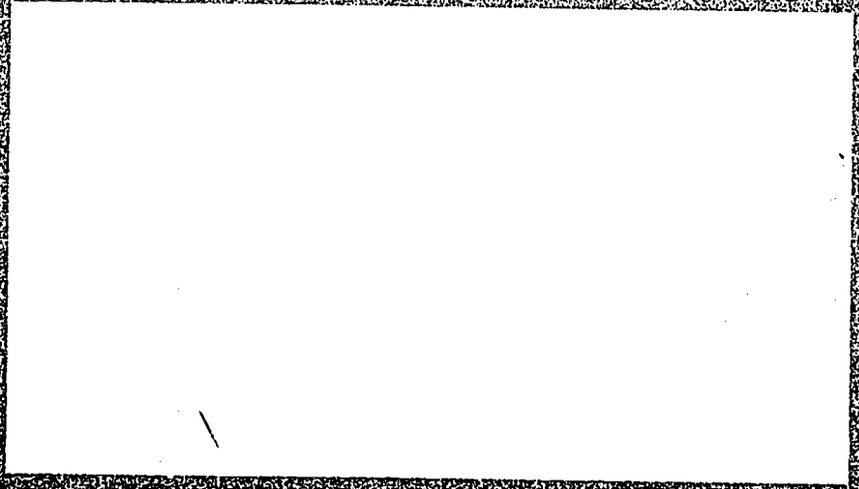
DOCUMENT RESUME

ED 068 535 TM 001 901

AUTHOR	Elashoff, Janet Dixon; Elashoff, Robert M.
TITLE	Missing Data Problems for Two Samples on a Dichotomous Variable.
INSTITUTION	Stanford Univ., Calif. Stanford Center for Research and Development in Teaching.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	SCRDT-RDM-73
BUREAU NO	BR-5-0252
PUB DATE	Apr 71
CONTRACT	OEC-6-10-078
NOTE	47p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Comparative Analysis; *Data Analysis; Educational Research; *Evaluation Techniques; *Mathematical Models; Probability Theory; *Research Methodology; *Statistical Analysis
IDENTIFIERS	*Monte Carlo Method

ABSTRACT

The problem of comparing proportions when some data are missing is investigated, and determination is made of what statistical techniques are appropriate under each of several probability models describing the observations likely to be missing. Monte Carlo methods were used to investigate the properties of standard estimators under each of the missing data models. Applying standard techniques which ignore the occurrence of missing observations may yield misleading conclusions. Some tests and estimators are fairly robust to the model for missing data, but others may be seriously affected. If the model for missing observations is complex, the sample information may be insufficient for adequate data analysis. This report is useful since the problem of missing data is recurrent in educational research and may present serious difficulties even for simple problems. (Author/LH)



Stanford Center for Research and Development in Teaching

SCHOOL OF EDUCATION STANFORD UNIVERSITY

STANFORD CENTER FOR RESEARCH AND DEVELOPMENT IN TEACHING

Publication Resume

Janet Dixon Elashoff and Robert M. Elashoff. Missing Data Problems for Two Samples on a Dichotomous Variable. Research and Development Memorandum No. 73. April 1971. 40 pp.

Purpose: To investigate the problem of comparing proportions when some data are missing, and to determine what statistical techniques are appropriate under each of several probability models describing the observations likely to be missing.

Method and sample: Monte Carlo methods were used to investigate the properties of standard estimators under each of the missing data models.

Conclusions: Applying standard techniques which ignore the occurrence of missing observations may yield misleading conclusions. Some tests and estimators are fairly robust to the model for missing data, others may be seriously affected. If the model for missing observations is complex, the sample information may be insufficient for adequate data analysis.

Usefulness: The problem of missing data is recurrent in educational research and may present serious difficulties even for the simple problem of comparing two proportions.

Target groups: Educational researchers, methodologists.

STANFORD CENTER
FOR RESEARCH AND DEVELOPMENT
IN TEACHING

Research and Development Memorandum No. 73

MISSING DATA PROBLEMS FOR TWO SAMPLES ON A DICHOTOMOUS VARIABLE

Janet Dixon Elashoff and Robert M. Elashoff

School of Education
Stanford University
Stanford, California

April 1971

Published by the Stanford Center for Research and Development in Teaching, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the Office of Education. (Contract No. OEC-6-10-078, Project No. 5-0252 0702.) Work on this study was also supported in part by National Institute of General Medical Science Grant No. GM 17182-02 SSS and by grant FR-3 of the Division of Research Resource Facilities of the National Institutes of Health.

Introductory Statement

The central mission of the Stanford Center for Research and Development in Teaching is to contribute to the improvement of teaching in American schools. Given the urgency of the times, technological developments, and advances in knowledge from the behavioral sciences about teaching and learning, the Center works on the assumption that a fundamental reformulation of the future role of the teacher will take place. The Center's mission is to specify as clearly, and on as empirical a basis as possible, the direction of that reformulation, to help shape it, to fashion and validate programs for training and retraining teachers in accordance with it, and to develop and test materials and procedures for use in these new training programs.

The Center is at work in three interrelated problem areas: (a) Heuristic Teaching, which aims at promoting self-motivated and sustained inquiry in students, emphasizes affective as well as cognitive processes, and places a high premium upon the uniqueness of each pupil, teacher, and learning situation; (b) The Environment for Teaching, which aims at making schools more flexible so that pupils, teachers, and learning materials can be brought together in ways that take account of their many differences; and (c) Teaching Students from Low-Income Areas, which aims to determine whether more heuristically oriented teachers and more open kinds of schools can and should be developed to improve the education of those currently labled as the poor and the disadvantaged.

The Methodology Unit developed Research and Development Memorandum No. 73, which follows, to deal with the problem of comparing proportions where some cases are missing. Such nonresponse problems are frequently encountered in the analysis of data gathered by Center projects.

Table of Contents

List of Tables

Abstract

Two-sample problems with dichotomous data are considered; some specific probability models are developed to describe which observations are missing and why; and the statistical techniques appropriate under each of the models are discussed.

MISSING DATA PROBLEMS FOR TWO SAMPLES ON A DICHOTOMOUS VARIABLE

Janet Dixon Elashoff and Robert M. Elashoff[1]

1. Introduction

Incomplete or missing data is a major problem in many fields. Data may be incomplete because of nonresponse, random loss, transcription errors, refusal to cooperate, and a variety of other reasons. In these instances, statistical techniques to deal with the incomplete data are necessary. One possibility is simply to delete and ignore the incomplete cases. To select the appropriate technique, however, some facts must be known about the kind of observations which are missing and which variables influence the loss of certain observations.

In this study two-sample problems with dichotomous data are considered; some specific probability models are developed to describe which observations are missing and why; and the statistical techniques appropriate under each of the models are discussed. Using techniques which assume that observations are missing at random may be extremely misleading. If the probability model governing the occurrence of missing data is complex, the only adequate solution may be to "find out what the missing observations are."

Section 2 discusses four probability models for the occurrence of missing observations. Section 3 introduces notation and lists the estimation and testing problems to be discussed. The succeeding three sections derive solutions under each of the first three probability

[1]Janet D. Elashoff is Assistant Professor of Education at Stanford University and a Research and Development Associate at SCRDT; Robert M. Elashoff is Associate Professor of Biostatistics at the University of California, San Francisco.

models proposed, while Section 7 indicates how headway might be made under Model 4. Then in Sections 8, 9, 10, and 11 the Model 1 and Model 3 estimators are compared using asymptotic and small sample results. Section 12 contains recommendations about procedures to use for each of the estimation and testing problems discussed and problems for further research.

2. Probability Models for Incomplete Data

This section discusses four general probability models proposed in the statistics literature to account for the occurrence of missing data.

Assume that one independent variable x and one dependent variable y are under study for each individual. Further assume that: (1) no x observations are missing, (2) for each value of x occurring in the study, a random sample of N_x individuals is drawn, and n_x individuals are observed on y and $N_x - n_x$ individuals are not observed on y (their y values are "missing" and so unknown), (3) no other variables have been measured.

Define

$$q(x,y) = \Pr(\text{an individual's } y \text{ is observed} | x,y) .$$

In other words, among individuals with values x and y of the independent and dependent variables, the probability that the value of the dependent variable is not observed is $1-q(x,y)$. Thus, the loss of particular observations may be influenced by the actual values of the dependent and independent variables.

Model 1: Randomly Missing Data

It is commonly assumed that missing observations have occurred at random or by chance. That is, neither the value of x nor the value of y influences whether an individual's y value is observed or not. Thus the random model states that $q(x,y)$, the probability that an individual's y value is observed, is independent of both x and y, or

$$q(x,y) = q \text{ for all } x \text{ and } y .$$

The random model is appropriate where factors completely independent of the variables under study are causing missing data or where a question y is asked of a random subsample of individuals surveyed.

The random model is the basis for the frequent practice of "ignoring" missing data, that is, analyzing only complete observations. The practice of ignoring missing data is appropriate if the random model holds, otherwise it may give misleading results (see Sections 8, 9, and 10).

Model 2: Independent Variable Influences Missing Data

Model 2 states that $q(x,y)$, the probability that an individual's y value is observed, is dependent on x but independent of the value of y, or

$$q(x,y) = q_x \text{ for all } y .$$

For example, suppose computer-assisted instruction is compared with a conventional teaching method. Let x denote the teaching method. A sample of N_x students is taught by method x, and each student attains a final score of y on material learned. Due to computer breakdowns final scores y are missing for some students. In this example, the

independent variable, teaching method, but not the dependent variable, final score, influences the probability that an observation is missing.

Model 3: Dependent Variable y Influences Missing Data

Model 3 states that $q(x,y)$ depends on the value of the dependent variable y but is independent of the value of the independent variable x

$$q(x,y) = q_y \quad \text{for all} \quad x .$$

For example, suppose patients with a certain disease are assigned either an active drug or a placebo x in a double blind study. The placebo has the same side effects as the active drug, but presumably it does not have the same curative or palliative effect as the active drug. A follow-up study is made and each patient is scored as improved or unimproved y. Lack of improvement may cause some patients to drop out of the study or refuse to cooperate further. Improvement also may give patients a reason to drop out or a chance to leave the area. In both cases the y measurements are unknown. Clearly, in these circumstances, missing y's may be influenced by whether or not the patient is improved but not directly by the drug the patient received.

Model 4: The Values of Both the Dependent and Independent Variable Influence Missing Data

Model 4 states that $q(x,y)$ depends on the value of the dependent variable y and the value of the independent variable x. Both an individual's y value and his x value affect the probability that his y value will be observed.

Suppose, for example, that a prospective panel study is undertaken to investigate differences in employment status y between the sexes x in New England over a ten-year period. Some people will be lost to follow-up in the course of the study because of emigration from the region. Clearly employment status is one factor influencing emigration--thus, employment status y influences whether an individual's employment status is observed. Furthermore, the sexes have differential mobility, so the independent variable x also influences whether an individual's employment status is observed or not.

3. Two-Sample Problems for y Dichotomous

This section outlines five statistical problems involving the comparison of two independent proportions [problems (a) through (e) below] and presents the notation used in describing samples with missing data.

Let p_i be the probability that y equals one in population i

$$p_i = \Pr (y = 1 \mid x = i) .$$

The five statistical problems to be discussed are:

(a) To estimate p_i for population i.

(b) To estimate the difference $d = p_1 - p_2$.

(c) To estimate the ratio $R = p_1/p_2$.

(d) To estimate the odds ratio $OR = \dfrac{p_1 (1 - p_2)}{p_2 (1 - p_1)}$.

(e) To test $H_0 : p_1 - p_2$ against the alternative $H_1 : p_1 \neq p_2$.

Random samples of N_1 and N_2 individuals are selected from the two infinite populations denoted by $x = 1$ and $x = 2$. Suppose that

n_i individuals are actually observed from each sample, $n_i \leq N_i$ $(i = i,2)$ so that $N_i - n_i$ observations are missing from each sample. Let r_i be the number of individuals for whom $y = 1$ out of the n_i actually observed in population i ; $r_i' = n_i - r_i$. Let u_i be the number of individuals with $y = 1$ in the $N_i - n_i$ individuals who weren't observed; $u_i' = N_i - n_i - u_i$. The number of missing observations $N_i - n_i$ is known but u_i is not known. This notation is summarized in Table 1.

TABLE 1

Notation

Population x	Value of y	P(y|x)	q(x,y)	Actual number in the sample	Observed number in the sample
1	1	p_1	q(1,1)	$r_1 + u_1$	r_1
1	0	$1-p_1$	q(1,0)	$r_1' + u_1'$	r_1'
1	Totals			N_1	n_1
2	1	p_2	q(2,1)	$r_2 + u_2$	r_2
2	0	$1-p_2$	q(2,0)	$r_2' + u_2'$	r_2'
2	Totals			N_2	n_2

Notice it is assumed that it is not feasible to make further efforts to obtain the y-values for individuals whose y-values are missing. Callbacks will not be carried out and further data on other measured variables will not enable us to obtain "good" predicted values of y . These stringent restrictions are relaxed only in the discussion of Model 4.

4. Randomly Missing Data: Statistical Techniques for Problems (a) Through (e) Under Model 1

When Model 1 is correct and missing observations occur at random, the $N_1 - n_1$ and $N_2 - n_2$ missing observations are ignored and the remaining observations are regarded as random samples of size n_1 and n_2 respectively. Standard statistical techniques are applied to these random samples. The maximum likelihood (ML) estimator of p_i under Model 1 is $\hat{p}_{1i} = r_i/n_i$ and the ML estimators of d , R , and OR are obtained by substituting $\hat{p}_{1i}$ for p_i in each of these expressions. The conditional and unconditional means and variances of the estimators of p_i , d , R , and OR are given in Tables 2, 3, and 4.

Alternative estimators for R and OR or simple functions of these quantities have been derived and studied under Model 1. For example, Haldane (1955) and Anscombe (1956) recommend that log OR should be estimated by substituting $\hat{p}_i + (1/2n_i)$ for $\hat{p}_i$ and $[(1-\hat{p}_i) + (1/2n_i)]$ for $(1-\hat{p}_i)$ in the expression $\hat{OR}_1$ to reduce bias (see Table 2). Since the primary focus of this study is comparison of estimators under different models for the missing data, such modifications were not investigated. For the conditional mean of an estimator the expectation of the estimator is taken conditional upon the observed n_i ; the unconditional mean is not conditioned upon the n_i . In the development of the asymptotic means and variances it is assumed that

(1) $\tau_i = \lim n_i/N_i > 0$

(2) $\lambda = \lim N_1/(N_1 + N_2) > 0$.

Occasionally, $\lambda_1 = \lambda$ and $\lambda_2 = (1-\lambda)$ will be used.

The statistics $\hat{p}_{1i}$, $\hat{d}_1$, $\hat{R}_1$, and $\hat{OR}_1$ have asymptotic normal distributions conditionally and unconditionally with the means and variances shown in Tables 2, 3, and 4.

TABLE 2

Conditional and Unconditional Means (Assuming Model 1)

Estimator	Mean	
$\hat{p}_{1i}$	p_i	
$\hat{d}_1 = \hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	
$\hat{R}_1 = \hat{p}_1/\hat{p}_2$	p_1/p_2	[asymptotic]
$\hat{OR}_1 = \hat{p}_1(1-\hat{p}_2)/\hat{p}_2(1-\hat{p}_1)$	$p_1(1-p_2)/p_2(1-p_1)$	[asymptotic]

TABLE 3

Asymptotic Conditional Variance Under Model 1

Estimator	Variance	
$\sqrt{N_1 + N_2}\ \hat{p}_{1i}$	$\frac{p_i(1-p_i)}{\lambda_i \tau_i}$	[exact]
$\sqrt{N_1 + N_2}\ \hat{d}_1$	$\frac{p_1(1-p_1)}{\lambda\tau_1} + \frac{p_2(1-p_2)}{(1-\lambda)\tau_2}$	[exact]
$\sqrt{N_1 + N_2}\ \hat{R}_1$	$\frac{1}{(p_2)^2}\left\{\frac{p_1(1-p_1)}{\tau_1\lambda} + \left(\frac{p_1}{p_2}\right)^2 \frac{p_2(1-p_2)}{\tau_2(1-\lambda)}\right\}$	
$\sqrt{N_1 + N_2}\ \hat{OR}_1$	$\frac{p_1(1-p_2)}{p_2^2(1-p_1)^2}\left\{\frac{1-p_2}{(1-p_1)\tau_1\lambda} + \frac{p_1}{p_2\tau_2(1-\lambda)}\right\}$	

TABLE 4

Asymptotic Unconditional Variance Under Model 1

Estimator	Variance
$\sqrt{N_1 + N_2}\ \hat{p}_{1i}$	$\frac{p_i(1-p_i)}{q\lambda_i}$
$\sqrt{N_1 + N_2}\ \hat{d}_1$	$\frac{1}{q\lambda(1-\lambda)}\ [(1-\lambda)p_1(1 - p_1) + \lambda p_2(1 - p_2)]$
$\sqrt{N_1 + N_2}\ \hat{R}_1$	$\frac{p_1}{qp_2^3(1-\lambda)\lambda}\ [p_2(1 - p_1) - \lambda(p_2 - p_1)]$
$\sqrt{N_1 + N_2}\ \hat{OR}_1$	$\frac{p_1(1-p_2)}{\lambda(1-\lambda)qp_2^3(1-p_1)^3}\ [p_2(1-p_2)(1-\lambda) + p_1(1-p_1)\lambda]$

A test of the $H_0 : p_1 = p_2$ against one or two-sided alternatives may be carried out using Fisher's exact test. Naturally, the power of the test based on sample sizes n_i will be less than that based on sample sizes N_i.

5. The Independent Variable Influences Missing Data (Model 2): Statistical Techniques for Problems (a) Through (e)

In this model the probability of observing the particular y score for a particular individual is independent of the value of y but does depend on the population sampled. The estimators defined under Model 1 for p_i, d, R, and OR are also the ML estimators assuming Model 2, and they have the same conditional means and variances under Model 2 as under Model 1 (see Tables 2 and 3). Moreover, the asymptotic unconditional means are also the same. However, the unconditional variances

under Model 2 are different from those under Model 1 (see Table 5).

TABLE 5

Asymptotic Unconditional Variance Under Model 2

Estimator	Variance
$\sqrt{N_1 + N_2}\ \hat{p}_{1i}$	$\frac{p_i(1 - p_i)}{q_i\lambda_i}$
$\sqrt{N_1 + N_2}\ \hat{d}_1$	$\frac{p_1(1 - p_1)}{\lambda q_1} + \frac{p_2(1 - p_2)}{(1 - \lambda)q_2}$
$\sqrt{N_1 + N_2}\ \hat{R}_1$	$\frac{p_1(1 - p_1)}{\lambda q_1 p_2^2} + \frac{p_1^2(1 - p_2)}{(1-\lambda)q_2 p_2^3}$
$\sqrt{N_1 + N_2}\ \hat{OR}_1$	$\frac{p_1(1 - p_2)}{\lambda(1-\lambda)p_2^3(1-p_1)^3 q_1 q_2}\left[p_2(1-p_2)q_2(1-\lambda) + p_1(1-p_1)q_1\lambda\right]$

It is possible to test whether Model 1 or Model 2 applies in a particular problem. The null hypothesis is $H_0 : q_x = q$ for $x = 1,2$; the alternative hypothesis is $H_1 : q_1 \neq q_2$. Fisher's Exact Test may be used to carry out a test conditional on the N_i and $(n_1 + n_2)$. To test $H_0 : p_1 = p_2$ against one or two-sided alternatives use the same tests as if Model 1 obtains.

6. The Dependent Variable Influences Missing Data (Model 3): Statistical Techniques for Problems (a) Through (e)

Under Model 3, the value of the dependent variable y influences the probability that an individual's y value will be observed. The

independent variable x does not influence the probability of a missing observation. Therefore

$$q(1,y) = q(2,y) = q_y \quad \text{for} \quad y = 0,1 . \tag{3}$$

The maximum likelihood equations for Model 3 have quadratic and cross product terms in the p's and q's . For example

$$\frac{\partial \ln L}{\partial q_0} = \frac{n_1 - r_1}{q_0} + \frac{(n_2 - r_2)}{q_0} + \frac{-(N_1 - n_1)(1-p_1)}{1-p_1q_1-(1-p_1)q_0} + \frac{-(N_2 - n_2)(1-p_2)}{1-p_2q_1-(1-p_2)q_0} = 0 .$$

Consequently, simple estimators are of interest. Eklund (1959) argues that if there were no missing observations, the p_i might be estimated by $\hat{p}_i = (r_i + u_i)/N_i$. Therefore, estimating the q's as

$$\hat{q}(i,1) = \frac{r_i}{r_i + u_i}$$

$$\hat{q}(i,0) = \frac{r_i'}{r_i' + u_i'} \tag{4}$$

and using relationship (3) yields equations

$$\frac{r_1}{r_1 + u_1} = \frac{r_2}{r_2 + u_2}$$

$$\frac{r_1'}{r_1' + u_1'} = \frac{r_2'}{r_2' + u_2'} \quad . \tag{5}$$

Solving for u_i and u_i' yields estimates

$$\hat{u}_1 = r_1 \left[\frac{N_2 r_1' - N_1 r_2'}{r_1' r_2 - r_2' r_1} - 1 \right]$$

$$\hat{u}_2 = \frac{r_2}{r_1} \hat{u}_1 . \tag{6}$$

This leads to estimating the q_y and p_i as

$$\hat{q}_1 = \frac{n_1 r_2 - n_2 r_1}{N_2 r_1' - N_1 r_2'}$$

(7)

$$\hat{q}_0 = \frac{n_1 r_2 - n_2 r_1}{N_1 r_2 - N_2 r_1}$$

$$\hat{p}_{3i} = \frac{r_i}{N_i} \frac{1}{\hat{q}_1}$$

(8)
$$= \frac{r_i}{N_i} \left[\frac{N_2 r_1' - N_1 r_2'}{n_1 r_2 - n_2 r_1}\right]$$

$$= \frac{r_i}{N_i} \frac{[N_2(n_1 - r_1) - N_1(n_2 - r_2)]}{n_1 r_2 - n_2 r_1} \quad .$$

It can be shown that (8) is indeed a consistent estimator of p_i. Using this estimator for p_i, possible estimators for d, R, and OR are $\hat{d}_3 = \hat{p}_{31} - \hat{p}_{32}$, $\hat{R}_3 = \hat{p}_{31}/\hat{p}_{32}$, and $\hat{OR}_3 = \hat{p}_{31}(1 - \hat{p}_{32})/\hat{p}_{32}(1 - \hat{p}_{31})$, respectively. Note that the estimator of OR, $\hat{OR}_3$, is identical to $\hat{OR}_1$. Under Model 3 these estimators have asymptotic normal distributions and are asymptotically unbiased and consistent--conditionally and unconditionally. The asymptotic conditional and unconditional variances are shown in Tables 6 and 7.

Notice that the Model 3 estimator for p_i fails for $p_1 = p_2$; both asymptotic variances are infinite for this case. Basically, for $p_1 = p_2 = p$ there is insufficient information in the samples to estimate p, q_1, and q_0. Thus we may not be able to obtain reasonable estimates of p_1 and p_2 using this procedure in cases where p_1 is close to p_2. To illustrate, consider the case $N_1 = N_2$. When $n_1 = n_2$, then

TABLE 6

Asymptotic Conditional Variance Under Model 3

Estimator	Variance
$\sqrt{N_1 + N_2}\ \hat{p}_{3i}$	$\frac{\theta_i(1-\theta_i)}{\tau_j^2(\theta_2-\theta_1)^4}\left\{\frac{1}{\lambda_i\tau_i}[\theta_j(\tau_i-\tau_j) + \theta_1^2\tau_1 + \theta_2^2\tau_2 - 2\tau_i\theta_1\theta_2]^2 + \frac{1}{\lambda_j\tau_j}[\theta_1\theta_2\ (1-\theta_1)(1-\theta_2)(\tau_1-\tau_2)^2]\right\}$
$\sqrt{N_1 + N_2}\ \hat{d}_3$	$\frac{1}{\tau_1^2\tau_2^2(\theta_2-\theta_1)^4}\left\{\frac{\theta_1(1-\theta_1)}{\lambda\tau_1}[\theta_2(1-\theta_2)(\tau_1-\tau_2)^2 + (\theta_2-\theta_1)^2\tau_1^2]^2 + \frac{\theta_2(1-\theta_2)}{(1-\lambda)\tau_2}[\theta_1(1-\theta_1)(\tau_1-\tau_2)^2 + (\theta_2-\theta_1)^2\tau_2^2]^2\right\}$
$\sqrt{N_1 + N_2}\ \hat{R}_3$	$(\frac{\tau_1}{\tau_2})^2\ \frac{1}{\theta_2^2}\left[\frac{\theta_1(1-\theta_1)}{\tau_1\lambda} + (\frac{\theta_1}{\theta_2})^2\ \frac{\theta_2(1-\theta_2)}{\tau_2(1-\lambda)}\right]$
$\sqrt{N_1 + N_2}\ \hat{OR}_3$	$\frac{\theta_1(1-\theta_2)}{\theta_2^2(1-\theta_1)^2}\left[\frac{1-\theta_2}{\tau_1\lambda(1-\theta_1)} + \frac{\theta_1}{\tau_2(1-\lambda)\theta_2}\right]$

where

$$\theta_i = E(\frac{r_i}{n_i}) = \frac{p_i q_1}{p_i q_1 + (1 - p_i)q_0}$$

$\hat{p}_{3i} = \hat{p}_{1i}$. Note, however, that if $r_2/n_2 = r_1/n_1$, $\hat{q}_1 = 0$ and $\hat{p}_i$ is undefined. If $n_1 - n_2 = r_1 - r_2$ then $\hat{q}_1 = (n_1 r_2 - n_2 r_1)/N{\cdot}0$ yielding $\hat{p}_i = 0$, another nonsense estimate. Even worse, $\hat{q}_1$ and $\hat{p}_i$ may both be negative; this will occur if $r_1 - (n_1 - n_2) < r_2 < n_2\ r_1/n_1$ or $\frac{n_2 r_1}{n_1} < r_2 < r_1 - (n_1 - n_2)$.

TABLE 7

Asymptotic Unconditional Variance Under Model 3

Estimator	Variance
$\sqrt{N_1 + N_2}\ \hat{p}_{3i}$	$\frac{p_i(1-p_i)}{\lambda(1-\lambda)q_1q_0(p_2-p_1)^2}\left\{\begin{array}{l} p_j(1-p_i)q_0\ [p_2 - \lambda(p_2-p_1)] \\ + p_i(1-p_j)q_1\ [(1-p_2) + \lambda(p_2-p_1)] \\ - p_i(1-p_i)q_1q_0 \end{array}\right\}$
$\sqrt{N_1 + N_2}\ \hat{d}_3$	$\frac{1}{q_0q_1\lambda(1-\lambda)}\left\{\begin{array}{l} p_1p_2\ q_0\ [p_2 - \lambda(p_2-p_1)] \\ + (1-p_1)(1-p_2)q_1\ [1-p_2 + \lambda(p_2-p_1)] \\ - q_0q_1\ (1-p_1-p_2)^2 \end{array}\right\}$
$\sqrt{N_1 + N_2}\ \hat{R}_3$	$\frac{p_1}{p_2^3q_1\lambda(1-\lambda)}\ [(1-\lambda)p_2 + \lambda p_1 - p_1p_2q_1]$
$\sqrt{N_1 + N_2}\ \hat{OR}_3$	$\frac{p_1(1-p_2)}{\lambda(1-\lambda)q_1q_0p_2^3(1-p_1)^3}\left[\begin{array}{l} p_2(1-p_2)[p_1q_1 + (1-p_1)\ q_0]\ (1-\lambda) \\ + p_1(1-p_1)\ [p_2q_1 + (1-p_2)\ q_0]\ \lambda \end{array}\right]$

This same problem is reflected in the behavior of the maximum likelihood estimators for Model 3. When $p_1 = p_2$, the information matrix is singular. For $p_1 \neq p_2$, numerical comparisons for parameter values listed below[2] indicate that the asymptotic variances of $\hat{p}_{3i}$, $\hat{d}_3$, $\hat{R}_3$, and $\hat{OR}_3$ are identical with those of the ML estimators of p_{3i}, d, R, and OR.

[2]Variance ratios were evaluated for p_1 = .1, .25, .50, .75, .90; p_2 = .1, .25, .50, .75, .90; q_1 = .5, .75, .90, 1.0; q_0 = .5, .75, .90, 1.0.

Detailed investigations of the behavior of the Model 3 estimators in large and small samples are reported in Sections 9, 10, and 11, while in Section 12 the testing of $H_0 : p_1 = p_2$ is discussed.

7. Both Variables Influence Missing Data (Model 4): Statistical Techniques for Problems (a) Through (e)

In Model 4, the probability that a particular observation is missing depends on both the value of x, the independent variable, and the value of y, the dependent variable. Therefore, the probability that a particular y observation is missing is different for each of the four x,y combinations. Without further assumptions or additional information, it is impossible to obtain consistent estimators of the p_i. No detailed studies of problems (a) through (e) were carried out for Model 4 since entirely new problems arise when this model holds. The following are four possible lines of attack.

(a) Assumptions can be made about relationships among the four probabilities $q(x,y)$ which would allow the use of techniques obtained for Model 2 or Model 3. For example, assume that missing observations are twice as likely in population 1 as in population 2.

(b) Estimates of the probabilities $q(x,y)$ may be obtained by a pilot study or intensive subsampling of nonrespondents (see e.g., Cochran, 1963).

(c) Use of some related variable z can be made. For instance, if a dichotomous variable z affects the probability distribution of y but does not influence $q(x,y)$, then Eklund (1959) has developed consistent estimators of the p_i.

(d) Estimators based upon Models 1, 2, and 3 could be employed if the magnitude of the biases when Model 4 holds were ascertained and the corresponding standard error formulae changed. That is, a robustness study could be made to find out the conditions under which these Model 1, Model 2, and Model 3 estimators give reasonable results. This point will be discussed in later sections.

8. Estimators of the p_i

In this section the concern is only with how well the p_i are estimated and not with how to estimate the variance of p_i. Since the Model 1 and Model 2 estimators of the p_i are the same, the estimation problem is reduced to a comparison of the behavior of $\hat{p}_{1i}$ and $\hat{p}_{3i}$ under Models 1 and 3. How much is lost if it is assumed observations were missing at random, if in fact $q_0 \neq q_1$? How much is lost by using the Model 3 estimators even though $q_0 = q_1$? To answer these questions it is necessary to examine asymptotic unconditional results for the bias, variance, and mean-squared error of the Model 1 and Model 3 estimators of p_i under Model 1 and Model 3. Since comparisons between p_1 and p_2 are the major interest, small sample work is reported only for d , R , and OR (see Sections 9, 10, 11, and 12).

The Model 1 and Model 3 estimators for p_1 are

$$\hat{p}_{11} = \frac{r_1}{n_1}$$

$$\hat{p}_{31} = \frac{r_1}{N_1}\left[\frac{N_2(n_1 - r_1) - N_1(n_2 - r_2)}{n_1 r_2 - n_2 r_1}\right] .$$

The estimator $\hat{p}_{31}$ is asymptotically unbiased with conditional and unconditional asymptotic variances given in Tables 6 and 7. Results for $\hat{p}_{11}$ under Model 3 are given in Table 8.

TABLE 8

Asymptotic Behavior of $\hat{p}_{11}$ Under Model 3

$E(\hat{p}_{11})$	θ_1	[exact]
Bias $(\hat{p}_{11})$	$\frac{(q_1-q_0)\, p_1(1-p_1)}{p_1q_1 + (1-p_1)q_0}$	[exact]
Var $\sqrt{N_1 + N_2}\, \hat{p}_{11}$		
conditional	$\frac{\theta_1\,(1-\theta_1)}{\tau_1\lambda}$	[exact]
unconditional	$\frac{\theta_1^2\,(1-\theta_1)}{p_1q_1\lambda}$	

Suppose Model 1 is true and $q_1 = q_0 = q$, how much is lost by using the Model 3 estimator of p_1? For simplicity, let $p_2 = p_1 + \Delta$ and $N_1 = N_2 = N$. Then under Model 1 both estimators are asymptotically unbiased and the conditional variance formulas for $\hat{p}_{11}$ and $\hat{p}_{31}$ become

$$\text{Var}\,(\hat{p}_{11}) = \frac{p_1(1-p_1)}{Nq}$$

$$\text{Var}\,(\hat{p}_{31}) = \frac{2}{N}\,\frac{p_1(1-p_1)}{q\Delta^2}\,[\Delta^2/2 + (1-q)p_1(1-p_1)]\,,$$

yielding

$$\frac{\text{Var } (\hat{p}_{31})}{\text{Var } (\hat{p}_{11})} = 1 + \frac{2(1-q)p_1(1-p_1)}{\Delta^2} .$$

Under Model 1 then, $\hat{p}_{31}$ always has a larger variance than $\hat{p}_{11}$ and gets worse in comparison with $\hat{p}_{11}$ as p_1 approaches 0.5 , as q approaches zero (the proportion of missing data increases) and as $\Delta = p_2 - p_1$ approaches zero.

Under Model 3, the asymptotic unconditional formulas for mean-squared errors are:

$$\text{MSE } (\hat{p}_{11}) = \frac{\theta_1^2}{q_1^2} \left[(q_1-q_0)^2(1-p_1)^2 + \frac{(1-\theta_1)q_1}{N_1 p_1}\right]$$

$$\text{MSE } (\hat{p}_{31}) = (N_1+N_2) \frac{p_1(1-p_1)}{N_1N_2q_1q_0(p_2-p_1)^2} \begin{cases} p_2(1-p_1)q_0\left[p_2 - \frac{N_1}{N_1+N_2}(p_2-p_1)\right] \\ + p_1(1-p_2)q_1\left[1-p_2 + \frac{N_1}{N_1+N_2}(p_2-p_1)\right] \\ - p_1(1-p_1)q_1q_0 \end{cases} .$$

As Δ approaches zero, MSE $(\hat{p}_{11})$ will be smaller than MSE $(\hat{p}_{31})$. However, for $p_1 \neq p_2$ and N large, the bias in $\hat{p}_{11}$, which increases with $|q_1-q_0|$ will make $\hat{p}_{31}$ preferable. In small samples, $\hat{p}_{31}$ is biased and may have a larger variance than asymptotic results indicate.

9. Comparisons of Model 1 and Model 3 Estimators of d

In this section the unconditional asymptotic and exact small sample behavior of estimators $\hat{d}_1$ and $\hat{d}_3$ under Models 1 and 3 are compared.

Model 1 and Model 3 estimators of $d = p_1 - p_2$ are:

$$\hat{d}_1 = \frac{r_1}{n_1} - \frac{r_2}{n_2}$$

$$\hat{d}_3 = \left(\frac{r_1}{N_1} - \frac{r_2}{N_2}\right)\left(\frac{N_2(n_1 - r_1) - N_1(n_2 - r_2)}{n_1 r_2 - n_2 r_1}\right) .$$

Results of the comparison indicate that the Model 1 estimator $\hat{d}_1$ will be preferable for $p_1 = p_2$, for $q_0 = q_1$ and for small N ($N \leq 50$). For $q_0 \neq q_1$, $|p_1 - p_2| \neq 0$, $\hat{d}_3$ will look better for large N. Next the three situations $p_1 = p_2$, $q_1 = q_0$, and the general case of Model 3 are discussed by comparing asymptotic results and by examining exact bias and mean square error for samples of $N_1 = N_2 = 20, 50$.

The Model 3 estimator, $\hat{d}_3$, is asymptotically unbiased with conditional and unconditional variances given in Tables 6 and 7. The behavior of $\hat{d}_1$ under Model 3 is given in Table 9.

TABLE 9

Asymptotic Behavior of $\hat{d}_1$ Under Model 3

$E(\hat{d}_1)$	$\theta_1 - \theta_2$	[exact]
Bias $(\hat{d}_1)$	$(q_1 - q_0)\left[\frac{p_1(1-p_1)}{p_1 q_1 + (1-p_1)q_0} - \frac{p_2(1-p_2)}{p_2 q_1 + (1-p_2)q_0}\right]$	[exact]
Var $\sqrt{N_1 + N_2}\,\hat{d}_1$		
conditional	$\frac{\theta_1(1-\theta_1)}{\tau_1 \lambda} + \frac{\theta_2(1-\theta_2)}{\tau_2(1-\lambda)}$	[exact]
unconditional	$\frac{\theta_1^2(1-\theta_1)}{\lambda p_1 q_1} + \frac{\theta_2^2(1-\theta_2)}{(1-\lambda)p_2 q_1}$	

Exact unconditional results for bias, variance and mean-square error were obtained for $\hat{d}_1$ and $\hat{d}_3$ for $N_1 = N_2 = 20, 50$, for 400 sets of parameter values $p_1, p_2 = .10, .25, .50, .75, .90$; $q_1, q_0 = .50, .75, .90, 1.0$. Results are summarized in Tables 10, 11, 12, and 13. Notice that except for sign changes in the bias, results for p_1, p_2 are identical to results for p_2, p_1 and, with q_0, q_1 reversed, to results for $1-p_1, 1-p_2$ and $1-p_2, 1-p_1$. Results were obtained conditional on $n_1 \neq 0$, $n_2 \neq 0$; for $n_1r_2 = n_2r_1$ $\hat{d}_3$ was defined to be 0.

When $p_1 = p_2$, both estimators are unbiased in large and small samples. The asymptotic unconditional variances of $\hat{d}_1$ and $\hat{d}_3$ respectively become

$$\frac{q_1q_0}{\lambda(1-\lambda)} \quad \frac{p(1-p)}{(pq_1 + (1-p)q_0)^3}$$

and $$\frac{1}{q_0q_1\lambda(1-\lambda)} [q_0p^3 + q_1(1-p)^3 - q_0q_1(1-2p)^2] .$$

Table 10a shows the ratio of the unconditional asymptotic variance formulas for several values of p, q_0, and q_1. (Note that the conditional variance of $\hat{d}_3$ is infinite for $p_1 = p_2$.) The ratio is always less than 1.0, indicating that for $p_1 = p_2$, d_1 is to be preferred. Table 12a shows the exact ratio; $\hat{d}_1$ is even more strongly preferable in small samples.

When $q_1 = q_0$, that is, when Model 1 obtains, $\hat{d}_1$ is unbiased in large and small samples; $\hat{d}_3$ is unbiased in large samples but has bias ranging from .001 to .075 in absolute value for samples of size 20 and from .001 to .045 for samples of size 50 (see Table 11c). The

bias ranges up to 39 and 26 percent of d for samples of size 20 and 50 respectively. The asymptotic variance formulas for $N_1 = N_2 = N$ are related by

$$\text{Var } \hat{d}_3 = \text{Var } \hat{d}_1 + \frac{4}{qN}(1-q)(1-p_1-p_2)^2 .$$

They are equal only for N infinite, $q = 1$ or $p_1 + p_2 = 1$; otherwise var $\hat{d}_3 >$ var $\hat{d}_1$ by an amount which increases as q decreases and as $p_1 + p_2$ differs from 1 . See Table 10b for ratios of the variances. Table 12b shows the ratio of exact mean-squared errors for N = 20, 50 . These results favor $\hat{d}_1$ more strongly than asymptotic comparisons would indicate.

For the general case of Model 3 when $p_1 \neq p_2$ and $q_1 \neq q_0$, $\hat{d}_1$ is biased and $\hat{d}_3$ unbiased in large samples. The asymptotic unconditional ratio of MSE $(\hat{d}_1)$ to var $\hat{d}_3$ is shown in Table 10. These asymptotic comparisons indicate that for small samples (N = 20) $\hat{d}_1$ is preferred for p_1 close to p_2 , $\hat{d}_3$ is preferred for $|p_1-p_2|$ large. For samples as large as 200 , the bias in $\hat{d}_1$ makes $\hat{d}_3$ appear preferable except for some cases where $|p_1-p_2|$ is small. The exact bias in $\hat{d}_1$ is independent of N and ranges up to .12 in absolute value and up to 45% of d for the cases considered; it increases in absolute value as $|q_0-q_1|$ increases. The absolute bias in $\hat{d}_3$ ranges up to .06 for N = 20 and .04 for N = 50 ; maximum percentage bias is 39 for N = 20 and 26 for N = 50 (see Table 11). For a given p_1, p_2 the bias in $\hat{d}_3$ is always one-sided while the bias in $\hat{d}_1$ may be either positive or negative. The bias in $\hat{d}_3$ decreases slowly with N , with increasing $|p_1-p_2|$, and wih increasing $q_0 + q_1$. The exact ratio of unconditional mean-squared errors (Table 12) generally favors

TABLE 10

Ratio of Asymptotic Unconditional Formulas for MSE $\hat{d}_1$ and MSE $\hat{d}_3$ [a]

$$\frac{\text{MSE } \hat{d}_1}{\text{MSE } \hat{d}_3}$$

a) When $p_1 = p_2$, the ratio is independent of N, both estimators are asymptotically unbiased. (For $q_0 = q_1 = 1$, the ratio is 1.0.)

p_1 [b]	p_2	Min	Max
.10	.10	.220	.926
.25	.25	.600	.962
.50	.50	.790	.994

b) When $q_0 = q_1$, the ratio is independent of N, both estimators are asymptotically unbiased. (For $q_0 = q_1 = 1$, the ratio is 1.0.)

$q_0 = q_1 \neq 1$

p_1	p_2	Min	Max
.10	.25	.396	.766
	.50	.680	.914
	.75	.925	.984
	.90	1.000	1.000
.25	.50	.875	.972
	.75	1.000	1.000

c) For $q_0 \neq q_1$, $\hat{d}_3$ is asymptotically unbiased.

		N = 20		N = 200	
p_1	p_2	Min	Max	Min	Max
.10	.25	.474	.961	.708	3.169
	.50	.696	1.712	1.033	8.636
	.75	.789	1.896	1.002	7.452
	.90	1.005	1.284	1.006	2.075
.25	.50	.757	.995	.812	2.155
	.75	.985	.999	.999	1.585

[a] Formulas evaluated for p_1, p_2 of .1, .25, .50, .75, .90; q_0, q_1 of .5, .75, .90, 1.0.

[b] Due to symmetries in the formulas, all other cases in p_1, p_2 reduce to those shown.

TABLE 11

Exact Unconditional Bias of $\hat{d}_1$, $\hat{d}_3$ for $N_1 = N_2$ [a]

a) For $p_1 = p_2$, both $\hat{d}_1$ and $\hat{d}_3$ are unbiased for all N .

b) The bias in $\hat{d}_1$ is independent of N . For $q_0 = q_1$, $\hat{d}_1$ is unbiased. For $q_0 \neq q_1$:

		Bias $\hat{d}_1$		$-\frac{100 \text{ Bias}}{d}$	
p_1	p_2	Min	Max	Min	Max
.10	.25	-.0681	.0597	-45	39
	.50	-.0848	.1191	-21	29
	.75	-.0271	.1025	4	15
	.90	.0008	.0344	.1	4
.25	.50	-.0179	.0594	-7	23
	.75	.0010	.0428	.2	9

c) For $\hat{d}_3$:

		Bias $\hat{d}_3$							
		N = 20				N = 50			
		$q_0 = q_1 \neq 1$		$q_0 \neq q_1$		$q_0 = q_1 \neq 1$		$q_0 \neq 1$	
p_1	p_2	Min	Max	Min	Max	Min	Max	Min	Max
.10	.25	.0094	.0557	.0022	.0593	.0052	.0391	.0009	.0393
	.50	.0059	.0579	.0014	.0511	.0021	.0203	.0005	.0170
	.75	.0027	.0288	.0009	.0222	.0010	.0094	.0003	.0074
	.90	.0014	.0143	.0007	.0093	--	--	--	--
.25	.50	.0115	.0746	.0042	.0598	.0050	.0453	.0019	.0352
	.75	.0051	.0542	.0026	.0359	.0018	.0176	.0009	.0116

[a] Exact unconditional results obtained for $p_1, p_2 = .10, .25, .50, .75, .90$; $q_0, q_1 = .50, .75, .90, .999$. Due to symmetries in the distribution, all other cases reduce to those shown with possible sign changes.

TABLE 11 (continued)

$-\frac{100\ \text{Bias}}{d}$

		N = 20				N = 50			
		$q_0 = q_1 \neq 1$		$q_0 \neq q_1$		$q_0 = q_1 \neq 1$		$q_0 \neq q_1$	
p_1	p_2	Min	Max	Min	Max	Min	Max	Min	Max
.10	.25	6.2	37	1.4	39	3.4	26	.6	26
	.50	1.4	14	.4	21	.5	5.0	.1	4.2
	.75	.4	4.4	.1	3.4	.2	1.4	.0	1.1
	.90	.2	1.7	.9	1.1	--	--	--	--
.25	.50	4.6	29	1.6	23	2.0	18	.8	14
	.75	1.0	10	.5	7.1	.4	3.5	.2	2.3

$\hat{d}_1$ except for some cases where $|p_1-p_2|$ is large and $N = 50$. Generally the ratio tends to increase as q_1, q_0 increase; that is, $\hat{d}_3$ looks worse as the proportion of missing data increases.

Table 13 gives the ratio of the exact to the asymptotic unconditional variances for $\hat{d}_1$ and $\hat{d}_3$ for $N = 20$ and $N = 50$. For N as small as 20, the asymptotic variance formula is quite close to the exact variance for $\hat{d}_1$; for $\hat{d}_3$ the asymptotic formula does not provide a reasonable approximation. For $p_1 = p_2$, the exact variance of $\hat{d}_3$ goes up with N, and for p_1 close to p_2, the exact variance does not decrease as fast as 1/N. Generally the ratio of exact to asymptotic variance is largest for q_0 or q_1 small as would be expected. Note that the ordinary estimator of the conditional variance of $\hat{d}_1$ should be a good estimate of its conditional variance under Model 3.

In summary, for $p_1 = p_2$ or $q_1 = q_0$, or N small to moderate, $\hat{d}_1$ is the preferred estimator. For N large, $p_1-p_2 \neq 0$, and q_1

TABLE 12

Exact Ratio of Unconditional Formulas for MSE $\hat{d}_1$ and MSE $\hat{d}_3$[a]

$$\frac{\text{MSE } \hat{d}_1}{\text{MSE } \hat{d}_3}$$

a) For $p_1 = p_2$, both $\hat{d}_1$ and $\hat{d}_3$ are unbiased. The ratio increases as q_0, q_1 increase. For $q_0 = q_1 = 1$, the ratio is 1.0.

		N = 20		N = 50	
p_1	p_2	Min	Max	Min	Max
.10	.10	.07	.95	.02	.97
.25	.25	.06	.93	.02	.67
.50	.50	.07	.73	.02	.43

b) For $q_0 = q_1 \neq 1$, $\hat{d}_1$ is unbiased.

		N = 20		N = 50	
p_1	p_2	Min	Max	Min	Max
.10	.25	.08	.53	.04	.40
	.50	.21	.83	.46	.90
	.75	.54	.96	.84	.98
	.90	.77	.98	--	--
.25	.50	.10	.60	.07	.70
	.75	.27	.94	.73	.99

c) For $q_0 \neq q_1$:

		N = 20		N = 50	
p_1	p_2	Min	Max	Min	Max
.10	.25	.09	.94	.05	.94
	.50	.17	1.00	.57	1.62
	.75	.42	1.16	.76	2.67
	.90	.88	1.03	--	--
.25	.50	.08	.85	.04	.89
	.75	.21	.97	.77	.99

[a]Exact unconditional results obtained for $p_1, p_2 = .10, .25, .50, .75, .90$; $q_0, q_1 = .50, .75, .90, .999$. Due to symmetries in the distributions, all other cases reduce to those shown. Generally speaking, the ratio increases as q_1, q_0 increase.

and q_0 known to be unequal, $\hat{d}_3$ may be employed. In other words, unless it is reasonably sure that Model 3 pertains and $p_1 \neq p_2$, more will be lost than gained by using $\hat{d}_3$.

TABLE 13

Ratio of Exact to Asymptotic Unconditional Variance of d^a

(Excluding $q_0 = q_1 = 1$ for Which Ratio Is 1.0)

		N = 20				N = 50			
		$\hat{d}_1$		$\hat{d}_3$		$\hat{d}_1$		$\hat{d}_3$	
p_1	p_2	Min	Max	Min	Max	Min	Max	Min	Max
.10	.10	1.00	1.06	.88	5.10	1.00	1.02	.98	17.55
	.25	1.00	1.06	1.02	6.27	1.00	1.02	1.03	12.41
	.50	1.00	1.06	1.02	4.32	1.00	1.02	1.00	1.77
	.75	1.00	1.06	1.01	1.94	1.00	1.02	1.00	1.12
	.90	1.00	1.06	1.01	1.37	1.00	1.02	--	--
.25	.25	1.00	1.06	1.04	9.82	1.00	1.02	1.43	35.4
	.50	1.00	1.06	1.17	9.29	1.00	1.02	1.12	18.4
	.75	1.00	1.06	1.04	4.90	1.00	1.02	1.01	1.39
.50	.50	1.00	1.06	1.37	15.21	1.00	1.02	2.32	61.50

[a] Exact unconditional results were obtained for $p_1, p_2 = .10, .25, .50, .75, .90$; $q_0, q_1 = .50, .75, .90, .999$. Due to symmetries in the distributions of $\hat{d}_1$, $\hat{d}_3$, all other cases reduce to those shown. [Note, variances were calculated conditional on $n_1 \neq 0$, $n_2 \neq 0$; for $n_1 r_2 = n_2 r_1$, define $\hat{d}_3 = 0$.]

10. Comparisons of $\hat{R}_1$ and $\hat{R}_3$

The estimators of the ratio p_1/p_2 are

$$\hat{R}_1 = \frac{r_1}{r_2}\frac{n_2}{n_1}$$

and $\hat{R}_3 = \frac{r_1}{r_2}\frac{N_2}{N_1}$.

In this section the unconditional asymptotic and exact small sample behavior of $\hat{R}_1$ and $\hat{R}_3$ under Models 1 and 3 are compared. Results show that for $p_1 = p_2$, $q_0 = q_1$, or N small to moderate, $\hat{R}_1$ is moderately preferable to $\hat{R}_3$.

The Model 3 estimator, $\hat{R}_3$, is asymptotically unbiased with conditional and unconditional variances given in Tables 6 and 7. The behavior of $\hat{R}_1$ under Model 3 is given in Table 14.

TABLE 14

Asymptotic Behavior of $\hat{R}_1$ Under Model 3

$E(\hat{R}_1)$	$\frac{\theta_1}{\theta_2}$
Bias $(\hat{R}_1)$	$\frac{(q_1-q_0)(p_2-p_1)\theta_1}{p_2 q_1}$
Var $\sqrt{N_1 + N_2}\,\hat{R}_1$	
conditional	$\frac{1}{\theta_2^2}\left[\frac{\theta_1(1-\theta_1)}{\tau_1\lambda} + \left(\frac{\theta_1}{\theta_2}\right)^2 \frac{\theta_2(1-\theta_2)}{\tau_2(1-\lambda)}\right]$
unconditional	$\frac{\theta_1^2 q_0}{\theta_2^2 q_1^2}\,\frac{[p_2^2(1-p_1)\theta_1\,(1-\lambda) + p_1^2(1-p_2)\theta_2\lambda]}{p_1^2 p_2^2\,(1-\lambda)\lambda}$

Under Model 3, the conditional variances of $\hat{R}_1$ and $\hat{R}_3$ have the ratio

$$\frac{\text{Var } \hat{R}_3}{\text{Var } \hat{R}_1} = \left(\frac{\tau_1}{\tau_2}\right)^2$$

which in large samples will be approximately

$$\left[\frac{q_0 + p_1\ (q_1-q_0)}{q_0 + p_2\ (q_1-q_0)}\right]^2$$

and consequently will be greater or less than 1.0 for p_1/p_2 greater or less than 1.0 .

Exact unconditional results for the bias, variance, and mean square error were obtained for $\hat{R}_1$ and $\hat{R}_3$ for $N_1 = N_2 = 20, 50$, and for 400 parameter sets in p_1, p_2, q_1, q_0 . Results were obtained conditional on $n_1 \neq 0$, $n_2 \neq 0$, and are summarized in Tables 15, 16, and 17.

For $p_1 = p_2$, both $\hat{R}_1$ and $\hat{R}_3$ are asymptotically unbiased. In small samples the range of the bias is generally comparable for the two estimators although always slightly less for $\hat{R}_1$ than for $\hat{R}_3$ (see Table 16a). The biases are generally positive and range up to 30% of R ; the biases decrease as p_1 , p_2 increase.

The ratio of the asymptotic unconditional variances is

$$\frac{\text{Var } \hat{R}_1}{\text{Var } \hat{R}_3} = \frac{1-\theta}{1-pq_1}$$

which is always less than 1.0 except for $q_1 = q_0 = 1$. The ratios have been evaluated in Table 15a. The exact ratio of mean-squared errors is shown in Table 17a and is quite similar to asymptotic results even for N = 20 . Therefore, for $p_1 = p_2$, the estimator $\hat{R}_1$ is clearly preferable to $\hat{R}_3$.

When Model 1 is true and $q_1 = q_0$ but $p_1 \neq p_2$, both $\hat{R}_1$ and $\hat{R}_3$ are asymptotically unbiased. The biases in $\hat{R}_1$ and $\hat{R}_3$ are usually positive and show very similar ranges. The percentage bias depends only on p_2 and decreases as p_2 increases (see Table 16). The ratio of the

asymptotic unconditional variances is

$$\frac{\text{Var } \hat{R}_1}{\text{Var } \hat{R}_3} \approx \frac{p_2(1-\lambda) + \lambda p_1 - p_1 p_2}{p_2(1-\lambda) + \lambda p_1 - q p_1 p_2} \leq 1.0 .$$

This ratio is evaluated in Table 15b. The exact unconditional ratio of MSE $\hat{R}_1$ to MSE $\hat{R}_3$ is shown in Table 17b. The small sample comparison favors $\hat{R}_1$ somewhat more than the asymptotic results. Therefore, under Model 1 $\hat{R}_1$ is to be preferred, although for p_2 small, the gain in using $\hat{R}_1$ may be relatively small.

Under Model 3, when $q_1 \neq q_0$ and $p_1 \neq p_2$, $\hat{R}_3$ is asymptotically unbiased and $\hat{R}_1$ is asymptotically biased. Except for p_2 small, $\hat{R}_3$ shows a smaller range for exact bias and its bias decreases with increasing N and increasing q_1 (it is almost unaffected by q_0). The ratio of asymptotic unconditional mean-squared errors is shown in Table 15c. For an N as small as 20 there is no clear-cut choice between $\hat{R}_1$ and $\hat{R}_3$; by $N = 200$ $\hat{R}_3$ is clearly preferable. The small sample results for $N = 20$ shown in Table 17c are quite similar to those obtained using asymptotic formulas. Although $\hat{R}_3$ improves with N, exact results do not clearly favor either estimator, even for N as large as 50.

The ratio of exact to asymptotic variance is quite similar for $\hat{R}_1$ and $\hat{R}_3$. The exact variance is generally larger except for $p_1 = p_2$ and $N = 20$. For $N = 50$, the ratios vary from 1.0 to 3.7, being close to 1.0 for $R < 1$ and larger for $R > 1$.

In conclusion, then, for $p_1 = p_2$, $q_0 = q_1$, or N small to moderate, $\hat{R}_1$ is moderately preferable to $\hat{R}_3$. For N large, $p_1 \neq p_2$, $q_1 \neq q_0$, $\hat{R}_3$ is preferable to $\hat{R}_1$. For other situations, the choice depends on the parameter values.

TABLE 15

Ratio of Asymptotic Unconditional Formulas for MSE $\hat{R}_1$ and MSE $\hat{R}_3$

$$\frac{\text{MSE } \hat{R}_1}{\text{MSE } \hat{R}_3}$$

a) When $p_1 = p_2$, both $\hat{R}_1$ and $\hat{R}_3$ are asymptotically unbiased and the ratio is independent of N. (For $q_0 = q_1 = 1$, the ratio is 1.0.)

p_1	p_2	Min	Max
.10	.10	.91	1.00
.25	.25	.80	.99
.50	.50	.65	.96
.75	.75	.40	.92
.90	.90	.18	.91

b) When $q_0 = q_1$, the ratio is independent of N. For $q_0 = q_1 = 1.0$, the ratio is 1.0. The ratio is symmetric in p_1, p_2.

$q_0 = q_1 \neq 1$

p_1	p_2	Min	Max
.10	.25	.92	.98
	.50	.91	.98
	.75	.90	.98
	.90	.90	.98
.25	.50	.80	.95
	.75	.77	.94
	.90	.76	.94
.50	.75	.57	.87
	.90	.53	.85
.75	.90	.31	.69

c) For $q_0 \neq q_1$, $\hat{R}_3$ is asymptotically unbiased.

		N = 20		N = 200	
p_1	p_2	Min	Max	Min	Max
.10	.25	.85	1.17	.89	1.45
	.50	.65	1.91	.88	4.29
	.75	.53	3.02	.95	9.75
	.90	.49	3.86	.96	14.31

TABLE 15 (continued)

p_1	p_2	N = 20 Min	N = 20 Max	N = 200 Min	N = 200 Max
.25	.10	.71	1.18	.84	1.23
	.50	.74	1.30	.90	3.10
	.75	.65	2.48	1.01	11.12
	.90	.62	3.56	1.08	19.21
.50	.10	.62	1.64	.91	2.22
	.25	.65	1.35	.91	1.92
	.75	.61	1.29	.90	5.04
	.90	.58	2.33	1.07	13.84
.75	.10	.61	2.50	.95	4.85
	.25	.64	2.16	1.03	5.48
	.50	.54	1.32	.94	3.53
	.90	.39	0.99	.64	3.97
.90	.10	.59	3.38	.98	8.09
	.25	.67	3.08	1.12	10.73
	.50	.58	1.98	1.16	8.38
	.75	.35	0.90	.71	3.28

TABLE 16

Exact Unconditional Bias for $\hat{R}_1$, $\hat{R}_3$ as a Percent of R^a

a) For $\hat{R}_3$, which is asymptotically unbiased, the percentage bias is independent of p_1 (the range is only slightly larger for $q_0 \neq q_1$ than for $q_0 = q_1$).

p_2	N = 20 Min	N = 20 Max	N = 50 Min	N = 50 Max
.10	-23	16	24	31
.25	22	28	7	20
.50	6	22	2	7
.75	2	11	1	4
.90	1	8	--	--

[a] Excluding $q_0 = q_1 = 1.0$

TABLE 16 (continued)

b) For $q_0 = q_1 \neq 1$, $\hat{R}_1$ is asymptotically unbiased and the percentage bias is independent of p_1.

	N = 20		N = 50	
p_2	Min	Max	Min	Max
.10	-23	10	23	30
.25	20	26	8	18
.50	7	15	2	5
.75	2	4	1	1
.90	1	1	--	--

c) For $q_0 \neq q_1$, $\hat{R}_1$ is asymptotically biased.

		N = 20		N = 50	
p_1	p_2	Min	Max	Min	Max
.10	.10	-23	15	24	30
	.25	13	33	8	20
	.50	-5	40	-15	40
	.75	-31	62	-31	62
	.90	-45	72	--	--
.25	.10	-23	13	9	36
	.25	18	27	6	20
	.50	4	24	-8	22
	.75	-24	42	-27	39
	.90	-37	55	--	--
.50	.10	-23	15	-9	57
	.25	-2	44	-12	40
	.50	4	20	1	6
	.75	-11	18	-15	15
	.90	-25	27	--	--
.75	.10	-29	20	-22	89
	.25	-16	72	-24	68
	.50	-11	43	-13	27
	.75	2	7	0	2
	.90	-10	8	--	--
.90	.10	-34	33		
	.25	-23	96		
	.50	-14	63		
	.75	-7	22		
	.90	0	3		

TABLE 17

Exact Unconditional Ratio of MSE $\hat{R}_1$ to MSE $\hat{R}_3$

a) For $p_1 = p_2$ (excluding $q_1 = q_0 = 1$) :

		N = 20		N = 50	
p_1	p_2	Min	Max	Min	Max
.10	.10	.93	1.04	.87	1.00
.25	.25	.73	.99	.76	.99
.50	.50	.56	.95	.61	.95
.75	.75	.32	.92	.37	.92
.90	.90	.15	.90	--	--

b) For $q_0 = q_1$ (for $q_0 = q_1 = 1$, the ratio is 1.0):

		N = 20		N = 50	
p_1	p_2	Min	Max	Min	Max
.10	.25	.91	.97	.85	.97
	.50	.75	.96	.85	.97
	.75	.73	.96	.85	.97
	.90	.76	.96	--	--
.25	.10	.98	1.01	.92	.98
	.50	.65	.93	.73	.94
	.75	.60	.92	.72	.94
	.90	.61	.92	--	--
.50	.10	.96	.98	.90	.98
	.25	.79	.95	.77	.95
	.75	.43	.85	.53	.86
	.90	.40	.83	--	--
.75	.10	.95	.98	.89	.98
	.25	.76	.94	.75	.94
	.50	.52	.86	.54	.87
	.90	.23	.67	--	--
.90	.10	.94	.97		
	.25	.74	.94		
	.50	.50	.85		
	.75	.26	.68		

TABLE 17 (continued)

c) For $q_0 \neq q_1$:

		N = 20		N = 50	
p_1	p_2	Min	Max	Min	Max
.10	.25	.83	1.07	.81	1.20
	.50	.56	1.86	.63	2.30
	.75	.41	3.05	.62	4.14
	.90	.39	3.94	--	--
.25	.10	.67	1.08	.62	1.23
	.50	.63	1.26	.69	1.61
	.75	.46	2.47	.68	3.91
	.90	.47	3.58	--	--
.50	.10	.52	1.26	.43	1.91
	.25	.46	1.43	.57	1.46
	.75	.45	1.29	.59	1.92
	.90	.41	2.32	--	--
.75	.10	.56	1.60	.38	3.24
	.25	.38	2.35	.51	2.45
	.50	.41	1.32	.53	1.62
	.90	.29	.99	--	--
.90	.10	.61	2.27		
	.25	.36	3.34		
	.50	.38	1.86		
	.75	.30	.87		

11. The Estimation of the Odds Ratio OR

The Model 1, Model 2, and Model 3 estimators of OR all reduce to

$$\hat{OR} = \frac{r_1 (n_2 - r_2)}{r_2 (n_1 - r_1)} .$$

This estimator is asymptotically unbiased under all three models with asymptotic unconditional variances under the three models given in Table 18.

TABLE 18

Asymptotic Unconditional Variance of $\sqrt{N_1 + N_2}\ \hat{OR}$

Model	Variance
1	$\frac{p_1(1-p_2)}{\lambda(1-\lambda)qp_2^3(1-p_1)^3}\,[p_2(1-p_2)(1-\lambda) + p_1(1-p_1)\lambda]$
2	$\frac{p_1(1-p_2)}{\lambda(1-\lambda)p_2^3(1-p_1)^3q_1q_2}\,[p_2(1-p_2)q_2(1-\lambda) + p_1(1-p_1)q_1\lambda]$
3	$\frac{p_1(1-p_2)}{\lambda(1-\lambda)q_1q_0p_2^3(1-p_1)^3}\,[p_2(1-p_2)[p_1q_1 + (1-p_1)q_0](1-\lambda) + p_1(1-p_1)[p_2q_1 + (1-p_2)q_0]\lambda]$

Alternatively, the asymptotic variances are given by

$$\frac{p_1(1-p_2)}{\lambda(1-\lambda)p_2^3(1-p_1)^3}\ f(i)$$

where

$$f(1) = \frac{p_1(1-p_1)\lambda}{q} + \frac{p_2(1-p_2)(1-\lambda)}{q},$$

$$f(2) = \frac{p_1(1-p_1)\lambda}{q_2} + \frac{p_2(1-p_2)(1-\lambda)}{q_1},$$

$$f(3) = \frac{p_1(1-p_1)\lambda}{q_0}\,\frac{p_2}{\theta_2} + \frac{p_2(1-p_2)(1-\lambda)}{q_0}\,\frac{p_1}{\theta_1}.$$

This independence of the form of the estimator from $q(x,y)$ suggests that the use of $\hat{OR}$ will be robust to $q(x,y)$. Further investigation of $\hat{OR}$ is then in order.

Under $p_1 = p_2$ the variance formulas reduce to

$$\frac{1}{\lambda(1-\lambda)p(1-p)} \left\{ \begin{array}{ll} \frac{1}{q} & \text{Model 1} \\ \frac{\lambda}{q_2} + \frac{(1-\lambda)}{q_1} & \text{Model 2} \\ \frac{p}{q_0\theta} & \text{Model 3} \end{array} \right.$$

Table 19 shows the exact bias in $\hat{OR}$ under Model 3 for $N_1 = N_2 = 20$; Table 20 shows the performance of the asymptotic variance formula for $N_1 = N_2 = 20$. Generally the bias is of the order of 20% to 50% of OR , although it does not contribute appreciably to MSE . This suggests a modification of $\hat{OR}$ to reduce bias along the lines suggested by Anscombe (1956) and Gart & Zweifel (1967) for estimating the logit. The exact behavior of OR does not seem to depend particularly on $|p_1 - p_2|$ or $|q_1 - q_0|$.

TABLE 19

Exact Unconditional Bias of $\hat{OR}$ for $N_1 = N_2 = 20$[a]

p_1[b]	p_2	Bias		100 Bias / OR	
		Min	Max	Min	Max
.10	.10	-.195	.331	-20	33
	.25	.116	.176	39	59
	.50	.0207	.0522	23	58
	.75	.00505	.0126	19	47
	.90	.00151	.00363	19	45

[a] For q_0, q_1 = .5, .75, .90, 1.0 .

[b] For the other cases in p_1, p_2 , note that p_1, p_2 is equivalent to $(1-p_2)$, $(1-p_1)$ with the q's reversed.

TABLE 19 (continued)

p_1	p_2	Bias Min	Bias Max	100 Bias/OR Min	100 Bias/OR Max
.25	.10	-.553	1.134	-18	28
	.25	.365	.578	37	58
	.50	.0672	.174	20	52
	.75	.0168	.0428	15	39
.50	.10	1.339	4.632	-15	51
	.25	1.275	2.206	43	75
	.50	.254	.670	25	67
.75	.10	-.866	15.427	-3	57
	.25	5.585	7.221	62	80
.90	.10	-34.198	30.053	-42	37

TABLE 20

Ratios of Exact to Asymptotic Variance and MSE[a]

for $\hat{OR}$ for $N_1 = N_2 = 20$

p_1[b]	p_2	Variance Min	Variance Max	MSE Min	MSE Max
.10	.10	.426	2.100	.444	2.190
	.25	2.304	4.550	2.384	4.795
	.50	1.860	4.420	1.906	4.566
	.75	1.610	3.399	1.632	3.469
	.90	1.660	3.299	1.674	3.338
.25	.10	.314	2.284	.337	2.435
	.25	2.219	5.541	2.362	5.966
	.50	1.984	5.300	2.071	5.593
	.75	1.647	3.879	1.690	4.018
.50	.10	.316	3.157	.332	3.448
	.25	2.520	7.673	2.756	8.417
	.50	2.439	7.389	2.601	7.950
.75	.10	.469	3.097	.470	3.444
	.25	3.080	7.930	3.441	8.852
.90	.10	.072	1.244	.152	1.368

[a] For $q_0, q_1 = .5, .75, .90, 1.0$.

[b] For the other cases in p_1, p_2 , note that p_1, p_2 is equivalent to $(1-p_2), (1-p_1)$ with the q's reversed. The extremes usually occur at $q_1, q_0 = (.5, .5), (1.0, 1.0), (.5, 1.0), (1.0, .5)$.

12. Conclusions: Tests and Confidence Intervals Under Models 1, 2, or 3

A test of $H_0 : p_1 = p_2$ may be carried out using the Irwin-Fisher exact test for the 2x2 table of r_i and $n_i - r_i$, conditional on n_1, n_2, and $r_1 + r_2$. The tabled significance values and the power function will be correct under all three models. If Model 4 obtains, an accurate test of $p_1 = p_2$ cannot be performed without additional information.

The major issues in point and interval estimation are the choice of an estimator and the calculation of a variance. For the estimation of d, $\hat{d}_1$ is the estimator of choice for Models 1 and 2, and, though biased may be useful for Model 3 unless N is larger than 50 and p_1 and p_2 are known to be widely different. The ordinary estimator of the conditional variance of $\hat{d}_1$ should perform well under all three models.

To estimate R, use $\hat{R}_1$ in Models 1 and 2; under Model 3 the choice between $\hat{R}_1$ and $\hat{R}_3$ depends strongly on the values of the parameters. Modification of these estimators to reduce bias is of interest. It is common to base confidence intervals on the large sample normal distributions of $\hat{R}$. In small samples the large sample standard error is biased. In addition, it may be sensible to estimate the large sample conditional variance formula for Model 3 by substituting r_i/n_i for the θ_i, but there is no good estimator for the p_i and q_j of the unconditional formula.

To estimate OR, $\hat{OR}$ (or a modification to reduce bias) can be used under all three models. Uniformly most-accurate confidence intervals can be constructed for OR using the noncentral distribution of r_1,

r_2 conditional on $(r_1 + r_2)$, n_1 , n_2 (see Lehmann, 1959). This non-central distribution is the same under all three models.

In conclusion, then, the effect of different models for missing data depends on the inference problem at hand. Choice of a test for H_0 : $p_1 = p_2$ and an estimator for OR is the same for Models 1, 2, and 3. Estimation of p and d and R is the same for Models 1 and 2 but may be difficult for Model 3. Under Model 4, additional information is necessary.

References

Anscombe, F. J. On estimating binomial response relations. Biometrika, 1956, 43, 461-464.

Cochran, W. G. Sampling techniques. New York: John Wiley & Sons, 1963.

Eklund, G. Selection bias in applied statistics. Uppsala: Almqvist & Wiksell boktr., 1959.

Gart, J. J. & Zweifel, J. R. On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika, 1967, 54, 181-187.

Haldane, J. B. S. The estimation and significance of the logarithm of a ratio of frequencies. Annals of Human Genetics, 1955, 20, 309-311.

Lehmann, E. L. Testing statistical hypotheses. New York: John Wiley & Sons, 1959.