# STATISTICAL ANALYSIS WITH MISSING DATA

SECOND EDITION

Roderick J. A. Little
& Donald B. Rubin

WILEY SERIES IN
PROBABILITY AND STATISTICS

# Statistical Analysis with Missing Data

Second Edition

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Peter Bloomfield, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Louis M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels*
Editors Emeriti: *Vic Barnet, J. Stuart Hunder, David G. Kendall*

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

# Statistical Analysis with Missing Data

## Second Edition

RODERICK J. A. LITTLE

DONALD B. RUBIN


A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada



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

Little, Roderick J. A.
Statistical analysis with missing data / Roderick J Little, Donald B. Rubin. - - 2nd ed.
p. cm. - - (Wiley series in probability and statistics)
"A Wiley-Interscience publication."
Includes bibliographical references and index.
ISBN 0-471-18386-5 (acid-free paper)
1. Mathematical statistics. 2. Missing observations (Statistics) I Rubin, Donald B. II.
Title. III. Series

QA276 .L57 2002
519.5- -dc21 2002027006

ISBN 0-471-18386-5

Printed in the United States of America.

20 19 18 17 16 15 14

# Contents

## PART II LIKELIHOOD-BASED APPROACHES TO THE ANALYSIS OF MISSING DATA

# Preface

The literature on the statistical analysis of data with missing values has flourished since the early 1970s, spurred by advances in computer technology that made previously laborious numerical calculations a simple matter. This book aims to survey current methodology for handling missing-data problems and present a likelihood-based theory for analysis with missing data that systematizes these methods and provides a basis for future advances. Part I of the book discusses historical approaches to missing-value problems in three important areas of statistics: analysis of variance of planned experiments, survey sampling, and multivariate analysis. These methods, although not without value, tend to have an ad hoc character, often being solutions worked out by practitioners with limited research into theoretical properties. Part II presents a systematic approach to the analysis of data with missing values, where inferences are based on likelihoods derived from formal statistical models for the data-generating and missing-data mechanisms. Part III presents applications of the approach in a variety of contexts, including ones involving regression, factor analysis, contingency table analysis, time series, and sample survey inference. Many of the historical methods in Part I can be derived as examples (or approximations) of this likelihood-based approach.

The book is intended for the applied statistician and hence emphasizes examples over the precise statement of regularity conditions or proofs of theorems. Nevertheless, readers are expected to be familiar with basic principles of inference based on likelihoods, briefly reviewed in Section 6.1. The book also assumes an understanding of standard models of complete-data analysis—the normal linear model, multinomial models for counted data—and the properties of standard statistical distributions, especially the multivariate normal distribution. Some chapters assume familiarity in particular areas of statistical activity—analysis of variance for experimental designs (Chapter 2), survey sampling (Chapters 3, 4, and 5), or loglinear models for contingency tables (Chapter 13). Specific examples also introduce other statistical topics, such as factor analysis or time series (Chapter 11). The discussion of these examples is self-contained and does not require specialized knowledge, but such knowledge will, of course, enhance the reader's appreciation