BOOK REVIEW

Discrete Statistical Models with Social Science Applications

Erling B. Andersen

According to the preface, this is the first book of a series of two. The second book will deal with continuous statistical models. The main subjects of the present book are based on three developments: (1) the theory of exponential families, (2) the theory of log-linear models, and (3) the theory of logistic models in psychometrics. More specifically, some of the models dealt with are log-linear models for contingency tables, latent trait models, Markov chains, and the Bradley-Terry model for paired comparisons. Each chapter contains a large number of examples and exercises. Solutions to the exercises are, however, not given. Before presenting general conclusion about the book, each chapter will be reviewed separately.

Overview

The first chapter (Statistical Models; 8 pages) gives an introduction to the use of statistical models in the social sciences. Examples from a variety of social disciplines, including medicine, are given.

The intention of chapter 2 (Preliminary Theory; 44 pages) is to refresh the reader's knowledge of probability theory and statistics, which are applied in the rest of the book. The given survey of subjects serves this goal well, except for the treatment of statistical optimality theory. Optimality concepts are introduced; but they are hardly, if ever, used in the rest of the book. A more serious drawback of the chapter is its large number of mathematical-statistical errors. Two examples are (1) the absence of the norming factor \( \sqrt{n} \) in the theory of asymptotic normality and (2) the entanglement of the concepts "convergence in probability" and "convergence with probability one."

The main results in chapter 3 (Statistical Inference; 38 pages) concern the asymptotic distributions of maximum-likelihood estimators and likelihood-ratio statistics for the case of exponential families. Attention is also drawn to models with nonidentically distributed random variables (which can be met, for example, in regression analyses) and to models containing so-called incidental parameters. The theory of incidental parameters is of great importance to the latent trait models discussed in chapter 6 because these models assume that every new person in the sample introduces one or more individual parameters in the model. Despite the good choice of subject matter, the pleasure in reading the chapter is to a large extent destroyed by the author's careless style. A close check of the proofs and lemmas by the reviewers often revealed that something was not (entirely) correct or that proofs, although promised by the author (theorems 3.2, 3.4, 3.9), were not given. Moreover, one may wonder whether the author should not have deleted some of the more involved proofs of standard results in this chapter and instead simply referred to statistical textbooks.
The main subjects of chapter 4 (Discrete Models; 54 pages) are (1) statistical inference for parametric multinomial distributions (i.e., multinomial distributions where the cell probabilities are functions of an underlying parameter vector $\theta$); (2) chi-square tests; (3) some discrete probability distributions (binomial, Poisson, Pascal, and hypergeometric distributions); and (4) the principle of conditional tests. The standard reference to statistical theory on parametric multinomial distributions is Rao (1973). It is a pity that Andersen felt the need to modify some of Rao’s proofs; for example, the proof given for the consistency of the maximum-likelihood estimator for the underlying parameter $\theta$ is erroneous. In the preface the author promises that the book offers “an introduction to model building and statistical analysis...” In this chapter it becomes clear, however, that the author often neglects the model-building aspect. To him statistics seem to start where data gathering ends. In the examples probabilistic models are generally not defended by referring to the underlying probabilistic experiment but only to the fact that they “fit” the data. It is to be regretted that by this approach the author reinforces the prevailing habit in the social sciences of forming the hypothesis only after the data are gathered and studied.

The analysis of discrete multivariate data given in the form of cross classifications is the subject of chapter 5 (Contingency Tables; 90 pages). A central place is given to the theory of loglinear models. In these models the theory of exponential families, treated in chapter 3, is made applicable by writing the expected frequency of each cell as a linear combination of parameters. The theory of loglinear models admits unified treatment of the statistical analyses of two-way and higher-order contingency tables. Furthermore, the analysis becomes to some extent independent of the sampling design. This last property may be attractive for the social researcher but also involves risks. If one doesn’t have to worry too much about the sampling design, one might even be tempted to think that loglinear models are applicable to every contingency table, even if the samples are not drawn randomly. We think Andersen underestimates this danger (see, e.g., page 152, line 1, or example 5.3). Possibly relevant hypotheses are obtained by setting one or more parameters of the fully saturated loglinear model equal to zero. The hypotheses or models thus obtained can be partially ordered such that a “later” model is obtained by setting some parameters of an “earlier” model equal to zero. Andersen recommends hierarchical testing techniques in which testing begins with an “early” hypothesis and proceeds with “later” hypotheses until a hypothesis has to be rejected. However, the use of this approach easily involves the danger of automatism in data analyses. It is of utmost importance that the social researcher is encouraged to formulate precisely the questions to be answered before gathering his or her data. The methodology of “fitting the simplest possible model” may be attractive only to the social researcher who explores a completely unknown field and is willing to cross validate his or her findings.

Chapter 6 (Latent Structure Models; 58 pages) treats some latent trait and latent class models. Among the latent trait models, most attention is paid to the Rasch model, primarily because it admits the use of the theory of exponential families. Estimators for the item parameters are obtained by the conditional maximum likelihood method. It is shown how (the distribution of) person parameters can be estimated. The likelihood-ratio test that had previously been introduced by the author (Andersen, 1973) is used to test the applicability of the Rasch model; also the possibility of a graphical model control is shown. Fischer (1974) has shown that “specific objectivity” of measurements is a sufficient condition for the Rasch model to hold. It is regretted that this important argument is omitted in Andersen’s book. Andersen also discusses the polychotomous Rasch model (incorrectly called by him a model for multiple-choice questionnaires; essential is only that the answers are scored polychotomously). Unfortunately, this model is introduced as a mathematical formalism; a satisfactory interpretation of the parameters is not given. Rightly, latent class models are only briefly discussed in the chapter. However, it is wrongly suggested that in using latent class models, it should be known beforehand to which class all individuals belong (p. 240). If this were true, the models would be useless.
After a short introduction to stochastic processes in chapter 7 (Discrete Time Models; 49 pages), attention is restricted to Markov processes and chains. The last two sections are devoted to statistical inference in Markov chains. Applications of Markov chains in demography are treated extensively in a separate section. The main problem with this chapter is that it is written too concisely to serve as an introduction to the field. Furthermore, loose wording and numerous typographical errors often impede a good understanding. To introduce the intended readers to the theory of chapter 7 in a careful and systematic way would require at least twice as many pages.

The final chapter (Some Special Models; 18 pages) deals with the Bradley-Terry model for paired comparisons, logistic bioassay models, and logistic models for binary regression. The treatment is concise and to the point.

Conclusion

This book gives a unified treatment of a great variety of important statistical models for the social sciences, the unifying element being the theory of exponential families. It is a pity, however, that we cannot be entirely enthusiastic about the book, mainly because its style is rather careless; we compiled a list of many pages containing inexact descriptions, errors in formulas and proofs, notational inconsistencies, and typing errors. These errors could have the effect that a reader of the book would lose the thread in a treatment of a subject or would even become irritated. This could have been avoided easily by more careful editing of the book. Furthermore, the composition of some chapters is somewhat unbalanced. Finally, we regret that metastatistical issues (such as, for example, the use of hypothesis-testing techniques in the social sciences) are largely ignored by the author. Sometimes the examples seem to be only numerical illustrations of mathematical formulas. It is hoped that the second edition will be thoroughly revised in these respects. In our opinion the subject matter of the book deserves it.

Theo J. H. M. Eggen, Wim J. van der Linden, and Sebie J. Oosterloo
Afdeling Toegepaste Onderwijskunde
Technische Hogeschool Twente
Postbus 217, 7500 AE Enschede
The Netherlands

References


Available on request from the reviewers.