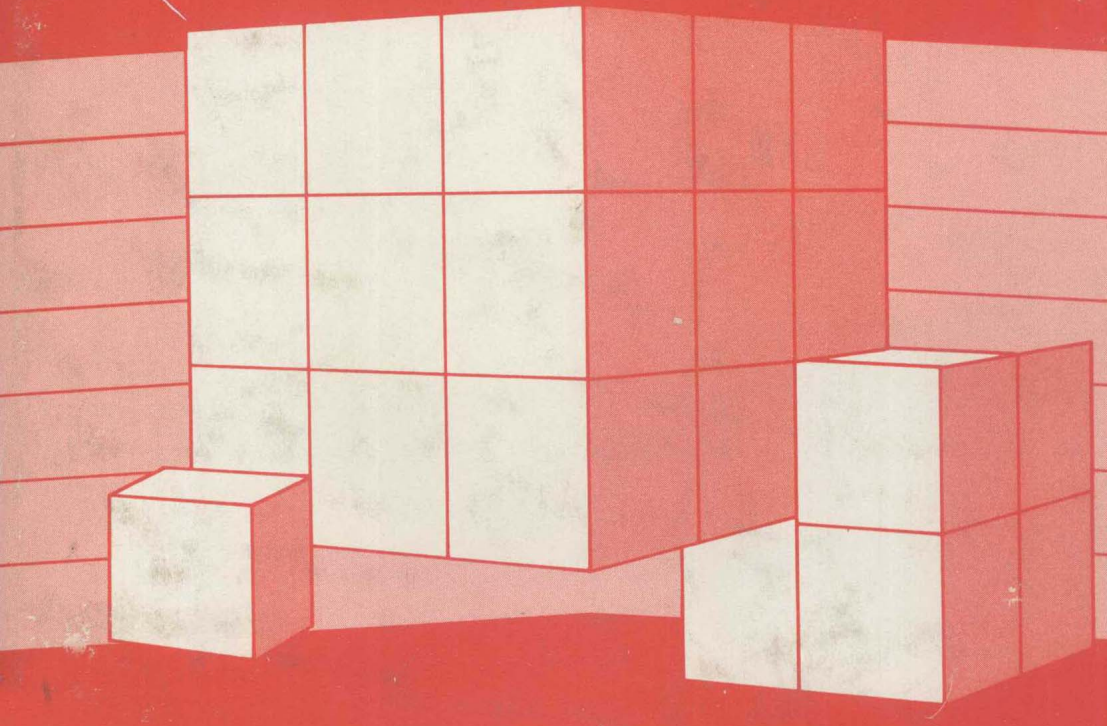
Statistical Methods for Rates and Proportions

Second Edition

Joseph L. Fleiss



A volume in the Wiley Series
in Probability and Mathematical Statistics;
Ralph A. Bradley, J. Stuart Hunter, David G. Kendall, and Geoffrey S. Watson—Advisory Editor

Statistical Methods for Rates and Proportions

Second Edition

JOSEPH L. FLEISS

Division of Biostatistics, School of Public Health, Columbia University

JOHN WILEY & SONS New York · Chichester · Brisbane · Toronto

Library of Congress Cataloging in Publication Data:

Fleiss, Joseph L.
Statistical methods for rates and proportions.

(Wiley series in probability and mathematical statistics)
Includes bibliographies and indexes.
1. Analysis of variance. 2. Sampling (Statistics) 3. Biometry. I. Title.

QA279.F58 1981 519.5'352 80-26382
ISBN 0-471-06428-9

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

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. Professors 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 Diseases* kindly gave me permission to use published data.

JOSEPH L. FLEISS

New York, 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 analyses 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. Rev.*, **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 Lichtenstaedter, 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. Barlow, D. J. Bartholomew, J. 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 *Psychiatric 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, New York
June 1972

Contents

CHAPTER

CHAPTER 1

An Introduction to Applied Probability

Some elements of applied probability theory are needed to appreciate fully and to manipulate the different kinds of rates that arise in research. Thus the clearest and most suggestive interpretation of a rate is as a probability —as a measure of the likelihood that a specified event occurs 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. The theory of Section 1.1 is applied in Section 1.2 to the evaluation of a screening test, and in Section 1.3 to the study of the bias possible in making inferences from selected samples. Section 1.4 is devoted to methods for testing hypotheses about and constructing confidence intervals for single probabilities or proportions.

1.1. NOTATION AND DEFINITIONS

In this book, the terms probability, relative frequency, proportion, and rate are used synonymously. If A denotes the event that a randomly selected individual 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, or, in the terminology of vital statistics, the case rate for arteriosclerotic heart disease.

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 $P(B) \neq 0$,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}. \tag{1.1}$$

Similarly, provided $P(A) \neq 0$,

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}. \tag{1.2}$$

By the *association* of two characteristics we mean that when a person has one of the characteristics, say B, his chances of having the other are affected. By the *independence* or lack of association of two characteristics we mean that the fact that a person has one of the characteristics does not affect his chances of having the other. Thus, if A and B are independent, then the rate at which A is present specific to people who possess B, $P(A|B)$, is equal to the overall rate at which A is present, $P(A)$. By (1.1), this implies that

$$\frac{P(A \text{ and } B)}{P(B)} = P(A),$$

or

$$P(A \text{ and } B) = P(A)P(B). \tag{1.3}$$

Equation 1.3 is often taken as the definition of independence, instead of

the equivalent statement

$$P(A|B) = P(A).$$

A heuristic justification of (1.1) is the following. Let N denote the total number of people in the population; N_A the number of people who have characteristic A; N_B the number of people who have characteristic B; and N_{AB} the number of people who have both characteristics. It is then clear that

$$P(A) = \frac{N_A}{N},$$

$$P(B) = \frac{N_B}{N},$$

and

$$P(A \text{ and } B) = \frac{N_{AB}}{N}.$$

By $P(A|B)$ we mean the proportion out of all people who have characteristic B who also have characteristic A, so that both the numerator and the denominator of $P(A|B)$ must be specific to B. Thus

$$P(A|B) = \frac{N_{AB}}{N_B}. \tag{1.4}$$

If we now divide the numerator and denominator of (1.4) by N, we find that

$$P(A|B) = \frac{N_{AB}/N}{N_B/N} = \frac{P(A \text{ and } B)}{P(B)}.$$

Equation 1.2 may be derived similarly:

$$P(B|A) = \frac{N_{AB}}{N_A} = \frac{N_{AB}/N}{N_A/N} = \frac{P(A \text{ and } B)}{P(A)}.$$

Equations 1.1 and 1.2 are connected by means of *Bayes' theorem:*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{1.5}$$

Equation 1.5 follows from the definition (1.2) of $P(B|A)$ and from the fact, seen by multiplying both sides of (1.1) by $P(B)$, that $P(A \text{ and } B) = P(A|B)P(B)$.

1.2. THE EVALUATION OF A SCREENING TEST

A frequent application of Bayes' theorem is in evaluating the performance of a diagnostic test intended for use in a screening program. Let B denote the event that a person has the disease in question; $\bar{B}$ the event that he does not have the disease; A the event that he gives a positive response to the test; and $\bar{A}$ the event that he gives a negative response. Suppose that the test has been applied to a sample of B's, that is, to a sample of people with the disease, and to a sample of $\bar{B}$'s, that is, to a sample of people without the disease.

The results of this trial of the screening test may by represented by the two conditional probabilities $P(A|B)$ and $P(A|\bar{B})$. $P(A|B)$ is the conditional probability of a positive response given that the person has the disease; the larger $P(A|B)$ is, the more *sensitive* the test is. $P(A|\bar{B})$ is the conditional probability of a positive response given that the person is free of the disease; the smaller $P(A|\bar{B})$ is [equivalently, the larger $P(\bar{A}|\bar{B})$ is], the more *specific* the test is. These definitions of a test's sensitivity and specificity are due to Yerushalmy (1947).

Of greater concern than the test's sensitivity and specificity, however, are the error rates to be expected if the test is actually used in a screening program. If a positive result is taken to indicate the presence of the disease, then the false positive rate, say P_{F+}, is the proportion of people, among those responding positive, who are actually free of the disease, or $P(\bar{B}|A)$. By Bayes' theorem,

$$P_{F+} = P(\bar{B}|A) = \frac{P(A|\bar{B})P(\bar{B})}{P(A)} = \frac{P(A|\bar{B})[1 - P(B)]}{P(A)}, \qquad (1.6)$$

since $P(\bar{B}) = 1 - P(B)$.

The false negative rate, say P_{F-}, is the proportion of people, among those responding negative on the test, who nevertheless have the disease, or $P(B|\bar{A})$. Again by Bayes' theorem,

$$P_{F-} = P(B|\bar{A}) = \frac{P(\bar{A}|B)P(B)}{P(\bar{A})} = \frac{[1 - P(A|B)]P(B)}{1 - P(A)}, \qquad (1.7)$$

since $P(\bar{A}|B) = 1 - P(A|B)$ and $P(\bar{A}) = 1 - P(A)$.

We still need the overall rates $P(A)$ and $P(B)$ in order to evaluate these two error rates. Actually, we only need $P(B)$, for the following reason. Note that

$$P(A) = \frac{N_A}{N} = \frac{N_{AB} + N_{A\bar{B}}}{N} = \frac{N_{AB}}{N} + \frac{N_{A\bar{B}}}{N}. \qquad (1.8)$$