Statistical Estimation of Epidemiological Risk

Kung-Jong Lui
Department of Mathematics and Statistics
San Diego State University, USA

John Wiley & Sons, Ltd

Statistical Estimation of Epidemiological Risk

STATISTICS IN PRACTICE

Advisory Editor

Stephen Senn
University College London, UK

Founding Editor

Vic Barnett
Nottingham Trent University, UK

Statistics in Practice is an important international series of texts, which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. The feedback of views from readers will be most valuable to monitor the success of this aim.

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

Statistical Estimation of Epidemiological Risk

Kung-Jong Lui
Department of Mathematics and Statistics
San Diego State University, USA

John Wiley & Sons, Ltd

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

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

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-85071-X

Typeset in 10/12pt Photina by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Biddles Ltd, Guildford and King's Lynn
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

In memory of my parents
Shung-Wu and Li-Ching

Contents

About the Author

KUNG-JONG LUI is a professor in the Department of Mathematics and Statistics at San Diego State University. Since he obtained his Ph.D. in biostatistics from UCLA in 1982, he has published more than 100 papers in peer-reviewed journals, including *Biometrics, Statistics in Medicine, Biometrical Journal, Psychometrika, Communications in Statistics: Theory and Methods, Science, Proceedings of National Academy of Sciences, Controlled Clinical Trials, Journal of Official Statistics, IEEE Transactions on Reliability, Environmetrics, Test, Computational Statistics and Data Analysis, American Journal of Epidemiology, American Journal of Public Health*, etc. He is a Fellow of the American Statistical Association, a life member of the International Chinese Statistical Association, and a member of the Western North American Region of the International Biometric Society.

Preface

The estimation of epidemiological indices plays an important role in epidemiological investigations. One aim of this book is to provide biostatisticians, epidemiologists, and medical researchers with a useful resource on the different estimators of the most commonly used measures of risk in a variety of designs. Through a systematic presentation and discussion, it is hoped that the reader will appreciate better the use and limitations of, and the relationships among, these indices. Because the material in each chapter is generally self-contained, readers may choose chapters according to their own interests without the need to read through all the preceding chapters. This may increase the utility of the book, although I must admit that some definitions are repeated between chapters to avoid ambiguities in the formulae.

This book is intended for postgraduates and researchers who have one year of training in biostatistics and possess some basic knowledge of epidemiological terms, such as prevalence, risk difference, odds ratio, relative risk, and attributable risk. It is also intended for students of biostatistics and epidemiology as a one-semester graduate course, focusing on statistical estimation of risk in epidemiology. Because research on estimation of epidemiological risk has been quite intensive in the last two decades, to provide readers with up-to-date information I have included many recently developed estimators and relevant references. Thus, this book may also be used as a desk reference for established researchers. Although the book is mainly directed at biostatisticians and epidemiologists, because measures such as the risk difference, relative difference (or relative risk reduction), and number needed to treat are often used to report clinical findings, the book should be useful for statisticians and clinicians working in pharmaceutical areas as well.

When the underlying disease is rare, the probability of obtaining only a few or zero cases in a sample under binomial sampling can be large or non-negligible. To ensure that a reasonable number of cases are obtained, we may consider use of inverse sampling, a fact which has not been widely familiar among practicing biostatisticians or epidemiologists. This may be the first book to attempt to systematically introduce in a unified manner statistical methods relevant to

inverse sampling in epidemiology. In contrast to binomial sampling, we show that the bias of estimators for the relative risk or the odds ratio in paired-sample data can easily be avoided by using inverse sampling. Furthermore, when the sample size is small, asymptotic interval estimators for the relative difference, the attributable risk in case–control studies when the underlying disease is rare, or the odds ratio in paired-sample data may be inappropriate. We note that under inverse sampling the derivation of exact confidence intervals for these indices is straightforward. The results and discussions on inverse sampling presented in this book can provide readers with an alternative way to design their studies.

When the response variable is on an ordinal scale with more than two categories, the odds ratio is inapplicable without arbitrarily collapsing the data. This book also includes a chapter (Chapter 6) focusing on the generalized odds ratio. This measure has an easy interpretation and should be useful for epidemiologists and clinicians when they wish to provide a quantitative measure of the strength of association for ordinal data between two comparison groups without assuming any parametric models.

The attributable risk (AR), representing the proportion of cases that may be prevented if the underlying risk factor under investigation is completely eliminated, is probably one of the most important indices for public health administrators to rank the relative importance of risk factors for intervention. Although there have been numerous recent publications that focus estimation on this useful measure in a variety of designs, many textbooks have touched this topic superficially by considering only the simplest cases in which there are no confounders. I discuss estimation of the AR from the simplest case – no confounders under a variety of designs – to the more complicated case with confounders. I also discuss estimation of the AR for paired-sample data. I further consider the situation in which the exposure variable has multiple levels, and the situation in which one applies the logistic regression model to adjust for the effects of confounding variables in case–control studies. A brief discussion on estimation of the AR under inverse sampling has also been included. The discussions on the AR presented in this book should be useful for researchers working in public health administration by providing relatively complete information on recent developments.

Upon the request of an anonymous reviewer, I have also included a chapter that discusses the use of the 'number needed to treat' (NNT). Because it can be easily understood by clinicians, this index has frequently been employed in randomized trials and evidence-based medicine. However, it has been subject to criticism by statisticians due to misuse and misunderstanding. For example, there are published papers that report the union of two disjoint open intervals as a confidence interval for the NNT, or provide a confidence interval that does not even contain the NNT point estimate. I have tried to present this index in such a way that these criticisms can be avoided. I sincerely hope that readers find the discussion presented here useful in clarifying the limitations of the NNT and in computing interval estimators for it.

I wish to express my indebtedness to my colleagues Drs. Duane Steffey, Colleen Kelly, and Richard Levine at San Diego State University, as well as to the three anonymous reviewers who generously provided valuable comments on an early draft of the manuscript. I also wish to thank the particular reviewer who spent valuable time on the revised draft and provided additional suggestions which led to improvements in the content of this book. I would like to thank Dr. N. Breslow at the University of Washington, the International Agency for Research on Cancer Center, the International Biometric Society, the American Medical Association, and Oxford University for their permission to include the data sets used to illustrate the methods discussed here. I would also like to express my gratitude to Drs. William G. Cumberland, A. A. Afifi, Sander Greenland, Frank Massey, Jr., Olive Dunn, Charles Stone, Robert Jennrich, Potter Chang, and Donald Ylvisaker for their teaching in biostatistics, epidemiology, and mathematical statistics when I was a student at UCLA, as well as Dr. Thomas Ferguson at UCLA and Dr. Daniel McGee at Florida State University for their encouragement and advice in the past. I wish especially to thank Mr. Rob Calver, Editor of Statistics and Mathematics at John Wiley & Sons, for his help and time during the preparation of this book. I also want to thank my wife Jen-Mei, whose patience and understanding have endured throughout so many years and made the work much more pleasant than it otherwise would have been. Finally, I want to express my deepest appreciation to my parents Shung-Wu and Li-Ching for their endless love, support, and guidance, which will live forever in my memory.

Kung-Jong Lui
San Diego, California

1

Population Proportion or Prevalence

To quantify the impact of a given disease on public health in a community, or in studying the variation of a disease distribution between geographical regions to locate the potential causes, we may wish to first estimate the prevalence of the disease, defined as the population proportion of subjects who have it. In this chapter, we start by discussing the estimation of population prevalence under the most commonly assumed case – binomial sampling, in which we take a random sample of n subjects and obtain X cases. For example, to estimate the prevalence of HIV-infected subjects, we may take a random sample of ($n =$) 1000 subjects in a local community and obtain ($x =$) 5 subjects with positive results from an HIV-antibody test. In practice, however, a complete list of the sampling population needed to employ binomial sampling may not be available. We therefore discuss estimation under cluster sampling, in which the sampled unit is the cluster itself rather than the individual subject. As an example, we take a random sample of households and estimate the proportion of people who went to see a doctor in the last 12 months (Cochran, 1977). In this case, the sampled units are households rather than individuals. Other examples of the use of cluster sampling include the study of the effect of an educational intervention program on the use of solar protection among children (Mayer *et al.*, 1997) and the effect of vitamin A supplementation on child mortality (Herrera *et al.*, 1992). As noted by Cochran (1977), the estimate of the population prevalence can be subject to a large relative error when the underlying population prevalence is small under binomial sampling. Furthermore, when the disease is rare, we may even obtain 0 cases in the sample. To alleviate these concerns, we discuss the use of inverse sampling (Haldane, 1945), in which we continue sampling subjects until we obtain a predetermined number x of cases. For example, we may decide to sample subjects until we obtain, say, 5 HIV-infected cases when estimating the prevalence of HIV-infected subjects in a community. In contrast to binomial sampling, the number of cases x under inverse sampling is fixed, but the total number of sampled subjects N needed to obtain these x cases is random. Except

Statistical Estimation of Epidemiological Risk K-J. Lui
 ISBN: 0-470-85071-X (HB)

for specifically referring to the incidence rate, calculated as the number of events divided by the number of person-years of follow-up time, we will generally use the terms probability, proportion, risk, and rate synonymously in this book (Fleiss, 1981). An excellent discussion on explicit definitions of these terms as used in epidemiology appears elsewhere (Selvin, 1996).

1.1 BINOMIAL SAMPLING

Suppose that a random sample of size n is taken from a very large population so that we can reasonably assume that the probability of a randomly selected subject being a case equals a constant π and the events for each randomly selected subject of being a case or a non-case are all mutually independent. Let X denote the random number of cases among these n sampled subjects. The random variable X then follows the binomial distribution with parameters n and π:

$$\mathrm{P}(X = x|\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \tag{1.1}$$

where $x = 0, 1, \ldots, n, 0 < \pi < 1$, and π denotes the underlying population proportion of cases. The most commonly used point estimator of the parameter π is simply the sample proportion of cases:

$$\hat{\pi} = X/n. \tag{1.2}$$

Note that under distribution (1.1), the point estimator $\hat{\pi}$ (1.2) has the expectation $\mathrm{E}(\hat{\pi}) = \pi$ (i.e., $\hat{\pi}$ is an unbiased estimator of the population proportion π) and the variance $\mathrm{Var}(\hat{\pi}) = \pi(1-\pi)/n$ (**Exercise 1.1**). In fact, the estimator $\hat{\pi}$ is the uniformly minimum variance unbiased estimator (UMVUE) of π under (1.1). By the central limit theorem, the random quantity $(\hat{\pi} - \pi)/\sqrt{\mathrm{Var}(\hat{\pi})}$ has the asymptotic standard normal distribution as $n \to \infty$. Thus, by Slutsky's theorem (Casella and Berger, 1990), we obtain an asymptotic $100(1-\alpha)$ percent confidence interval for π using Wald's statistic (Agresti and Coull, 1998),

$$[\max\{\hat{\pi} - \mathrm{Z}_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}, 0\}, \quad \min\{\hat{\pi} + \mathrm{Z}_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}, 1\}]. \tag{1.3}$$

Note that when $\hat{\pi} = 0$ or $\hat{\pi} = 1$, the estimated variance $\hat{\pi}(1-\hat{\pi})/n$ equals 0. Obviously, this underestimates the true variance. Therefore, whenever $\hat{\pi} = 0$ or $\hat{\pi} = 1$, we recommend use of $\hat{\pi}^*(1-\hat{\pi}^*)/n$ to estimate the variance, where $\hat{\pi}^* = (X + 0.5)/(n+1)$. Note also that although interval estimator (1.3) is easy to use, it is well known that when n is not so large that both $n\hat{\pi} \geq 5$ and $n(1-\hat{\pi}) \geq 5$ hold, (1.3) is not expected to perform well due to the possibly skewed sampling distribution of $\hat{\pi}$. To improve the performance of (1.3), we consider the probability $\mathrm{P}([(\hat{\pi} - \pi)/\sqrt{\mathrm{Var}(\hat{\pi})}]^2 \leq \mathrm{Z}^2_{\alpha/2}) \doteq 1 - \alpha$ as n is large. This leads us to obtain the