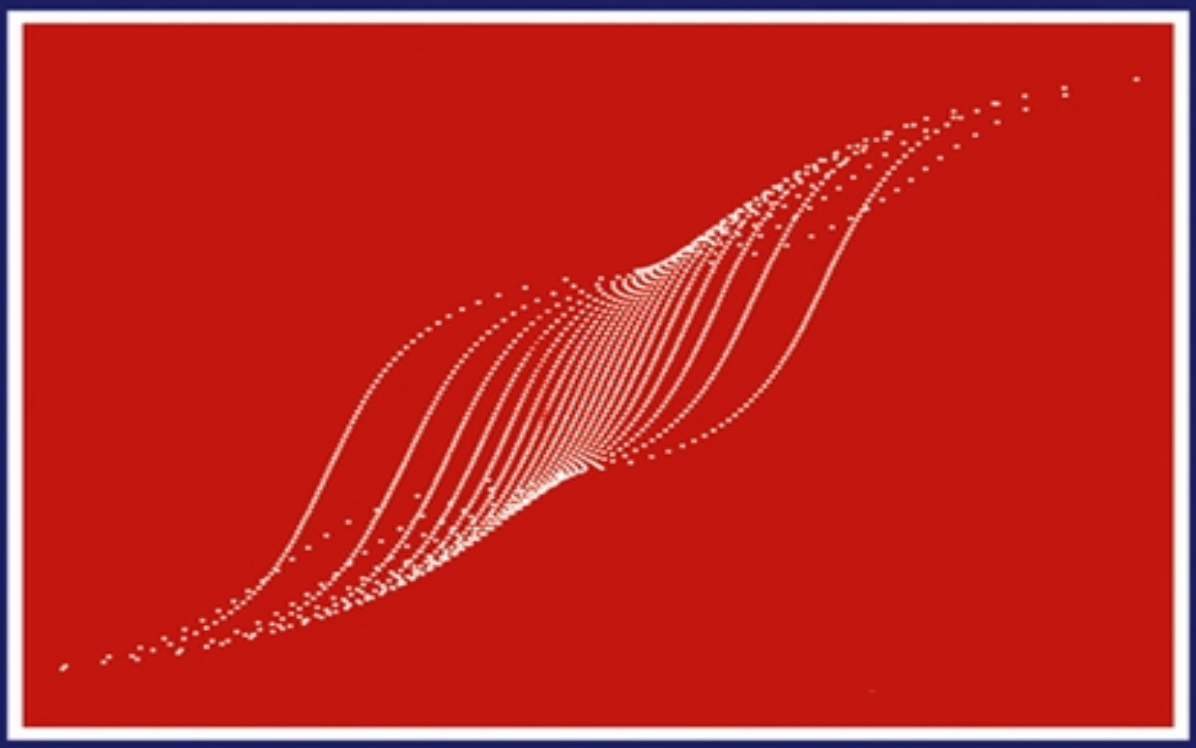
WILEY

Statistical Methods for Rates and Proportions

Third Edition



Joseph L. Fleiss
Bruce Levin
Myunghee Cho Paik

WILEY SERIES IN PROBABILITY AND STATISTICS

Contents

WILEY SERIES IN PROBABILITY AND STATISTICS
Established by WALTER A. SHEWART and SAMUEL S. WILKS
Editors: *David J. Balding, Noel A. C. Cressie, Nicholas I. Fisher,*
Iain M. Johnstone, J. B. Kadane, Louise M. Ryan, David W. Scott,
Adrian F. M. Smith, Jozef L. Teugels
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*
A complete list of the titles in this series appears at the end of this volume.

Statistical Methods for Rates and Proportions

Third Edition

JOSEPH L. FLEISS
BRUCE LEVIN
MYUNGHEE CHO PAIK

Department of Biostatistics
Mailman School of Public Health
Columbia University

A JOHN WILEY & SONS, INC., PUBLICATION

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

Published simultaneously in Canada.

U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data

Fleiss, Joseph L.

Statistical methods for rates and proportions.-- 3rd ed. / Joseph L. Fleiss, Bruce Levin,

Myunghee Cho Paik.

p. cm.-- (Wiley series in probability and statistics)

Includes bibliographical references and indexes.

ISBN 0-471-52629-0 (cloth : acid-free paper)

1. Analysis of variance. 2. Sampling (Statistics) 3. Biometry. I. Levin, Bruce. II. Paik, Myunghee Cho. III. Title. IV. Series.

QA279.F58 2003

519.5′38--dc21

2002191005

10 9 8 7 6 5 4 3 2 1

To Isabel, Betty, and Yi Hyon
and
To Joe, who passed away as this book went to press

Preface

Much has happened in the twenty-two years since the publication of the second edition. The explosive development of personal computing and statistical software has removed the main impediment to sophisticated analyses of data. Indeed, these developments have brought the ability to carry out such analyses out of the sole possession of the specialist and into the hands of every researcher. Logistic, Poisson, and other generalized linear regression models have taken their rightful place as standard analytic methods. Our clinical and public health colleagues no longer view the odds ratio as an inscrutable version of the rate ratio—they understand and use odds ratios all the time. Generalized estimating equations and empirical Bayes methods have become powerful tools to deal with complex data. Exact methods and other computational challenges like conditional likelihood analysis have gone way beyond the Fisher-Irwin exact test for 2×2 tables. Correct ways to deal with missing data can no longer be ignored. The randomized clinical trial and the special role statisticians play in safeguarding the validity of the trial's conduct and findings have come of age in dramatic ways.

This means we can no longer content ourselves with methods that require only a desktop or pocket calculator, which was a hallmark of the second edition. Anyway, the limitations that those devices once represented no longer exist. Yet the elegance of simple, clear, and common-sense methods, which was another hallmark of the previous editions, must never be allowed to take second place to useless complexity. Meeting these requirements has been something of a challenge because, tragically, time has also brought a disabling form of Parkinson's disease to the first author. Joe's inimitable writing style—direct, friendly,

honest, sensible, authoritative, and prescriptive in the best sense of the word—hopefully has been allowed to shine through our attempts to bring the book up to date for both students and researchers engaged in the analysis of their data.

Our approach has been to leave much of the original material intact with clarifications, corrections, and streamlining only as necessary, while covering new material at two levels. The first introduces methods with only as much complexity as is required to give a clear presentation. The mathematical prerequisites continue to be a knowledge of high school algebra and an ability to take logarithms and extract square roots. A familiarity with computers in everyday use is now assumed.

The second level is, admittedly, aimed at students of biostatistics and specialists in data analysis, and requires a level of mathematical preparation equivalent to a first and second course in statistics. These sections have been marked with an asterisk to indicate that full comprehension may require some familiarity with matrix algebra, multivariate statistical concepts, or asymptotic methods. Our suggestion to novice readers is to skim these sections in order to get the lay of the land without getting lost in the thicket of details.

We have added many new problems, some numerical and some theoretical. The numerical problems all have answers at the back of the book, as in the second edition. Many of the theoretical problems have cogent hints to guide students to a successful solution and, we hope, an increase in their understanding and level of expertise. We have tried to bear in mind, and cannot resist paraphrasing here, Stanislaw Lem’s humbling definition of an expert, given in *His Masters Voice*: An *expert* is a barbarian with an uneven level of ignorance. Our hope is that we can move the level without exacerbating Lem’s characterization.

The statistical analysis of single-sample data, such as a prevalence study, now occupies an entire early chapter. We took this as an opportunity to introduce a few technical definitions of a two-tailed p-value for asymmetrical discrete distributions, notions which arise in the exact analysis of categorical data. Armed with such tools, we have fully reinstated exact and approximate-yet-accurate confidence intervals as appropriate statements of statistical uncertainty, notwithstanding Joe's initial reluctance to promulgate their routine use in the first edition. Again, modern computing enables us to recommend them while respecting Joe's warning that a properly constructed confidence interval is frequently more complicated than simply the point estimate plus or minus a multiple of its standard error.

Regarding other foundational issues of statistical inference, we have continued Joe's unabashed preference for frequentist methods. The reader will see, however, that in key places we take an empirical Bayes approach. This occurs, for example, in the new sections on the analysis of many proportions with an element of randomness (Section 9.6*), random effects meta-analysis (Section 10.9.1), tests of odds ratio homogeneity in the large-sparse case (Section 10.9.2), overdispersion in Poisson regression (Section 12.3*), and extensions of logistic regression for correlated binary data (Section 15.5*). In these applications, the empirical Bayes approach provides the most natural analytic framework while avoiding the abuses of subjectivism that underpinned Joe's original distaste for Bayesian methods. Of course, Bayes' theorem, being the fundamental means to pass between conditional probabilities, is used throughout the book. The reader will also notice the likelihood ratio highlighted for its fundamental role in the weighing of statistical evidence, in addition to its frequentist use as a

hypothesis test statistic. This, however, is as close as we come to a frank Bayesian approach.

There is new material on sample size calculations in Chapter 4 (formerly Chapter 3) and some insights about why the sample size tables in the second edition work as well as they do. As mentioned above, exact statistical methods are now presented in a formal and routine manner throughout the first half of the book, where they can be feasibly applied. There is new material on randomization in clinical trials (Chapter 5) and factors relating to statistical power in randomized clinical trials (Section 8.3). The Mantel–sHaenszel procedure and its generalizations for combining the evidence across several cross-classification tables plays a prominent role starting in Chapter 10 and recurs in several subsequent chapters. We have included entirely new chapters on logistic regression (both binary and polytomous, in Chapter 11), Poisson regression (Chapter 12), regression models for matched samples (Chapter 14), the analysis of correlated binary data (Chapter 15), and methods for analyzing fourfold tables with missing data (Chapter 16). The chapters on the effects of, control of, and adjustment for errors of misclassification from the previous editions have been consolidated into one (Chapter 17), with new material on these issues in logistic regression. Chapter 18 on the measurement of interrater agreement has a new section connecting this topic with that of Chapter 15.

We are most grateful to the colleagues and students who helped us with critical review and constructive suggestions. We especially want to thank Melissa Begg for helpful comments; Ann Kinney for her massive and masterful editing; Boanerges Dominguez and Amy Murphy, who assisted us in teaching “The Analysis of Categorical Data” course at Columbia University with preliminary versions of this edition; Cheng-Shiun Leu and Mei-Yin Cheng for reading and computing; Jennie Kline for allowing us to use her

spontaneous abortion data and her epidemiologic guidance as to their interpretation; and Ralph Sacco for the NOMASS data on risk factors for stroke. We thank James Walkup for kindly updating the citations to the psychiatric literature—Joe's knowledge of this field in the earlier editions was immense and no doubt contributed to their success. We are also indebted to Michael Finkelstein for his insight into applied statistics and for stimulating us to think about how to present technical material to nontechnical audiences during the writing of *Statistics for Lawyers*. We thank Steve Quigley and Heather Haselkorn for their utterly endless patience during this project. And we are forever grateful to our spouses, Betty and Yi Hyon, and our families, who, like Joe's wife Isabel, now departed, have been a constant source of inspiration and forbearance during the writing of this book.

BRUCE LEVIN

MYUNGHEE CHO PAIK

New York, New York and
Bear River, Nova Scotia
August 2002

Preface to the Second Edition

The need for a revised edition became apparent a few years after the publication of the first edition. Reviewers, researchers, teachers, and students cited some important topics that were absent, treated too briefly, or not presented in their most up-to-date form. In the meantime, the field of applied statistics continued to develop, and new results were obtained that deserved citation and illustration.

Of the several topics I had omitted from the first edition, the most important was the construction of confidence intervals. In the revision, interval estimation is treated almost as extensively as hypothesis testing. In fact, the close connection between the two is pointed out in the new Section 1.4. The reader will find there, in the new Section 5.6, and elsewhere realizations of the warning I gave in the Preface to the first edition that a properly constructed confidence interval is frequently more complicated than simply the point estimate plus or minus a multiple of its standard error.

Another important topic missing from the first edition was the planning of comparative studies with unequal sample sizes. This is treated in the new Section 3.4.

Several other topics not covered in the first edition are covered here. The Fisher-Irwin “exact” test for a fourfold table is described in the new Section 2.2. Attributable risk, an important indicator of the effect of exposure to a risk factor, is discussed in the new Sections 5.7 and 6.4. The Cornfield method for making inferences about the odds ratio is presented in the new Sections 5.5 and 5.6.

A number of topics touched on superficially or not dealt with adequately in the first edition have, I hope, now been covered properly. The analysis of data from a two-period

crossover study is described in an expansion of Section 7.2. A more appropriate method for analyzing data from a study of matched pairs when the response variable is qualitatively ordered is presented in Section 8.2. The comparison of proportions from several matched samples in Section 8.4 has been expanded to include the case of quantitatively ordered samples. A method for comparing data from several fourfold tables that has been found capable of yielding erroneous results has been relegated to the section (now Section 10.7) on methods to be avoided.

Developments in statistics since the appearance of the first edition are reflected in most sections and every chapter of the revision. The determination of sample sizes is brought up to date in Section 3.2; the corresponding table in the Appendix (Table A.3) has been revised accordingly. Some recently proposed alternatives to simple randomization in clinical studies are discussed in two new sections, 4.3 and 7.3. The presentation of ridit analysis in Section 9.4 has been revised in the light of recent research. The effects and control of misclassification in both variables in a fourfold table are considered in Sections 11.3 and 12.2. The new Chapter 13, which is an expansion and updating of the old Section 12.2, presents recent results on the measurement of interrater agreement for categorical data. Some recent insights into indirect standardization are cited in Sections 14.3 and 14.5.

The emphasis continues to be on, and the examples continue to be from, the health sciences. The selection of illustrative material was determined by the field I know best, not by the field I necessarily consider the most important.

The revision is again aimed at research workers and students who have had at least a year’s course in applied statistics, including chi square and correlation. Many of the problems that conclude the chapters have been revised. Several new problems have been added.

Several of my colleagues and a few reviewers urged me to include the solutions to at least some of the numerical problems. I have decided to provide the solutions to all of them. Teachers who wish to assign these problems for homework or on examinations may do so simply by changing some of the numerical values.

The mathematical prerequisites continue to be a knowledge of high school algebra and an ability to take logarithms and extract square roots. All methods presented can be applied using only a desktop or pocket calculator. As a consequence, the book does not present the powerful but mathematically complicated methods of log-linear or logistic regression analysis for high order cross-classification tables. The texts by D. R. Cox (*The analysis of binary data*, Methuen, London, 1970) and by Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland (*Discrete multivariate analysis*: *Theory and practice*, M.I.T. Press, Cambridge, Mass., 1975). are excellent references at a somewhat advanced mathematical level. Two more recent short monographs (B. S. Everitt, *The analysis of contingency tables*, Halsted Press, New York, 1977 and S. E. Fienberg, *The analysis of cross-classified categorical data*, M.I.T. Press, Cambridge, Mass., 1977). provide less mathematically advanced reviews of these topics.

Professors Agnes Berger, John Fertig, Bruce Levin, and Patrick Shrout of Columbia University and Professor Gary Simon of the State University of New York at Stony Brook reviewed draft copies of the revision and made many helpful suggestions. Professor Berger, Fertig, and Simon were especially critical, and offered advice that I took seriously but did not always follow.

Most helpful of all were the students who took my course on the analysis of categorical data the last couple of years at the Columbia University School of Public Health, and the students who took my course on advanced statistical

methods in epidemiology in the 1978 Graduate Summer Session in Epidemiology at the University of Minnesota School of Public Health. They served as experimental subjects without informed consent as I tried out various approaches to the presentation of the new (and old) material. Students who took my course in the 1980 Graduate Summer Session in Epidemiology saw draft copies of the revision and pointed out several typographical errors I had made. I thank them all.

Ms. Blanche Agdern patiently and carefully typed the several drafts of the revision. Ms. Beatrice Shube, my editor at Wiley, was always supportive and a ready source of advice and encouragement. My wife Isabel was a constant source of inspiration and reinforcement when the going got tough.

The new table of sample sizes was generated by a program run at the computer center of the New York State Psychiatric Institute. The publishers of the *American Journal of Epidemiology*, *Biometrics*, and the *Journal of Chronic Disease* kindly gave me permission to use published data.

JOSEPH L. FLEISS

New York
December 1980

Preface to the First Edition

This book is concerned solely with comparisons of qualitative or categorical data. The case of quantitative data is treated in the many books devoted to the analysis of variance. Other books have restricted attention to categorical data (such as A. E. Maxwell, *Analysing qualitative data*. Methuen, London, 1961, and R. G. Francis, *The rhetoric of science: A methodological discussion of the two-by-two table*, University of Minnesota Press, Minneapolis, 1961), but an updated monograph seemed overdue. A recent text (D. R. Cox, *The analysis of binary data*, Methuen, London, 1970) is at once more general than the present book in that it treats categorical data arising from more complicated study designs and more restricted in that it does not treat such topics as errors of misclassification and standardization of rates.

Although the ideas and methods presented here should be useful to anyone concerned with the analysis of categorical data, the emphasis and examples are from the disciplines of clinical medicine, epidemiology, psychiatry and psychopathology, and public health. The book is aimed at research workers and students who have had at least a year's course in applied statistics, including a thorough grounding in chi square and correlation. Most chapters conclude with one or more problems. Some call for the proof of an algebraic identity. Others are numerical, designed either to have the reader apply what he has learned or to present ideas mentioned only in passing in the text.

No more complicated mathematical techniques than the taking of logarithms and the extraction of square roots are required to apply the methods described. This means that anyone with only high school algebra, and with only a desktop calculator, can apply the methods presented. It also

means, however, that analysis requiring matrix inversion or other complicated mathematical techniques (e.g., the analysis of multiple contingency tables) are not described. Instead, the reader is referred to appropriate sources.

The estimation of the degree of association or difference assumes equal importance with the assessment of statistical significance. Except where the formulas are excessively complicated, I present the standard error of almost every measure of association or difference given in the text. The standard errors are used to test hypotheses about the corresponding parameters, to compare the precision of different methods of estimation, and to obtain a weighted average of a number of independent estimates of the same parameter.

I have tried to be careful in giving both sides of various arguments that are still unresolved about the proper design of studies and analysis of data. Examples are the use of matched samples and the measurement of association. Inevitably, my own biases have probably affected how I present the opposing arguments.

In two instances, however, my bias is so strong that I do not even permit the other side to be heard. I do not find confidence intervals to be useful, and therefore do not discuss interval estimation at all. The reader who finds a need for confidence intervals will have to refer to some of the cited references for details. He will find, by the way, that the proper interval is almost always more complicated than simply the point estimate plus or minus a multiple of its standard error.

The second instance is my bias against the Bayesian approach to statistical inference. See W. Edwards, H. Lindman, and L. J. Savage. Bayesian statistical inference for psychological research, *Psychol. Rec.*, **70**, 193–242, 1963, for a description of the Bayesian approach to data in psychology; and J. Cornfield, A Bayesian test of some

classical hypotheses—with applications to sequential clinical trials, *J. Am. Stat. Assoc.*, **61**, 577–594, 1966, for a description of that approach to data in medicine. I believe that the kind of thinking described in Chapter 3, especially in Section 3.1, provides an adequate alternative to the Bayesian approach.

It is with gratitude that I acknowledge the advice, criticism, and encouragement of Professors John Fertig, Mervyn Susser, and Andre Varma of Columbia University and of Dr. Joseph Zubin of the Biometrics Research unit of the New York State Department of Mental Hygiene. Dr. Gary Simon of Princeton University and Professor W. Edwards Deming of New York University reviewed the manuscript and pointed out a number of errors I had made in an earlier draft. Needless to say, I take full responsibility for any and all errors that remain.

My wife Isabel was a constant source of inspiration as well as an invaluable editorial assistant.

The major portion of the typing was admirably performed by Vilma Rivieccio. Additional typing, collating, and keypunching were ably carried out by Blanche Agdern, Rosalind Fruchtman, Cheryl Keller, Sarah Lichten-staedter, and Edith Pons.

My work was supported in part by grant DR 00793 from the National Institute of Dental Research (John W. Fertig, Ph.D., Principal Investigator and in part by grant MH 08534 from the National Institute of Mental Health (Robert L. Spitzer, M.D., Principal Investigator). Except when noted otherwise, the tables in the Appendix were generated by programs run on the computers of the New York State Psychiatric Institute and of Rockland State Hospital.

I thank Professor E. S. Pearson and the Biometrika Trustees for permission to quote from Tables 1, 4, and 8 of *Biometrika tables for statisticians*, *Vol. I*, edited by E. S. Pearson and H. O. Hartley; John Wiley & Sons for permission to use Tables

A.1 to A.3 Of *Statistical inference under order restrictions* by R. E. Bartlow, D. J. Bartholomew, L. M. Bremner and H. D. Brunk; Van Nostrand Reinhold Co. for permission to quote data from *Smoking and Health*; the Institute of Psychiatry of the University of London for permission to quote data from *Psychiatry diagnosis in New York and London* by J. E. Cooper et al.; and Sir Austin Bradford Hill and Oxford University Press for permission to quote from *Statistical methods in clinical and preventive medicine*.

I also thank the editors of the following journals for permission to use published data: the *American Journal of Public Health*, the *American Statistician*, *Biometrics*, the *Journal of Laboratory and Clinical Medicine*, the *Journal of the National Cancer Institute*, the *Journal of Psychiatric Research*, and *Psychometrika*.

JOSEPH L. FLEISS

New York
June 1972

CHAPTER 1

An Introduction to Applied Probability

Some elements of applied probability theory are needed fully to appreciate and work with the different kinds of rates and proportions that arise in research. Thus the clearest and most suggestive interpretation of a proportion is as a probability—as a measure of the chance that a specified event happens to, or that a specified characteristic is possessed by, a typical member of a population. An important use of probabilities is in estimating the number of individuals, out of a sample of size n, who have the characteristic under consideration. If P is the probability that an individual possesses the characteristic, the expected number having the characteristic is simply nP.

Section 1.1 presents notation and some important definitions, and Section 1.2 presents the rule of total probability, along with an application. The theory in Sections 1.1 and 1.2 is applied in Section 1.3 to the evaluation of a screening test, and in Section 1.4 to the bias possible when making inferences from selected samples.

1.1. NOTATION AND DEFINITIONS

In this book, the terms *probability, relative frequency*, and *proportion* are used synonymously. If A denotes the event

that a randomly selected person from a population has a defined characteristic (e.g., has arteriosclerotic heart disease), then $P(A)$ denotes the proportion of all people who have the characteristic. For the given example, $P(A)$ is the probability that a randomly selected individual has arteriosclerotic heart disease. The term *rate* has two meanings: one is as a simple synonym for probability, whereas the other attaches a notion of time to the expression of probability. The time setting may be over a given interval such as a year, or may refer to a particular point in time, and may or may not be stated explicitly. For convenience we use *rate* mostly in its first sense, but where the second sense is important we so indicate. For the given example, in the terminology of vital statistics, $P(A)$ is the case rate for arteriosclerotic heart disease (at a particular point in time).

One can go only so far with overall rates, however. Of greater usefulness usually are so-called *specific rates:* the rate of the defined characteristic specific for age, race, sex, occupation, and so on. What is known in epidemiology and vital statistics as a specific rate is known in probability theory as a *conditional probability*. The notation is

$P(A|B)$ = probability that a randomly selected individual has characteristic A, given that he has characteristic B, or *conditional* on his having characteristic B.

If, in our example, we denote by B the characteristic of being aged 65–74, then $P(A/B)$ is the conditional probability that a person has arteriosclerotic heart disease, given that he is aged 65–74. In the terminology of vital statistics, $P(A/B)$ is the rate of arteriosclerotic heart disease specific to people aged 65–74.

Let $P(B)$ represent the proportion of all people who possess characteristic B, and let $P(A \text{ and } B)$ represent the proportion of all people who possess both characteristic A and characteristic B. Then, by definition, provided $P(B) \neq 0$,