Complex Surveys

Parimal Mukhopadhyay

Complex Surveys

Analysis of Categorical Data

Springer

Parimal Mukhopadhyay
Indian Statistical Institute
Kolkata, West Bengal
India

ISBN 978-981-10-0870-2 ISBN 978-981-10-0871-9 (eBook)
DOI 10.1007/978-981-10-0871-9

Library of Congress Control Number: 2016936288

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media Singapore Pte Ltd.

To
The Loving Memory of My Wife

Manju

Preface

Most of the data a statistician uses is categorical in nature. In the realms of biomedicine and social sciences, ecology and demography, voting pattern and marketing, to name a few, categorical data dominate. They are the major raw materials for analysis for testing different hypotheses relating to the populations which generate such data. Such data are generally obtained from surveys carried under a complex survey design, generally a stratified multistage design. In analysis of data collected through sample surveys, standard statistical techniques are often routinely employed.

We recall the well-recited phrase which we chanted in our undergraduate days: Let $x_1,\ldots, x_n$ be a random sample of size n drawn from a population with probability density function $f(x)$ or probability mass function $p_M(x)$, etc. This means that the sampled units whose observations are $x_1, x_2,\ldots, x_n$, are drawn by simple random sampling with replacement (*srswr*). This also implies that observed variables $x_1,\ldots, x_n$ are independently and identically distributed (IID). In fact, most of the results in theoretical statistics, including those in usual analysis of categorical data, are based on these assumptions.

However, survey populations are often complex with different cell probabilities in different subgroups of the population, and this implies a situation different from the IID setup. Longitudinal surveys—where sample subjects are observed over two or more time points—typically lead to dependent observations over time. Moreover, longitudinal surveys often have complex survey designs that involve clustering which results in cross-sectional dependence among samples.

In view of these facts, it is, therefore, necessary to modify the usual tests of goodness of fit, like Pearsonian chi-square, likelihood ratio and Wald statistic to make them suitable for use in the context of data from complex surveys. A host of ardent researchers have developed a number of such tests for this purpose over more than last four decades.

There are already a myriad number of textbooks and research monographs on analysis of categorical data. Then why is another book in the area required? My humble answer is that all those treatise provide excellent description of the

categorical data analysis under the classical setup (usual *srswr* or IID assumption), but none addresses the problem when the data are obtained through complex sample survey designs, which more often than not fail to satisfy the usual assumptions. The present endeavor tries to fill in the gap in the area.

The idea of writing this book is, therefore, to make a review of some of the ideas that have blown out in the field of analysis of categorical data from complex surveys. In doing so, I have tried to systematically arrange the results and provide relevant examples to illuminate the ideas. This research monograph is a review of the works already done in the area and does not offer any new investigation. As such I have unhesitatingly used a host of brilliant publications in this area. A brief outline of different chapters is indicated as under:

(1) Chapter 1: Basic ideas of sampling; finite population; sampling design; estimator; different sampling strategies; design-based method of making inference; superpopulation model; model-based inference
(2) Chapter 2: Effects of a true complex design on the variance of an estimator with reference to a *srswr* design or an IID-model setup; design effects; misspecification effects; multivariate design effect; nonparametric variance estimation
(3) Chapter 3: Review of classical models of categorial data; tests of hypotheses for goodness of fit; log-linear model; logistic regression model
(4) Analysis of categorical data from complex surveys under full or saturated models; different goodness-of-fit tests and their modifications
(5) Analysis of categorical data from complex surveys under log-linear model; different goodness-of-fit tests and their modifications
(6) Analysis of categorical data from complex surveys under binomial and polytomous logistic regression model; different goodness-of-fit tests and their modifications
(7) Analysis of categorical data from complex surveys when misclassification errors are present; different goodness-of-fit tests and their modifications
(8) Some procedures for obtaining approximate maximum likelihood estimators; pseudo-likelihood approach for estimation of finite population parameters; design-adjusted estimators; mixed model framework; principal component analysis
(9) Appendix: Asymptotic properties of multinomial distribution; asymptotic distribution of different goodness-of-fit tests; Neyman's (1949) and Wald's (1943) procedures for testing general hypotheses relating to population proportions

I gratefully acknowledge my indebtedness to the authorities of PHI Learning, New Delhi, India, for kindly allowing me to use a part of my book, *Theory and Methods of Survey Sampling*, 2nd ed., 2009, in Chap. 2 of the present book. I am thankful to Mr. Shamin Ahmad, Senior Editor for Mathematical Sciences at

Springer, New Delhi, for his kind encouragement. The book was prepared at the Indian Statistical Institute, Kolkata, to the authorities of which I acknowledge my thankfulness. And last but not the least, I must acknowledge my indebtedness to my family for their silent encouragement and support throughout this project.

January 2016 Parimal Mukhopadhyay

Contents

About the Author

Parimal Mukhopadhyay is a former professor of statistics at the Indian Statistical Institute, Kolkata, India. He received his Ph.D. degree in statistics from the University of Calcutta in 1977. He also served as a faculty member at the University of Ife, Nigeria, Moi University, Kenya, University of South Pacific, Fiji Islands and held visiting positions at the University of Montreal, University of Windsor, Stockholm University and the University of Western Australia. He has written more than 75 research papers in survey sampling, some co-authored and eleven books on statistics. He was a member of the Institute of Mathematical Statistics and elected member of the International Statistical Institute.