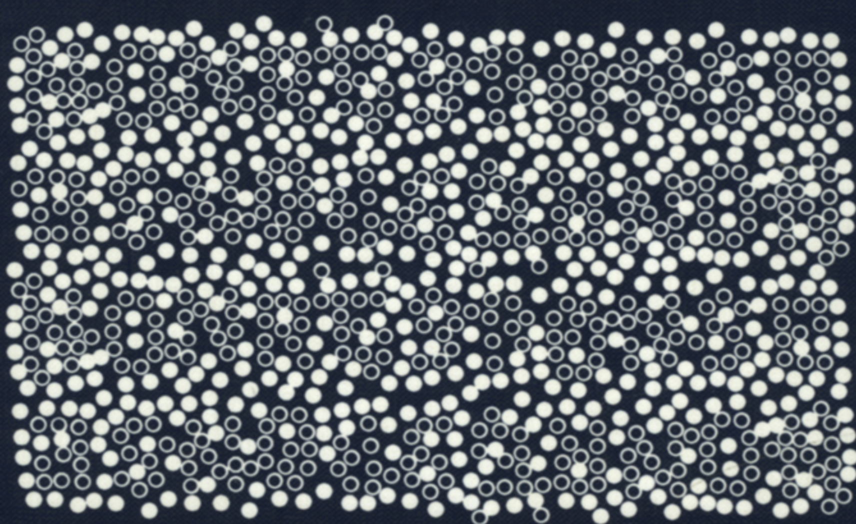
# Nonresponse in Household Interview Surveys

Robert M. Groves
Mick P. Couper



WILEY SERIES IN
PROBABILITY AND STATISTICS:
SURVEY METHODOLOGY SECTION

# Nonresponse in Household Interview Surveys

WILEY SERIES IN PROBABILITY AND STATISTICS
*SURVEY METHODOLOGY SECTION*

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *Robert M. Groves, Graham Kalton, J. N. K. Rao, Norbert Schwarz, Christopher Skinner*

A complete list of the titles in this series appears at the end of this volume.

# Nonresponse in Household Interview Surveys

ROBERT M. GROVES
MICK P. COUPER
University of Michigan
Ann Arbor, Michigan
and
Joint Program in Survey Methodology
College Park, Maryland

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This text is printed on acid-free paper. ⊗



***Library of Congress Cataloging in Publication Data:***

Groves, Robert M.
Nonresponse in household interview surveys / Robert M. Groves, Mick P. Couper.
p. cm. — (Wiley series in probability and statistics. Survey methodology section)
"Wiley-Interscience publication."
Includes bibliographical references and index.
ISBN 0-471-18245-1 (cloth : alk. paper)
1. Household surveys. I. Couper, Mick. II. Title. III. Series.
HB849.49.G757 1998

97-39223
CIP

Printed in the United States of America

10 9 8 7 6 5 4 3

# Contents

# Preface

This book was written out of frustration. Its genesis came in 1986–1988 when a review of the then extant research literature of survey nonrepsonse yielded few answers to the question, "How important is nonresponse to surveys?"

In teaching courses in survey methodology, it was common for us to emphasize that once a probability sample had been drawn, full measurement of the sample was crucial for proper inference to apply. Bright students would sometimes question, "How do we know when nonresponse implies error and when it doesn't? Is it cheaper and more effective to reduce nonresponse error by decreasing nonresponse rates or by adjusting for it *post hoc*? Is it more important to reduce nonresponse due to noncontact or nonresponse due to refusals? Why, after all, do people choose not to cooperate with survey requests?" We felt unprepared for such questions and, indeed, grew to believe that the lack of answers was a pervasive weakness in the field, not just a result of our ignorance.

Gathering information *post hoc* about nonrespondents from diverse surveys, which formed one of the central databases of this book, was an attempt to address a critical weakness in the area—the lack of common information about nonresponse across several surveys. (This was an idea stolen from Kemsley, who in 1971, mounted such a study in Great Britain.) Around 1988, the major U.S. federal household surveys were beginning to consider redesign efforts related to incorporating new population distribution data from the 1990 decennial census. We approached Maria Gonzalez, of the Statistical Policy Office of the Office of Management and Budget, who was leading an interagency group developing those redesign research plans. Our idea was to draw samples of nonrespondents and respondents from household surveys conducted about the time of the decennial census, and match their records to the decennial census data. We would thus have at our disposal all variables on the census form to describe the nonrespondents.

This was an idea whose time had clearly not come to the interagency group. We received tough criticism on what practical lessons would be learned; how would those surveys be improved because of the work, and so on. Maria should be credited with quietly listening to the criticism, but forcefully arguing the merits of our case to the survey sponsors. We dedicate this book to her memory.

What appear to the reader as 11 chapters of theory and analysis are based on many person-years of effort, during which the perspectives on the conceptual foundations of survey participation evolved. Some history of the project may provide a sense of that process.

We sought to develop a diverse set of surveys to match to the decennial records. Ideally, we wanted to represent all major survey design variations among the matched surveys. However, the match tool was to be a unit's address, so we were limited to area frame surveys, most often conducted in face-to-face mode. We failed to get cooperation from the commercial surveys we approached. We failed to get extra funds to add some academic surveys to the set.

In the end we established a consortium of funders including the Bureau of the Census, Bureau of Justice Statistics (BJS), Bureau of Labor Statistics (BLS), National Center for Health Statistics (NCHS) and the National Institute on Drug Abuse [NIDA, later called the Substance Abuse and Mental Health Services Administration (SAMSHA)]. Research Triangle Institute and the National Opinion Research Center also provided documentation on surveys sponsored by NIDA and Census, respectively, to facilitate the match and administered questionnaires to their interviewers. At each agency there were key contact people, who facilitated our work. These were William Nicholls, Robert Tortora, and Jay Waite (Census Bureau), Cathryn Dippo and Clyde Tucker (BLS), Michael Rand (BJS), Steve Botman (NCHS), and Joseph Gfroerer (SAMHSA).

Completely independent of this research program, in early 1990, Groves took on a temporary post as an Associate Director at the Bureau of the Census, as the project was nearing its implementation. Couper simultaneously took a post as visiting researcher at Census. This permitted Couper to focus full time on the project between 1990 and 1994.

In 1989, samples were drawn from the Census Bureau surveys, mostly by staff in the Statistical Methods Division of the Bureau, under the direction of Jay Waite. John Paletta coordinated the selection of match cases. After the Census, in 1991–1992, Couper began a commuting life between Washington, DC, and Jeffersonville, Indiana, the vast processing complex for the U.S. Census Bureau. There he worked with a team headed by Judith Petty. The leader of the match team, Maria Darr, collaborated in defining the match methods, training and supervising staff, and implementing quality control procedures. Couper directed the match effort, living out of a suitcase, eating too many meals at the Waffle House in Jeffersonville (whose broken sign read "affle House"). Matching survey and census records was a tedious, slow process, but the care and professionalism of the Jeffersonville staff produced a match data set that we believe is as complete and accurate as possible. Acquiring the completed survey data, cleaning data, merging files, determining weighting schemes, variance estimators, and appropriate modeling techniques took some time after the completion of the match in 1993.

We are both indebted to the executive staff of the Census Bureau, which provided a research cocoon at Suitland, permitting Couper to focus entirely on the research activities at crucial times during the match process, and Groves to join him after ending his stint as associate director in 1992.

However, the work of the decennial match project was not our only focus during the years 1988–1992. Even while the match project was being discussed, two other lines of research were developing. The first was a refinement of conceptual thinking on the process of survey participation. This was partially funded by the Census Bureau and was a collaborative effort with Robert Cialdini, a social psychologist who has made important contributions to understanding helping behavior and compliance. We collaborated in a series of focus groups with interviewers from different organizations, seeking insights from their expertise in gaining the cooperation of persons in surveys. This led to a basic framework of influences on survey participation that forms the structure of this book. Cialdini provided important insights about how survey participation decisions might resemble to other decisions about requests and, more broadly, to attitude change. We are in his debt, especially for the insight one Saturday morning that most decision making in the survey context must be heuristically based, ill-informed by the central features of the respondent's job in a survey.

When our interest grew concerning the effect of the social environment of survey participation, we joined with Lars Lyberg, our friend and colleague, to organize a set of international workshops on household survey nonresponse, starting in 1990. These gave researchers in different countries a chance to compare notes on survey participation across societies. The workshops have stimulated the replication of nonresponse research across countries. Our own research has benefitted from such replication. We have also learned much from the interactions and enjoyed the camaraderie. We thank the regulars at the meetings, including Lars, Bab Barnes, Sandy Braver, Pam Campanelli, Cathy Dippo, Wim de Heer, Lilli Japec, Seppo Laaksonen, Clyde Tucker, and many others.

The other line of research that arose in 1990 involved chances to test empirically our ideas with new data collection efforts. Through the good graces of our colleague Ron Kessler, we smuggled into the National Comorbidity Survey a set of interviewer observations that permitted key initial tests of our notions of the influence of contact-level interactions. This survey was supported by the National Institutes on Mental Health (Grants MH46376 and MH 49098). Later we recieved support from the National Institute on Aging (Grant RO1 AG31059) to add similar measures to the AHEAD survey, which permitted tests of the ideas on a survey of the elderly. Bill Rodgers and Tom Juster were very supportive of including these in AHEAD. Both of these grants were important to Chapters 8, 9, and Chapter 11 of this text.

After the match project data were available, Trivellore Raghunathan became a collaborator when he joined the Survey Methodology Program at Michigan. He collaborated in translating our ideas and findings into a statisitcal modeling strategy for postsurvey adjustment. Raghu deserves full credit for the two-stage adjustment procedures in Chapter 11.

***Audience.*** We've written the book for students of survey methodology: those in school, practicing in the field, and teaching the subject. We assume basic knowledge of survey design, at a level comparable to most initial undergraduate survey methods courses. The statistical models are kept simple deliberately

Those readers with limited time should read Chapters 1 and 2 in order to understand the conceptual framework. Then they should read the summary sections of each chapter, as well as Chapter 11.

Those readers most interested in the practical implications of the work should read the last sections of Chapters 4–9, labeled "Practical Implications for Survey Implementation" as well as Chapters 10 and 11.

In using the book as a text in a course on survey nonresponse we have used Chapters 2 and 4–10.

***Collaborators.*** In addition to those mentioned above, other stimulating colleagues helped shape the research. These include Toni Tremblay and Larry Altmayer at the U.S. Census Bureau, and Joe Parsons, Ashley Bowers, Nancy Clusen, Jeremy Morton, and Steve Hanway at the Joint Program in Survey Methodology. Lorraine McCall was responsible for the interviewer surveys at the Census Bureau. Teresa Parsley Edwards and Rachel Caspar at Research Triangle Institute worked with us on parts of the analysis of the National Household Survey on Drug Abuse. Brian Harris-Kojetin, John Eltinge, Dan Rope, and Clyde Tucker examined various features of nonreponse in the Current Population Survey. Judith Clemens, Darby Miller-Steiger, Stacey Erth, and Sue Ellen Hansen provided assistance at various points during the work, especially on the Michigan Survey Research Center surveys. We appreciate the criticisms of a set of students in a summer course on survey nonresponse in 1994 offered through the SRC Summer Institute in Survey Research Techniques. Finally, the administrative staff of the Joint Program in Survey Methodology, including Jane Rice, Pam Ainsworth, Nichole Ra'uf, Christie Nader, and Heather Campbell, provided help at many crucial points.

We are members of the Survey Methodology Program (SMP) at the University of Michigan's Institute for Social Research, a research environment that stimulates theoretical questions stemming from applied problems. We thank our SMP colleagues for helping us think through much of the material we present in this book. Jim House, as director of the Survey Research Center, has been a consistent supporter of bringing science to survey methodology and we thank him for being there.

We have profited from critical reviews by Paul Biemer, John Eltinge, Robert Fay, Brian Harris-Kojetin, Lars Lyberg, Nancy Mathiowetz, Beth-Ellen Pennell, Stanley Presser, Eleanor Singer, Seymour Sudman, Roger Tourangeau, and Clyde Tucker. Errors remaining are our responsibility.

We are especially indebted to Northwest Airlines, whose many delayed and cancelled flights between Detroit Metro and Washington National airports permitted long and uninterrupted discussions of the research.

Finally, we thank our editor at Wiley, Steve Quigley, for making the publication process as trouble-free as possible.

ROBERT M. GROVES
MICK P. COUPER

*Ann Arbor, Michigan*
*College Park, Maryland*

# Acknowledgments

We are grateful to various copyright holders for permission to reprint or present adaptations of material previously published. These include the University of Chicago Press, on behalf of the American Association for Public Opinion Research, for adaptation of material from Groves, Cialdini, and Couper (1992) and Couper (1992), appearing in Chapters 2 and 10, and for reprinting a table from Dillman, Gallegos, and Frey (1976), as Table 10.1; the Minister of Industry of Canada, through Statistics Canada for adaptation of Couper and Groves (1992) in Chapter 7; Statistics Sweden, for adaptation of Groves, R.M., and Couper, M.P. (1995) "Theoretical Motivation for Post-Survey Nonresponse Adjustment in Household Surveys," 11, 1, 93–106, in Chapter 9; and "Contact-Level Influences on Cooperation in Face-to-Face Surveys," 12, 1, 63–83, in Chapter 8; Kluwer Academic Publishers, for adaptations of Couper, M.P., and Groves, R.M. (1996) "Social Environmental Impacts on Survey Cooperation," 30, 173–188, in Chapter 6; and Jossey-Bass Publishers for adaptation of Couper and Groves (1996) in Chapter 5.

Nonresponse in
Household Interview Surveys

CHAPTER ONE

# An Introduction to Survey Participation

## 1.1 INTRODUCTION

This is a book about error properties of statistics computed from sample surveys. It is also a book about why people behave the way they do.

When people are asked to participate in sample surveys, they are generally free to accept or reject that request. In this book we try to understand the several influences on their decision. What influence is exerted by the attributes of survey design, the interviewer's behavior, the prior experiences of the person faced with the request, the interaction between interviewer and householder, and the social environment in which the request is made? In the sense that all the social sciences attempt to understand human thought and behavior, this is a social science question. The interest in this rather narrowly restricted human behavior, however, has its roots in the effect these behaviors have on the precision and accuracy of statistics calculated on the respondent pool resulting in the survey. It is largely because these behaviors affect the quality of sample survey statistics that we study the phenomenon.

This first chapter sets the stage for this study of survey participation and survey nonresponse. It reviews the statistical properties of survey estimates subject to nonresponse, in order to describe the motivation for our study, then introduces key concepts and perspectives on the human behavior that underlies the participation phenomenon. In addition, it introduces the argument that will be made throughout the book—that attempts to increase the rate of participation and attempts to construct statistical adjustment techniques to reduce nonresponse error in survey estimates achieve their best effects when based on sound theories of human behavior.

## 1.2 STATISTICAL IMPACTS OF NONRESPONSE ON SURVEY ESTIMATES

Sample surveys are often designed to draw inferences about finite populations, by measuring a subset of the population. The classical inferential capabilities of the survey rest on probability sampling from a frame covering all members of the population. A probability sample assigns known, nonzero chances of selection to every member of the population. Typically, large amounts of data from each member of the population are collected in the survey. From these variables, hundreds or thousands of different statistics might be computed, each of which is of interest to the researcher only if it describes well the corresponding population attribute. Some of these statistics describe the population from which the sample was drawn; others stem from using the data to test causal hypotheses about processes measured by the survey variables (e.g., how education and work experience in earlier years affect salary levels).

One example statistic is the sample mean, an estimator of the population mean. This is best described by using some statistical notation, in order to be exact in our meaning. Let one question in the survey be called "$Y$," and the answer to that question for a sample member, say the $i$th member of the population, be designated by $Y_i$. Then we can describe the population mean by

$$\overline{Y} = \sum_{i=1}^{N} Y_i/N$$

where $N$ is the number of units in the target population. The estimator of the population mean is often

$$\overline{y} = \left(\sum_{i=1}^{r} w_i y_i\right) \Bigg/ \left(\sum_{i=1}^{r} w_i\right)$$

where $r$ is the number of respondents in the sample and $w_i$ is the reciprocal of the probability of selection of the $i$th respondent. (For readers accustomed to equal probability samples, as in a simple random sample, the $w_i$ is the same for all cases in the sample and the computation above is equivalent to $\Sigma y_i/n$.)

One problem with the sample mean as calculated above is that is does not contain any information from the nonrespondents in the sample. However, all the desirable inferential properties of probability sample statistics apply to the statistics computed on the *entire* sample. Let's assume that in addition to the $r$ respondents to the survey, there are $m$ (for "missing") nonrespondents. Then the total sample size is $n = r + m$. In the computation above we miss information on the $m$ missing cases.

How does this affect our estimation of the population mean, $\overline{Y}$? Let's first make a simplifying assumption. Assume that everyone in the target population is either, permanently and forevermore, a respondent or a nonrespondent. Let the entire target population, thereby, be defined as $N = R + M$, where the capital letters denote numbers in the total population.

Assume that we are unaware at the time of the sample selection about which stratum each person belongs to. Then, in drawing our sample of size $n$, we will likely select some respondents and some nonrespondents. They total $n$ in all cases but the actual number of respondents and nonrespondents in any one sample will vary. We know that, in expectation, the fraction of *sample* cases that are respondent should be equal to the fraction of population cases that lie in the respondent stratum, but there will be sampling variability about that number. That is, $E(r) = fR$, where $f$ is the sampling fraction used to draw the sample from the population. Similarly $E(m) = fM$.

For each possible sample we could draw, given the sample design, we could express a difference between the full sample mean, $\bar{y}_n$, and the respondent mean, in the following way:

$$\bar{y}_n = \left(\frac{r}{n}\right)\bar{y}_r + \left(\frac{m}{n}\right)\bar{y}_m$$

which, with a little manipulation becomes

$$\bar{y}_r = \bar{y}_n + \left(\frac{m}{n}\right)[\bar{y}_r - \bar{y}_m]$$

that is,

Respondent Mean = Total Sample Mean + (Nonresponse Rate)
× (Difference between Respondent and
Nonrespondent Means)

This shows that the deviation of the respondent mean from the full sample mean is a function of the nonresponse rate ($m/n$) and the difference between the respondent and nonrespondent means.

Under this simple expression, what is the expected value of the respondent mean, over all samples that could be drawn given the same sample design? The answer to this question determines the nature of the *bias* in the respondent mean, where "bias" is taken to mean the difference between the expected value (over all possible samples given a specific design) of a statistic and the statistic computed on the target population. That is, in cases of equal probability samples of fixed size the bias of the respondent mean is approximately

$$B(\bar{y}_r) = \left(\frac{M}{N}\right)(\bar{Y}_r - \bar{Y}_m)$$

or

Bias(Respondent Mean) = (Nonresponse Rate in Population)
× (Difference in Respondent and
Nonrespondent Population Means)

where the capital letters denote the population equivalents to the sample values. This shows that the larger the stratum of nonrespondents, the higher the bias of the respondent mean, other things being equal. Similarly, the more distinctive the nonrespondents are from the respondents, the larger the bias of the respondent mean.

These two quantities, the nonresponse rate and the differences between respondents and nonrespondents on the variables of interest, are key to the studies reported in this book. Because the literature on survey nonresponse does not directly reflect this fact (an important exception is the work of Lessler and Kalsbeek, 1992), it is important for the reader to understand how this affects nonresponse errors.

Figure 1.1 shows four alternative frequency distributions for respondents and nonrespondents on a hypothetical variable, *y*, measured on all cases in some target

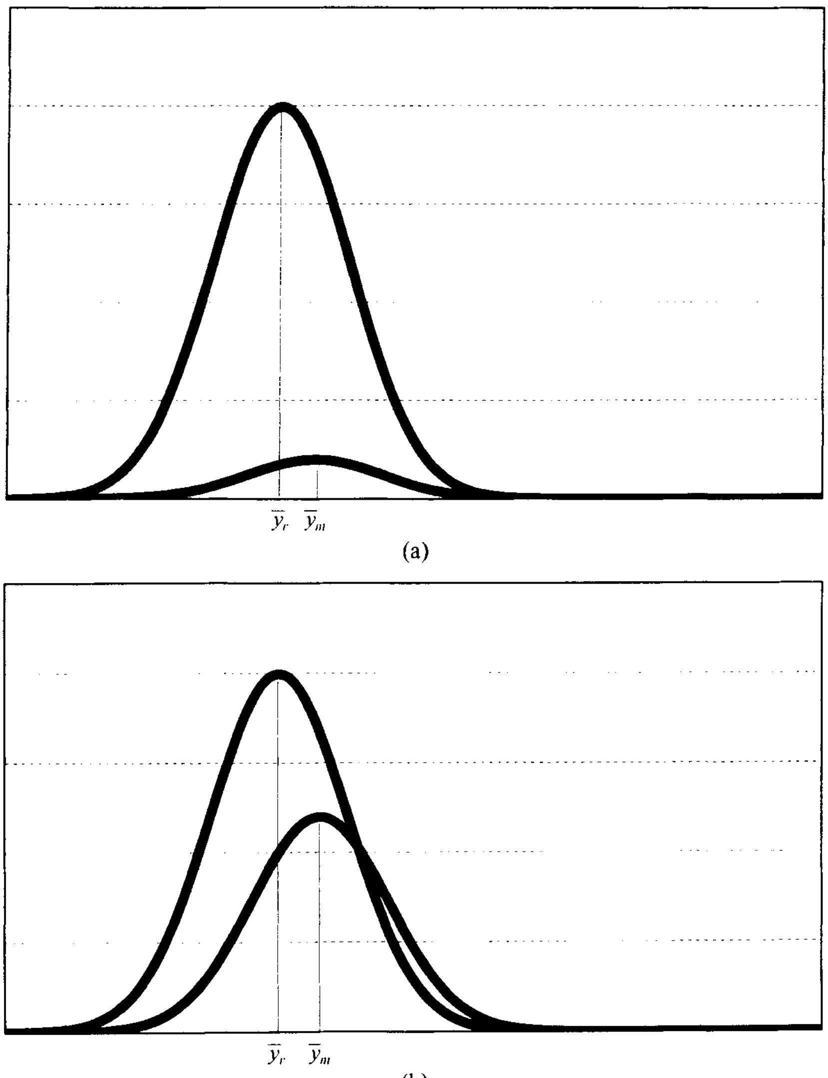



**Figure 1.1.** Hypothetical frequency distributions of respondents and nonrespondents. (a) High response rate, nonrespondents similar to respondents. (b) Low response rate, nonrespondents similar to respondents.

population. The area under the curves is proportional to the size of the two groups, respondents and nonrespondents.

Case (a) in the figure reflects a high response rate survey and one in which the nonrespondents have a distribution of $y$ values quite similar to that of the respondents. This is the lowest-bias case—both factors in the nonresponse bias are small. For example, assume the response rate is 95%, the respondent mean for reported ex-

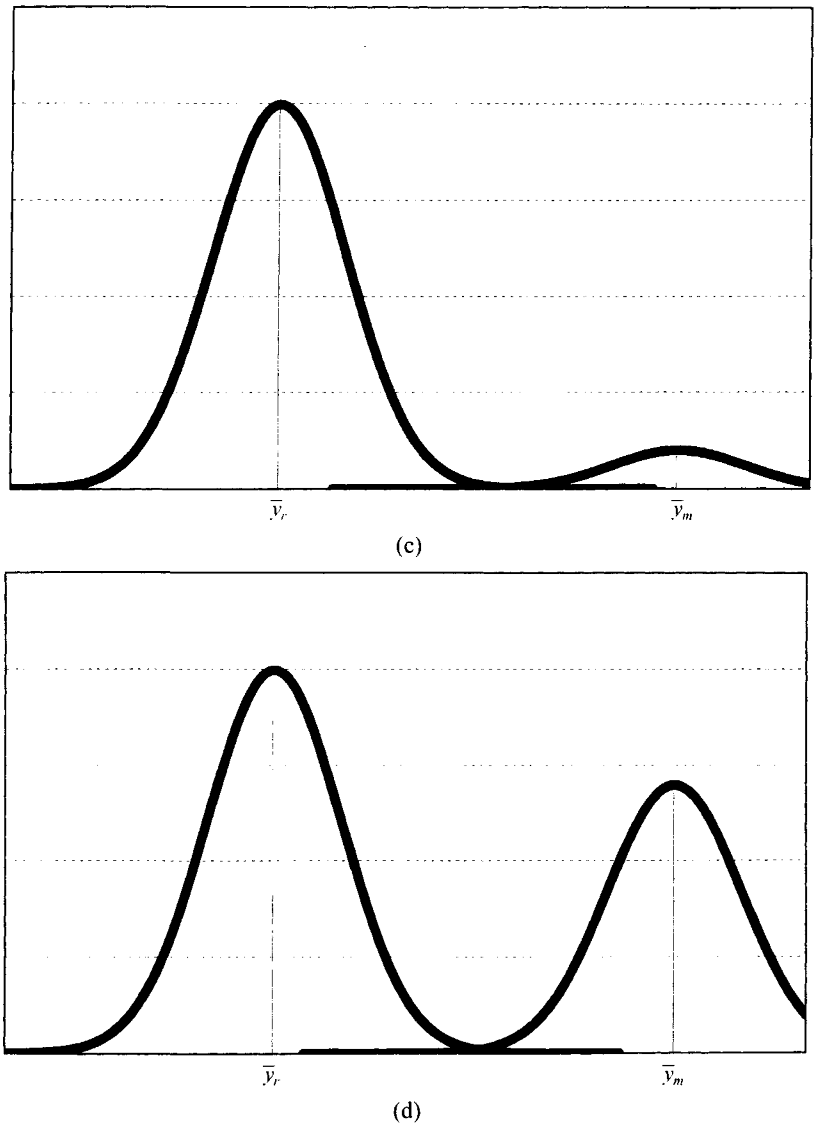



**Figure 1.1.** (c) High response rate, nonrespondents different from respondents. (d) Low response rate, nonrespondents different from respondents

penditures on clothing for a quarter was \$201.00, and the mean for nonrespondents was \$228.00. Then the nonresponse error is 0.05(\$201.00 – \$228.00) = –\$1.35.

Case (b) shows a very high nonresponse rate (the area under the respondent distribution is about 50% greater than that under the nonrespondent—a nonresponse rate of 40%). However, as in (a), the values on *y* of the nonrespondents are similar to those of the respondents. Hence, the respondent mean again has low bias due to nonresponse. With the same example as in (a), the bias is 0.40(\$201.00 – \$228.00) = –\$10.80.

Case (c), like (a), is a low nonresponse survey, but now the nonrespondents tend to have much higher values than the respondents. This means that the difference term, $[\bar{y}_r - \bar{y}_m]$, is a large negative number—the respondent mean underestimates the full population mean. However, the size of the bias is small because of the low nonresponse rate, about 5% or so. Using the same example as in (a), with a nonrespondent mean now of \$501.00, the bias is 0.05(\$201.00 – \$501.00) = –\$15.00.

Case (d) is the most perverse, exhibiting a large group of nonrespondents, who have much higher values in general on *y* than the respondents. In this case, $m/n$ is large (judging by the area under the nonrespondent curve) and $[\bar{y}_r - \bar{y}_m]$ is large in absolute terms. This is the case of large nonresponse bias. Using the example above, the bias is 0.40(\$201.00 – \$501.00) = –\$120.00, a relative bias of 60% of the respondent-based estimate!

To provide another concrete illustration of these situations, assume that the statistic of interest is a proportion, say, the number of adults who intend to save some of their income in the coming month. Figure 1.2 illustrates the level of nonresponse bias possible under various circumstances. In all cases, the survey results in a respondent mean of 0.50; that is, we are led to believe that half of the adults plan to

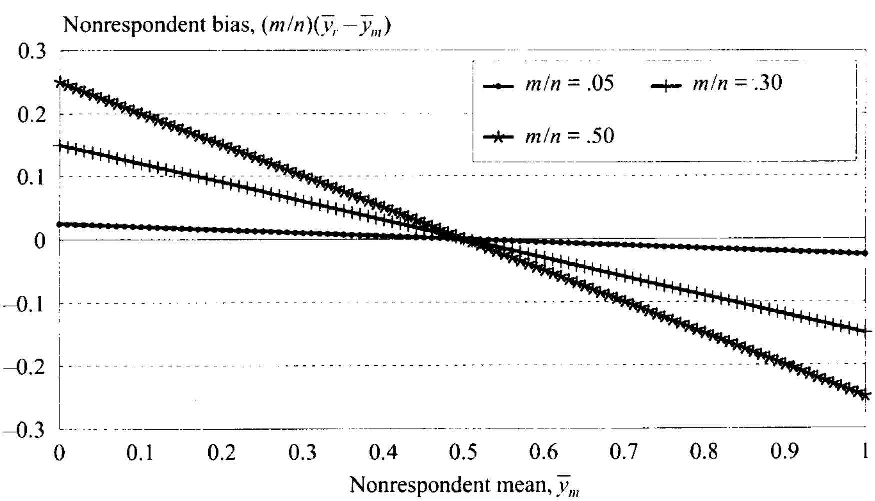



**Figure 1.2.** Nonresponse bias for a proportion, given a respondent mean of 0.50, various response rates, and various nonresponse means.

save in the coming month. The $x$-axis of the figure displays the proportion of nonrespondents who plan to save in the coming month. (This attribute of the sample is not observed.) The figure is designed to illustrate cases in which the nonrespondent proportion is less or equal to the respondent proportion. Thus, the nonrespondent proportions range from 0.50 (the no bias case) to 0.0 (the largest bias case). There are three lines in the figure, corresponding to different nonresponse rates: 5%, 30%, and 50%.

The figure gives a sense of how large a nonresponse bias can be for different nonresponse rates. For example, in a survey with a low nonresponse rate, 5%, the highest bias possible is 0.025. That is, if the survey respondent mean is 0.50, then one is assured that the full sample mean lies between 0.475 and 0.525.

In the worst case appearing in Figure 1.2, a survey with a nonresponse rate of 50%, the nonresponse bias can be as large as 0.25. That is, if the respondent mean is 0.50, then the full sample mean lies between 0.25 and 0.75. This is such a large range that it offers very little information about the statistic of interest.

The most important feature of Figure 1.2 is its illustration of the dependence of the nonresponse bias on both response rates and the difference term. The much larger slope of the line describing the nonresponse bias for the survey with a high nonresponse rate shows that high nonresponse rates increase the likelihood of bias even with relatively small differences between respondents and nonrespondents on the survey statistic.

### 1.2.1 Nonresponse Error on Different Types of Statistics

The discussion above focused on the effect of nonresponse on estimates of the population mean, using the sample mean. This section briefly reviews effects of nonresponse on other popular statistics. We examine the case of an estimate of a population total, the difference of two subclass means, and a regression coefficient.

***The Population Total.*** Estimating the total number of some entity is common in government surveys. For example, most countries use surveys to estimate the total number of unemployed persons, the total number of new jobs created in a month, the total retail sales, the total number of criminal victimizations, etc. Using notation similar to that in Section 1.2, the population total is $\Sigma Y_i$, which is estimated by a simple expansion estimator, $\Sigma w_i y_i$, or by a ratio-expansion estimator, $X(\Sigma w_i y_i / \Sigma w_i x_i)$, where $X$ is some auxiliary variable, correlated with $Y$, for which target population totals are known. For example, if $y$ were a measure of the number of criminal victimizations experienced by a sample household, and $x$ were a count of households, $X$ would be a count of the total number of households in the country.

For variables that have nonnegative values (such as count variables), simple expansion estimators of totals based only on respondents always underestimate the total. This is because the full sample estimator is

$$\sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{r} w_i y_i + \sum_{i=r+1}^{n} w_i y_i$$

that is,

Full Sample Estimate of Population Total = Respondent-Based Estimate
+ Nonrespondent-Based Estimate

Hence, the bias in the respondent-based estimator is

$$-\sum_{i=r+1}^{n} w_i y_i$$

It is easy to see, thereby, that the respondent-based total (for variables that have non-negative values) will always underestimate the full sample total, and thus, in expectation, the full population total.

***The Difference of Two Subclass Means.*** Many statistics of interest from sample surveys estimate the difference between the means of two subpopulations. For example, the Current Population Survey often estimates the difference in the unemployment rate for Black and nonBlack men. The National Health Interview Survey estimates the difference in the mean number of doctor visits in the last 12 months between males and females.

Using the expressions above, and using subscripts 1 and 2 for the two subclasses, we can describe the two respondent means as

$$\bar{y}_{1r} = \bar{y}_{1n} + \left(\frac{m_1}{n_1}\right)[\bar{y}_{1r} - \bar{y}_{1m}]$$

$$\bar{y}_{2r} = \bar{y}_{2n} + \left(\frac{m_2}{n_2}\right)[\bar{y}_{2r} - \bar{y}_{2m}]$$

These expressions show that each respondent subclass mean is subject to an error that is a function of a nonresponse rate for the subclass and a deviation between respondents and nonrespondents in the subclass. The reader should note that the nonresponse rates for individual subclasses could be higher or lower than the nonresponse rates for the total sample. For example, it is common that nonresponse rates in large urban areas are higher than nonresponse rates in rural areas. If these were the two subclasses, the two nonresponse rates would be quite different.

If we were interested in $\bar{y}_1 - \bar{y}_2$ as a statistic of interest, the bias in the difference of the two means would be approximately

$$B(\bar{y}_1 - \bar{y}_2) = \left(\frac{M_1}{N_1}\right)[\bar{Y}_{1r} - \bar{Y}_{1m}] - \left(\frac{M_2}{N_2}\right)[\bar{Y}_{2r} - \bar{Y}_{2m}]$$

Many survey analysts are hopeful that the two terms in the bias expression above cancel. That is, the bias in the two subclass means is equal. If one were dealing with two subclasses with equal nonresponse rates that hope is equivalent to a hope that the difference terms are equal to one another. This hope is based on an assumption that nonrespondents will differ from respondents in the same way for both subclasses. That is, if nonrespondents tend to be unemployed versus respondents, on average, this will be true for all subclasses in the sample.

If the nonresponse rates were not equal for the two subclasses, then the assumptions of canceling biases is even more complex. But to simplify, let's continue to assume that the difference between respondent and nonrespondent means is the same for the two subclasses. That is, assume $[\bar{y}_{r1} - \bar{y}_{m1}] = [\bar{y}_{r2} - \bar{y}_{m2}]$. Under this restrictive assumption, there can still be large nonresponse biases.

For example, Figure 1.3 examines differences of two subclass means where the statistics are proportions (e.g., the proportion planning to save some of their income next month). The figure treats the case in which the proportion planning to save among respondents in the first subclass (say, high-income households) is $\bar{y}_{r1} = 0.5$ and the proportion planning to save among respondents in the second subclass (say, low-income households) is $\bar{y}_{r2} = 0.3$. This is fixed for all cases in the figure. We examine the nonresponse bias for the entire set of differences between respondents and nonrespondents. That is, we examine situations where the differences between respondents and nonrespondents lie between –0.5 and 0.3. (This difference applies to both subclasses.) The first case of a difference of 0.3 would correspond to

$$[\bar{y}_{r1} - \bar{y}_{m1}] = 0.5 - 0.2 = 0.3$$

$$[\bar{y}_{r2} - \bar{y}_{m2}] = 0.3 - 0.0 = 0.3$$

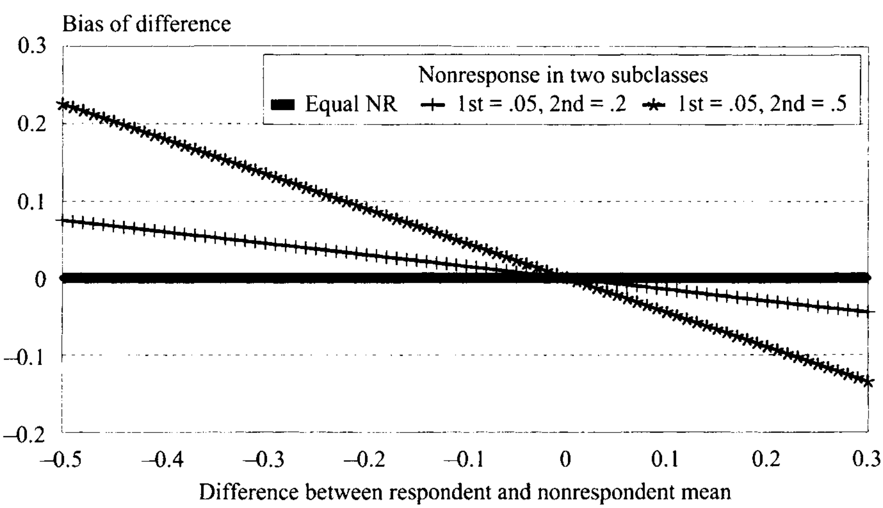



**Figure 1.3.** Nonresponse bias for a difference of subclass means, for the case of two respondent subclass means (0.5, 0.3) by various response rate combinations, by differences between respondent and nonrespondent means.

The figure shows that when the two nonresponse rates are equal to one another, there is no bias in the difference of the two subclass means. However, when the response rates of the two subclasses are different, large biases can result. Larger biases in the difference of subclass means arise with larger differences in nonresponse rates in the two subclasses (note the higher absolute value of the bias for any given $[\bar{y}_r - \bar{y}_m]$ value for the case with a 0.05 nonresponse rate in subclass 1 and a 0.5 in subclass 2 than for the other cases).

***A Regression Coefficient.*** Many survey data sets are used by analysts to estimate a wide variety of statistics measuring the relationship between two variables. Linear models testing causal assertions are often estimated on survey data. Imagine, for example, that the analysts were interested in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

which, using the respondent cases to the survey, would be estimated by

$$\hat{y}_{ri} = \hat{\beta}_{r0} + \hat{\beta}_{r1} x_{ri}$$

The ordinary least squares estimator of $\beta_{r1}$ is

$$\hat{\beta}_{r1} = \frac{\sum_{i=1}^{r}(x_i - \bar{x}_r)(y_i - \bar{y}_r)}{\sqrt{\sum_{i=1}^{r}(x_i - \bar{x}_r)^2}}$$

Both the numerator and denominator of this expression are subject to potential nonresponse bias. For example, the bias in the covariance term in the numerator is approximately

$$B(s_{rxy}) = \frac{M}{N}(S_{rxy} - S_{mxy}) - \left(\frac{M}{N}\right)\left(1 - \frac{M}{N}\right)(X_r - X_m)(Y_r - Y_m)$$

This bias expression can be either positive or negative in value. The first term in the expression has a form similar to that of the bias of the respondent mean. It reflects a difference in covariances for the respondents ($S_{rxy}$) and nonrespondents ($S_{mxy}$). It is large in absolute value when the nonresponse rate is large. If the two variables are more strongly related in the respondent set than in the nonrespondent, the term has a positive value (that is the regression coefficient tends to be overestimated). The second term has no analogue in the case of the sample mean; it is a function of cross-products of difference terms. It can be either positive or negative depending on these deviations.

As Figure 1.4 illustrates, if the nonrespondent units have distinctive combinations of values on the $x$ and $y$ variables in the estimated equation, then the slope of the regression line can be misestimated. The figure illustrates the case when the pat-

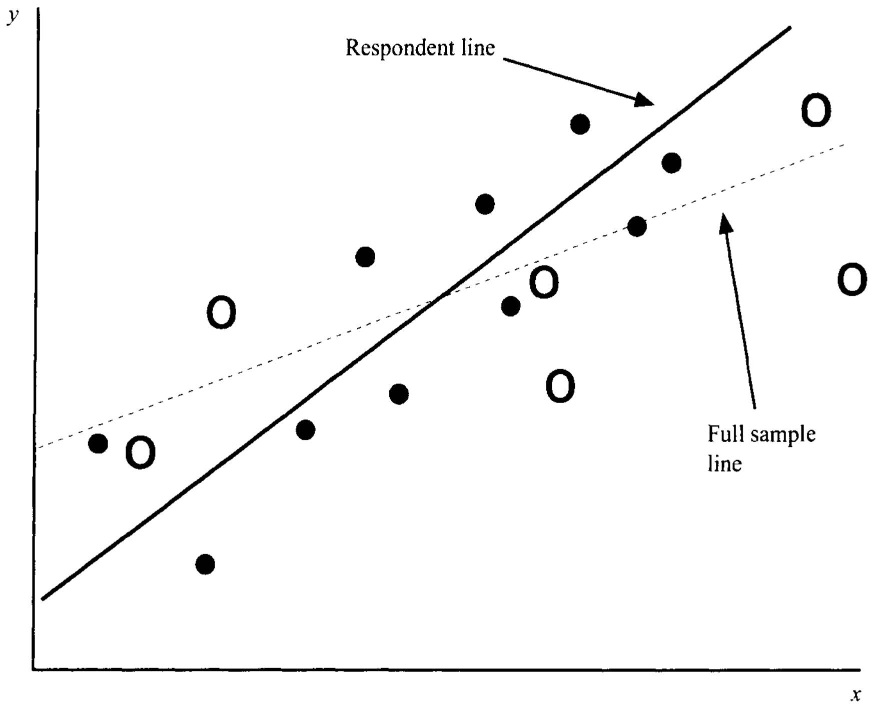



**Figure 1.4.** Illustration of the effect of unit nonresponse on estimated slope of regression line.

tern of nonrespondent cases (designated by "O") differ from that of respondent cases (designated by "●"). The result is that the fitted line on the respondents only has a larger slope than that for the full sample. In this case, the analyst would normally find more support for an hypothesized relationship than would be true for the full sample.

### 1.2.2 Considering Survey Participation a Stochastic Phenomenon

The discussion above made the assumption that each person (or household) in a target population either is a respondent or a nonrespondent for all possible surveys. That is, it assumes a fixed property for each sample unit regarding the survey request. They will always be a nonrespondent or they will always be a respondent, in all realizations of the survey design.

An alternative view of nonresponse asserts that every sample unit has a probability of being a respondent and a probability of being a nonrespondent. It takes the perspective that each sample survey is but one realization of a survey design. In this case, the survey design contains all the specifications of the research data collection. The design includes the definition of the sampling frame, the sample design, the questionnaire design, choice of mode, hiring, selection, and training regimen for interviewers, data collection period, protocol for contacting sample units, callback rules, refusal conversion rules, and so on. Conditional on all these fixed properties of the sample survey, sample units can make different decisions regarding their participation.

In this view, the notion of a nonresponse rate must be altered. Instead of the nonresponse rate merely being a manifestation of how many nonrespondents were sampled from the sampling frame, we must acknowledge that in each realization of a survey different individuals will be respondents and nonrespondents. In this perspective the nonresponse rate above ($m/n$) is the result of a set of Bernoulli trials; each sample unit is subject to a "coin flip" to determine whether it is a respondent or nonrespondent on a particular trial. The coins of various sample units may be weighted differently; some will have higher probabilities of participation than others. However, all are involved in a stochastic process of determining their participation in a particular sample survey.

The implications of this perspective on the biases of respondent means, respondent totals, respondent differences of means, and respondent regression coefficients is minor. The more important implication is on the variance properties of unadjusted and adjusted estimates based on respondents.

### 1.2.3 The Effects of Different Types of Nonresponse

The discussion above considered all sources of nonresponse to be equivalent to one another. However, this book attempts to dissect the process of survey participation into different components. In household surveys it is common to classify outcomes of interview attempts into the following categories: interviews (including complete and partial), refusals, noncontacts, and other noninterviews. The other noninterview category consists of those sample units in which whoever was designated as the respondent is unable to respond, for physical and mental health reasons, for language reasons, or for other reasons that are not a function of reluctance to be interviewed. Various survey design features affect the distribution of nonresponse over these categories. Surveys with very short data collection periods tend to have proportionally more noncontacted sample cases. Surveys with long data collection periods or intensive contact efforts tend to have relatively more refusal cases. Surveys with weak efforts at accommodation of nonEnglish speakers tend to have somewhat more "other noninterviews." So, too, may surveys of special populations, such as the elderly or immigrants.

If we consider separately the different types of nonresponse, many of the expressions above generalize. For example, the respondent mean can be described as a function of various nonresponse sources, as in

$$\bar{y}_r = \bar{y}_n + \frac{m_{rf}}{n}(\bar{y}_r - \bar{y}_{rf}) + \frac{m_{nc}}{n}(\bar{y}_r - \bar{y}_{nc}) + \frac{m_{nio}}{n}(\bar{y}_r - \bar{y}_{nio})$$

where the subscripts $rf$, $nc$, and $nio$ refer to refusals, noncontacts, and other noninterviews, respectively.