The Stata Journal (2005)
5, Number 3, pp. 461–464

Review of Statistics for Epidemiology by Jewell

Rino Bellocco
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
rino@meb.ki.se

Abstract. The new book by Jewell (2004) is reviewed.

Keywords: gn0029, epidemiology, biostatistics

1 Introduction

I met the author at a statistical workshop in Germany several years ago when I was a young researcher. I was delighted by his ability to explain complicated statistical models of HIV dynamics with clarity and precision. I feel honored to review his new book; it has been a wonderful learning experience. I have already recommended this text to doctoral students in epidemiology in my logistic regression class at the Karolinska Institutet. I appreciate Jewell's talent for explaining some difficult epidemiological and statistical concepts in clear language. Statistical thinking is an important component in developing and interpreting epidemiological studies. "Although epidemiologists do not need a highly mathematical background in statistical theory to effectively conduct and interpret such studies, they do need more than an encyclopedia of 'recipes'" (Jewell 2004, back cover). In my opinion, this book reaches an almost perfect balance between including very advanced concepts and keeping the discussion at a acceptable level for intermediate graduate students. I consider it to set a new standard for texts on statistics in epidemiology, especially in its treatment of causation and bias.

Jewell's goal is to introduce current statistical techniques used to analyze binary outcome data arising from observational epidemiological studies, although this is not directly implied by his title. Such a goal leaves less space for analysis of rates and times to event, which are also important statistical tools in epidemiological analyses. However, this restriction keeps the book to a reasonable length as intended by the author and encouraged by his lucid overview. The material is based on a one-semester graduate course that he has taught for more than 20 years in the School of Public Health at the University of California, Berkeley. It is assumed that readers are familiar with concepts of random variables, sampling, population parameters, estimation, and hypothesis testing. In particular, familiarity with the binomial, normal, and chi-square distributions is expected, while previous knowledge of linear regression models would make some of the material covered in the last part of the book easier to follow. I like Jewell's statement that "the overall goal here is to give some basic driving lessons, not to get under the hood and tinker with the mechanics of the internal combustion engine" (Jewell 2004, 5). I think that the ideal reader is someone who appreciates doing some mathematics to get an insight into the mathematical formulation without, however, wishing to read formal mathematical proofs. Partial or no previous knowledge of epidemiology is assumed, but emphasis here is more on statistical analysis and interpretation of the results.

I found both innovative and outstanding the introduction to basic causal inference, counterfactuals, causal graphs, and their relationship with the concepts of confounding and selection bias. Researchers aiming to establish possible causal pathways from their data will find this introduction extremely useful.

There are a few simple case studies to lead readers from simple stratified analyses towards more complex regression modeling. Using these examples through several chapters makes it easy to compare the interpretations that emerge from varying approaches.

2 Book content

The text begins with an introductory chapter in which the author outlines the key features in disease processes, study designs, and statistical models for binary data, emphasizing the importance of causality in drawing inferential statements. This chapter can be considered an extended overview, and like every subsequent chapter, it ends with a "Comments and further reading" section. These I found extremely interesting as a way to summarize the most recent material on each topic.

Chapters 2 to 4 review measures of disease occurrence, the role of probability in observational studies, and measures of association. I particularly like Jewell's way of introducing relative risk and its relationship to odds ratios. A reader familiar with these topics could well skip these chapters, but they do provide a wonderful introduction to the clear and elegant writing style of the author.

Chapter 5 provides an introduction to study design. When I first read it, I was surprised by the distinction between population-based studies and cohort studies; I had to think more carefully, but then I found it innovative to think of a cohort study, where we select subjects based on exposure, as the opposite of a case–control study, where we select subjects based on disease status. In both types of studies, it follows that only conditional probabilities are estimable. Nested case–control studies based on risk set sampling and case–cohort studies are introduced as key variants of the case–control design, and their applicability is well discussed with real-life examples.

Chapters 6 and 7 introduce the reader to the study of 2 × 2 tables, how to perform statistical testing, and more specifically, how to do estimation and inference for the measures of association presented in previous chapters. I appreciate the statement towards the end of chapter 7 on sample size, emphasizing the importance of quality of information, such as a precise assessment of exposure information. The chapter ends with a nicely written section on measurement error, with a short introduction to SIMEX (Carroll, Ruppert, and Stefanski 1995).

Chapters 8 through 11 describe confounding and interaction in the context of causal inference using recent ideas of counterfactuals and causal graphs. These chapters also introduce stratification-based methods growing from Mantel–Haenszel procedures, which the author views as an important step of data exploration before moving towards regression modeling. I enjoyed reading these chapters, even if it takes some effort to follow carefully the formal, but important, definitions needed to define counterfactual variables and direct acyclic graphs. The discussion on additive and multiplicative interaction through definition of proper counterfactuals is appealing. I agree with Jewell when he says that without further understanding of a true biological mechanism describing how two factors act, together and alone, to produce the disease outcome, it is misleading to focus on a 'specified interaction scale', whether additive, multiplicative, or any alternative. The focus then moves quite naturally to the second part, mainly on logistic regression.

Chapter 12 introduces the rationale for regression models, briefly touching on linear, probit, and log-linear models before presenting the logistic regression model and the interpretation of parameters in simple and multivariate versions for both categorical and continuous variables.

Chapters 13 through 15 cover logistic regression in greater detail, starting from a clear and sharp description of the likelihood function underlying the logistic model and of Wald, score, and likelihood-ratio methods. The treatment continues with adjustment of logistic regression to case–control studies. Confounding and interactions are nicely framed now within logistic regression. Detailed advice follows on handling continuous variables (with good paragraphs on centering and multicollinearity), interpreting results, model building, and finally assessing goodness of fit.

Chapter 16 considers the application of statistical techniques to matched studies (frequency and paired data), again starting with classical methods before introducing conditional logistic regression and ending with a useful discussion of the effects of breaking the match during the analysis and some final comments on the pros and cons of matched designs.

The penultimate chapter of the book presents alternatives and extensions to the logistic regression model (generalized additive models, classification trees, methods for clustered and longitudinal data) before giving a fairly brief description of the Cox proportional hazards model, with emphasis on situations in which logistic regression and the Cox model give similar estimates. The author does not seem to consider the substantial gain in precision available from the Cox model when the outcome event is observed in most subjects by the end of follow-up. Finally a short chapter discusses the literature pertaining to three of the case studies that are used in many of the example analyses in the book.

3 Assessment

Teachers, students, and many others will benefit from this new book. I did so myself. It represents a valuable addition to *Statistical Models in Epidemiology* (Clayton and Hills 1993). The book is not linked exclusively to any statistical software. However, most solutions use Stata, and datasets are available in both Excel and Stata on the publisher's web site. Solutions of problems, as acknowledged by Jewell, are provided by teaching assistants and need not be considered as standard answers. I think that this could be improved in any future editions. Sometimes I found it cumbersome jumping from one chapter to another to find a cited table or figure. But these are details that do not affect the generally good flow of the book. Chapter 8 is an outstanding description of basic concepts in counterfactual and causal inference. I am glad that Jewell considered this a key chapter, placing it right in the middle of the book.

Scientific eminence and superior teaching skills are needed for a book to become essential reading. Let the author have the final words: "I love teaching this material, and whenever I do, I take some quiet time at the beginning of the term to remind myself about what lies beneath the numbers and the formulas. Many of the studies I use as examples, and many of the epidemiological studies I have had the privilege of being a part of, investigate diseases that are truly devastating. Making a contribution, however small, to understanding these human conditions is at the core of every epidemiological investigation. As a statistician, it is easy to be immersed in the numbers churned up by data and the tantalizing implications of their interpretation. But behind every data point there is a human story, there is a family, and there is suffering" (Jewell 2004, 8).

4 References

Carroll, R., D. Ruppert, and L. Stefanski. 1995. *Measurement Error in Nonlinear Models*. New York: Chapman & Hall.

Clayton, D. and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Jewell, N. P. 2004. *Statistics for Epidemiology*. New York: Chapman & Hall /CRC.

About the Author

Rino Bellocco is an associate professor of Biostatistics at the Department of Medical Epidemiology and Biostatistics, at the Karolinska Institutet, Stockholm, Sweden.