Springer Proceedings in Mathematics & Statistics

Pankaj K. Choudhary
Chaitra H. Nagaraja
Hon Keung Tony Ng *Editors*

Ordered Data Analysis, Modeling and Health Research Methods

HNN (60)

Springer

Springer Proceedings in Mathematics & Statistics

Volume 149

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Pankaj K. Choudhary · Chaitra H. Nagaraja
Hon Keung Tony Ng
Editors

Ordered Data Analysis, Modeling and Health Research Methods

In Honor of H.N. Nagaraja's 60th Birthday

HNN (60)

Springer

Editors
Pankaj K. Choudhary
University of Texas at Dallas
Richardson, TX
USA

Hon Keung Tony Ng
Southern Methodist University
Dallas, TX
USA

Chaitra H. Nagaraja
Fordham University
New York, NY
USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-25431-9 ISBN 978-3-319-25433-3 (eBook)
DOI 10.1007/978-3-319-25433-3

Library of Congress Control Number: 2015953804

Mathematics Subject Classification (2010): 62G30, 62N01, 62F10, 62P10, 62L12, 62G86, 92D10, 92D20, 92D40, 94A17, 62-06

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

H.N. Nagaraja

Group photo taken at the conference in honor of H.N. Nagaraja (March 8, 2014)

Preface

This volume emerged from Ordered Data Analysis, Models and Health Research Methods: An International Conference in Honor of H.N. Nagaraja for His 60th Birthday that was held from March 7 to 9, 2014 at the University of Texas at Dallas. Over 200 participants from 14 countries attended the conference which broadly focused on the areas in which Prof. H.N. Nagaraja has made significant contributions. The papers in this book are arranged in accordance with the conference themes, starting with order statistics, followed by stochastic modeling and estimation, and concluding with developments in statistical methods for health research.

Biography

Haikady Navada Nagaraja was born in 1954 in a small village in Karnataka, India. He received his Bachelor's degree in Mathematics and Statistics in 1972, and his Master's degree in Statistics in 1974, both from the University of Mysore. After teaching at the university as a lecturer for 3 years, he began his doctoral studies at Iowa State University in 1977, where he received his Ph.D. in Statistics in 1980. His dissertation, entitled *Contributions to the Theory of the Selection Differential and Order Statistics,* was completed under the supervision of Prof. H.A. David. He joined the Department of Statistics at The Ohio State University after graduation, where he remained ever since. H.N. Nagaraja became a biostatistician for the General Clinical Research Center at The Ohio State University College of Medicine in 1993, resulting in a joint appointment with the College of Medicine. In 2010, he moved from the Department of Statistics to the Division of Biostatistics, College of Public Health, as the Chair of the division. After 35 years at Ohio State, he retired in 2015 and is now Professor Emeritus of Biostatistics. He remains a prolific researcher (and dedicated teacher) with over 180 research publications, split among theoretical, methodological, and applied topics.

He started his career by studying order statistics and related subjects, including record values, concomitants of order statistics, stochastic modeling, and characterizations of distributions. For example, Nagaraja [1], the first paper drawn from his doctoral thesis, focuses on asymptotics for order statistics. Nagaraja [2] is one of his fundamental contributions to order statistics for discrete populations. Bunge and Nagaraja [3], one of his first papers with a doctoral student, is on record values. Nagaraja and David [4] is on distributions of the maximum of concomitants. Abo-Eleneen and Nagaraja [5] provides an important contribution to Fisher Information for censored samples. Even after his move to the College of Public Health, he has been able to find time for theoretical work such as his recent paper on spacings of neighboring order statistics, Nagaraja, Bharath, and Zhang [6].

H.N. Nagaraja is among the few statisticians who can work with equal mastery on problems involving statistical theory and methodology as well as on applications. His collaborative work at The Ohio State University College of Medicine, and later at the College of Public Health, generated new types of statistical problems to solve for a variety of applications such as cognition and nephrology. For example, Choudhary and Nagaraja [7] provides a methodology for agreement studies, motivated by a question about comparing devices for measuring daily caloric expenditure in exercise physiology. Berntson et al. [8] develops a framework to standardize methods on how to study heart rate variability. One of the most cited papers in genetics, Yang et al. [9] is an important contribution to the study of lupus. Zhang et al. [10] develops an equation to predict kidney injury. Most recently, Scharre et al. [11] validates a self-administered cognitive test called SAGE.

H.N. Nagaraja has coauthored three books and coedited two volumes. Arnold, Balakrishnan, and Nagaraja [12] and David and Nagaraja [13] are both on order statistics while Arnold, Balakrishnan, and Nagaraja [14] focuses on record values. In 1996, he edited a book with Profs. P.K. Sen and D.F. Morrison [15], a compilation in honor of his doctoral advisor, Prof. H.A. David. In addition, he edited a second compilation volume honoring Prof. S. Panchapakesan with Profs. N. Balakrishnan and N. Kannan, published in 2005 [16].

In recognition of his influential work in statistics, H.N. Nagaraja was selected to be a Fellow of both the American Statistical Association (2000) and the American Association for the Advancement of Science (2012). He is also an elected member of the International Statistical Institute (1993). H.N. Nagaraja has contributed substantially to the service of the statistical profession, including as President of the International Indian Statistical Association (2010–2011).

In addition to being a distinguished researcher, H.N. Nagaraja is an outstanding teacher and mentor, winning the Powers Award for Excellence in the Teaching of Statistics at Ohio State in 1993. He often tells students, “Statistics is never a ‘Love at first sight’ subject. If you only take one statistics class, you most likely would have hated it. Take another one, and you will sign up for many more.” His intellect, humility, integrity, and humor, along with his caring nature, inspire his students,

bringing out their best. While at Ohio State, H.N. Nagaraja has supervised 18 Ph.D. students and co-supervised two others. He has also developed many courses, especially for non-statisticians, to improve statistical literacy and statistical practice. For the past 3 years, he has taken undergraduate students from Ohio State to Karnataka, India, to teach them about public health issues in developing countries.

At home in Columbus, he lives with Jyothi, his wife of 34 years, and has two daughters, Chaitra and Smitha. He enjoys traveling, reading about history, and cheering on the Ohio State Buckeyes.

Outline of This Volume

This volume brings together 15 invited research papers written by authors drawn from the conference participants. These papers are categorized as follows: Ordered Data Analysis (Part I), Stochastic Modeling and Estimation (Part II), and Statistical Methods for Health Research (Part III). All papers underwent a rigorous peer-review process.

Part I of this book contains six papers on both theoretical and applied topics in ordered data analysis. Arnold and Villaseñor investigate the direction of bias of the estimated sample Lorenz curve. Balakrishnan, Davies, Keating, and Mason discuss the Pitman closest estimators based on convex linear combinations of two contiguous order statistics. Hosking studies different methods of constructing joint confidence regions for *L*-skewness and *L*-kurtosis. The last three papers in this part discuss some recent results on progressively censored order statistics. Ng, Duan, and Chan provide simple computational methods of the conditional single and product moments of progressively censored order statistics under a time constraint. Cramer and Iliopoulos generalize adaptive progressive censoring to a scheme that allows for arbitrary inspection times and possible removals of units. Finally, Abo-Eleneen and Almohaimeed discuss computational approaches of the Renyi entropy in sets of consecutive progressively Type-II censored order statistics.

The theme for Part II is stochastic models and estimation methods, and the papers are inspired by a range of applications from contagion in financial networks to estimating the number of species in a population. We start with Burkschat, Kamps, and Kateri, who develop three approaches for estimating hazard rates of sequential order statistics within the context of connected systems. Next, Barlevy and Nagaraja examine a framework to study banking networks using a discrete spacings model. Bunge then develops a framework to improve estimates of the number of classes in a population using distributions based on generalized hypergeometric functions. The next contribution is by Serhiyenko, Ravishanker, and Venkatesan, who model multivariate counts data over time using a level correlated, zero-inflated Poisson model. They apply their method to prescription drug sales data. This part concludes with a paper by Sengupta, Choudhary, and Cassey, who propose a larger class of models that enables one to model skewed and heavy-tailed data. They illustrate their robust mixed model on crab claw measurements.

Part III is composed of four papers focusing on health research methods. Rettiganti and Nagaraja study a model for analyzing brain lesion counts in multiple sclerosis patients. Mukhopadhyay and Banerjee focus on constructing confidence intervals for the probability of success in a Bernoulli trial using a sequential approach. Papachristou follows by developing a method to detect a subset of single nucleotide polymorphisms on a genetic map that may be associated with a quantitative phenotype. Finally, Park and Lin propose a model to reconstruct the underlying three-dimensional spatial structure of a species' genome.

July 2015

Pankaj K. Choudhary
Chaitra H. Nagaraja
Hon Keung Tony Ng

References

1. Nagaraja, H.N. 1982. Some asymptotic results for the induced selection differential. *Journal of Applied Probability* 19:253–261.
2. Nagaraja, H.N. 1986. Structure of discrete order statistics. *Journal of Statistical Planning and Inference* 13:165–177.
3. Bunge, J.A., and H.N. Nagaraja. 1992. Exact distribution theory for some point process record models. *Advances in Applied Probability* 24:20–44.
4. Nagaraja, H.N., and H.A. David. 1994. Distribution of the maximum of concomitants of selected order statistics. *Annals of Statistics* 22:478–494.
5. Nagaraja, H.N., and Z. Abo-Eleneen. 2008. Fisher Information in order statistics and their concomitants in bivariate censored samples, *Metrika* 67:327–347.
6. Nagaraja, H.N., K. Bharath, and F. Zhang. 2015. Spacings around an order statistic. *Annals of the Institute of Statistical Mathematics* 67:515–540.
7. Choudhary, P.K., and H.N. Nagaraja. 2005. Selecting the instrument closest to a gold standard. *Journal of Statistical Planning and Inference* 129:229–237.
8. Berntson, G.G., J.T. Bigger, Jr., D.L. Eckberg, P. Grossman, P.G. Kaufmann, M. Malik, H.N. Nagaraja, S.W. Porges, J.P. Saul, P.H. Stone, and M.W. van der Molen. 1997. Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology* 34:623–648.
9. Yang, Y., E.K. Chung, Y.L. Wu, S.L. Savelli, H.N. Nagaraja, B. Zhou, M. Hebert, K.N. Jones, Y. Shu, K. Kitzmiller, K.A. Blanchong, K.A. McBride, G.C. Higgins, R.M. Rennebohm, R.R. Rice, K.V. Hackshaw, R.A.S. Roubey, J.M. Grossman, B.P. Tsao, D.J. Birmingham, B.H. Rovin, L.A. Hebert, and C.Y. Yu. 2007. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *American Journal of Human Genetics* 80:1037–1054.
10. Zhang, X., H.N. Nagaraja, T. Nadasdy, H. Song, A. McKinley, J. Prosek, S. Kamadana, and B.H. Rovin. 2012. A composite urine biomarker reflects interstitial inflammation in lupus nephritis kidney biopsies. *Kidney International,* 81:401–406.
11. Scharre, D.W., S.I. Chang, H.N. Nagaraja, J. Yager-Schweller, and R.A. Murden. 2014. Community cognitive screening using the self-administered gerocognitive examination (SAGE). *Journal of Neuropsychiatry and Clinical Neurosciences* 26:369–375.

12. Arnold, B.C., N. Balakrishnan, and H.N. Nagaraja. 1992. *A first course in order statistics*. John Wiley & Sons. 279 p. (SIAM Reprint 2008).
13. David, H.A., and H.N. Nagaraja. 2003. *Order Statistics,*3rd Edition. John Wiley. 458 p.
14. Arnold, B.C., N. Balakrishnan, and H.N. Nagaraja. 1998. *Records*. John Wiley. 312 p.
15. Nagaraja, H.N., P.K. Sen, and D.F. Morrison, (eds.). 1996. *Statistical theory and applications —Papers in Honor of Herbert A. David*. Springer-Verlag, 334 p.
16. Balakrishnan, N., N. Kannan, and H.N. Nagaraja, (eds.). 2005. *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*. Birkhäuser, 412 p.

Acknowledgements

We thank the sponsors of the conference—Cytel Inc., SAS JMP, The Ohio State University, Southern Methodist University, and The University of Texas at Dallas—for their support. In addition, we thank Prof. N. Balakrishnan for his valuable advice as we prepared this book. We also want to recognize the contributors to this volume for their enthusiastic participation. Special thanks are due to the referees as well for their careful and constructive reviews that enhanced the quality of the papers. The logo for the conference, which also appears on the cover of this book, was designed by Smitha H. Nagaraja. We are grateful to Springer, in particular, to Marina Reizakis for guiding the volume from conception to print, and to Frank Holzwarth for assistance with the Online Conference System.

On behalf of the contributors, we dedicate this volume to Prof. H.N. Nagaraja. We are all fortunate to be able to call him our mentor, colleague, and friend.

July 2015

Pankaj K. Choudhary
Chaitra H. Nagaraja
Hon Keung Tony Ng

Contents

List of Referees

We would like to thank the following individuals for serving as reviewers and providing thoughtful and thorough evaluations of the papers contained in this volume:

Barry C. Arnold (University of California, Riverside, USA)
Huiman Xie Barnhart (Duke University, Durham, USA)
Ping Shing Chan (The Chinese University of Hong Kong, Hong Kong SAR, China)
Katherine Davies (University of Manitoba, Winnipeg, Canada)
Shyamal K. De (National Institute of Science Education and Research—Bhubaneswar, India)
Joseph L. Gastwirth (George Washington University, Washington, USA)
Subharup Guha (University of Missouri—Columbia, USA)
George Iliopoulos (University of Piraeus, Piraeus, Greece)
Victor X. Jin (University of Texas Health Science Center at San Antonio, USA)
Yuhlong Lio (University of South Dakota, Vermillion, USA)
Sangun Park (Yonsei University, Seoul, South Korea)
Asuman Turkmen (The Ohio State University, Columbus, USA)
Dongliang Wang (SUNY Upstate Medical University, Syracuse, USA)

Part I
Ordered Data Analysis

Bias of the Sample Lorenz Curve

Barry C. Arnold and Jose A. Villaseñor

Abstract Inequality is often underestimated using sample data. For several parent distributions it is possible to prove that the sample Lorenz curve is a positively biased estimate of the population Lorenz curve. In this paper, several sufficient conditions for such positive bias are investigated. An example shows that negative bias is not impossible, though apparently not common.

Keywords Inequality · Majorization · Exponential distribution · Gini index · Pietra index · Amato index

1 Introduction

The Lorenz curve is defined for each member of the class of all non-negative random variables with positive finite expectations. It has been noted that the sample Lorenz curve often exhibits less inequality than does the population Lorenz curve. This suggests that the sample curve is a positively biased estimate of the population curve. The possibility of such positive bias is investigated in this paper. Some sufficient conditions for positive bias are presented. A simple example is presented to confirm that negative bias is actually possible, though it is not expected to be encountered frequently in practice.

B.C. Arnold (✉)
Department of Statistics, University of California, Riverside, USA
e-mail: barry.arnold@ucr.edu

J.A. Villaseñor
Department of Statistics, Colegio de Postgraduados, Montecillo, Mexico
e-mail: jvillasr@colpos.mx

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_1

2 The Lorenz Curve and the Sample Lorenz Curve

Further discussion of the topics in this section may be found in Chap. 17 of Marshall et al. [6].

Gastwirth [3] proposed the following definition of the Lorenz curve defined on the class $\mathscr{L}_+$ of non-negative random variables with positive finite expectations.

Definition 1 The Lorenz curve L_X of a random variable $X \in \mathscr{L}_+$ is defined as

$$L_X(u) = \frac{\int_0^u F_X^{-1}(y)dy}{\int_0^1 F_X^{-1}(y)dy} = \frac{\int_0^u F_X^{-1}(y)dy}{E(X)}, \quad 0 \le u \le 1, \tag{1}$$

where

$$\begin{aligned} F_X^{-1}(y) &= \sup\{x : F_X(x) \le y\}, \qquad 0 \le y < 1, \\ &= \sup\{x : F_X(x) < 1\}, \qquad y = 1, \end{aligned}$$

is the right continuous inverse distribution function of the random variable X.

Observe that if X is the random variable associated with the vector $\underset{\sim}{x}$ defined by $P(X = x_i) = 1/n,\ i = 1, 2, \ldots, n$, then the corresponding Gastwirth Lorenz curve $L_X(u)$ and the curve suggested by Lorenz [5], $L_{\underset{\sim}{x}}(u)$, are identical. The Lorenz order can then be extended to allow comparison of random variables as follows.

Definition 2 For $X, Y \in \mathscr{L}_+$, with corresponding Lorenz curves L_X and L_Y, X is less than Y in the Lorenz order, written as $X \le_L Y$, if $L_X(u) \ge L_Y(u)$ for all $u \in [0, 1]$.

It is possible to prove a natural extension of a theorem due to Hardy, Littlewood and Polya [4], which makes use of the Lorenz order rather than being restricted to majorization. Thus:

Theorem 1 *For $X, Y \in \mathscr{L}_+$, $X \le_L Y$ if and only if $E(g(X/E(X))) \le E(g(Y/E(Y)))$ for every continuous convex function g such that the expectations exist.*

Another relevant result (due to Strassen [8]), restated in terms of the Lorenz order, is as follows.

Theorem 2 *For non-negative random variables with $E(X) = E(Y)$ we have $Y \le_L X$ if and only if there exist jointly distributed random variables X', Z' such that $X \overset{d}{=} X'$ and $Y \overset{d}{=} E(X'|Z')$. (The notation $X \overset{d}{=} X'$ is to be read "X and X' are identically distributed.")*

If we have a sample $X_1, X_2, \ldots, X_n$ of size n from a distribution $F_X(x)$, the corresponding sample Lorenz curve is defined to be a linear interpolation of the points $(0, 0)$ and $(j/n, \sum_{i=1}^{j} X_{i:n} / \sum_{i=1}^{n} X_{i:n})$, $j = 1, 2, \ldots, n$. Denote the sample Lorenz curve by $L_n(u)$. This notation parallels the popular notation for a

sample distribution function. As the sample size n increases, it is well known that the sequence of sample Lorenz curves $L_n(u)$ converges almost surely uniformly to the population Lorenz curve $L_X(u)$. In this paper interest is focused on the possible bias of $L_n(u)$ as an estimate of $L_X(u)$ for a fixed value of n.

The function $L^*(u) = E(L_n(u))$ is a valid Lorenz curve corresponding to a discrete random variable $\widetilde{X}$ defined by $P(\widetilde{X} = E(X_{j:n}/\sum_{i=1}^n X_{i:n})) = 1/n$. Note that the function $L^*(u)$ is a linear interpolation of the points $(0,0)$ and $(j/n, E(\sum_{i=1}^j X_{i:n}/\sum_{i=1}^n X_{i:n}))$, $j = 1, 2, \ldots, n$. Consequently if we wish to show that $X \leq_L X$, it will be sufficient to verify that $L^*(j/n) \geq L_X(j/n)$, $j = 1, 2, \ldots, n$. This follows since $L_X(u)$ is a convex function and $L^*(u)$ is piecewise linear.

In general, it is quite difficult to obtain an analytic expression for

$$L^*(j/n) = E(\sum_{i=1}^{j} X_{i:n}/\sum_{i=1}^{n} X_{i:n}).$$

One case where the computation is straightforward is that in which F_X corresponds to an exponential distribution. In this case, not only can we obtain an analytic expression for $L^*(j/n)$ but also we can verify that it is a biased estimate of $L_X(j/n)$. This result, to be confirmed in the next section, suggested that such bias might be commonly encountered for a variety of possible distributions F_X. Note that the statement that the sample Lorenz curve, $L^*(u)$, is a positively biased estimate of the population Lorenz curve, $L_X(u)$, is equivalent to the statement $\widetilde{X} \leq_L X$.

3 The Exponential Case

Suppose that we have a sample $X_1, X_2, \ldots, X_n$ of i.i.d. random variables with a common $Exponential(\lambda)$ distribution. As usual denote the corresponding order statistics by $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$. If we denote the common distribution function of the X's by F_X, it may be verified that the corresponding Lorenz curve is of the form

$$\begin{aligned} L_X(u) &= \frac{\int_0^u F_X^{-1}(v)dv}{\int_0^1 F_X^{-1}(v)dv} = \frac{\int_0^u \frac{-1}{\lambda}\log(1-v)dv}{\int_0^1 \frac{-1}{\lambda}\log(1-v)dv} \\ &= u + (1-u)\log(1-u). \end{aligned} \tag{2}$$

To verify that, in this case, the sample Lorenz curve is positively biased, we need to show that for any $j \in \{1, 2, \ldots, n-1\}$ it is the case that

$$\begin{aligned} L^*(j/n) &= E\left(\frac{\sum_{i=1}^j X_{i:n}}{\sum_{i=1}^n X_{i:n}}\right) \\ &> \frac{j}{n} + \frac{n-j}{n}\log\left(\frac{n-j}{n}\right) = L_X(j/n). \end{aligned}$$

We thus need to evaluate $E(\sum_{i=1}^{j} X_{i:n}/\sum_{i=1}^{n} X_{i:n})$. Two key observations are

- The complete minimal sufficient statistic for λ is $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_{i:n}$. By Basu's [1] Lemma, any function of the X_i's whose distribution does not depend on λ is independent of this minimal sufficient statistic. In particular, the statistic $(\sum_{i=1}^{j} X_{i:n}/\sum_{i=1}^{n} X_{i:n})$ is independent of $\sum_{i=1}^{n} X_{i:n}$. Consequently

$$E(\sum_{i=1}^{j} X_{i:n}) = E\left(\frac{\sum_{i=1}^{j} X_{i:n}}{\sum_{i=1}^{n} X_{i:n}} \times \sum_{i=1}^{n} X_{i:n}\right) = E\left(\frac{\sum_{i=1}^{j} X_{i:n}}{\sum_{i=1}^{n} X_{i:n}}\right) E(\sum_{i=1}^{n} X_{i:n})$$

  so that

$$E\left(\frac{\sum_{i=1}^{j} X_{i:n}}{\sum_{i=1}^{n} X_{i:n}}\right) = \frac{E(\sum_{i=1}^{j} X_{i:n})}{E(\sum_{i=1}^{n} X_{i:n})}.$$

  Alternatively, it may be argued that the independence of $(\sum_{i=1}^{j} X_{i:n}/\sum_{i=1}^{n} X_{i:n})$ and $\sum_{i=1}^{n} X_{i:n} = \sum_{i=1}^{n} X_i$, in this case, follows from an earlier result of Pitman [7]. Pitman proved that if $X_1, X_2, \ldots, X_n$ are independent gamma variables with the same scale parameter (which includes the case of an i.i.d. exponential sample), then any scale invariant function of the X_i's is independent of the sum of the X_i's.
- For each i, $X_{i:n}$ can be written as a sum of independent exponentially distributed spacings. Thus $X_{i:n} = \sum_{k=1}^{i} \left(\frac{Y_k}{n-k+1}\right)$, where the Y_k's are i.i.d. $Exponential(\lambda)$ random variables.

Using these observations we have, for each $j \in \{1, 2, \ldots, n-1\}$,

$$\begin{aligned}
L^*(j/n) &= E\left(\frac{\sum_{i=1}^{j} X_{i:n}}{\sum_{i=1}^{n} X_{i:n}}\right) = \frac{E(\sum_{i=1}^{j} X_{i:n})}{E(\sum_{i=1}^{n} X_{i:n})} = \frac{E\left(\sum_{i=1}^{j}\sum_{k=1}^{i} \frac{Y_k}{n-k+1}\right)}{E\left(\sum_{i=1}^{n} X_i\right)} \\
&= \frac{1}{n/\lambda}\left(\sum_{i=1}^{j}\sum_{k=1}^{i} \frac{1}{\lambda(n-k+1)}\right) = \frac{1}{n}\left(\sum_{k=1}^{j}\sum_{i=k}^{j} \frac{1}{n-k+1}\right) \\
&= \frac{1}{n}\sum_{k=1}^{j}\left(\frac{j-k+1}{n-k+1}\right) = \frac{1}{n}\sum_{k=1}^{j}\left(\frac{n-k+1+j-n}{n-k+1}\right) \\
&= \frac{j}{n} + \frac{n-j}{n}\left[-\sum_{k=1}^{j} \frac{1}{n-k+1}\right] > \frac{j}{n} + \frac{n-j}{n}\left[-\int_{n-j}^{n} \frac{dx}{x}\right] \\
&= \frac{j}{n} + \frac{n-j}{n}\left[\ln\left(\frac{n-j}{n}\right)\right] = L_X(j/n),
\end{aligned}$$

confirming the positive bias of the sample Lorenz curve in the exponential case.

Note 1. Numerical computations indicate that, for a fixed value of n, the bias $L^*(j/n) - L_X(j/n)$ is maximized when $j = n-1$. In addition, for each j, the bias appears to decrease monotonically as n increases.

Note 2. Gail and Gastwirth [2] proposed a test of goodness-of-fit to the exponential distribution based on the sample Lorenz curve. They provided a large sample normal approximation to the distribution of $L^*(u)$ for a particular value of $u \in (0, 1)$. Small sample behavior of the test was evaluated via simulation. They do not focus attention on the bias of the sample Lorenz curve, since it is asymptotically negligible as sample size increases. However, if the asymptotic normal approximation is used for intermediate values of n, say between 20 and 50, the bias of the sample Lorenz curve may be expected to degrade the performance of the test.

4 A Simple Counterexample

Numerical investigation using simulated samples from a variety of distributions suggested that sample Lorenz curves might always be positively biased. A proof of this conjecture seemed to be elusive. It was elusive, with good reason, since the general result is not true as is confirmed by consideration of the following simple example.

Consider a random variable X with two possible values such that

$$\begin{aligned} X &= 1 \text{ w.p. } p \\ &= c \text{ w.p. } (1-p), \end{aligned}$$

where $p \in (0, 0.5)$ and $c > 1$. The corresponding quantile function is of the form

$$\begin{aligned} F_X^{-1}(u) &= 1, \quad 0 \le u \le p, \\ &= c, \quad p \le u \le 1. \end{aligned}$$

Now consider a sample of size 2 from this distribution and focus on the values taken on by the Lorenz curve and the expected value of the sample Lorenz curve at the point $u = 1/2$. Elementary computations yield

$$L_X(1/2) = \frac{p + c(0.5 - p)}{p + c(1 - p)}$$

and

$$\begin{aligned} L_2^*(1/2) &= E\left(\frac{X_{1:2}}{X_{1:2} + X_{2:2}}\right) \\ &= \frac{1}{c+1} 2p(1-p) + \frac{1}{2}[p^2 + (1-p)^2]. \end{aligned}$$

It may then be verified that, for a variety of choices of values for c and p, it is the case that $L_X(1/2) > L_2^*(1/2)$. For example, if $c = 3$ and $p = 0.1$, $L_X(1/2) -$

$L_2^*(1/2) = 0.09$, showing that in this case the sample Lorenz curve is negatively biased rather than positively biased.

5 Sufficient Conditions for a Positive Bias

If we look back at the analysis in the exponential case it would appear that a key property of the Exponential distribution was that, for it, it is the case that

$$E\left(\frac{\sum_{i=1}^{j} X_{i:n}}{\sum_{i=1}^{n} X_{i:n}}\right) = \frac{E(\sum_{i=1}^{j} X_{i:n})}{E(\sum_{i=1}^{n} X_{i:n})} = \frac{\sum_{i=1}^{j} \mu_{i:n}}{\sum_{i=1}^{n} \mu_{i:n}}, \tag{3}$$

where the notation $\mu_{i:n} = E(X_{i:n})$ has been introduced.

The property (3) holds for many distributions in addition to the exponential distribution. For example, if the X's have a common density in a one parameter exponential family of the form

$$f(x;\sigma) = r(x)\beta(\sigma)e^{-x/\sigma},$$

where $\sigma \in (0, \infty)$, then Basu's Lemma can be invoked to conclude that property (3) holds in this case also.

In addition, it would appear that replacement of the equality sign in (3) by a greater than or equal sign might still be sufficient to ensure positive bias of the sample Lorenz curve. Such is the case, but a proof of this claim cannot involve analytic computation of the $\mu_{i:n}$'s since this is usually not possible. Instead one can make use of Strassen's theorem, Theorem 2.

Theorem 3 *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with a common distribution F_X with the property that*

$$E\left(\frac{\sum_{i=1}^{j} X_{i:n}}{\sum_{i=1}^{n} X_{i:n}}\right) \geq \frac{\sum_{i=1}^{j} \mu_{i:n}}{\sum_{i=1}^{n} \mu_{i:n}}, \quad j = 1, 2, \ldots, n. \tag{4}$$

Then the corresponding sample Lorenz curve is a positively biased estimate of the population Lorenz curve, i.e., $L^*(u) = E(L_n(u)) \geq L_X(u)$ for all $u \in (0, 1)$.

Proof Without loss of generality, since Lorenz curves are scale invariant, we may assume that $E(X) = 1$ so that $\sum_{i=1}^{n} \mu_{i:n} = n$. There will be three Lorenz curves involved in the following discussion. The first is $L^*(u)$, the expected sample Lorenz curve. It corresponds to a discrete random variable X^* with n equally likely possible values $E[X_{i:n}/(\sum_{i=1}^{n} X_{i:n})], \quad i = 1, 2, \ldots, n$. The second Lorenz curve is $\widetilde{L}(u)$, which corresponds to a discrete random variable $\widetilde{X}$ with n equally likely possible values $\mu_{i:n}, \ i = 1, 2, \ldots, n$. The third Lorenz curve is $L_X(u)$, which corresponds to the random variable X. Since the Lorenz order is defined in terms of nested Lorenz curves, our goal is to prove that $X^* \leq_L X$.

In fact, we will verify that this is true since $X^* \leq_L \widetilde{X} \leq_L X$. Property (4) assures us that $X^* \leq_L \widetilde{X}$, so that it remains only to prove that $\widetilde{X} \leq_L X$.

By Strassen's theorem, it will suffice to show that there exist random variables X' and Z' such that

$$X' \stackrel{d}{=} X$$

and

$$\widetilde{X} \stackrel{d}{=} E(X'|Z').$$

Note that Z' must take on only n values in such a representation.

Consider $X' = X_1$ and a random variable Z' defined by

$$Z' = i \text{ if } X_1 = X_{i:n}, \ \ i = 1, 2, \ldots, n.$$

Note that $P(Z' = i) = 1/n, \ \ i = 1, 2, \ldots, n$. We then have for a fixed value of i,

$$\begin{aligned}
E(X_1|Z = i) &= E(X_1|X_1 = X_{i:n}) \\
&= \int_0^\infty P(X_1 > x|X_1 = X_{i:n})\, dx \\
&= \int_0^\infty \frac{P(X_1 > x, X_1 = X_{i:n})}{P(X_1 = X_{i:n})}\, dx \\
&= \int_0^\infty n P(X_{i:n} > x, X_1 = X_{i:n})\, dx \\
&= \int_0^\infty n \frac{1}{n} \sum_{j=1}^{n} P(X_{i:n} > x, X_j = X_{i:n})\, dx \\
&= \int_0^\infty P(X_{i:n} > x)\, dx = \mu_{i:n}.
\end{aligned}$$

Thus $\widetilde{X} \stackrel{d}{=} E(X'|Z')$ and consequently $\widetilde{X} \leq_L X$.

6 Thoughts on the Sufficient Condition (4)

The sufficient condition (4) can be reinterpreted as a majorization result. Define two vectors $\underline{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$ and $\underline{\tau} = (\tau_1, \tau_2, \ldots, \tau_n)$ by

$$\delta_i = E\left(\frac{X_{i:n}}{\sum_{i=1}^n X_{i:n}}\right), \quad i = 1, 2, \ldots, n$$

and

$$\tau_i = \frac{\mu_{i:n}}{\sum_{i=1}^n \mu_{i:n}}, \quad i = 1, 2, \ldots, n.$$

Note that $\sum_{i=1}^n \delta_i = \sum_{i=1}^n \tau_i = 1$. With this notation condition (4) can be rewritten as simply a majorization, thus

$$\underline{\delta} \prec \underline{\tau}. \tag{5}$$

There are many equivalent definitions of majorization but it does not seem possible to use them to enhance our understanding of the importance of condition (4).

There is a possibility that a simpler sufficient condition than (4) can be identified. For example, one that is easily described, but not easily checked, is a consequence of the following lemma.

Lemma 1 *If X and Y are absolutely continuous and negative dependent such that $P(0 < X < Y) = 1$ then*

$$E\left(\frac{X}{Y}\right) \geq \frac{E(X)}{E(Y)}.$$

Proof $E\left(\frac{X}{Y}\right) = E\left(E\left(\frac{X}{Y}|X\right)\right) = E\left(XE\left(\frac{1}{Y}|X\right)\right)$.

By Jensen's inequality, $E\left(\frac{1}{Y}|X\right) \geq \frac{1}{E(Y|X)}$. Define for $x > 0$, $g(x) = E(Y|X = x)$, hence

$$E\left(\frac{X}{Y}\right) \geq E\left(\frac{X}{g(X)}\right). \tag{6}$$

X and Y are negative dependent if and only if $f_{X,Y}(x, y) \leq f_X(x) f_Y(y)$. Then for $x > 0$, $f_{Y|X}(y|x) \leq f_Y(y)$.

On the other hand, $P(0 < X < Y) = 1$ implies for $x > 0$, $h(x) = \int_x^\infty y f_Y(y)\, dy \geq g(x)$. Hence, by (6)

$$E\left(\frac{X}{Y}\right) \geq E\left(\frac{X}{h(X)}\right). \tag{7}$$

Now notice that by the Fundamental Theorem of Calculus $h'(x) = -x f_Y(x) < 0$ for all $x > 0$. That is, $h(x)$ is non-increasing, which implies $Cov(X, 1/h(X)) \geq 0$. It follows that $E\left(\frac{X}{h(X)}\right) \geq E(X)E\left(\frac{1}{h(X)}\right)$. Also, by Jensen's inequality

$$E\left(\frac{1}{h(X)}\right) \geq \frac{1}{E(h(X))}.$$

Therefore,

$$E\left(\frac{X}{h(X)}\right) \geq \frac{E(X)}{E(h(X))} \geq \frac{E(X)}{E(Y)} \tag{8}$$

since $h(x) \leq h(0) = E(Y)$.

Thus, by (7) and (8), the lemma follows.

Using this result we conclude that a sufficient condition for (4) is that $\sum_{i=1}^{j} X_{i:n}$ and $\sum_{i=1}^{n} X_{i:n}$ are negative dependent variables for each j. However, it is not easy to find such an example.

An alternative approach is described in the following lemma.

Lemma 2 *Let X and Y have finite means and be such that $P(0 < X < Y) = 1$. Define a real valued function $g(x) = E(Y|X = x)$. Consider the following two conditions:*

(i) $g(x)$ is non-decreasing and is such that $x/g(x)$ is non-increasing.
(ii) $g(x)$ is non-increasing.

If either (i) or (ii) holds then

$$E\left(\frac{X}{Y}\right) \geq \frac{E(X)}{E(Y)}. \tag{9}$$

Proof First note that $E\left(\frac{X}{Y}\right) = E\left(E\left(\frac{X}{Y}|X\right)\right) = E\left(XE\left(\frac{1}{Y}|X\right)\right)$.

By Jensen's inequality, $E\left(\frac{1}{Y}|X\right) \geq \frac{1}{E(Y|X)}$. Define for $x > 0$, $g(x) = E(Y|X = x)$. With this notation we have

$$E\left(\frac{X}{Y}\right) \geq E\left(\frac{X}{g(X)}\right). \tag{10}$$

Assume (i) holds. Then

$$0 \geq cov(g(X), X/g(X)) = E(X) - E(g(X))E(X/g(X)).$$

Hence

$$E(X/g(X)) \geq E(X)/E(g(X)) = E(X)/E(Y). \tag{11}$$

Therefore, by Eq. (10), expression (9) follows.

Assume (ii) holds. Then by Jensen's inequality we have

$$\begin{aligned} 0 \leq cov(X, 1/g(X)) &= E(X/g(X)) - E(X)E(1/g(X)) \\ &\leq E(X/g(X)) - E(X)/E(g(X)). \end{aligned}$$

Hence, inequality (11) holds and by (10), (9) holds.

Corollary 1 *Let X and Y have finite means and be such that $P(0 < X < Y) = 1$. If for any $y > x$, $f_{Y|X}(y|x)$ is a non-increasing function in x, then*

$$E\left(\frac{X}{Y}\right) \geq \frac{E(X)}{E(Y)}. \tag{12}$$

Proof Let $g(x) = E(Y|X = x)$. Suppose $0 < x_1 < x_2$. Since $f_{Y|X}(y|x)$ is non-increasing in x, we have

$$\begin{aligned} g(x_1) &= \int_{x_1}^{\infty} y f_{Y|X}(y|x_1)dy \geq \int_{x_2}^{\infty} y f_{Y|X}(y|x_1)dy \\ &\geq \int_{x_2}^{\infty} y f_{Y|X}(y|x_2)dy = g(x_2). \end{aligned}$$

That is, $g(x)$ is a non-increasing function. Therefore, by condition (ii) of Lemma 2, the result follows.

In order to use Lemma 2 to verify the bias of a Lorenz curve we will need to identify a parent distribution for the X_i's for which condition (i) or condition (ii) of the lemma is satisfied for $X = \sum_{i=1}^{j} X_{i:n}$ and $Y = \sum_{i=1}^{n} X_{i:n}$ for $j = 1, 2, \ldots, n-1$. An example in the case in which $n = 2$ may be developed as follows.

Example 1 Let $F(x)$ be a cdf with support on $(0, \infty)$ with density $f(x)$ such that $\mu_F = \int_0^{\infty} u f(u)du < \infty$. Let $0 < Y_1 < Y_2$ be the corresponding order statistics of a random sample of size two from $F(x)$. Notice that

$$E(Y_2|Y_1 = y) = y + \frac{\int_y^{\infty} \bar{F}(u)du}{\bar{F}(y)}, \quad y > 0. \tag{13}$$

Hence, $E(Y_1 + Y_2|Y_1 = y) = 2y + B(y)$, where $B(y) = \frac{\int_y^{\infty} \bar{F}(u)du}{\bar{F}(y)}$.

Therefore, using notation of Lemma 2 and letting $X = Y_1$ and $Y = Y_1 + Y_2$, $g(x) = 2x + B(x)$.

Define $\bar{F}(x) = \left(\exp\{-\sqrt{\log x}\}\right)/x$, $x \geq 1$. Then

$$\int_x^{\infty} \bar{F}(y)dy = \int_x^{\infty} \sqrt{\log y}\frac{\exp\{-\sqrt{\log y}\}}{y\sqrt{\log y}}dy. \tag{14}$$

Using integration by parts, with $u = \sqrt{\log y}$ and $dv = \frac{\exp\{-\sqrt{\log y}\}}{y\sqrt{\log y}}$, the right-hand side of (14) equals

$$2\left(\sqrt{\log x} + 1\right)\exp\{-\sqrt{\log x}\}.$$

Therefore,

$$B(x) = 2x\left(\sqrt{\log x} + 1\right). \tag{15}$$

It follows that $B(x)$ is non-decreasing, besides $B(x)/x = 2\left(\sqrt{\log x} + 1\right)$ is also non-decreasing.

Then $g(x)/x = 2 + B(x)/x$ is non-decreasing. That is, $x/g(x)$ is non-increasing and $g(x)$ is non-decreasing. Therefore, by (i) of Lemma 2, we have

$$E\left\{Y_1/\left(Y_1 + Y_2\right)\right\} \geq E\left\{Y_1\right\}/E\left\{\left(Y_1 + Y_2\right)\right\}.$$

7 Related Inequality Indices

There are several popular inequality indices that are intimately related to the Lorenz curve. Three examples are:

- The Gini index, G, which is defined to be twice the area between the Lorenz curve and the egalitarian line, $L^*(u) = u$.
- The Pietra index, P, which is the maximum vertical distance between the Lorenz curve and the egalitarian line.
- The Amato index, A, which is the length of the Lorenz curve.

Since sample Lorenz curves are typically positively biased, it is natural to expect that sample versions, G_n, P_n and A_n, of these inequality indices will be negatively biased. In this paper we will focus attention on the Gini index. A more detailed study of the bias of various sample inequality indices (including the Gini index) will appear in a separate report.

In the case of the Gini index, the issue is complicated by the fact that there are two distinct frequently used versions of the sample Gini index:

$$G_n = \frac{\sum_{i=1}^{n}(2i - n - 1)X_{i:n}}{\sum_{i=1}^{n}(n-1)X_{i:n}}$$

and

$$G_n' = \frac{\sum_{i=1}^{n}(2i - n - 1)X_{i:n}}{\sum_{i=1}^{n} nX_{i:n}} = \frac{n-1}{n}G_n.$$

The second version G_n' actually corresponds to the Gini index of the sample Lorenz curve. We would consequently expect that the sample Gini index, G_n', will be negatively biased. Thus we would expect to have

$$E(G_n') \leq G = \frac{E(X_{2:2}) - E(X_{1:2})}{E(X_{2:2}) + E(X_{1:2})},$$

i.e., that

$$E\left(\frac{\sum_{i=1}^{n}(2i - n - 1)X_{i:n}}{\sum_{i=1}^{n} nX_{i:n}}\right) \leq \frac{E(X_{2:2}) - E(X_{1:2})}{E(X_{2:2}) + E(X_{1:2})}.$$

In the exponential (1) case we have, using Basu's lemma,

$$E(G'_n) = \frac{\sum_{i=1}^{n}(2i - n - 1)E(X_{i:n})}{n^2}$$

and

$$G = \frac{E(X_{2:2}) - E(X_{1:2})}{E(X_{2:2}) + E(X_{1:2})} = \frac{1}{2}.$$

Substituting the well-known expressions for the expectations of exponential order statistics and simplifying, it may be verified that

$$E(G'_n) = \left(\frac{n-1}{n}\right)\frac{1}{2},$$

which indeed is less than $1/2$.

Note that in this exponential case we have:

$$E(G_n) = \frac{n}{n-1}E(G'_n) = \frac{n}{n-1}\left(\frac{n-1}{n}\right)\frac{1}{2} = \frac{1}{2},$$

i.e., G_n is unbiased. Note that, since G_n is not the Gini index of the sample Lorenz curve, we did not have strong justification for expecting it to be negatively biased in most cases.

In contrast, we do have reason to expect that G'_n will often be negatively biased. Just as was the case for the bias of the sample Lorenz curve, consideration of samples of size 2 from a two point distribution will be instructive.

Suppose that X_1 and X_2 are i.i.d. with $P(X_i = 1) = p$ and $P(X_i = c) = 1 - p$, where $c > 1$. We have in this case the sample Gini index, G'_2, given by

$$G'_2 = \frac{1}{2}\frac{X_{2:2} - X_{1:2}}{X_{2:2} + X_{1:2}}$$

and the population Gini index given by

$$G = \frac{E\,(X_{2:2}) - E\,(X_{1:2})}{E\,(X_{2:2}) + E\,(X_{1:2})}.$$

For this sample we have:

Probability\Variable	$X_{1:2}$	$X_{2:2}$	G_2	G'_2
p^2	1	1	0	0
$2p(1-p)$	1	c	$\frac{c-1}{c+1}$	$\frac{c-1}{2(c+1)}$
$(1-p)^2$	c	c	0	0

Consequently

$$G - E(G_2) = \frac{(c-1)2p(1-p)}{2[p+c(1-p)]} - 2p(1-p)\frac{c-1}{c+1} = \frac{p(1-p)(c-1)^2}{[p+c(1-p)]\,(c+1)}\,[2p-1]\,.$$

The sign of the difference $G - E(G_2)$ thus is determined by the value of p. It is positive if $p > 0.5$, negative if $p < 0.5$ and equal to 0 if $p = 0.5$. However, if we consider the bias of G_2', we have

$$\begin{aligned} G - E(G_2') &= \frac{(c-1)2p(1-p)}{2[p+c(1-p)]} - p(1-p)\frac{c-1}{c+1} \\ &= \frac{p(1-p)(c-1)}{[p+c(1-p)]\,(c+1)}\,[1+(c-1)p] \\ &> 0 \end{aligned}$$

for every $p \in (0, 1)$. Thus the expected negative bias is present, for values of p for which the corresponding sample Lorenz curve is positively biased, and even for values of p for which the curve is negatively biased.

Acknowledgments We are grateful to an anonymous referee for helpful suggestions, including drawing our attention to the early paper by Pitman, which have resulted in an improved manuscript.

References

1. Basu, D. 1955. On statistics independent of a complete sufficient statistic. *Sankhya* 15: 377–380.
2. Gail, M.H., and J.L. Gastwirth. 1978. A scale-free goodness-of-fit test for the exponential distribution based on the Lorenz curve. *Journal of the American Statistical Association* 73: 787–793.
3. Gastwirth, J.L. 1971. A general definition of the Lorenz curve. *Econometrica* 39: 1037–1039.
4. Hardy, G.H., J.E. Littlewood, and G. Polya. 1952. *Inequalities*, 2nd ed. London: Cambridge University Press.
5. Lorenz, M.O. 1905. Methods of measuring the concentration of wealth. *Publication of the American Statistical Association* 9: 209–219
6. Marshall, A.W., I. Olkin, and B.C. Arnold. 2011. *Inequalities: Theory of majorization and its applications*, 2nd ed. Springer: New York.
7. Pitman, E.J.G. 1937. The "closest" estimates of statistical parameters. *Proceedings of the Cambridge Philological Society* 33: 212–222.
8. Strassen, V. 1965. The existence of probability measures with given marginals. *Annals of Mathematical Statistics* 36: 423–439.

Pitman Closest Estimators Based on Convex Linear Combinations of Two Contiguous Order Statistics

N. Balakrishnan, Katherine F. Davies, Jerome P. Keating and Robert L. Mason

Abstract Comparisons of best linear unbiased estimators with some other prominent estimators have been carried out over the last six decades since the ground breaking work of Lloyd [13]; see Arnold et al. [1] and David and Nagaraja [9] for elaborate details in this regard. Recently, Pitman closeness comparison of order statistics as estimators for population parameters, such as medians and quantiles, and their applications have been carried out by Balakrishnan et al. [3–5, 7]. In this paper, we discuss the Pitman closest estimators based on convex linear combinations of two contiguous order statistics, which sheds additional insight with regard to the estimation of the population median in the case of even sample sizes. We finally demonstrate the proposed method for the uniform, exponential, power function and Pareto distributions.

Keywords Order statistics · Pitman closeness · Probabilities of closeness · Convex linear estimator · Location-scale family

N. Balakrishnan (✉)
Department of Mathematics and Statistics, McMaster University,
Hamilton, ON L8S 4K1, Canada
e-mail: bala@mcmaster.ca

K.F. Davies
Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
e-mail: Katherine.Davies@UManitoba.CA

J.P. Keating
Department of Management Science and Statistics, University of Texas at San Antonio,
78249-0704 San Antonio, Texas, USA
e-mail: jerome.keating@utsa.edu

R.L. Mason
Southwest Research Institute, 78228-0510 San Antonio, Texas, USA
e-mail: robert.mason@swri.org

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_2

1 Introduction

The comparison of estimators under the Pitman closeness criterion has a long history since it was introduced by Pitman [17] and further discussed by Rao [18]. For estimation based on order statistics, Nagaraja [15] considered Pitman closeness of estimators and predictors for the two-parameter exponential distribution. In a similar light, Balakrishnan et al. [6] and Balakrishnan and Davies [2] considered Pitman comparison of estimators for the one-parameter exponential distribution based on Type-I and II censored samples, respectively. Recently, Balakrishnan et al. [3] carried out Pitman closeness comparisons between pairs of order statistics arising from a random sample of size n with regard to the estimation of population quantiles ξ_p. Specifically, with $X_1, \ldots, X_n$ denoting a random sample taken from a continuous population with probability density function (pdf) $f(x)$ and cumulative distribution function (cdf) $F(x)$, and $X_{1:n}, \ldots, X_{n:n}$ denoting the corresponding order statistics, Balakrishnan et al. [3] derived formulas for the comparison of any two contiguous order statistics as estimators of population quantiles.

It is well known that (see David and Nagaraja [9] and Arnold et al. [1])

$$F\left(X_{i:n}\right) = U_{i:n} \sim \mathscr{B}\left(i, n-i+1\right), \tag{1}$$

where $\mathscr{B}(\alpha, \beta)$ denotes a beta random variable with shape parameters α and β; here, $U_{i:n}$ denotes the ith order statistic from a sample of size n from the Uniform(0, 1) distribution. Mean ranks in quantile-quantile plots are based on the relation

$$\boldsymbol{E}\left[F\left(X_{i:n}\right)\right] = \frac{i}{n+1} = e_{i:n}.$$

Similarly, if $m_{i:n}$ denotes the median of the beta random variable in (1), it is referred to as the median rank.

Definition 1 An estimator $\hat{\theta}$ will be said to overestimate a parameter θ if

$$\Pr\left(\hat{\theta} > \theta\right) > \frac{1}{2}.$$

This definition of *overestimation* is in the sense that the estimator $\hat{\theta}$ more frequently overestimates θ than it underestimates θ, or equivalently, the median of the distribution of $\hat{\theta}$ is less than θ.

In this paper, we discuss the Pitman closest estimation based on a convex linear combination of two contiguous order statistics. We then demonstrate the established results with uniform, exponential, power function and Pareto distributions.

2 Narrowing Down the Choices Among Order Statistics

In some cases, one may want to improve on the choice of order statistics given by Balakrishnan et al. [3], which provides the probability that a given order statistic is Pitman-closer to a specific population quantile ξ_p than any other order statistic from the same sample. The natural question that arises in this regard is whether one can improve on the estimation of ξ_p by using a linear combination of two contiguous order statistics. In some cases, no improvement can be made (e.g., the sample median in odd sample sizes as an estimator of the population median of a symmetric distribution is the Pitman-closest linear equivariant estimator of $\xi_{0.50}$), as shown in Balakrishnan et al. [4]. If we restrict our attention to convex linear combinations of two order statistics, then we can reduce the number of pairs to be considered to produce a Pitman-closer estimator and the following two lemmas facilitate this. For this specific purpose, we therefore want to bracket ξ_p so that we find the largest order statistic that underestimates ξ_p and the smallest order statistic that overestimates ξ_p, in the sense of Definition 1.

Lemma 1 *Let $X_1, \ldots, X_n$ be a random sample from a continuous population and $X_{1:n}, \ldots, X_{n:n}$ be the corresponding order statistics. For $p \geq 1 - 2^{-1/n}$, let $m_{j:n}$ be the largest median rank less than p. Then, the largest order statistic that does not overestimate ξ_p is $X_{j:n}$.*

Proof If $p < 1 - 2^{-1/n} = m_{1:n}$, then

$$\Pr\left(X_{1:n} < \xi_p\right) = \Pr\left[F\left(X_{1:n}\right) < F\left(\xi_p\right)\right] = \Pr\left(U_{1:n} < p\right) < \Pr\left(U_{1:n} < m_{1:n}\right) = \frac{1}{2}.$$

Thus all order statistics overestimate ξ_p whenever $p < 1 - 2^{-1/n}$. Next, let us consider the case when $p \in [m_{1:n}, 1)$. Since

$$m_{1:n} < m_{2:n} < \cdots < m_{n:n} < m_{n+1:n} = 1,$$

the spacings between the median ranks form a partition of the interval which immediately implies that there exists a j such that

$$m_{j:n} \leq p < m_{j+1:n}.$$

Thus, $U_{j:n}$ is the largest order statistic that underestimates p and consequently $X_{j:n}$ is the largest order statistic that underestimates ξ_p.

Note that j in the previous lemma does not depend on the underlying continuous distribution function $F(x)$, but only on the medians of order statistics from the Uniform(0,1) distribution, which can be determined numerically.

Lemma 2 *Let $X_1, \ldots, X_n$ be a random sample from a continuous population and $X_{1:n}, \ldots, X_{n:n}$ be the corresponding order statistics. For $p \leq 2^{-1/n}$, let $m_{\ell:n}$ be the largest median rank less than p. Then, the smallest order statistic that overestimates ξ_p is $X_{\ell:n}$.*

Notice that all order statistics underestimate ξ_p whenever $p > 2^{-1/n} = m_{n:n}$. The proof of this lemma proceeds in a similar way to that of Lemma 1. Determination of $X_{j:n}$ or $X_{j+1:n}$ does not require knowledge of the underlying distribution $F(x)$, but only needs the solution for j as a function of n and p through the medians of the beta distributions. We can combine Lemmas 1 and 2 to form the following theorem.

Theorem 1 *Let $X_1, \ldots, X_n$ be a random sample from a continuous population with pdf $f(x)$ and cdf $F(x)$, and $X_{1:n}, \ldots, X_{n:n}$ be the corresponding order statistics. Then, there exists a largest order statistic $X_{j:n}$ that does not overestimate ξ_p and a smallest order statistic $X_{j+1:n}$ that overestimates ξ_p (in the sense of Definition 1) for $m_{1:n} \leq p < m_{n:n}$.*

3 Pitman Closeness Criterion

We now introduce the comparison criterion known as Pitman closeness or Pitman nearness.

Definition 2 Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be univariate estimators of a real-valued parameter θ based on a sample of size n. Then, Pitman Closeness (PC) is defined as

$$\boldsymbol{P}\left(\hat{\theta}_1, \hat{\theta}_2 | \theta, n\right) = \Pr\left(|\hat{\theta}_1 - \theta| < |\hat{\theta}_2 - \theta|\right).$$

Interested readers may refer to the monograph by Keating et al. [12] for pertinent details. The measure in Definition 2 quantifies the frequency with which one estimator is closer to the value of the parameter θ than a competing estimator; see, for example, [6, 10, 14, 16–18].

Definition 3 Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be univariate estimators of a real-valued parameter θ based on a sample of size n. Then, $\hat{\theta}_1$ is said to be Pitman-closer to θ, for a given value of θ, than $\hat{\theta}_2$ provided

$$\boldsymbol{P}\left(\hat{\theta}_1, \hat{\theta}_2 | \theta, n\right) \geq \boldsymbol{P}\left(\hat{\theta}_2, \hat{\theta}_1 | \theta, n\right).$$

Definition 4 The estimator $\hat{\theta}_1$ is said to be uniformly Pitman-closer than $\hat{\theta}_2$ if $\boldsymbol{P}(\hat{\theta}_1, \hat{\theta}_2 | \theta, n) \geq \boldsymbol{P}(\hat{\theta}_2, \hat{\theta}_1 | \theta, n)$ for all θ in the parameter space Θ, with strict inequality holding for at least one $\theta \in \Theta$. The estimator $\hat{\theta}_1$ is uniformly Pitman-closest among the estimators in a class $\mathscr{C}$ provided

$$\boldsymbol{P}(\hat{\theta}_1, \hat{\theta}_j | \theta, n) \geq \boldsymbol{P}(\hat{\theta}_j, \hat{\theta}_1 | \theta, n)$$

for all $\hat{\theta}_j$ in $\mathscr{C}$ and for all $\theta \in \Theta$, with strict inequality holding for at least one $\theta \in \Theta$.

Lemma 3 *Let $X_1, \ldots, X_n$ be a random sample from a continuous population with pdf $f(x)$ and cdf $F(x)$, and $X_{1:n}, \ldots, X_{n:n}$ be the corresponding order statistics.*

If $m_{1:n} < p$, then if j is such that $X_{j:n}$ is the largest order statistic that does not overestimate ξ_p, $X_{j:n}$ is Pitman-closer to ξ_p than any of $X_{1:n}, \ldots, X_{j-1:n}$.

Proof We have

$$\begin{aligned}
\boldsymbol{P}\left(X_{j:n}, X_{\ell:n}|\xi_p\right) &= \Pr\left[|X_{j:n} - \xi_p| < |X_{\ell:n} - \xi_p|\right] \\
&= \Pr\left[\left(X_{j:n} - \xi_p\right)^2 < \left(X_{\ell:n} - \xi_p\right)^2\right] \\
&= \Pr\left[X_{j:n}^2 - X_{\ell:n}^2 < 2\xi_p\left(X_{j:n} - X_{\ell:n}\right)\right] \\
&= \Pr\left[\left(X_{j:n} - X_{\ell:n}\right)\left(X_{j:n} + X_{\ell:n}\right) < 2\xi_p\left(X_{j:n} - X_{\ell:n}\right)\right] \\
&= \Pr\left[X_{j:n} + X_{\ell:n} < 2\xi_p\right] \\
&< \Pr\left[X_{\ell:n} < \xi_p\right] < 1/2.
\end{aligned}$$

Thus, it follows that $X_{j:n}$ is Pitman-closer to ξ_p than $X_{\ell:n}$ for all $\ell = 1, \ldots, j-1$.

In an analogous manner, we can establish the following lemma.

Lemma 4 *Let $X_1, \ldots, X_n$ be a random sample from a continuous population with pdf $f(x)$ and cdf $F(x)$, and $X_{1:n}, \ldots, X_{n:n}$ be the corresponding order statistics. If $p < m_{n:n}$, then if j is such that $X_{j:n}$ is the largest order statistic that does not overestimate ξ_p, $X_{j+1:n}$ is Pitman-closer to ξ_p than any of $X_{j+2:n}, \ldots, X_{n:n}$.*

Now, let $p \in (m_{1:n}, m_{n:n})$. Then, due to Lemmas 3 and 4, it is evident that there exists a largest integer j such that $\Pr\left(X_{j:n} < \xi_p\right) \leq 1/2$ and $\Pr\left(X_{j+1:n} < \xi_p\right) > 1/2$, which is formally stated in the following theorem.

Theorem 2 *Let $X_1, \ldots, X_n$ be a random sample from a continuous population with pdf $f(x)$ and cdf $F(x)$, and $X_{1:n}, \ldots, X_{n:n}$ be the corresponding order statistics. Then, there exists a largest order statistic $X_{j:n}$ such that $X_{j:n}$ is Pitman-closer to ξ_p than $X_{\ell:n}$ for $\ell = 1, \ldots, j-1$, and $X_{j+1:n}$ is Pitman-closer to ξ_p than $X_{\ell:n}$ for $\ell = j+2, \ldots, n$, when $m_{1:n} \leq p < m_{n:n}$.*

Consequently, in terms of comparisons of individual order statistics, the Pitman-closest one to ξ_p, for a given p, will depend on the comparison of $X_{j:n}$ and $X_{j+1:n}$. The better of these two in the sense of Pitman closeness will depend on the underlying distribution $F(x)$. For this reason, it will be reasonable to compare the largest order statistic that underestimates ξ_p with the smallest order statistic that overestimates ξ_p.

In fact, one can generalize the use of contiguous order statistics, $X_{j:n}$ and $X_{j+1:n}$, to any pair $X_{i:n}$ and $X_{k:n}$, where $1 \leq i \leq j$ and $j+1 \leq k \leq n$. These results imply that if we are to find a Pitman-closer estimator than any individual order statistic from a convex class based on two order statistics, then one order statistic must underestimate ξ_p and the other must overestimate ξ_p. Of course, a single order statistic may outperform any convex linear combination of all other order statistics.

4 Use of a Convex Class

Based on Theorem 2, we may consider some linear combination of these contiguous order statistics. The use of a convex linear combination, i.e.,

$$\hat{\xi}_p = wX_{j:n} + (1-w)X_{j+1:n},\ w \in [0,1],$$

produces a class of ordered estimators in the closed bounded interval $\left[X_{j:n}, X_{j+1:n}\right]$. Furthermore, in location-scale families, the individual order statistics are location invariant estimators of the location parameter and so convex linear combinations of order statistics are location invariant estimators as well. So, we wish to find a median unbiased estimator of ξ_p within the convex class given above. While the value of w, for which the convex linear combination has a median of ξ_p, may not be independent of the unknown parameters in the distributions of $X_{j:n}$ and $X_{j+1;n}$, certain special and important cases do exist in which the choice only depends on n and p. However, it should be kept in mind there is no certainty that this median unbiased convex linear combination of $X_{j:n}$ and $X_{j+1:n}$ will be the Pitman-closest median unbiased convex linear combination of any pair of order statistics $X_{i:n}$ and $X_{k:n}$, where $1 \le i \le j$ and $j+1 \le k \le n$.

In order to assess the median unbiased estimator within the class of convex linear combinations of two contiguous order statistics, we need the joint density of $X_{j:n}$ and $X_{j+1:n}$ given by (see Arnold et al. [1] and David and Nagaraja [9])

$$f(u,v) = \frac{n!}{(j-1)!(n-j-1)!}[F(u)]^{j-1}[1-F(v)]^{n-j-1}f(u)f(v),\ \text{if } u < v, \tag{2}$$

for $j = 1, \ldots, n-1$. Since we have reduced our consideration to just contiguous order statistics mentioned in the sense of overestimating and underestimating, we can now consider a new class of estimators based on them.

Definition 5 Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics from a random sample from $F(x)$, which is strictly monotone on the support of X. Let ξ_p be the pth quantile of the distribution. Let j be the largest integer for which $m_{j:n} \le p$. Define the class $\mathscr{Q}$ as the collection of all convex linear combinations of $X_{j:n}$ and $X_{j+1:n}$, i.e.,

$$\mathscr{Q} = \left\{\hat{\xi}_p(w) | \hat{\xi}_p(w) = wX_{j:n} + (1-w)X_{j+1:n},\ w \in [0,1]\right\}. \tag{3}$$

In general, determining the Pitman-closest estimator in the class $\mathscr{Q}$ can be difficult, and also can produce a random variable that depends on the unknown parameters of the distribution and consequently not an estimator. But, the determination of a best choice within the class is guaranteed for a location-scale family as shown below.

4.1 Location-Scale Family

Let us consider the location-scale family of distributions with the density function of X given by

$$f(x \mid \mu, \sigma) = (1/\sigma) g\left[(x-\mu)/\sigma\right], \tag{4}$$

where $g(z)$ is a continuous parameter-free density. The parameter space for these families is the upper half-plane $\Omega = \{-\infty < \mu < \infty, \sigma > 0\}$.

The cdf of X is

$$F(x \mid \mu, \sigma) = G\left[(x-\mu)/\sigma\right], \text{ where } G(t) = \int_{-\infty}^{t} g(u)du.$$

The $100p$ percentage point (or percentile) of the random variable X, denoted by ξ_p, is defined as $\xi_p = \inf\{x \in \mathbf{R} : F(x) \geq p\}$. The distribution function $G(t)$ is usually taken to be a parameter-free cdf. If the essential range, $\mathbf{R}$, of X is an open connected subset of $\mathbf{R}$, then ξ_p is unique for each $p \in (0, 1)$ with

$$\xi_p = \mu + G^{-1}(p)\sigma, \tag{5}$$

where $G^{-1}(\cdot)$ is the inverse function of $G(\cdot)$. One can see from (5) that within this family, percentiles are linear combinations of the parameters μ and σ. This family includes many well-known distributions such as normal, extreme-value, exponential, Laplace, Cauchy, uniform and logistic as members, but also includes several other distributions such as lognormal, log-uniform, inverse Gaussian, Pareto and Weibull through suitable transformations.

Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics from a random sample of size n from a location-scale parameter density $f(x)$ in (4). The estimation of location and scale parameters as well as percentiles have been discussed quite extensively based on order statistics; see, for example, Balakrishnan and Cohen [8]. First define $Z_{1:n}, \ldots, Z_{n:n}$ as

$$Z_{i:n} = \frac{X_{i:n} - \mu}{\sigma}, \text{ for } i = 1, \ldots, n. \tag{6}$$

Theorem 3 *Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics from a random sample from a continuous location-scale parameter density $f(x)$ in (4). Let $Z_{1:n}, \ldots, Z_{n:n}$ be the corresponding order statistics as defined in (6). Then, we have*

$$\boldsymbol{P}\left(X_{j:n}, X_{\ell:n} | \xi_p\right) = \boldsymbol{P}\left(Z_{j:n}, Z_{\ell:n} | G^{-1}(p)\right).$$

Thus, within the class of location-scale families of distributions, the Pitman closeness of any two order statistics in the estimation of ξ_p is independent of the unknown parameters.

Theorem 4 *Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics from a random sample from $F(x)$, which is strictly monotone on the support of X. Let $p \in (m_{1:n}, m_{n:n})$. With*

$w \in [0, 1]$, let us consider the class $\mathcal{Q}$ in (3) of estimators ξ_p. Then, $\Pr\left(\hat{\xi}_p(w) < \xi_p\right)$ *is a continuous non-increasing function of w.*

Proof For $w \in [0, 1]$, let us define

$$Q_{n,p}(w) = \Pr\left(\hat{\xi}_p(w) < \xi_p\right).$$

Then, by Definition 5,

$$Q_{n,p}(1) = \Pr\left(X_{j:n} < \xi_p\right) > \frac{1}{2}, \quad Q_{n,p}(0) = \Pr\left(X_{j+1:n} < \xi_p\right) < \frac{1}{2}.$$

Since $F(x)$ is continuous, $Q_{n,p}(w)$ is continuous and so there exists a value $0 \leq w_0 \leq 1$ such that $Q_{n,p}(w_0) = 1/2$. For $0 \leq w_1 < w_2 \leq 1$, we have

$$\begin{gathered} w_1 X_{j:n} + (1 - w_1) X_{j+1:n} < w_2 X_{j:n} + (1 - w_2) X_{j+1:n}, \\ \Pr\left\{w_1 X_{j:n} + (1 - w_1) X_{j+1:n} < x\right\} > \Pr\left\{w_2 X_{j:n} + (1 - w_2) X_{j+1:n} < x\right\}, \\ Q_{n,p}(w_1) > Q_{n,p}(w_2). \end{gathered}$$

Therefore, $Q_{n,p}(w)$ is a continuous non-increasing function of w.

Corollary 1 *Under the conditions of Theorem 4, a median unbiased estimator of ξ_p exists within the considered convex class. If $F(x)$ is strictly increasing over its support, the median unbiased estimator is unique within this class.*

Proof Since $Q_{n,p}(1) > \frac{1}{2}$ and $Q_{n,p}(0) \leq \frac{1}{2}$, by the continuity of $Q_{n,p}(w)$, there exists a value $0 \leq w_0 \leq 1$ such that $Q_{n,p}(w_0) = 1/2$. Further, if $F(x)$ is strictly increasing over its support, $Q_{n,p}(w)$ will be strictly decreasing on [0,1) and so the solution w_0 will be unique.

Corollary 2 *Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics from a random sample from $F(x)$, which is strictly monotonically increasing on the support of X. Let $F(x)$ be a member of the location-scale family of distributions, and $Z_{1:n}, \ldots, Z_{n:n}$ be as defined in (6). Let $p \in (m_{1:n}, m_{n:n})$. With $w \in [0, 1)$, let us consider the class $\mathcal{Q}$ in (3) for the estimation of ξ_p. Then, there exists a unique Pitman-closest estimator of ξ_p within $\mathcal{Q}$.*

Proof The proof follows directly from Corollary 1. Within an ordered class of estimators of some parameter, say θ, the median unbiased estimator within the class will be the Pitman-closest estimator of θ. We are guaranteed that the class $\mathcal{Q}$, by its very construction, produces some estimators that overestimate ξ_p and some that underestimate ξ_p such that

$$\begin{aligned} Q_{n,p}(w) &= \Pr\left(\hat{\xi}_p(w) < \xi_p\right) \\ &= \Pr\left(wX_{j:n} + (1 - w)X_{j+1:n} < \xi_p\right) \\ &= \Pr\left(wZ_{j:n} + (1 - w)Z_{j+1:n} < G^{-1}(p)\right). \end{aligned}$$

The value of w such that $Q_{n,p}(w) = 1/2$ is unique does not involve the unknown parameters, and is only a function of n, p and the corresponding value of j. This choice of w produces an estimator that is median unbiased and is therefore Pitman-closer than all other estimators within $\mathscr{Q}$ since the class is completely ordered, and is therefore the Pitman-closest estimator of ξ_p in $\mathscr{Q}$. While this result guarantees the existence and uniqueness of a median unbiased estimator in $\mathscr{Q}$, it would be good to have a Rao-Blackwell type result that would provide a method for its construction.

4.2 Transformation

In this case, consider the following transformation $R = X_{j:n}$ and $T = X_{j+1:n} - X_{j:n}$, for $1 \leq j \leq n-1$. It follows that

- R is a location invariant statistic and as noted before $Z_{j:n} = (X_{j:n} - \mu)/\sigma$ has a parameter-free distribution;
- $T = X_{j+1:n} - X_{j:n}$ is a scale invariant statistic and T/σ is a pivotal quantity for σ;
- $Z_{j:n}/T$ is a pivotal quantity for μ;
- $\left(R - \xi_p\right)/T$ is a pivotal quantity for ξ_p with a distribution that depends on $G^{-1}(p)$ and the sample size n.

Under this transformation, one can rewrite any convex class for which $X_{j:n}$ and $X_{j+1:n}$, respectively, underestimate and overestimate ξ_p, in the following way:

$$\begin{aligned}\mathscr{Q} &= \left\{\hat{\xi}_p(w) | \hat{\xi}_p(w) = X_{j+1:n} - w\left(X_{j+1:n} - X_{j:n}\right),\ w \in (0,1]\right\} \\ &= \left\{\hat{\xi}_p(c) | \hat{\xi}_p(c) = X_{j:n} + c\left(X_{j+1:n} - X_{j:n}\right),\ c = 1 - w \in (0,1]\right\}.\end{aligned}$$

One can now derive the median unbiased estimator within $\mathscr{Q}$ according to the following theorem.

Theorem 5 *In the context of Lemma 3, consider two order statistics $X_{j:n}$ and $X_{j+1:n}$ in a location-scale family where $X_{j:n}$ underestimates ξ_p and $X_{j+1:n}$ overestimates ξ_p (in the sense of Definition 1). Then, a median unbiased estimator within $\mathscr{Q}$ is given by*

$$\hat{x}_p = X_{j:n} + \boldsymbol{M}_{(0,1)}\left(\frac{G^{-1}(p) - R}{T}\right) T, \tag{7}$$

where $T = X_{j+1:n} - X_{j:n}$ and $\boldsymbol{M}_{(0,1)}(U)$ denotes the median of the random variable U when $\mu = 0$ and $\sigma = 1$.

Proof We have

$$\begin{aligned}\Pr\left\{\hat{\xi}_p(c) < \xi_p\right\} &= \Pr\left\{X_{j:n} + c\left(X_{j+1:n} - X_{j:n}\right) < \xi_p\right\} \\ &= \Pr\left\{c < \tfrac{\xi_p - X_{j:n}}{X_{j+1:n} - X_{j:n}}\right\} \\ &= \Pr\left\{c < \tfrac{G^{-1}(p) - Z_{j:n}}{Z_{j+1:n} - Z_{j:n}}\right\}.\end{aligned}$$

If the estimator is median unbiased, then the probability content of the interval is 1/2, and so

$$c = \boldsymbol{M}_{(0,1)}\left(\frac{G^{-1}(p) - R}{T}\right).$$

It follows that the median unbiased estimator within $\mathscr{Q}$ is the estimator in (7). Solving for c will require numerical methods and the fact that c must be in the interval (0, 1] would facilitate the use of the secant method. Incidentally, a naive and nonparametric estimate of c can be obtained as

$$\hat{c} = \frac{p - m_{j:n}}{m_{j+1:n} - m_{j:n}}.$$

4.3 Examples

Uniform distribution

In the case of the Uniform(0,1) distribution, consider $\Pr\left[R + cT < \xi_p\right]$, i.e., $\Pr\left[R + cT < p\right]$, which we need to set to 1/2. First, we note that we can express

$$R + cT = X_{j:n} + c(X_{j+1:n} - X_{j:n}) = X_{j+1:n}\left(\frac{X_{j:n}}{X_{j+1:n}} + c\left(1 - \frac{X_{j:n}}{X_{j+1:n}}\right)\right) = VW,$$

where $U = \frac{X_{j:n}}{X_{j+1:n}}$, $V = X_{j+1:n}$ and $W = U + c(1 - U)$. It is known that $U \sim Beta(j, 1)$ and $V \sim Beta(j + 1, n - j)$ and that the two random variables are independent; see Arnold et al. [1]. Using the distribution of U, it can be shown that the pdf of W is given by

$$f_W(w) = \frac{j}{(1 - c)^j}(w - c)^{j-1} \text{ if } w \in (c, 1).$$

We then have

$$\Pr\left[R + cT < p\right] = \Pr(VW < p) = \Pr\left(V < \frac{p}{W}\right) = \int_c^1 \Pr\left(V < \frac{p}{w}\right) f_W(w)dw,$$

where

$$\Pr\left(V < \frac{p}{w}\right) = \begin{cases} I_{\frac{p}{w}}(j + 1, n - j) & \text{if } \frac{p}{w} < 1 \\ 1 & \text{if } \frac{p}{w} \geq 1, \end{cases}$$

and $I_q(a,b)$ is the incomplete beta ratio defined by $I_q(a,b) = \frac{1}{B(a,b)} \int_0^q t^{a-1}(1-t)^{b-1}dt$ and $B(a,b)$ is the complete beta function defined by $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Thus, we get

$$\Pr(VW < p) = \begin{cases} \int_c^1 I_{\frac{p}{w}}(j+1, n-j) f_W(w)dw & \text{if } \frac{p}{c} < 1 \\ \int_p^1 I_{\frac{p}{w}}(j+1, n-j) f_W(w)dw + \left(\frac{p-c}{1-c}\right)^j & \text{if } \frac{p}{c} \geq 1. \end{cases} \tag{8}$$

Equation (8) can be solved for various p to find c when p is away from the bounds 0 and 1. For p close to 0, c is found such that $\Pr(cX_{1:n} \leq \xi_p) = 1/2$, and similarly, for p close to 1, c is found such that $\Pr(cX_{n:n} \leq \xi_p) = 1/2$. However, in order to use $cX_{n:n}$, we need to need to check the validity of the determined c since it is possible that the estimator may exist outside the support. Since we choose c such that $\Pr(X_{n:n} \leq \frac{\xi_p}{c})$ = 1/2, i.e., $\Pr(X_{n:n} \leq \frac{p}{c})$ = 1/2, the desired c turns out to be $c = 2^{1/n}p$. Now, let $W = 2^{1/n}pX_{n:n}$. In this case, we find

$$\Pr(W \leq 1) = \Pr(2^{1/n}pX_{n:n} \leq 1) = \Pr\left(X_{n:n} \leq \frac{1}{2^{1/n}p}\right) = \left(\frac{1}{2}\right)\left(\frac{1}{p}\right)^n.$$

This is a valid probability if and only if $\left(\frac{1}{p}\right)^n \leq 2$, i.e., $-\log(p) \leq \frac{1}{n}\log(2)$. So, the corresponding entries in Table 1 have been checked accordingly.

Exponential distribution

We again consider $\Pr(R + cT < \xi_p)$, which we can rewrite as

$$\Pr(R + cT < \xi_p) = \Pr[X_{j:n} + c(X_{j+1:n} - X_{j:n}) < \xi_p].$$

We then have this probability as

$$\begin{aligned} \Pr(R + cT < \xi_p) &= \Pr\left((n-j)(X_{j+1:n} - X_{j:n}) < (n-j)\left(\tfrac{\xi_p - X_{j:n}}{c}\right)\right) \\ &= \int_0^\infty \Pr\left(S_{j+1} < (n-j)\left(\tfrac{\xi_p - x_j}{c}\right)\right) f_{X_{j:n}}(x_j)dx_j \\ &= \int_0^{\xi_p} \left(1 - e^{-\frac{(n-j)(\xi_p - x_j)}{c}}\right) \tfrac{n!}{(j-1)!(n-j)!}(1-e^{-x_j})^{j-1}(e^{-x_j})^{n-j} \\ &\qquad \times e^{-x_j}dx_j \\ &= I_p(j, n-j+1) - e^{-\frac{(n-j)\xi_p}{c}} \int_0^{\xi_p} e^{\frac{(n-j)x_j}{c}} \tfrac{1}{B(j,n-j+1)}(1-e^{-x_j})^{j-1} \\ &\qquad \times (e^{-x_j})^{n-j}e^{-x_j}dx_j \\ &= I_p(j, n-j+1) - \tfrac{e^{-\frac{(n-j)\xi_p}{c}}}{B(j,n-j+1)} \textstyle\sum_{k=0}^{j-1}(-1)^k\binom{j-1}{k} \int_{1-p}^1 u^{k+n-j-\frac{n-j}{c}}du, \end{aligned}$$

where S_{j+1} is the normalized spacing defined as $S_{j+1} = (n-j)(X_{j+1:n} - X_{j:n}) \sim Exp(1)$, and since it is known to be independent of $X_{j:n}$ (see Arnold et al. [1]), we have

$$\Pr(S_{j+1} < s) = \begin{cases} 0 & \text{if } s \leq 0, \\ 1 - e^{-s} & \text{if } s > 0. \end{cases}$$

Table 1 Values of j and c for the uniform distribution when $n = 10$ for various choices of p

p	j	c	p	j	c	p	j	c
0.01	1	0.1493	0.34	3	0.8336	0.67	7	0.2728
0.02	1	0.2987	0.35	3	0.9430	0.68	7	0.3763
0.03	1	0.4480	0.36	4	0.0528	0.69	7	0.4774
0.04	1	0.5973	0.37	4	0.1583	0.70	7	0.5770
0.05	1	0.7466	0.38	4	0.2612	0.71	7	0.6762
0.06	1	0.8960	0.39	4	0.3622	0.72	7	0.7765
0.07	1	0.0320	0.40	4	0.4621	0.73	7	0.8790
0.08	1	0.1312	0.41	4	0.5622	0.74	7	0.9847
0.09	1	0.2238	0.42	4	0.6635	0.75	8	0.0982
0.10	1	0.3169	0.43	4	0.7668	0.76	8	0.2094
0.11	1	0.4142	0.44	4	0.8726	0.77	8	0.3170
0.12	1	0.5163	0.45	4	0.9813	0.78	8	0.4211
0.13	1	0.6234	0.46	5	0.0899	0.79	8	0.5221
0.14	1	0.7354	0.47	5	0.1954	0.80	8	0.6206
0.15	1	0.8517	0.48	5	0.2986	0.81	8	0.7180
0.16	1	0.9722	0.49	5	0.3998	0.82	8	0.8166
0.17	2	0.0817	0.50	5	0.0000	0.83	8	0.9183
0.18	2	0.1834	0.51	5	0.6002	0.84	9	0.0278
0.19	2	0.2820	0.52	5	0.7014	0.85	9	0.1483
0.20	2	0.3794	0.53	5	0.8046	0.86	9	0.2646
0.21	2	0.4779	0.54	5	0.9101	0.87	9	0.3766
0.22	2	0.5789	0.55	6	0.0187	0.88	9	0.4837
0.23	2	0.6830	0.56	6	0.1274	0.89	9	0.5858
0.24	2	0.7906	0.57	6	0.2332	0.90	9	0.6831
0.25	2	0.9018	0.58	6	0.3365	0.91	9	0.7762
0.26	3	0.0153	0.59	6	0.4378	0.92	9	0.8688
0.27	3	0.1210	0.60	6	0.5379	0.93	9	0.9680
0.28	3	0.2235	0.61	6	0.6378	0.94	10	1.0075
0.29	3	0.3238	0.62	6	0.7388	0.95	10	1.0182
0.30	3	0.4230	0.63	6	0.8417	0.96	10	1.0289
0.31	3	0.5226	0.64	6	0.9472	0.97	10	1.0396
0.32	3	0.6237	0.65	7	0.0570	0.98	10	1.0503
0.33	3	0.7272	0.66	7	0.1664	0.99	10	1.0611

Furthermore, we have

$$\int_{1-p}^{1} u^{k+n-j-\frac{n-j}{c}} du$$

$$= \begin{cases} \frac{1}{k+n-j+1-\frac{n-j}{c}} \left(1-(1-p)^{k+n+j+1-\frac{n-j}{c}}\right) & \text{if } \frac{n-j}{c} - (k+n-j) \neq 1 \\ -log(1-p) & \text{if } \frac{n-j}{c} - (k+n-j) = 1. \end{cases}$$

Values of c and j for various choices of p were numerically determined in this case and are presented in Table 2.

Pareto and power function distributions
Let $X \sim$ Power Function(θ), i.e.,

$$f_X(x) = \theta x^{\theta-1} \text{if } 0 < x < 1 \tag{9}$$

for $\theta > 0$. In this case, by proceeding as in the uniform case, it can be shown that

$$\Pr(VW < \xi_p) = \Pr\left(V < \frac{\xi_p}{W}\right) = \int_w \Pr\left(V < \frac{\xi_p}{w}\right) f_W(w)dw, \tag{10}$$

where $F_V(v) = \Pr(V \leq v) = I_{v^\theta}(j+1, n-j)$ and

$$f_W(w) = \theta j \frac{(w-c)^{\theta j-1}}{(1-c)^{\theta j}} \text{if } c < w < 1. \tag{11}$$

Next, let us consider $X \sim$ Pareto(θ), i.e.,

$$f_X(x) = \nu x^{-\nu-1} \text{if } x \geq 1$$

for $\nu > 0$. Then, the joint density of $X_{j:n}$ and $X_{j+1:n}$ is obtained from (2) as

$$f(x_j, x_{j+1}) = \frac{n!}{(j-1)!(n-j-1)!}(1-x_j^{-\nu})^{j-1}(x_{j+1}^{-\nu})^{n-j-1}\nu x_j^{-\nu-1}\nu x_{j+1}^{-\nu-1},$$
$$\text{if } 1 < x_j < x_{j+1} < \infty.$$

Let $U = \frac{X_{j+1:n}}{X_{j:n}}$ and $V = X_{j:n}$. In this case, it is known that U and V are independent with $U \sim$ Pareto$((n-j)\nu)$ and the pdf of V is

$$f_V(v) = \frac{n!}{(j-1)!(n-j)!}(1-v^{-\nu})^{j-1}(v^{-\nu})^{n-j}\nu v^{-\nu-1} \text{if } v \geq 1; \tag{12}$$

Table 2 Values of j and c for the exponential distribution when $n = 10$ for various choices of p

p	j	c	p	j	c	p	j	c
0.01	1	0.1450	0.34	3	0.8183	0.67	7	0.2167
0.02	1	0.2915	0.35	3	0.9373	0.68	7	0.3079
0.03	1	0.4394	0.36	4	0.0455	0.69	7	0.4035
0.04	1	0.5889	0.37	4	0.1382	0.70	7	0.5047
0.05	1	0.7400	0.38	4	0.2314	0.71	7	0.6123
0.06	1	0.8927	0.39	4	0.3264	0.72	7	0.7270
0.07	1	0.0288	0.40	4	0.4243	0.73	7	0.8495
0.08	1	0.1185	0.41	4	0.5258	0.74	7	0.9806
0.09	1	0.2057	0.42	4	0.6316	0.75	8	0.0669
0.10	1	0.2967	0.43	4	0.7422	0.76	8	0.1471
0.11	1	0.3935	0.44	4	0.8579	0.77	8	0.2308
0.12	1	0.4967	0.45	4	0.9789	0.78	8	0.3196
0.13	1	0.6062	0.46	5	0.0757	0.79	8	0.4149
0.14	1	0.7219	0.47	5	0.1673	0.80	8	0.5181
0.15	1	0.8434	0.48	5	0.2603	0.81	8	0.6304
0.16	1	0.9705	0.49	5	0.3557	0.82	8	0.7530
0.17	2	0.0730	0.50	5	0.4545	0.83	8	0.8873
0.18	2	0.1655	0.51	5	0.5575	0.84	9	0.0140
0.19	2	0.2580	0.52	5	0.6653	0.85	9	0.0775
0.20	2	0.3528	0.53	5	0.7784	0.86	9	0.1451
0.21	2	0.4513	0.54	5	0.8970	0.87	9	0.2188
0.22	2	0.5543	0.55	6	0.0150	0.88	9	0.3011
0.23	2	0.6624	0.56	6	0.1039	0.89	9	0.3944
0.24	2	0.7757	0.57	6	0.1940	0.90	9	0.5013
0.25	2	0.8942	0.58	6	0.2864	0.91	9	0.6252
0.26	3	0.0134	0.59	6	0.3820	0.92	9	0.7701
0.27	3	0.1070	0.60	6	0.4819	0.93	9	0.9416
0.28	3	0.2002	0.61	6	0.5867	0.94	10	1.0406
0.29	3	0.2944	0.62	6	0.6972	0.95	10	1.1081
0.30	3	0.3911	0.63	6	0.8138	0.96	10	1.1906
0.31	3	0.4912	0.64	6	0.9371	0.97	10	1.2970
0.32	3	0.5955	0.65	7	0.0432	0.98	10	1.4470
0.33	3	0.7044	0.66	7	0.1289	0.99	10	1.7034

see Arnold et al. [1]. So, the probability of interest is

$$\begin{aligned}\Pr[(1-c)X_{j:n} + cX_{j+1:n} \le \xi_p] &= \Pr\left[X_{j:n}\left\{(1-c) + c\left(\frac{X_{j+1:n}}{X_{j:n}}\right)\right\} \le \xi_p\right] \\ &= \Pr\left[V\left\{(1-c) + cU\right\} \le \xi_p\right].\end{aligned}$$

The pdf of $W = (1-c) + cU$ is

$$f_W(w) = (n-j)\nu \frac{(w-(1-c))^{-\nu(n-j)-1}}{c^{-\nu(n-j)}} \text{if } 1 < w < \infty. \tag{13}$$

Consequently, the probability becomes

$$\Pr(VW < \xi_p) = \Pr\left(V < \frac{\xi_p}{W}\right) = \int_w \Pr\left(V < \frac{\xi_p}{w}\right) f_W(w)dw,$$

where $F_V(v) = \Pr(V \le v) = I_{1-v^{-\nu}}(j, n-j+1)$ from (12) and $f_W(w)$ is as given in (13).

5 Some Heuristic Attempts

One may be tempted to estimate the value of j in the preceding discussions without inspecting of the underlying median ranks. It certainly seems plausible to attempt to estimate j by the largest integer less than or equal to $(n+1)p$. Such approximations can lead to order statistics that are upper and lower bounds, just as $X_{j:n}$ and $X_{j+1:n}$ were in the preceding discussion. However, the order statistics are no longer contiguous. All the methodology developed in the preceding sections can be reapplied here except that the order statistics, used to form the convex class, are no longer contiguous.

Lemma 5 *Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics from a random sample from $F(x)$, which is strictly monotone on the support of X. Let p be a real-number in the interval (0, 1) such that $p \in \left(\frac{1}{n+1}, \frac{n}{n+1}\right)$, and ξ_p be the pth quantile of $F(x)$. Then, there exists a $j \in \{1, \ldots, n\}$ such that $\Pr\left(X_{j:n} < \xi_p\right) > \frac{1}{2}$ and $\Pr\left(X_{n-j+1:n} < \xi_p\right) < \frac{1}{2}$.*

Proof Define j as $j = [\![(n+1)p]\!]$, where $[\![x]\!]$ denotes the largest integer less than or equal to x. Observe that

$$\Pr\left(X_{j:n} < \xi_p\right) = \Pr\left[F\left(X_{j:n}\right) < F\left(\xi_p\right)\right] = \Pr\left(U_{j:n} < p\right),$$

where $U_{j:n}$ is the jth order statistic from a random sample of size n from the Uniform(0,1) distribution. As mentioned in Sect. 1, $U_{j:n} \sim \mathscr{B}(j, n-j+1)$, where $\mathscr{B}(\alpha, \beta)$ denotes a beta distribution with shape parameters α and β. Without loss of generality, we assume that $p \le \frac{1}{2}$ and so $j \le \frac{n+1}{2}$. If $j < \frac{n+1}{2}$, then $U_{j:n}$ is unimodal and positively skewed and therefore satisfies the mode-median-mean inequality. Hence, it follows that

$$\Pr\left(U_{j:n} < p\right) \ge \Pr\left(U_{j:n} < \frac{j}{n+1}\right) \ge \frac{1}{2}.$$

This means that the order statistic $X_{j:n}$ underestimates ξ_p. Notice that this result is nonparametric in the sense that it does not depend on the form of the distribution function $F(x)$, only that it be strictly monotone on its support.

Using symmetry arguments, we can then prove that

$$\Pr\left(X_{n-j+1:n} > \xi_p\right) \geq \frac{1}{2},$$

which means that $X_{n-j+1:n}$ overestimates ξ_p. Hence, the required result.

Lemma 6 *Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics from a random sample from $F(x)$, which is strictly monotone on the support of X. Let $p \in \left(\frac{1}{n+1}, \frac{n}{n+1}\right)$. With $w \in (0, 1)$, let us consider a class of estimators $\hat{\xi}_p(w) = wX_{j:n} + (1-w)X_{n-j+1:n}$ for the pth quantile ξ_p, where $j = [(n+1)p]$ and $j < n - j + 1$. Then, $\Pr\left(\hat{\xi}_p(w) < \xi_p\right)$ is a continuous increasing function of w.*

Proof With $j = [(n+1)p]$, let us define

$$Q_{n,p}(w) = \Pr\left(\hat{\xi}_p(w) < \xi_p\right). \tag{14}$$

By Lemma 5, we have $Q_{n,p}(1) = \Pr\left(X_{j:n} < \xi_p\right) > \frac{1}{2}$ and $Q_{n,p}(0) = \Pr(X_{n-j+1:n} < \xi_p) < \frac{1}{2}$. Since $F(x)$ is continuous, $Q_{n,p}(w)$ is continuous. Hence, for $0 < w_1 < w_2 < 1$, we have

$$\begin{gathered} w_1 X_{j:n} + (1-w_1) X_{n-j+1:n} < w_2 X_{j:n} + (1-w_2) X_{n-j+1:n}, \\ \Pr\left\{w_1 X_{j:n} + (1-w_1) X_{n-j+1:n} < x\right\} < \Pr\left\{w_2 X_{j:n} + (1-w_2) X_{n-j+1:n} < x\right\}, \\ Q_{n,p}(w_1) < Q_{n,p}(w_2). \end{gathered}$$

Thus $Q_{n,p}(w)$ is a continuous increasing function, as desired.

Theorem 6 *Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics from a random sample from $F(x)$, which is strictly monotone on the support of X. Let $p \in \left(\frac{1}{n+1}, \frac{n}{n+1}\right)$. With $w \in (0, 1)$, let us consider a class of estimators $\hat{\xi}_p(w) = wX_{j:n} + (1-w)X_{n-j+1:n}$ for the pth quantile ξ_p, where $j = [(n+1)p]$. Then, there exists a unique value w_0 $(0 < w_0 < 1)$ such that $\Pr\left(\hat{\xi}_p(w_0) < \xi_p\right) = \frac{1}{2}$.*

6 Applications

In this section, we illustrate the results of the last section for the special cases of uniform and exponential distributions.

6.1 Uniform Distribution

Use of $|(n+1)p|$ for j

In order to evaluate $Q_{n,p}(w)$, we must develop an expression for the cdf of the convex linear combination of $X_{j:n}$ and $X_{n-j+1:n}$. Of special interest is the Uniform(0,1) distribution since in this case the subsequent estimator will be an L-estimator of ξ_p.

Suppose $U_{1:n}, \ldots, U_{n:n}$ are the order statistics from the uniform Uniform(0,1) distribution. Then, the joint density function of $U_{j:n}$ and $U_{n-j+1:n}$ is given by

$$f(u,v) = \frac{n!}{[(j-1)!]^2(n-2j)!} u^{j-1}(v-u)^{n-2j}(1-v)^{j-1} \text{ if } 0 < u < v < 1.$$

Performing the transformation $\bar{w} = 1 - w$ and $q = wu + (1-w)v$, we arrive at

$$f(u,q;w) = \frac{n!}{[(j-1)!]^2(n-2j)!\bar{w}^{n-j}} u^{j-1}(q-u)^{n-2j}(\bar{w}+wu-q)^{j-1},$$

if $0 < u < q < wu + \bar{w} < 1$. By noting the ranges of integration as $0 < u < q$ for $0 \le q \le 1-w$ and $\frac{q-\bar{w}}{w} < u < q$ for $1-w \le q \le 1$ and once again making use of binomial expansions, we find the density of q to be:

$$f(q;w) = \begin{cases} \sum_{r=0}^{j-1}(-1)^r \binom{n-2j+r}{r} \frac{w^r}{\bar{w}^{n-2j+r+1}} \frac{q^{n-j+r}(1-q)^{j-1-r}}{B(n-j+r+1,j-r)} & \text{if } 0 \le q \le 1-w, \\ \sum_{r=0}^{j-1} \binom{2j-r-2}{j-r-1} \frac{\bar{w}^{j-r-1}}{w^{n-j}} \frac{(q-\bar{w})^r(1-q)^{n-r-1}}{B(r+1,n-r)} & \text{if } 1-w < q \le 1. \end{cases}$$

Therefore, the distribution function of q is

$$F(q;w) = \begin{cases} \sum_{r=0}^{j-1} \left\{ (-1)^r \binom{n-2j+r}{r} \frac{w^r}{\bar{w}^{n-2j+r+1}} \right. \\ \qquad \left. \times I_q(n-j+r+1, j-r) \right\} & \text{if } 0 \le q \le 1-w, \\ \sum_{r=0}^{j-1}(-1)^r \binom{n-2j+r}{r} \frac{w^r}{\bar{w}^{n-2j+r+1}} I_{\bar{w}}(n-j+r+1, j-r) \\ \quad + \sum_{r=0}^{j-1} \binom{2j-r-2}{j-r-1} w^j \bar{w}^{j-1-r} I_{\frac{q-\bar{w}}{w}}(r+1, n-r) & \text{if } 1-w < q \le 1, \end{cases}$$

respectively, where, as defined earlier, $I_q(a,b)$ is the incomplete beta ratio and $B(a,b)$ is the complete beta function. For fixed n and p, there exists a unique value of w for which

$$F(p;w) = Q_{n,p}(w) = \frac{1}{2}, \tag{15}$$

where $Q_{n,p}(w)$ is as given in (14). Thus, we can regard the corresponding $Q_{n,p}(w)$ as a nonparametric competitor to the Harrell-Davis [11] estimator, which is a robust L_1-estimator. Of course, for this purpose, we need to solve (15) for w, for given values of n and p.

Table 3 Values of w satisfying (15), for different choices of n and p

		p				
n	j	$\frac{j}{2j+1}$	$\frac{j+\frac{1}{4}}{2j+1}$	$\frac{j+\frac{1}{2}}{2j+1}$	$\frac{j+\frac{3}{4}}{2j+1}$	$\frac{j+1}{2j+1}$
4	2	0.9313	0.7056	0.5000	0.2944	0.0687
6	3	0.9509	0.7178	0.5000	0.2822	0.0491
8	4	0.9619	0.7248	0.5000	0.2752	0.0381
10	5	0.9689	0.7293	0.5000	0.2707	0.0311
12	6	0.9738	0.7325	0.5000	0.2675	0.0262

Special Case

If $n = 2m$ and $p \in \left(\frac{m}{2m+1}, \frac{m+1}{2m+1}\right)$, then $j = [(n+1)p] = [(2m+1)/2] = \left[m + \frac{1}{2}\right] = m$. In this case, by solving (15) for varying n, we found the values of w for different choices of p, and these are presented in Table 3.

Use of the largest order statistic that underestimates p

Suppose $U_{1:n}, \ldots, U_{n:n}$ are the order statistics from the uniform $\mathscr{U}(0, 1)$ distribution. Earlier, we considered $i = |(n+1)p|$, but it is possible that $U_{i:n}$ may underestimate p with a probability of at least 1/2. Yet, that does not guarantee that $U_{i+1:n}$ overestimates p with probability of at least 1/2. However, such an i does exist, but it just may not correspond to $|(n+1)p|$.

Consider the joint density of $U_{i:n}$ and $U_{i+1:n}$ given by

$$f(u, v) = \frac{n!}{(i-1)!)(n-i-1)!} u^{i-1}(1-v)^{n-i-1}, \ 0 < u < v < 1.$$

Letting $U = u$ and $Q = wU + (1-w)V$, then the joint density becomes

$$f(u, q; w) = \frac{n!}{(i-1)!(n-i-1)!\bar{w}} u^{i-1} (\bar{w} - q + wu)^{n-i-1} \text{ if } 0 < u < q < wu + \bar{w} < 1$$

where again $\bar{w} = 1 - w$. As in the previous case, noting the ranges of integration as $0 < u < q$ for $0 \le q \le 1 - w$ and $\frac{q-\bar{w}}{w} < u < q$ for $1 - w < q \le 1$, we find the corresponding density and distribution functions of q to be:

$$f(q; w) = \begin{cases} \sum_{r=0}^{n-i-1} (-1)^r \frac{w^r}{\bar{w}^{r+1}} \frac{q^{r+i}(1-q)^{n-i-1-r}}{B(r+i+1, n-i-r)}) & \text{if } 0 \le q \le 1-w, \\ \frac{n!}{(i-1)!(n-i-1)!} \sum_{r=0}^{i-1} \binom{i-1}{r} \frac{\bar{w}^{i-1-r}}{w^i} \frac{(q-\bar{w})^r (1-q)^{n-r-1}}{n-r-1} & \text{if } 1-w < q \le 1, \end{cases}$$

and so

$$F(q;w)=\begin{cases}\sum_{r=0}^{n-i-1}(-1)^r\frac{w^r}{\bar{w}^{r+1}}I_q(r+i+1,n-i-r) \text{ if } 0\le q\le 1-w,\\ \sum_{r=0}^{n-i-1}\frac{w^r}{\bar{w}^{r+1}}I_{1-w}(r+i+1,n-i-r)\\ +\sum_{r=0}^{i-1}\sum_{s=0}^{n-r-1}\left\{(-1)^s\binom{n}{s}\binom{n-r-2}{i-1-r}\frac{\bar{w}^{i-1-r+s}}{w^i}\right.\\ \left.\quad\times I_{q-\bar{w}}(r+1,n-r-s)\right\} \qquad \text{if } 1-w<q\le 1.\end{cases}$$

We can now solve for w, using successive order statistics, instead of the earlier approach when the two order statistics are determined by the mean rank approach.

6.2 Exponential Distribution

Use of $|(n+1)p|$ for j

In the case of exponential distribution, by proceeding in a manner analogous to the uniform case, we can show that the cdf of q is

$$F(q)=\frac{n!}{[(j-1)!]^2(n-2j)!\bar{w}}\sum_{r=0}^{j-1}\sum_{s=0}^{n-2j}\frac{(-1)^{r-s}\binom{j-1}{r}\binom{n-2j}{s}}{r+n-2j-s-\frac{w}{\bar{w}(s+j-1)}}\times\left[\frac{\bar{w}}{s+j-1}\left(1-e^{-\left(\frac{s+j-1}{\bar{w}}\right)q}\right)-\frac{1}{r+n-j-1}\left(1-e^{-(r+n-j-1)q}\right)\right],\quad 0<q<\infty.$$

Use of the largest order statistic that underestimates p

Here again, for the case of exponential distribution, by proceeding in a manner analogous to the uniform case, we can show that the cdf of q is

$$F(q)=\frac{n!}{(n-j)!\bar{w}}\frac{\Gamma\left(\frac{(n-j)w}{\bar{w}}+1\right)}{\Gamma(j+\frac{(n-j)w}{\bar{w}}+1)}\left\{1-I_{e^{-q}}\left(j,\frac{(n-j)w}{\bar{w}}+1\right)\right\}e^{-\frac{(n-j)}{\bar{w}}q},\quad \text{if } 0<q<\infty, \tag{16}$$

where $I_q(a,b)$ is the incomplete beta ratio defined earlier. We may use (16) to determine w such that

$$F(p;w)=\frac{1}{2} \text{ and } j \text{ is the greatest integer such that } m_{j:n}\le p.$$

7 Concluding Remarks

Pitman closeness of order statistics to population parameters such as quantiles have been discussed in the literature. Here, we have discussed Pitman closest estimation based on convex linear combinations of two contiguous order statistics. We have then illustrated the developed results for the uniform, exponential, power function and Pareto distributions. As done in the case of quantile estimation, one may also propose convex linear combinations of two contiguous order statistics as Pitman closest predictors of a future failure time. This work is currently under progress and we hope to report these findings in a future paper.

Acknowledgments The authors thank the editors, Drs. Pankaj Choudhary, Chaitra H. Nagaraja and Tony Ng, for their kind invitation to present this paper for the volume. Our sincere thanks go to an anonymous reviewer whose valuable comments and suggestions on an earlier version of this manuscript led to this significantly improved version. We also take this opportunity to congratulate Dr. H.N. Nagaraja for his accomplishments so far and hopefully for many more productive years in the future!

References

1. Arnold, B.C., N. Balakrishnan, and H.N. Nagaraja. 1992. *A first course in order statistics*. New York, NY: Wiley.
2. Balakrishnan N., K. Davies. 2013. Pitman closeness results for type-I censored data from exponential distribution. *Statistics and Probability Letters*, 2693–2698.
3. Balakrishnan, N., K. Davies, and J.P. Keating. 2009. Pitman closeness of order statistics to population quantiles. *Communications in Statistics - Simulation and Computation* 38: 802–820.
4. Balakrishnan, N., G. Iliopoulos, J.P. Keating, and R.L. Mason. 2009. Pitman closeness of sample median to population median. *Statistics and Probability Letters* 79: 1759–1766.
5. Balakrishnan, N., K.F. Davies, J.P. Keating, and R.L. Mason. 2010. Simultaneous closeness among order statistics to population quantiles. *Journal of Statistical Planning and Inference* 140: 2408–2415.
6. Balakrishnan, N., K.F. Davies, J.P. Keating, and R.L. Mason. 2011. Pitman closeness, monotonicity and consistency of best linear unbiased and invariant estimators for exponential distribution under Type-II censoring. *Journal of Statistical Computation and Simulation* 81: 985–999.
7. Balakrishnan, N., K. Davies, J.P. Keating, and R.L. Mason. 2012. Computation of optimal plotting points based on Pitman closeness with an application to goodness-of-fit for location-scale families. *Computational Statistics and Data Analysis* 56: 2637–2649.
8. Balakrishnan, N., and A.C. Cohen. 1991. *Order statistics and inference: Estimation methods*. San Diego, CA: Academic Press.
9. David, H.A., and H.N. Nagaraja. 2003. *Order statistics*, 3rd ed. Hoboken, NJ: Wiley.
10. Fountain, R.L. 1991. Pitman closeness comparison of linear estimators: A canonical form. *Communications in Statistics - Theory and Methods* 20: 3535–3550.
11. Harrell, F.E., and C.E. Davis. 1982. A new distribution-free quantile estimator. *Biometrika* 69: 635–640.
12. Keating, J.P., R.L. Mason, and P.K. Sen. 1983. *Pitman's measure of closeness*. Philadelphia: Society for Industrial and Applied Mathematics.

13. Lloyd, E.H. 1952. Least squares estimation of location and scale parameters using order statistics. *Biometrika* 39: 88–95.
14. Mason, R.L., J.P. Keating, P.K. Sen, and N.W. Blaylock. 1990. Comparison of linear estimators using Pitman's measure of closeness. *Journal of the American Statistical Association* 85: 579–581.
15. Nagaraja, H.N. 1986. Comparison of estimators and predictors from two-parameter exponential distribution. *Sankyha Series B*, 10–18.
16. Peddada, S.D., and R. Khattree. 1991. Comparison of estimators of the location parameter using Pitman's closeness criterion. *Communications in Statistics - Theory and Methods* 20: 3525–3534.
17. Pitman, E.J.G. 1937. The closest estimates of statistical parameters. *Proceedings of the Cambridge Philosophical Society* 33: 212–222.
18. Rao, C.R. 1981. Some comments on the minimum mean square error as a criterion in estimation. *Statistics and Related Topics*, 123–143. Amsterdam: North-Holland.

Nonparametric Confidence Regions for L-Moments

J.R.M. Hosking

Abstract Methods for constructing joint confidence regions for L-skewness and L-kurtosis are compared by Monte Carlo simulation. Exact computations can be based on variance estimators given by Elamir and Seheult (2003, *Journal of Statistical Planning and Inference*) and by Wang and Hutson (2013, *Journal of Applied Statistics*). Confidence regions can also be constructed using the bootstrap; several variants are considered. The principal conclusions are that all methods perform poorly for heavy-tailed distributions, and that even for light-tailed distributions a sample size of 200 may be required in order to achieve good agreement between nominal and actual coverage probabilities. A bootstrap method based on estimation of the covariance matrix of the sample L-moment ratios is overall the best simple choice. Among the practical results is an L-moment ratio diagram on which confidence regions for sample L-moment statistics are plotted. This gives an immediate visual indication of whether different samples can be regarded as having been drawn from the same distribution, and of which distributions are appropriate for fitting to a given data sample.

Keywords Bootstrap · Kurtosis · Maximum entropy · Order statistics · Skewness

1 Order Statistics and L-Moments

The order statistic $X_{j:n}$ is a random variable distributed as the jth smallest element of a random sample of size n drawn from the distribution of some random variable X. The theory of order statistics has provided many insights into the properties of, and relations between, probability distributions, and has generated many effective methods for inference from a data sample about the probability distribution from which the sample was drawn. Many key results are in the authoritative book of David and Nagaraja [3].

J.R.M. Hosking (✉)
New York, USA
e-mail: jrmhosking@gmail.com

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_3

A practically useful offshoot of the theory of order statistics concerns measures of location, scale, and shape of distributions based on expectations of linear combinations of order statistics. These measures, called L-moments [9], are defined by

$$\lambda_r = r^{-1} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \mathrm{E}\, X_{r-j:r}. \tag{1}$$

L-moments can be used as summary statistics for data samples, and to identify probability distributions and fit them to data. L-moments have been used in many areas of application: recent examples include the environmental sciences [2], finance [14], and reliability [17, Chap. 6]. Recent theoretical developments include extension of L-moments to multivariate distributions [21], bias-reduced estimates of L-moments [23], and derivations of L-moments for power-transformed normal and logistic distributions [8] and for the symmetric triangular distribution [18].

L-moments can be estimated from a sample of data by a linear combination of the ordered data. An unbiased estimator of λ_r is conveniently computed as the weighted sum $\ell_r = \sum_{i=1}^{n} w_{k:n}^{(r)} X_{k:n}$. The weights can be computed by the recursion

$$w_{k:n}^{(1)} = 1\,, \qquad w_{k:n}^{(2)} = (2k - n - 1)/(n - 1), \tag{2}$$

and, for $r \geq 2$,

$$r(n-r)w_{k:n}^{(r+1)} = (2r-1)(2k-n-1)w_{k:n}^{(r)} - (r-1)(n+r-1)w_{k:n}^{(r-1)} \tag{3}$$

[11, Eqs. (14)–(15)].

For inference about the shape of a probability distribution, independently of its scale, the dimensionless quantities called L-moment ratios are useful measures. They are defined by $\tau_r = \lambda_r/\lambda_2$, r=3, 4, ... Particularly useful L-moment ratios are τ_3 and τ_4, which respectively measure the skewness and kurtosis of a distribution. The analogous sample estimators are the sample L-moment ratios $t_r = \ell_r/\ell_2$, r=3, 4, ...

2 L-Moment Ratio Diagram

A convenient tool for use with L-moments is the L-moment ratio diagram, which shows the L-skewness and L-kurtosis of probability distributions and data samples, and enables judgment of which distributions may give a good fit to a given data sample. For an example we use 8 sets of annual maximum streamflow data for sites in Texas. The data are given in [1], data sets `USGSsta01515000peaks` etc. Table 1 shows the sample size and L-moments of each data set. Figure 1 shows the sample L-skewness and L-kurtosis of the 8 data sets on an L-moment ratio diagram.

Closeness of a point on an L-moment ratio diagram to a distribution curve suggests that the distribution may give a good fit to the data. For example, in Fig. 1 the point

Table 1 L-moments of stream gaging sites in Texas

Site	n	ℓ_1	ℓ_2	t_3	t_4
01515000	71	69406	13384	0.1889	0.0993
02366500	76	39697	13062	0.4149	0.3603
05405000	73	3135	894	0.1786	0.0989
08151500	67	51156	28880	0.3925	0.1701
08167000	69	27586	17395	0.4914	0.2596
08190000	84	33406	23443	0.5669	0.3209
09442000	85	8875	4305	0.4970	0.3423
14321000	100	101866	26787	0.1798	0.1621

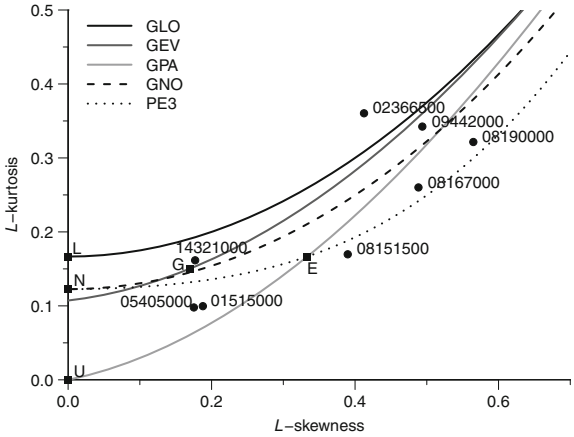


Fig. 1 L-moment ratio diagram. The graph shows sample L-skewness and L-kurtosis for 8 data sets of annual maximum streamflows for sites in Texas, and the relations between L-skewness and L-kurtosis for several families of distributions. Curves represent distribution families: generalized logistic (GLO), generalized extreme-value (GEV), generalized Pareto (GPA), generalized normal (GNO), and Pearson type III (PE3). Labelled *square dots* represent individual distributions: exponential (E), Gumbel (G), logistic (N), normal (N), and uniform (U)

for Site 14321000 lies very close to the Gumbel distribution. But some further natural questions do not have obvious answers:

- How accurate are the points on the L-moment ratio diagram?
- How can we assess the statistical significance of distances between points?
- How can we assess which values of (τ_3, τ_4) are plausible candidates for the distribution from which the data were drawn?

These questions can be addressed by constructing a confidence region for sample L-skewness and L-kurtosis. This provides an immediate indication of which values of population L-skewness and L-kurtosis are consistent with the data. In particular

the absence of overlap between a distribution curve and the confidence region is a strong indicator that the data were not drawn from that family of distributions.

3 Confidence Regions for *L*-Moment Ratios

Construction of confidence regions for L-moment ratios requires estimation of the variability of the sample L-moment ratios from a given data sample $x_1 \le x_2 \le \cdots \le x_n$. First we state some key theoretical results concerning the variability of sample L-moment ratios. If the random variable X has finite variance then the following results hold.

Result 1 (adapted from [7]). If $r + s \le n$ then

$$\operatorname{cov}(\ell_r, \ell_s) = \Lambda_{rs}^{(n)} = \sum_{1\le i<j\le r+s}\sum c_{ij}\,\mathrm{E}(X_{j:r+s} - X_{i:r+s})^2 = \sum_{1\le i<j\le n}\sum c_{ij}^*\,\mathrm{E}(X_{j:n} - X_{i:n})^2. \tag{4}$$

Result 2 [9, Theorem 3(a)]. Asymptotically as $n \to \infty$, for any integer $R > 0$ the quantities $n^{1/2}(\ell_r - \lambda_r)$, $r = 1, \ldots, R$, are jointly normally distributed with

$$n\operatorname{cov}(\ell_r, \ell_s) \sim \Lambda_{rs} = \iint_{x<y} \{P_{r-1}^*(F(x))\,P_{r-1}^*(F(y)) + P_{s-1}^*(F(x))\,P_{r-1}^*(F(y))\}\,F(x)\,\{1 - F(y)\}\,dx\,dy, \tag{5}$$

where P_m^* denotes the mth shifted Legendre polynomial, defined by

$$P_m^*(u) = \sum_{k=0}^{m}(-1)^{m-k}\binom{m}{k}^2 u^k(1-u)^{m-k}. \tag{6}$$

Result 3 [9, Theorem 3(b)]. Asymptotically as $n \to \infty$, for any integer $R > 0$ the quantities $n^{1/2}(t_r - \tau_r)$, $r = 3, \ldots, R$, are jointly normally distributed with

$$n\operatorname{cov}(t_r, t_s) \sim T_{rs} = (\Lambda_{rs} - \tau_r\Lambda_{2s} - \tau_s\Lambda_{2r} + \tau_r\tau_s\Lambda_{22})/\lambda_2^2. \tag{7}$$

Result 1 immediately provides an estimator of $\Lambda_{rs}^{(n)}$: substitute x_j for $\mathrm{E}\,X_{j:n}$ in (4). This estimator is distribution-free, i.e., it is unbiased for all distributions (with finite variance) from which the sample may have been drawn. Substituting these estimators of $\Lambda_{rs}^{(n)}$ for Λ_{rs} in (7), we obtain estimators of T_{rs}, the asymptotic covariances of the L-moment ratios. Assuming joint normality of the L-moment ratios (a valid approximation for large n, by Result 3), we can construct a confidence region for (τ_3, τ_4) by computing probabilities for the bivariate normal distribution. We call this distribution-free procedure for construction of confidence regions **Method DF**.

Other approaches for constructing confidence regions involve estimating the variability of sample L-moment ratios for samples from a particular distribution. A nat-

ural choice is the empirical distribution, i.e., the distribution that assigns probability mass $1/n$ to each of the points $x_1, x_2, \ldots, x_n$.

We consider four possibilities.

- The covariance $\Lambda_{rs}^{(n)}$ in (4) can be computed exactly for the empirical distribution, in effect as a weighted sum of squared differences of the sample data points. This is the "exact bootstrap" of Wang and Hutson [22]. Substituting these estimators of $\Lambda_{rs}^{(n)}$ for Λ_{rs} in (7) and using Result 3 as in Method DF, we can construct a confidence region based on bivariate normal probabilities. We call this **Method EB**.
- The covariance Λ_{rs} in (5) can also be computed exactly for the empirical distribution. Straightforward algebra shows that the result is

$$\tilde{\Lambda}_{rs} = \sum_{1 \le i < j \le n-1} \sum \{P_{r-1}^*(i/n)\, P_{r-1}^*(j/n) + P_{s-1}^*(i/n)\, P_{r-1}^*(j/n)\} \times (i/n)\,(1 - j/n)\,(x_{i+1} - x_i)\,(x_{j+1} - x_j)\,. \tag{8}$$

  Again substituting $\tilde{\Lambda}_{rs}$ for Λ_{rs} in (7) and using Result 3 as in Method DF, we can construct a confidence region based on bivariate normal probabilities. We call this **Method AB** (for "asymptotic bootstrap").
- The bootstrap [4, 5] is a familiar method for assessing the variability of sample statistics by simulation from the empirical distribution. It is straightforward to use it for L-moment ratios.

1. Generate B samples of size n from the empirical distribution.
2. Compute (t_3, t_4) from each bootstrap sample.
3. Compute the sample covariance matrix T^* of the (t_3, t_4) values.

  Again assuming a joint normal distribution of the L-moment ratios, a confidence region for (t_3, t_4) can be computed from bivariate normal probabilities. We call this **Method BEN** (Bootstrap the Empirical distribution, then assume a Normal distribution).
- After steps 1 and 2 of the bootstrap we can construct a confidence region by peeling the convex hull of the (t_3, t_4) points. The peeling procedure is described in [24] and illustrated in Fig. 2. This approach does not assume a joint normal distribution of the L-moment ratios. We call this **Method BEP** (Bootstrap the Empirical distribution, then Peel the convex hull).

Bootstrap approaches typically sample from the empirical distribution, using it as a proxy for the true distribution from which the data were sampled. In this role the empirical distribution has some disadvantages. It is discrete and bounded, even when the true distribution is not, and samples from it tend to have less dispersion than those from the true distribution. Several authors have remarked on this phenomenon. For example, Kysely [15] found that "the nonparametric bootstrap [i.e., a bootstrap of the empirical distribution] should be interpreted with caution because it leads to confidence intervals that are too narrow and underestimate the real uncertainties involved in the frequency models".

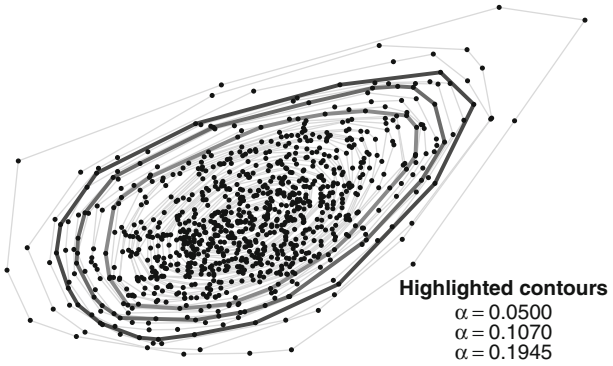


Fig. 2 Illustration of peeling a convex hull. *Dots* indicate (t_3, t_4) points from bootstrap samples. Successive convex hulls (*gray polygons*) are removed ("peeled") from the set of points. When some specified proportion α of the original points has been removed, the convex hull of the remaining points encloses proportion $1 - \alpha$ of the original points and constitutes a $100(1 - \alpha)$ % confidence region for the population L-moment ratios (τ_3, τ_4)

Table 2 Recommended relation between sample size n and the number, m, of L-moments used in maximum-entropy estimation

n	25	50	100	200	500	1000
m	6	8	10	12	16	20

An alternative is to generate bootstrap samples from some other "parent" distribution than the empirical distribution. Some possibilities include the smoothed kernel quantile estimator [20] and the empirical distribution extended with exponentially decreasing tails [13]. Here we propose the distribution that has maximum entropy subject to having the same first m L-moments as the data. This gives a flexible form for the parent distribution; it is unbounded, so is less likely than the empirical distribution to generate underdispersed samples; it can closely approximate, for large enough m, essentially any distribution with finite mean; it avoids the need to specify a fixed parametric form for the parent distribution; and its close connection to L-moments makes it a natural choice for the present application.

Using the maximum-entropy distribution requires a choice of m. The optimal choice is unclear; we use $m = \lceil 2n^{1/3} \rceil$, where $\lceil x \rceil$ denotes the smallest integer that is not less than x, which appears to work reasonably well in practice. Some examples of (n, m) pairs are given in Table 2. Fitting the L-moment maximum-entropy distribution to data requires, in general, solution of a convex optimization problem [10, Remark 2.3]. This optimization can be achieved by standard numerical procedures. In this application we used Newton-Raphson iteration; in all the computations described in Sect. 4 the iterations never failed to converge.

A bootstrap based on the L-moment maximum-entropy distribution proceeds as follows.

1. Set $m = \lceil 2n^{1/3} \rceil$.
2. Fit the maximum-entropy distribution to the first m L-moments of the data.
3. Generate B samples of size n from the maximum-entropy distribution.
4. Compute (t_3, t_4) from each bootstrap sample.

At this point there is the same choice as before for constructing confidence regions: assume a joint normal distribution for the L-moment ratios (**Method BMN**) or peel the convex hull of the (t_3, t_4) values (**Method BMP**).

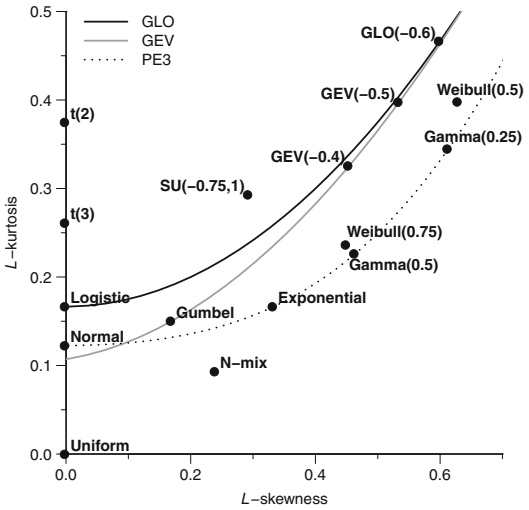
We have defined seven methods of confidence region construction for L-moment ratios. We note some other possibilities. Wang and Hutson [22, Sect. 3.2] defined a method based on characteristic functions. This uses an Edgeworth-type correction to the characteristic function of a joint normal distribution of sample L-moment ratios, and involves sums of third and fourth powers of the sample data values. This seems likely to be unstable for samples from distributions for which these higher moments may not exist. Peng [19] defined a confidence interval for λ_2 using empirical likelihood. This is a promising approach, but extending it to L-moment ratios τ_3 and τ_4 seems to involve complex algebra and challenging numerical optimizations.

4 Evaluation of Methods of Confidence Region Construction

4.1 Distributions

The seven methods of confidence region construction were tested on samples from 16 probability distributions, chosen to cover a range of population L-skewness and L-kurtosis often encountered in practice. The distributions are as follows. Here and in Figs. 4, 5 and 6 they are listed in increasing order of tail weight, defined as the ratio of the distance between the 0.0001 and 0.9999 quantiles to the distance between the 0.001 and 0.999 quantiles.

- Uniform
- N-mix: location mixture of normal distributions—N(0, 1) with probability 0.75, N(5, 1) with probability 0.25.
- Normal.
- Gumbel (extreme-value type I).
- Logistic.
- Exponential.
- Gamma(0.5): gamma distribution with shape parameter 0.5.
- Weibull(0.75): Weibull distribution [12, Appendix A.6] with shape parameter $\delta = 0.75$.
- Gamma(0.25): gamma distribution with shape parameter 0.25.
- Weibull(0.5): Weibull distribution with shape parameter 0.5.



	τ_3	τ_4
GLO(−0.6)	0.6000	0.4667
Weibull(0.5)	0.6296	0.3981
GEV(−0.5)	0.5347	0.3978
t(2)	0.0000	0.3750
Gamma(0.25)	0.6135	0.3449
GEV(−0.4)	0.4543	0.3258
SU(−0.75,1)	0.2937	0.2931
t(3)	0.0000	0.2612
Weibull(0.75)	0.4505	0.2365
Gamma(0.5)	0.4641	0.2265
Logistic	0.0000	0.1667
Exponential	0.3333	0.1667
Gumbel	0.1699	0.1504
Normal	0.0000	0.1226
N−mix	0.2404	0.0933
Uniform	0.0000	0.0000

Fig. 3 L-moment ratio diagram showing the 16 distributions used in the simulations in Sect. 4

- SU(−0.75, 1): Johnson SU distribution with parameters $\gamma = -0.75$, $\delta = 1$, i.e., cumulative distribution function $F(x) = \sinh(0.75 + \Phi(x))$ where Φ is the standard normal cumulative distribution function.
- t(3): Student t distribution, 3 degrees of freedom.
- GEV(−0.4): generalized extreme-value distribution [12, Appendix A.6] with shape parameter $k = -0.4$.
- t(2): Student t distribution, 2 degrees of freedom.
- GEV(−0.5): generalized extreme-value distribution, shape parameter −0.5.
- GLO(−0.6): generalized logistic distribution [12, Appendix A.7] with shape parameter $k = -0.6$.

The distributions and their L-moment ratios are shown on an L-moment ratio diagram in Fig. 3.

4.2 Computational Cost

The methods described in Sect. 3 have very different computational costs. Methods DF and AB involve sums across all pairs of data points, and have complexity $O(n^2)$. Method EB, as described in [22], has complexity $O(n^5)$ (in [22, Eq. (15)], quantities $w_{ij(rs)}$ must be computed for each (i, j, r, s) combination with $1 \le i < j \le n$ and $1 \le r < s \le n$, and each is a sum of $s - r$ terms). Bootstrap methods have complexity $O(nB)$ once the data have been sorted; the number of bootstrap samples B can be set independently of the sample size n. A common choice of B for variance estimation

Table 3 Proportion of invalid results obtained with method DF

	Sample size					Sample size			
	25	50	100	200		25	50	100	200
N-mix	0.70	0.16	0.00	0	GEV(−0.5)	0.30	0.05	0.00	0
Gamma(0.25)	0.54	0.14	0.00	0	GEV(−0.4)	0.28	0.04	0.00	0
Weibull(0.5)	0.44	0.12	0.00	0	SU(−0.75,1)	0.23	0.02	0.00	0
Gamma(0.5)	0.38	0.04	0.00	0	Gumbel	0.22	0.00	0.00	0
Uniform	0.36	0.01	0.00	0	t(3)	0.20	0.00	0.00	0
Weibull(0.75)	0.35	0.03	0.00	0	t(2)	0.20	0.01	0.00	0
Exponential	0.33	0.02	0.00	0	Normal	0.19	0.00	0.00	0
GLO(−0.6)	0.31	0.07	0.01	0	Logistic	0.19	0.00	0.00	0

is in the range 100–1000. In the simulations in Sect. 4.4 we used $B = 1000$. However, Methods BMN and BMP can be several times slower than Methods BEN and BEP, owing to the need to fit the maximum-entropy distribution to the data and to generate random samples from the maximum-entropy distribution.

In practice, for sample sizes greater than 100 Method EB becomes very slow (taking more than 1 s in our R implementation) and results for it for larger samples are not included in Sect. 4.4. The other methods remain feasible (typically taking 0.1 s or less) for sample sizes up to at least 1000.

4.3 Invalid Results

Method DF has the disadvantage that it can sometimes produce invalid results. This occurs when the computed covariance matrix of the L-moments (ℓ_2, ℓ_3, ℓ_4) is not positive definite; in some cases the estimated variance of an L-moment is negative. As shown in Table 3, invalid results occur quite frequently for samples of size 25 and occasionally for samples of size 50.

4.4 Coverage

The effectiveness of a method of confidence region construction depends on how closely the generated regions achieve their nominal coverage level. This issue was addressed by Monte Carlo simulation. From each of the 16 parent distributions defined in Sect. 4.1, and for each of six sample sizes between 25 and 1000, 1000 samples were generated. For each of the 1000 samples confidence regions were generated for confidence levels 80, 90, and 95 %, using each of the seven methods described in Sect. 3. For each method and confidence level, the empirical coverage

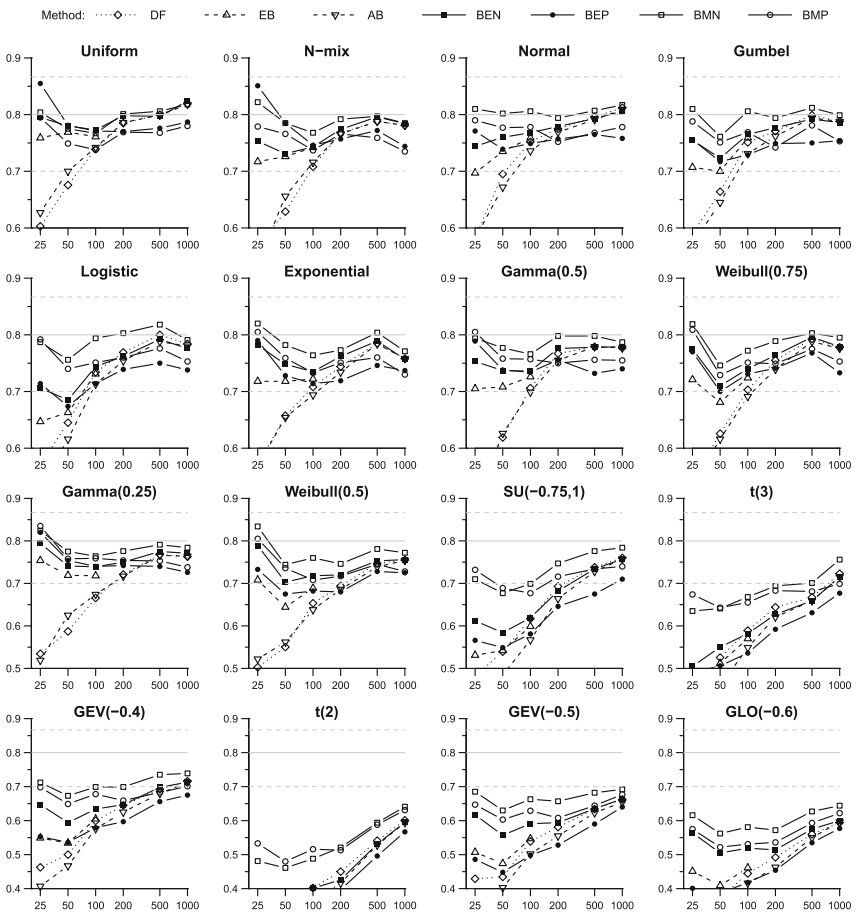


Fig. 4 Empirical coverage probability (*vertical axis*) versus sample size (*horizontal axis*) for nominal coverage probability 80 %. Each *panel* shows the results for one of the distributions defined in Sect. 4.1. *Gray horizontal lines* indicate (*solid line*) the nominal coverage level and (*dotted lines*) non-coverage levels ±50 % different from the nominal level (i.e., 30 and 13.33 %)

level is the proportion of the 1000 confidence regions that contained the (τ_3, τ_4) values of the parent distribution. The results are shown in Figs. 4, 5 and 6.

The results overall are mixed. In the majority of cases the empirical coverage probability is less than the nominal coverage level, meaning that the confidence regions are anticonservative, or "overconfident". Methods DF and AB in particular are very anticonservative for sample sizes $n \leq 100$. For the lighter-tailed distributions (the first two rows of panels in Figs. 4, 5 and 6), reasonable coverage—noncoverage within 50 %, sometimes much less, of the nominal value—is attained for sample sizes $n \geq 100$ at coverage level 80 % and for $n \geq 200$ at coverage level 90 %. For distributions with the heaviest tails the coverage is typically overestimated. This is

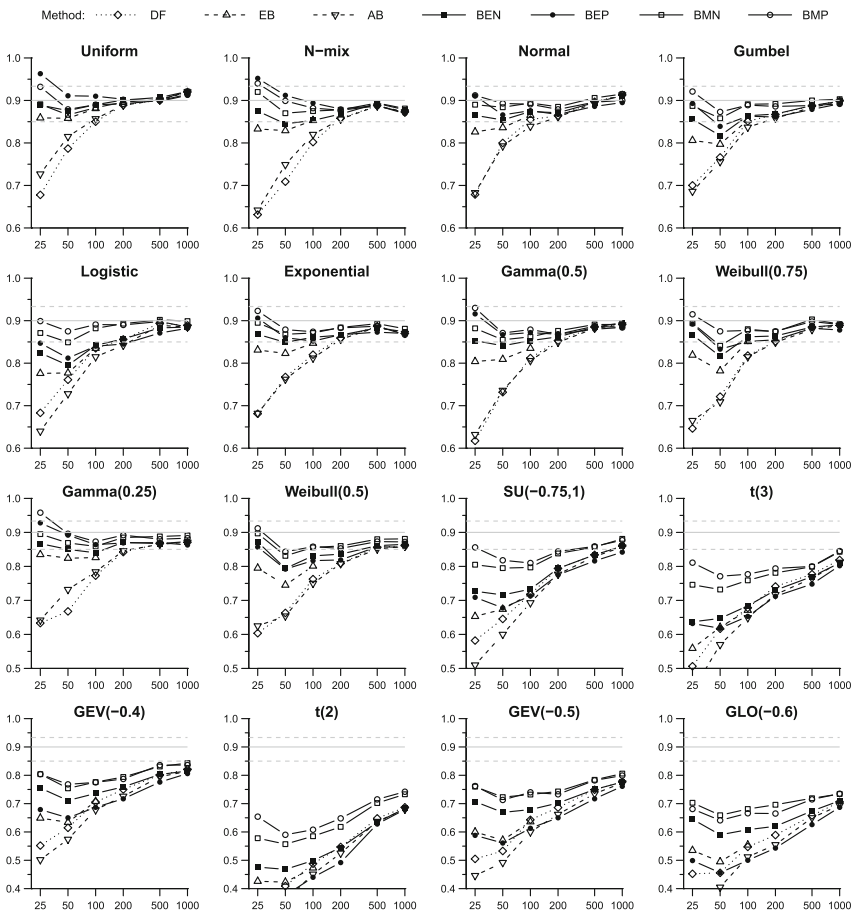


Fig. 5 Empirical coverage probability (*vertical axis*) versus sample size (*horizontal axis*) for nominal coverage probability 90 %. Each *panel* shows the results for one of the distributions defined in Sect. 4.1. *Gray horizontal lines* indicate (*solid line*) the nominal coverage level and (*dotted lines*) non-coverage levels ±50 % different from the nominal level (i.e., 15 and 6.67 %)

particularly true for the t(2), GEV(−0.5) and GLO(−0.6) distributions, which have infinite variance. For these distributions, Results 1–3 cannot be assumed to hold and only the bootstrap regions with convex hull peeling can be expected to perform well, but even these seem to require sample sizes in excess of 1000 if they are to achieve accurate coverage.

Comparing the different methods, bootstrap methods give the best overall performance. For light-tailed distributions they often achieve close to the nominal coverage even for $n = 25$. Methods BMN and BMP, which generate bootstrap samples from the maximum-entropy distribution, generally have higher coverage than Methods BEN and BEP, which use the empirical distribution. This means that Methods BMN

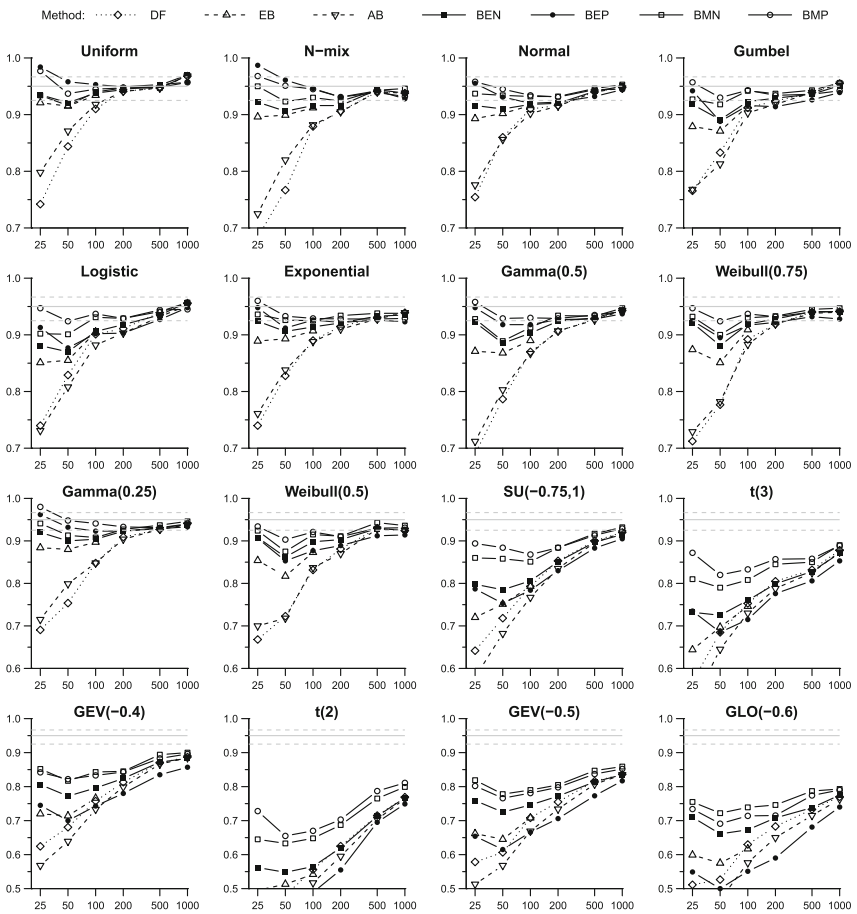


Fig. 6 Empirical coverage probability (*vertical axis*) versus sample size (*horizontal axis*) for nominal coverage probability 95 %. Each *panel* shows the results for one of the distributions defined in Sect. 4.1. *Gray horizontal lines* indicate (*solid line*) the nominal coverage level and (*dotted lines*) non-coverage levels ±50 % different from the nominal level (i.e., 7.5 and 3.33 %)

and BMP are generally closer to the nominal coverage level, though for heavy-tailed distributions this merely means that their undercoverage is less severe than that of the other methods. There is no clear advantage for the "peeling" methods BEP and BMP over the normality-based methods BEN and BMN: peeling gives generally more accurate coverage at nominal coverage 95 % but normality-based methods appear more accurate at coverage level 80 %.

4.5 Summary

The main conclusions of the comparison of the methods of confidence region construction are as follows.

- The distribution-free "unbiased" estimator fails often for $n < 50$.
- The exact bootstrap is computationally taxing for $n > 100$.
- Confidence regions are generally anticonservative (or "overconfident").
- All methods give poor results for heavy-tailed (power-law) distributions, even those with finite variance.
- For light-tailed distributions, almost all methods give respectable results for $n \geq 200$ at coverage levels 0.8 and 0.9, but need $n \geq 500$ at coverage level 0.95.
- The best methods use bootstrapping and estimate $\mathrm{var}(t_r)$ (or a confidence region for (τ_3, τ_4)) directly rather than estimating $\mathrm{var}(\ell_r)$.
- Bootstrapping with the maximum-entropy distribution appears preferable to the empirical distribution, except perhaps for distributions with very light tails.
- There is some evidence that bootstrap-and-peel is preferable to bootstrapping the covariance matrix of (t_3, t_4), particularly at coverage level 95 %. But at coverage level 80 % use of the covariance matrix seems to give more accurate coverage.

5 Texas Streamflow Data Revisited

We can now return to the L-moment ratio diagram in Fig. 1 and add confidence regions around the (t_3, t_4) points plotted there. We use Method BMN, which we judge to give the best overall results. The result is shown in Fig. 7.

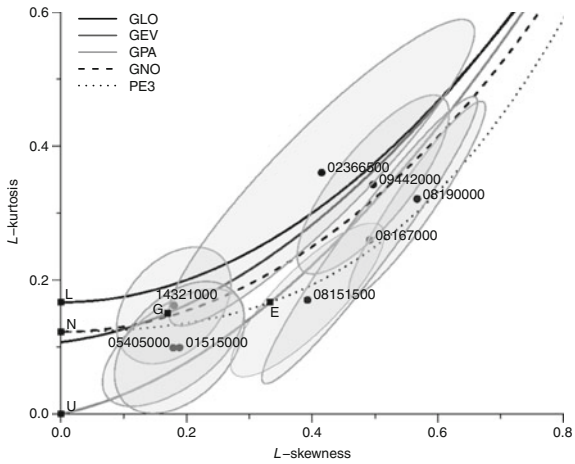


Fig. 7 L-moment ratio diagram for Texas streamflow data, with 90 % confidence regions constructed by Method BMN

The graph confirms the similarity of the low-skewness data sets, from sites 01515000, 054054000, and 14321000. These are all consistent with one another, and with having been drawn from a Gumbel distribution. These sites are distinct from the other sites, which, though well separated, have considerable overlap in their confidence regions. Even the seemingly distant points for sites 08151500 and 08190000 have some overlap in their confidence regions, suggesting that they could have been sampled from the same distribution. The large confidence region associated with the high-L-kurtosis site 02366500 is particularly striking, and emphasizes the high uncertainty associated with large L-moment ratios computed from samples of small or moderate size.

6 Conclusions

We have compared seven methods for constructing confidence regions for the L-moment ratios (τ_3, τ_4) from a sample of data. These initial investigations suggest that, at least for light-tailed distributions, regions with reasonably accurate coverage can be obtained for sample sizes of 200 or more at coverage levels up to 90 %. Bootstrap methods gave the best results, and generating bootstrap samples from the L-moment maximum-entropy distribution gives better coverage accuracy than using the empirical distribution. Overall, though, coverage accuracy for heavy-tailed distributions is disappointing for the sample sizes considered here.

Further research may enable improvements to the methods described here. Refinement of bootstrap confidence regions using iterated bootstrap methods [16] is worthy of investigation. Sample L-moment ratios can have significant bias for heavy-tailed distributions, and this will affect the coverage accuracy of confidence regions. Possible solutions include the use of distribution-free bias corrections for sample L-moments [23] and the use of bias-corrected and accelerated (BCa) bootstrap methods [6] for confidence region construction.

References

1. Asquith, W.H. 2014. `lmomco`—L-moments, trimmed L-moments, L-comoments, censored L-moments, and many distributions. R package version 2.1.1. http://www.cran.r-project.org/package=lmomco. Accessed 31 Dec 2014
2. Baratti, E., A. Montanari, A. Castellarin, J.L. Salinas, A. Viglione, and A. Bezzi. 2012. Estimating the flood frequency distribution at seasonal and annual time scale. *Hydrology and Earth System Sciences Discussions* 9: 7947–7967.
3. David, H.A., and H.N. Nagaraja. 2003. *Order statistics*, 3rd ed. New York: Wiley.
4. Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7: 1–26.
5. Efron, B. 1981. *The jackknife, the bootstrap, and other resampling plans*. CBMS Monograph, vol. 38. Philadelphia: SIAM
6. Efron, B. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association* 82: 171–200.

7. Elamir, E.A.H., and A.H. Seheult. 2004. Exact variance structure of sample L-moments. *Journal of Statistical Planning and Inference* 124: 337–359.
8. Headrick, T.C. 2011. A characterization of power method transformations through *L*-moment. *Journal of Probability and Statistics* Article ID 497463
9. Hosking, J.R.M. 1990. L-moment: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society B* 52: 105–124.
10. Hosking, J.R.M. 2007. Distributions with maximum entropy subject to constraints on their L-moment or expected order statistics. *Journal of Statistical Planning and Inference* 137: 2870–2891.
11. Hosking, J.R.M., and N. Balakrishnan. 2014. A uniqueness result for L-estimators, with applications to L-moment. *Statistical Methodology* 24: 69–80.
12. Hosking, J.R.M., and J.R. Wallis. 1997. *Regional frequency analysis: An approach based on L-moment*. Cambridge: Cambridge University Press.
13. Hutson, A.D. 2001. A semi-parametric quantile function estimator for use in bootstrap estimation procedures. *Statistical Computing* 12: 331–338.
14. Kerstens, K., A. Mounir, and I. Van de Woestyne. 2011. Non-parametric frontier estimates of mutual fund performance using C- and L-moments: Some specification tests. *Journal of Banking and Finance* 35: 1190–1201.
15. Kysely, J. 2008. A cautionary note on the use of nonparametric bootstrap for estimating uncertainties in extreme-value models. *Journal of Applied Meteorology and Climatology* 47: 3236–3251.
16. Lee, S.M.S., and G.A. Young. 1995. Asymptotic iterated bootstrap confidence intervals. *Annals of Statistics* 23: 1301–1330.
17. Nair, N.U., P.G. Sankaran, and N. Balakrishnan. 2013. *Quantile-based reliability analysis*. Basel: Birkhäuser.
18. Nagaraja, H.N. 2013. Moments of order statistics and L-moment for the symmetric triangular distribution. *Statistics and Probability Letters* 83: 2357–2363.
19. Peng, L. 2011. Empirical likelihood methods for the Gini index. *Australian and New Zealand Journal of Statistics* 53: 131–139.
20. Silverman, B.W., and G.A. Young. 1987. The bootstrap: To smooth or not to smooth? *Biometrika* 74: 469–479.
21. Serfling, R., and P. Xiao. 2007. A contribution to multivariate L-moments: L-comoment matrices. *Journal of Multivariate Analysis* 98: 1765–1781.
22. Wang, D., and A.D. Hutson. 2013. Joint confidence region estimation of L-moment ratios with an extension to censored data. *Journal of Applied Statistics* 40: 368–379.
23. Withers, C.S., and S. Nadarajah. 2011. Bias-reduced estimated for skewness, kurtosis, L-skewness and L-kurtosis. *Journal of Statistical Planning and Inference* 141: 3839–3861.
24. Yeh, A.B., and K. Singh. 1997. Balanced confidence regions based on Tukey's depth and the bootstrap. *Journal of the Royal Statistical Society B* 59: 639–652.

On Conditional Moments of Progressively Censored Order Statistics with a Time Constraint

Hon Keung Tony Ng, Fang Duan and Ping Shing Chan

Abstract Different hybrid progressive censoring schemes, which are mixtures of Type-I censoring and Type-II progressively censoring schemes, have been proposed in the literature. These censoring schemes impose a time constraint on the life-testing experiment and the number of progressively censored order statistics observed before this time constraint is recorded. Conditional on the number of progressively censored order statistics being observed before the time constraint, a computational method for the conditional moments of progressively censored order statistics is discussed. Simple computational formulae are presented and these formulae are illustrated with examples when the underlying distributions are uniform and exponential. These results will be useful for the development of estimation methods such as the least squares estimation, best linear unbiased estimation and approximate maximum likelihood estimation methods and for deriving asymptotic distributions of the estimates of model parameters for Type-I hybrid progressively censored data.

Keywords Approximate maximum likelihood estimation · Best linear unbiased estimation · Hybrid censoring · Type-I censoring · Type-II censoring

1 Introduction

There are many situations in life-testing experiments where only partial information on the failure times of the experimental units is available. Conventional Type-I

H.K.T. Ng (✉)
Department of Statistical Science, Southern Methodist University,
Dallas, TX 75275-0332, USA
e-mail: ngh@mail.smu.edu

F. Duan
PayPal, 2211 N 1st Street, San Jose, CA 95131, USA
e-mail: duanfang@gmail.com

P.S. Chan
Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China
e-mail: benchan@cuhk.edu.hk

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_4

and Type-II censoring schemes are designed to save total time on the test, to save experimental units for future use, and to save on the corresponding cost of the experiment. However, these censoring schemes do not have the flexibility of allowing the removal of experimental units before the termination of the experiment. Thus, progressive Type-II censoring was introduced as a more general censoring scheme. The progressive Type-II censoring can be described as follows: suppose that n units are placed on a life test, at the time of the first failure, R_1 units from the remaining $(n-1)$ surviving units are randomly selected and removed from the experiment immediately. Then, the life test continues and at the time of the second failure, R_2 units from the remaining $(n-2-R_1)$ surviving units are randomly selected and removed from the test, and so on. The life test continues until m failures are observed, and all the remaining $R_m = n - m - \sum_{i=1}^{m-1} R_i$ units are removed. The removal pattern (also called censoring scheme) $\mathbf{R} = (R_1, R_2, \ldots, R_m)$ is pre-fixed prior to the experiment. Extensive reviews of the literature on progressive censoring are provided by Balakrishnan and Aggarwala [4], Balakrishnan [3], Balakrishnan and Cramer [5].

To further control the total time of the experiment, different kinds of hybrid censoring schemes are proposed by imposing a time constraint T such that the experiment is terminated immediately after a random time $\min\{X_{m:m:n}, T\}$, where $T \in (0, \infty)$ and the values of n and m, $1 \le m \le n$, are fixed prior to the experiment. Here, $X_{1:m:n} \le X_{2:m:n} \le \cdots \le X_{m:m:n}$ are the ordered failure times resulting from the experiment. This censoring scheme is called the Type-I progressive hybrid censoring scheme which was first proposed by Kundu and Joarder [10]. This scheme ensures that the experiment time will not exceed time T. If the mth progressively censored ordered failure occurs before time T, the experiment will end at $X_{m:m:n}$. Otherwise, the experiment will stop at time T with $X_{J:m:n} \le T < X_{J+1:m:n}$, where J is the number of observed failures up to time T. All the remaining $\left(n - J - \sum_{i=1}^{J} R_i\right)$ surviving items are censored at time T. The drawback of this Type-I hybrid progressive censoring scheme is that it is possible to observe no failure before time T or the observed number of failures is not large enough to make an efficient statistical inference. Therefore, another hybrid censoring scheme called the adaptive progressive censoring scheme was proposed to address this issue [12]. For adaptive progressive censoring, the experiment continues until the mth failure is observed, but the censoring scheme is changed after time T to make the mth failure occur as soon as possible. Specifically, suppose there are J progressively censored order statistics observed before time T, then after the experiment passes time T, we set $R_{J+1} = \cdots = R_{m-1} = 0$ and $R_m = \left(n - m - \sum_{i=1}^{J} R_i\right)$. This formulation leads us to terminate the experiment as soon as possible if the $(J+1)$th failure time is greater than T for $(J+1) < m$. It is hoped that by using this hybrid censoring scheme, the total experiment time will be greatly reduced while the effective sample size is always m. For more recent developments and statistical inference based on different adaptive progressive censoring schemes, one can refer to Bairamov and Parsi [2], Cramer and Iliopoulos [8], Lin et al. [11], Park et al. [13], and Ye et al. [14].

For the aforementioned hybrid progressive censoring schemes, the computation of the moments of the progressively censored order statistics with the time constraint have not been discussed. These moments will be useful for the development of estimation methods such as the least squares estimation, best linear unbiased estimation and approximate maximum likelihood estimation methods and for deriving asymptotic distributions of the estimators of model parameters; see, for example, Balakrishnan and Aggarwala [4], Balakrishnan et al. [6, 7], Lin et al. [11]. For instance, the development of conditional least squares estimators or regression-type estimators, given the number of observed failures before time constraint T, will require the expected values of the progressively censored order statistics.

This paper aims to provide simple computational methods for the conditional single and product moments of the progressively censored order statistics, given the number of progressively censored order statistics being observed before the time constraint. This paper is organized as follows. In Sect. 2, we provide the computational method for the conditional moments of progressively censored order statistics. Then, in Sect. 3, we illustrate these computational formulae for uniform and exponential distributions and present some numerical results. In Sect. 4, an application of these computational formulae in least squares estimation is discussed and a numerical example is used to illustrate the methodology.

2 Computation of Conditional Moments

In this section, we will develop a computational method for the conditional moments of the progressively censored order statistics with a time constraint. The experiment considered here can be described as follows: suppose that n items are placed on a life test and let $X_1, X_2, \ldots, X_n$ be the corresponding lifetimes. We assume that X_i, $i = 1, 2, \ldots, n$, are independently and identically distributed (i.i.d.) with probability density function (PDF) $f(x; \boldsymbol{\theta})$ and cumulative distribution function (CDF) $F(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the vector of parameters and $x \in [0, \infty)$. For notational convenience, the parameter $\boldsymbol{\theta}$ in $f(x; \boldsymbol{\theta})$ and $F(x; \boldsymbol{\theta})$ is suppressed in this section. Prior to the experiment, the number of observed failures $m < n$ is determined and the progressive Type-II censoring scheme $(R_1, R_2, \ldots, R_m)$ with $R_i \geq 0$ and $\sum_{i=1}^{m} R_i + m = n$ is also specified. Suppose that there is a time constraint T, then the number of progressively censored order statistics being observed before time T is a discrete random variable, denoted as J, with support $\{0, 1, \ldots, m\}$.

Let us denote the ith progressively censored order statistic by $X_{i:m:n}$ and the a_ith order statistic by $X_{a_i:n}$, $i = 1, 2, \ldots, m$ [1]. Then, the observed progressively censored order statistics $(X_{1:m:n}, X_{2:m:n}, \ldots, X_{m:m:n})$ can be represented by the usual order statistics as $\left(X_{a_1:n}, X_{a_2:n}, \ldots, X_{a_m:n}\right)$, where $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ is the index vector that indicates the ith progressively censored order statistic corresponds to the a_ith order statistic in the random sample $X_1, X_2, \ldots, X_n$. For example, consider a progressively censored experiment with non-random removal whereas the items with the smallest lifetimes are being censored at each stage, then we have

$$X_{1:m:n} = X_{1:n}, X_{2:m:n} = X_{(R_1+2):n}, \ldots, X_{i:m:n} = X_{(\sum_{\ell=1}^{i-1} R_\ell + i):n}, \ldots, X_{m:m:n} = X_{n:n},$$

i.e., $a_i = \sum_{\ell=1}^{i-1} R_\ell + i$, for $i = 1, 2, \ldots, m$. Here, we are interested in the conditional moments of $X_{i:m:n}$ given $J = j$, $i = 1, 2, \ldots, m$, $j = 0, 1, 2, \ldots, m$.

Conditional on $J = j$, the (η, δ)th conditional product moment of $X_{i:m:n}$ and $X_{l:m:n}$, $i < l$, can be expressed as:

$$\begin{aligned} & E\left(X_{i:m:n}^{\eta} X_{l:m:n}^{\delta} \middle| J = j\right) \\ &= \sum_{\mathbf{a}} E\left(X_{a_i:n}^{\eta} X_{a_l:n}^{\delta} \middle| \mathbf{a}, J = j\right) \Pr\left(\mathbf{a} = (a_1, a_2, \ldots, a_m) \middle| J = j\right). \end{aligned} \tag{1}$$

This expression involves two parts: $\Pr(\mathbf{a} = (a_1, a_2, \ldots, a_m)|J = j)$ which depends on the progressive censoring scheme $(R_1, R_2, \ldots, R_m)$ and $F(T; \boldsymbol{\theta})$ and $E\left(X_{a_i:n}^{\eta} X_{a_l:n}^{\delta} \middle| \mathbf{a}, J = j\right)$ which depends on the underlying parametric distribution $F(x; \boldsymbol{\theta})$ only. We will derive the computational formulae of $\Pr(\mathbf{a} = (a_1, a_2, \ldots, a_m)|J = j)$ and $E\left(X_{a_i:n}^{\eta} X_{a_l:n}^{\delta} \middle| \mathbf{a}, J = j\right)$ separately in the following sections.

2.1 *Computation of* $\Pr(\mathbf{a} = (a_1, a_2, \ldots, a_m)|J = j)$

First, we compute the conditional probability of the observed progressively censored order statistics $(X_{1:m:n}, X_{2:m:n}, \ldots, X_{m:m:n})$ correspond to the $(a_1, a_2, \ldots, a_m)$ order statistics, given $J = j$. This conditional probability can be expressed as $\Pr(\mathbf{a} = (a_1, a_2, \ldots, a_m)|J = j)$. We observe the following relationship in Result 1.

Result 1. Given that the ith progressively censored order statistics $(X_{i:m:n})$ corresponds to the a_ith order statistic of the random sample $X_1, X_2, \ldots, X_n$, the probability that the $(i + 1)$th progressively censored order statistic, $X_{i+1:m:n}$ corresponds to the a_{i+1}th order statistic is

$$\begin{aligned} q_{i,a_i,a_{i+1}}^{(R_1,R_2,\ldots,R_m)} &= \Pr\left(X_{i+1:m:n} = X_{a_{i+1}:n} | X_{i:m:n} = X_{a_i:n}\right) \\ &= \frac{\binom{n - a_{i+1}}{\sum_{k=1}^{i} R_k - a_{i+1} + i + 1}}{\binom{n - a_i}{\sum_{k=1}^{i} R_k - a_i + i}}. \end{aligned}$$

If we consider the situation in which the first order statistic must be observed, then $a_1 = 1$, $a_{i+1} = a_i + 1, \ldots, \min\left(a_i + \sum_{k=1}^{i} R_k + 1, n\right)$, $i = 1, \ldots, m - 1$.

Since the selection of the items being censored at the time of each failure is random, Result 1 can be obtained from combinatorial arguments. From Result 1, given the progressive censoring scheme, we can compute the probability of observing $\left(X_{a_1:n}, X_{a_2:n}, \ldots, X_{a_m:n}\right)$ as the actual observations (denoted as $P_{\mathbf{a}}$), i.e.,

$$P_{\mathbf{a}} = \Pr\left(X_{i:m:n} = X_{a_i:n}, i = 1, \ldots, m\right) = \prod_{i=1}^{m-1} q_{i,a_i,a_{i-1}}^{(R_1,\ldots,R_m)}.$$

For given values of n, m and $(R_1, R_2, \ldots, R_m)$, we can compute the probability of $J = j$, $j = 0, 1, \ldots, m$. For $J = 0$, all lifetimes $X_1, X_2, \ldots, X_n$ are larger than T, then $\Pr(J = j) = [1 - F(T)]^n$. For $J = 1, 2, \ldots, m$,

$$\begin{aligned}
\Pr(J = j) &= \sum_{\mathbf{a}} \Pr(\mathbf{a} = (1, a_2, \ldots, a_m), J = j) \\
&= \sum_{\mathbf{a}} \Pr\left(\mathbf{a} = (1, a_2, \ldots, a_m), X_{a_j:n} < T < X_{a_{j+1}:n}\right) \\
&= \sum_{\mathbf{a}} \Pr\left(X_{a_j:n} < T < X_{a_{j+1}:n} | \mathbf{a} = (1, a_2, \ldots, a_m)\right) P_{\mathbf{a}} \\
&= \sum_{\mathbf{a}} \left\{ P_{\mathbf{a}} \sum_{l=a_j}^{a_{j+1}-1} \binom{n}{l} [F(T)]^l [1 - F(T)]^{n-l} \right\},
\end{aligned}$$

where $a_{m+1} \equiv n + 1$. Thus, we have

$$\begin{aligned}
\Pr(\mathbf{a} = (1, a_2, \ldots, a_m) | J = j) &= \frac{\Pr(\mathbf{a} = (1, a_2, \ldots, a_m), J = j)}{Pr(J = j)} \\
&= \frac{P_{\mathbf{a}} \sum_{l=a_j}^{a_{j+1}-1} \binom{n}{l} [F(T)]^l [1 - F(T)]^{n-l}}{Pr(J = j)}. \qquad (2)
\end{aligned}$$

2.2 *Single Moments: Computation of $E\left[X_{a_i:n}^{\eta} | \mathbf{a}, J = j\right]$ for $i \le j$*

We can write the conditional expectation as

$$\begin{aligned}
E\left[X_{a_i:n}^{\eta} | \mathbf{a}, J = j\right] = \sum_{k=a_j}^{a_{j+1}-1} & E\left(X_{a_i:n}^{\eta} | \text{exactly } k \ X\text{'s less than } T, J = j\right) \\
& \times \Pr(\text{exactly } k \ X\text{'s less than } T | \mathbf{a}, J = j). \qquad (3)
\end{aligned}$$

The following result in order statistics will provide a simple way to compute the required probabilities and expected values.

Result 2. Let $X_1, X_2, \ldots, X_n$ be a random sample from an absolutely continuous population with CDF $F(x)$ and PDF $f(x)$, and let $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ denote the order statistics obtained from this sample. Then, the conditional distribution of $X_{i:n}$, given that $X_{k:n} \le T < X_{k+1:n}$ (i.e., exactly k order statistics are smaller than T) for $i \le k$, is the same as the distribution of the ith order statistics in a sample of size k from a population whose distribution is simply $F(x)$ truncated on the right at T.

Note that Result 2 can be obtained directly from Theorem 1 of Iliopoulos and Balakrishnan [9] on the conditional independence of blocked ordered data. Hence, we can express the conditional distribution of $X_{i:n}$, given that $X_{k:n} \le T < X_{k+1:n}$ (i.e., exactly k order statistics are smaller than T) for $i \le k$ as

$$f_{i:n}(x|X_{k:n} \le T < X_{k+1:n}) = \frac{k!}{(i-1)!(k-i)!}\left[\frac{F(x_i)}{F(T)}\right]^{i-1}\left[1-\frac{F(x_i)}{F(T)}\right]^{k-i}\left[\frac{f(x_i)}{F(T)}\right], \quad -\infty < x_i < T < \infty.$$

The ηth conditional moment of the a_ith order statistic, given that exactly k of the lifetimes are smaller than T, $a_i < k$, can be expressed as

$$\begin{aligned} & E\left[X_{a_i:n}^{\eta} \middle| \text{exactly } k \ X\text{'s less than } T, \mathbf{a}, J=j\right] \\ &= \frac{k!}{(a_i-1)!\,(k-a_i)!}\int_0^T x^{\eta}\frac{f(x)}{F(T)}\left[\frac{F(x)}{F(T)}\right]^{a_i-1}\left[1-\frac{F(x)}{F(T)}\right]^{k-a_i}dx \\ &= \frac{k!}{(a_i-1)!\,(k-a_i)!}\frac{1}{[F(T)]^k}\int_0^T x^{\eta} f(x)\,[F(x)]^{a_i-1}\,[F(T)-F(x)]^{k-a_i}\,dx. \end{aligned} \quad (4)$$

The probability that exactly k of the X's are smaller than T is

$$\Pr(\text{exactly } k \ X\text{'s less than } T|\mathbf{a}, J=j) = \frac{\binom{n}{k}[F(T)]^k[1-F(T)]^{n-k}}{\sum_{l=a_j}^{a_{j+1}-1}\binom{n}{l}[F(T)]^l[1-F(T)]^{n-l}}.$$

Then, from Eq. (3), we can obtain

$$\begin{aligned} & E\left[X_{a_i:n}^{\eta} \middle| \mathbf{a}, J=j\right] \\ &= \sum_{k=a_j}^{a_{j+1}-1}\left[\frac{k!}{(a_i-1)!\,(k-a_i)!}\frac{1}{[F(T)]^k}\right. \\ &\qquad \left.\times\int_0^T x^{\eta} f(x)\,[F(x)]^{a_i-1}\,[F(T)-F(x)]^{k-a_i}\,dx\right] \end{aligned}$$

$$\times \left[\frac{\binom{n}{k} [F(T)]^k [1 - F(T)]^{n-k}}{\sum_{l=a_j}^{a_{j+1}-1} \binom{n}{l} [F(T)]^l [1 - F(T)]^{n-l}} \right]. \tag{5}$$

2.3 Product Moments: Computation of $E\left[X^{\eta}_{a_i:n} X^{\delta}_{a_l:n} | \mathbf{a}, J = j\right]$ for $i < l \leq j$

Following the idea used in Eq. (3), the product moments of the a_ith and the a_lth order statistics can be expressed as

$$\begin{aligned} & E\left[X^{\eta}_{a_i:n} X^{\delta}_{a_l:n} | \mathbf{a} = (a_1, a_2, \ldots, a_m), J = j\right] \\ &= \sum_{k=a_j}^{a_{j+1}-1} E\left(X^{\eta}_{a_i:n} X^{\delta}_{a_l:n} | \text{exactly } k \ X\text{'s less than } T, J = j\right) \\ &\quad \times \Pr\left(\text{exactly } \ k \ X\text{'s less than } T | \mathbf{a}, J = j\right). \end{aligned} \tag{6}$$

For a fixed value of k, where k is the number of X's less than T with $a_i < a_l \leq k$, we have the conditional product moment of the a_ith and the a_lth order statistic, given that exactly k of the lifetimes are smaller than T,

$$\begin{aligned} & E\left(X^{\eta}_{a_i:n} X^{\delta}_{a_l:n} \middle| \text{ exactly } k \ X\text{'s less than } T, \mathbf{a}, J = j\right) \\ &= \frac{k!}{(a_i - 1)!\,(a_i - a_l - 1)!\,(k - a_l)!} \frac{1}{[F(T)]^k} \\ &\quad \times \left\{ \int_0^T \int_0^{x_l} x_i^{\eta} x_l^{\delta} f(x_i) f(x_l) [F(x_i)]^{a_i - 1} \right. \\ &\qquad \left. \times [F(x_l) - F(x_i)]^{a_l - a_i - 1} [F(T) - F(x_l)]^{k - a_l} \, dx_i dx_l \right\}. \end{aligned}$$

Therefore, for $a_i < a_l \leq k$,

$$\begin{aligned} & E\left[X^{\eta}_{a_i:n} X^{\delta}_{a_l:n} \middle| \mathbf{a} = (a_1, a_2, \ldots, a_m), J = j \right] \\ &= \sum_{k=a_j}^{a_{j+1}-1} \frac{k!}{(a_i - 1)!\,(a_i - a_l - 1)!\,(k - a_l)!} \frac{1}{[F(T)]^k} \\ &\quad \times \left\{ \int_0^T \int_0^{x_l} x_i^{\eta} x_l^{\delta} f(x_i) f(x_l) [F(x_i)]^{a_i - 1} \right. \end{aligned}$$

$$\times \left[F(x_l) - F(x_i)\right]^{a_l - a_i - 1} \left[F(T) - F(x_l)\right]^{k - a_l} dx_i dx_l \Bigg\}$$

$$\times \left[\frac{\binom{n}{k} \left[F(T)\right]^k \left[1 - F(T)\right]^{n-k}}{\sum_{l=a_j}^{a_{j+1}-1} \binom{n}{l} \left[F(T)\right]^l \left[1 - F(T)\right]^{n-l}} \right]. \tag{7}$$

For a fixed value of k and $a_i \leq k < a_l$, it can be shown that $X_{a_i:n}$ and $X_{a_l:n}$ are independent [9] and hence the covariance of $X_{a_i:n}$ and $X_{a_l:n}$ is equal to 0. Then, the product moment of the a_ith and the a_lth order statistics can be written as the product of two single moments for a fixed value of k where $a_i \leq k < a_l$, i.e.,

$$\begin{aligned} & E\left[X_{a_i:n}^{\eta} X_{a_l:n}^{\delta} \middle| \text{exactly } k \text{ } X\text{'s less than } T, \mathbf{a}, J = j\right] \\ &= E\left[X_{a_i:n}^{\eta} \middle| \text{exactly } k \text{ } X\text{'s less than } T, \mathbf{a}, J = j\right] \\ &\quad \times E\left[X_{a_l:n}^{\delta} \middle| \text{exactly } k \text{ } X\text{'s less than } T, \mathbf{a}, J = j\right]. \end{aligned} \tag{8}$$

Therefore, the product moments of an order statistic observed before time T and an order statistic observed after time T can be obtained from Eq. (5).

2.4 Main Results

Substituting Eqs. (2), (5) and (7) into Eq. (1), given $J = j$, the conditional single and product moments of progressively censored order statistics with a time constraint for $i < l \leq j$, $j = 1, 2, \ldots, m$, are

$$\begin{aligned} & E\left(X_{i:m:n}^{\eta} | J = j\right) \\ &= \sum_{\mathbf{a}} \Pr\left(\mathbf{a} = (a_1, a_2, \ldots, a_m) | J = j\right) E\left[X_{a_i:n}^{\eta} | \mathbf{a}, J = j\right] \\ &= \frac{1}{\Pr\left(J = j\right)} \sum_{\mathbf{a}} P_{\mathbf{a}} \Bigg\{ \sum_{k=a_j}^{a_{j+1}-1} \frac{k!}{(a_i - 1)!\,(k - a_i)!} \frac{1}{\left[F(T)\right]^k} \\ &\quad \times \int_0^T x^{\eta} f(x) \left[F(x)\right]^{a_i - 1} \left[F(T) - F(x)\right]^{k - a_i} dx \\ &\quad \times \binom{n}{k} \left[F(T)\right]^k \left[1 - F(T)\right]^{n-k} \Bigg\} \end{aligned}$$

and

$$E\left(X_{i:m:n}^{\eta} X_{l:m:n}^{\delta} | J = j\right)$$

$$= \sum_{\mathbf{a}} \Pr(\mathbf{a} = (a_1, a_2, \ldots, a_m) \,|J = j)\, E\left[X^{\eta}_{a_i:n} X^{\eta}_{a_l:n} | \mathbf{a}, J = j\right]$$

$$= \frac{1}{\Pr(J = j)} \sum_{\mathbf{a}} P_{\mathbf{a}} \Bigg\{ \sum_{k=a_j}^{a_{j+1}-1} \frac{k!}{(a_i - 1)!\,(a_i - a_l - 1)!\,(k - a_l)!} \frac{1}{[F(T)]^k}$$

$$\times \Bigg[\int_0^T \int_0^{x_l} x_i^{\eta} x_l^{\delta} f(x_i) f(x_l) [F(x_i)]^{a_i - 1}$$

$$\times [F(x_l) - F(x_i)]^{a_l - a_i - 1} [F(T) - F(x_l)]^{k - a_l} \, dx_i dx_l \Bigg]$$

$$\times \binom{n}{k} [F(T)]^k [1 - F(T)]^{n-k} \Bigg\}, \quad i < l \le j,$$

respectively. For $i > j$, the conditional single and product moments of the progressively censored order statistics can be obtained from the results for progressive censored order statistics (see, e.g., Balakrishnan and Aggarwala [4]) with the fact that the distribution of $X_{i:m:n}$ $(i > j)$ is the same as the distribution of the $(i - j)$th progressively censored order statistic with sample size $n^* = n - j - \sum_{l=1}^{j} R_l$, effective sample size $m^* = m - j$ and progressive censoring scheme $(R_{j+1}, R_{j+2}, \ldots, R_m)$ from a left-truncated distribution at T.

3 Illustrations

We now illustrate the calculation of the conditional moments when the underlying distribution of $X_1, X_2, \ldots, X_n$ is the uniform or exponential distribution.

3.1 Uniform Distribution

Suppose that the underlying distribution of the i.i.d. variables $X_1, X_2, \ldots, X_n$ is the uniform(0, 1) distribution (denoted $U(0, 1)$) with PDF $f(x) = 1$ and CDF $F(x) = x$ for $0 < x < 1$. $X_{a_i:n}$ will follow a beta distribution with parameter a_i and $n - a_i + 1$. With the time constraint T, Eq. (4) can be simplified as:

$$E\left(X^{\eta}_{a_i:n} | \text{exactly } k\ X'\text{s} < T, \mathbf{a}, J = j\right) = \left[\frac{\prod_{\nu=0}^{\eta-1} (a_i + \nu)}{\prod_{\nu=0}^{\eta-1} (k + 1 + \nu)} \right] T^{\eta}.$$

Hence, the conditional ηth moment of the ith progressively censored order statistic, given $J = j$ $(i \le j)$, is

$$
\begin{aligned}
&E\left(X_{i:m:n}^{\eta}|J=j\right)\\
&=\frac{1}{\Pr\left(J=j\right)}\\
&\quad\times\sum_{\mathbf{a}}P_{\mathbf{a}}\left\{\left[\sum_{k=a_j}^{a_{j+1}-1}\left(\frac{\prod_{\nu=0}^{\eta-1}\left(a_i+\nu\right)}{\prod_{\nu=0}^{\eta-1}\left(k+1+\nu\right)}\right)\binom{n}{k}T^{k+\eta}\left(1-T\right)^{n-k}\right]\right\}. \qquad (9)
\end{aligned}
$$

For $a_i < a_l \leq k$, we have

$$
\begin{aligned}
&E\left(X_{a_i:n}^{\eta}X_{a_l:n}^{\delta}|\text{exactly } k\ X'\text{s} < T, \mathbf{a}, J=j\right)\\
&=\left[\frac{\prod_{\nu=0}^{\eta-1}\left(a_i+\nu\right)\prod_{\nu=0}^{\delta-1}\left(a_l+\eta+\nu\right)}{\prod_{\nu=0}^{\eta+\delta-1}\left(k+1+\nu\right)}\right]T^{\eta+\delta}.
\end{aligned}
$$

Hence, the conditional (η, δ)th moment of the ith and lth progressively censored order statistics, given $J = j$, is

$$
\begin{aligned}
&E\left(X_{i:m:n}^{\eta}X_{l:m:n}^{\delta}|J=j\right)\\
&=\frac{1}{\Pr\left(J=j\right)}\\
&\quad\sum_{\mathbf{a}}P_{\mathbf{a}}\left\{\sum_{k=a_j}^{a_{j+1}-1}\left(\frac{\prod_{\nu=0}^{\eta-1}\left(a_i+\nu\right)\prod_{\nu=0}^{\delta-1}\left(a_l+\eta+\nu\right)}{\prod_{\nu=0}^{\eta+\delta-1}\left(k+1+\nu\right)}\right)\binom{n}{k}T^{k+\eta+\delta}\left(1-T\right)^{n-k}\right\},
\end{aligned}
$$

for $i < l \leq j$.

3.2 Exponential Distribution

Consider the exponential distribution with PDF $f(x;\theta) = \theta^{-1}e^{-x/\theta}$ and CDF $F(x;\theta) = 1 - e^{-x/\theta}$ for $0 < x < \infty$, $\theta > 0$. With the time constraint T and given $J = j$, the conditional ηth moment of the ith progressively censored order statistic when the underlying distribution is exponential with mean θ can be written as

$$
\begin{aligned}
&E\left(X_{i:m:n}^{\eta}|J=j\right)\\
&=\frac{1}{\Pr\left(J=j\right)}\sum_{\mathbf{a}}\left\{\frac{k!}{(a_i-1)!(k-a_i)!}\right.\\
&\quad\int_0^1\left\{-\theta\ln\left[1-F(T;\theta)u\right]\right\}^{\eta}u^{a_i-1}(1-u)^{k-a_i}du\\
&\quad\left.\times\binom{n}{k}\left[F(T;\theta)\right]^{k+\eta}\left[1-F(T;\theta)\right]^{n-k}P_{\mathbf{a}}\right\},
\end{aligned}
\tag{10}
$$

for $i \le j$. Similarly, Eq. (1) can be expressed as

$$
\begin{aligned}
&E\left(X_{i:m:n}^{\eta}X_{l:m:n}^{\delta}|J=j\right)\\
&=\frac{1}{\Pr\left(J=j\right)}\sum_{\mathbf{a}}\left\{P_{\mathbf{a}}\left[\sum_{k=a_j}^{a_{j+1}-1}\frac{k!}{(a_i-1)!\,(a_i-a_l-1)!\,(k-a_l)!}\right.\right.\\
&\quad\times\int_0^1\int_0^1\left\{-\theta\ln\left[1-F(T;\theta)uw\right]\right\}^{\eta}\left\{-\theta\ln\left[1-F(T;\theta)u\right]\right\}^{\delta}\\
&\qquad\times u^{a_i-1}(1-u)^{a_l-a_i-1}w^{a_l-1}(1-w)^{k-a_l}dudw\\
&\quad\left.\left.\times\binom{n}{k}\left[F(T;\theta)\right]^{k}\left[1-F\left(T;\theta\right)\right]^{n-k}\right]\right\},
\end{aligned}
\tag{11}
$$

for $i < l \le j$. Note that the integrals involved in Eqs. (10) and (11) depend on the parameter θ. Since these integrals are finite with range in between 0 and 1, for specific values of θ, these integrals can be accurately approximated by using numerical algorithms which are available in commonly used statistical or mathematical software such as R, SAS and Matlab.

For illustrative purposes, we present the conditional means, variances and covariances of the progressively censored order statistics from $U(0, 1)$ given $J = j$, $j = 0, 1, \ldots, 5$ with $n = 10$, $m = 5$, $T = 0.6$ and censoring scheme $(R_1, R_2, R_3, R_4, R_5) = (1, 1, 1, 1, 1)$ in Table 1. For the conditional means and variances of $X_{i:m:n}$, $i > j$, and the covariances of $X_{i:m:n}$ and $X_{l:m:n}$, $j < i < l$, the values can be computed based on the formulae in Balakrishnan and Aggarwala ([4], Sect. 2.3.3). Note that from Eq. (8), the covariance of $X_{i:m:n}$ and $X_{l:m:n}$ $(i \le j < l)$ is 0. For the sake of comparison, the unconditional means, variances and covariances, as well as the probability of $J = j$ $(j = 0, 1, \ldots, 5)$, are also presented in Table 1. From Table 1, we can observe that the conditional and unconditional means, variances and covariances can be very different. Specifically, with the condition that $J = j$ progressively censored order statistics are observed before time $T = 0.6$, the conditional means of $X_{i:m:n}$ must be smaller than $T = 0.6$ for $i < j$ and the conditional means of $X_{i:m:n}$ must be greater than $T = 0.6$ for $i > j$. Moreover, the conditional variances of the progressively censored order statistics are smaller than those unconditional variances, as expected.

Table 1 Conditional means, variances and covariances of the progressively censored order statistics given $J = j$ with $n = 10$, $m = 5$, $T = 0.6$ and censoring scheme $(R_1, R_2, R_3, R_4, R_5) = (1, 1, 1, 1, 1)$

	$J = 0$	$J = 1$	$J = 2$	$J = 3$	$J = 4$	$J = 5$	Without constraint
$E(X_{1:5:10})$	0.636364	0.257143	0.160816	0.116601	0.091369	0.075088	0.090909
$E(X_{2:5:10})$	0.676768	0.644444	0.353469	0.249246	0.192298	0.156495	0.191919
$E(X_{3:5:10})$	0.722944	0.695238	0.657143	0.405581	0.306194	0.246004	0.307359
$E(X_{4:5:10})$	0.778355	0.756191	0.725714	0.680000	0.438710	0.346321	0.445888
$E(X_{5:5:10})$	0.852237	0.837460	0.817143	0.786667	0.733333	0.461807	0.630592
$Var(X_{1:5:10})$	0.001102	0.028163	0.015771	0.009607	0.006390	0.004250	0.006887
$Var(X_{2:5:10})$	0.002188	0.001580	0.021998	0.016369	0.011721	0.008499	0.013672
$Var(X_{3:5:10})$	0.003240	0.003120	0.002449	0.016666	0.014633	0.011490	0.020249
$Var(X_{4:5:10})$	0.004207	0.004557	0.004767	0.004267	0.012954	0.012703	0.026293
$Var(X_{5:5:10})$	0.004833	0.005581	0.006563	0.007822	0.008889	0.010349	0.030204
$Cov(X_{1:5:10}, X_{2:5:10})$	0.000979	0	0.009279	0.007222	0.005268	0.003923	0.006122
$Cov(X_{1:5:10}, X_{3:5:10})$	0.000840	0	0	0.004197	0.003934	0.003221	0.005247
$Cov(X_{1:5:10}, X_{4:5:10})$	0.000672	0	0	0	0.002263	0.002391	0.004198
$Cov(X_{1:5:10}, X_{5:5:10})$	0.000448	0	0	0	0	0.001363	0.002799
$Cov(X_{2:5:10}, X_{3:5:10})$	0.001875	0.001354	0	0.009504	0.008673	0.006977	0.011719
$Cov(X_{2:5:10}, X_{4:5:10})$	0.001500	0.001084	0	0	0.004986	0.005179	0.009375
$Cov(X_{2:5:10}, X_{5:5:10})$	0.001000	0.000722	0	0	0	0.002952	0.006250
$Cov(X_{3:5:10}, X_{4:5:10})$	0.002592	0.002496	0.001959	0	0.008406	0.008525	0.016199
$Cov(X_{3:5:10}, X_{5:5:10})$	0.001728	0.001664	0.001306	0	0	0.004857	0.010799
$Cov(X_{4:5:10}, X_{5:5:10})$	0.002805	0.003038	0.003178	0.002844	0	0.007232	0.017528
$\Pr(J = j)$	0.000105	0.002753	0.028901	0.151732	0.398297	0.418212	

4 Applications

4.1 Least Squares Estimation

Based on the model described in Sect. 2, if $X_{1:m:n} < X_{2:n:m} < \cdots < X_{m:m:n}$ and $J = j$ are the progressively censored order statistics observed and the number of observed failures before time constraint T, respectively, then

$$E[F(X_{i:m:n};\boldsymbol{\theta})|J=j] = E(U_{i:m:n}|J=j)$$

and

$$Var[F(X_{i:m:n};\boldsymbol{\theta})|J=j] = E(U^2_{i:m:n}|J=j) - [E(U_{i:m:n}|J=j)]^2,$$

for $i = 1, 2, \ldots, m$, where $U_{i:m:n}$ is the ith progressively censored order statistic from $U(0, 1)$. These values can be obtained from Eqs. (9) and (10) with the time constraint in the (0, 1) scale as $T^* = F(T;\boldsymbol{\theta})$. Therefore, these expected values and variances depend on the parameter $\boldsymbol{\theta}$ via T^*. The least squares estimator of $\boldsymbol{\theta}$ can then be obtained by minimizing

$$G_{LS}(\boldsymbol{\theta}) = \sum_{i=1}^{m} [F(X_{i:m:n};\boldsymbol{\theta}) - E(U_{i:m:n}|J=j)]^2 \tag{12}$$

with respect to $\boldsymbol{\theta}$. The weighted least squares estimator of $\boldsymbol{\theta}$ can be obtained in a similar manner by using the weight function $w_i = 1/Var[F(X_{i:m:n};\boldsymbol{\theta})|J=j]$ for $i = 1, 2, \ldots, m$, i.e., minimizing

$$G_{WLS}(\boldsymbol{\theta}) = \sum_{i=1}^{m} w_i\,[F(X_{i:m:n};\boldsymbol{\theta}) - E(U_{i:m:n}|J=j)]^2 \tag{13}$$

with respect to $\boldsymbol{\theta}$. Since an analytical solution cannot be obtained, a numerical method such as the Nelder-Mead algorithm is required to compute the least squares estimates.

Here, a numerical example is used to illustrate the least squares estimation method. Adaptive progressive censored order statistics with $n = 10$, $m = 5$ and censoring scheme (1, 1, 1, 1, 1) from the exponential distribution with mean $\theta = 2$ and time constraint $T = 1.8$ are generated. The simulated data is $(X_{1:5:10}, X_{2:5:10}, \ldots, X_{5:5:10})$ = (0.3411787, 0.5997713, 0.7748804, 0.8471616, 1.9351048) with $J = 4$. Conditional on $J = 4$, by using the least squares estimation method, the least squares estimate obtained by minimizing Eq. (12) is 2.02614 and the weighted least squares estimate obtained by minimizing Eq. (13) is 2.24817.

4.2 Approximate Best Linear Unbiased Estimation

Suppose the distribution of the random variable X belongs to the location-scale family of distributions with PDF

$$f(x;\mu,\sigma)=\frac{1}{\sigma}f^*\left(\frac{x-\mu}{\sigma}\right),\quad -\infty<\mu<\infty,\sigma>0,$$

and CDF

$$F(x;\mu,\sigma)=F^*\left(\frac{x-\mu}{\sigma}\right),\quad -\infty<\mu<\infty,\sigma>0,$$

where μ is the location parameter, σ is the scale parameter, and $f^*(\cdot)$ and $F^*(\cdot)$ are respectively the PDF and CDF of the standard ($\mu=0$, and $\sigma=1$) distribution in the location-scale family. Based on a progressively censored sample $\mathbf{X}=(X_{1:m:n},X_{2:m:n},\ldots,X_{m:m:n})$ without time constraint, the best linear unbiased estimators (BLUEs) of μ and σ can be obtained by minimizing the generalized variance (see, for example, Balakrishnan and Aggarwala [4], Sect. 6.2),

$$Q(\mu,\sigma)=(\mathbf{X}-\mu\mathbf{1}-\sigma\boldsymbol{\gamma})'\Gamma^{-1}(\mathbf{X}-\mu\mathbf{1}-\sigma\boldsymbol{\gamma}),\tag{14}$$

with respect to μ and σ, where $\mathbf{1}$ is an $m\times 1$ vector with components all 1's, $\boldsymbol{\gamma}$ is the mean vector of $\mathbf{X}$, and Γ is the variance-covariance matrix of $\mathbf{X}$. The resulting BLUEs are

$$\begin{pmatrix}\tilde{\mu}\\ \tilde{\sigma}\end{pmatrix}=(\mathbf{Y}'\Gamma^{-1}\mathbf{Y})^{-1}(\mathbf{Y}'\Gamma^{-1}\mathbf{X}),\tag{15}$$

where $\mathbf{Y}=[\mathbf{1},\boldsymbol{\gamma}]$. The variance-covariance matrix of $\tilde{\mu}$ and $\tilde{\sigma}$ can be approximated as

$$\begin{pmatrix}Var(\tilde{\mu}) & Cov(\tilde{\mu},\tilde{\sigma})\\ Cov(\tilde{\mu},\tilde{\sigma}) & Var(\tilde{\sigma})\end{pmatrix}\approx\tilde{\sigma}^2(\mathbf{Y}'\Gamma^{-1}\mathbf{Y})^{-1}.$$

Here, we aim to develop the first-order approximation of the conditional BLUEs of μ and σ based on a progressively censored sample with time constraint. Consider the progressively censored order statistics $X_{1:m:n}<X_{2:n:m}<\cdots<X_{m:m:n}$ with censoring scheme $(R_1,R_2,\ldots,R_m)$ and $J=j$ observed progressively censored order statistics before time T (i.e., $X_{j:m:n}\le T<X_{j+1:n:m}$). Let $F^{*-1}(\cdot)$ be the inverse CDF of the standard location-scale distribution. Then, we have

$$X_{i:m:n}=\mu+\sigma F^{*-1}(U_{i:m:n}),\quad i=1,2,\ldots,m,$$

where $U_{1:m:n} < U_{2:n:m} < \cdots < U_{m:m:n}$ are the progressively censored order statistics from $U(0, 1)$ with the censoring scheme $(R_1, R_2, \ldots, R_m)$ and $U_{j:m:n} \leq T^* < U_{j+1:m:n}$ with $T^* = F^*\left(\frac{T-\mu}{\sigma}\right) \in (0, 1)$. Let $W = F^{*-1}(U)$, where $U \sim U(0, 1)$. Given $J = j$, we can express the conditional expectations and conditional covariances of $X_{1:m:n}, X_{2:n:m}, \ldots, X_{m:m:n}$ as

$$E(X_{i:m:n}|J = j) = \mu + \sigma E(W_{i:m:n}|J = j), i = 1, 2, \ldots, m,$$
$$Cov(X_{i:m:n}, X_{l:m:n}|J = j) = \sigma^2 Cov(W_{i:m:n}, W_{l:m:n}|J = j), i = 1, 2, \ldots, m, l = 1, 2, \ldots, m.$$

Using the first-order Taylor series expansion, given $J = j$, we can approximate $W_{i:m:n}$ as

$$W_{i:m:n} \approx F^{*-1}[E(U_{i:m:n}|J = j)] + [U_{i:m:n} - E(U_{i:m:n}|J = j)]\left[\frac{dW_{i:m:n}}{dU_{i:m:n}}\Big|_{U_{i:m:n}=E(U_{i:m:n}|J=j)}\right],$$

$i = 1, 2, \ldots, m$. Then, the conditional expectations and conditional variances and covariances of $X_{1:m:n}, X_{2:n:m}, \ldots, X_{m:m:n}$ can be approximated by

$$E(W_{i:m:n}|J = j) \approx F^{*-1}[E(U_{i:m:n}|J = j)] \triangleq \alpha_i, \tag{16}$$

$$Var(W_{i:m:n}|J = j) \approx Var(U_{i:m:n}|J = j)\left[\frac{dW_{i:m:n}}{dU_{i:m:n}}\Big|_{U_{i:m:n}=E(U_{i:m:n}|J=j)}\right]^2 \triangleq s_{ii}, \tag{17}$$

$$Cov(W_{i:m:n}, W_{l:m:n}|J = j) \approx Cov(U_{i:m:n}, U_{l:m:n}|J = j)\left[\frac{dW_{i:m:n}}{dU_{i:m:n}}\Big|_{U_{i:m:n}=E(U_{i:m:n}|J=j)}\right] \times \left[\frac{dW_{l:m:n}}{dU_{l:m:n}}\Big|_{U_{l:m:n}=E(U_{l:m:n}|J=j)}\right] \triangleq s_{il}, \tag{18}$$

for $i = 1, 2, \ldots, m$ and $l = 1, 2, \ldots, m$. The parameters μ and σ can be estimated by the values that minimize the conditional generalized variance

$$Q^*(\mu, \sigma|J = j) = (\mathbf{X} - \mu\mathbf{1} - \sigma\boldsymbol{\alpha})' \Sigma^{-1} (\mathbf{X} - \mu\mathbf{1} - \sigma\boldsymbol{\alpha}),$$

where

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)' \text{and } \Sigma = \begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1m} \\ s_{21} & s_{22} & \ldots & s_{2m} \\ \vdots & \vdots & & \vdots \\ s_{m1} & s_{m2} & \ldots & s_{mm} \end{pmatrix}.$$

Because the conditional expectations, conditional variances and covariances in Eqs. (16)–(18), respectively, depend on the parameters μ and σ through $T^* = F^*\left(\frac{T-\mu}{\sigma}\right)$, the conditional first-order approximate BLUEs of μ and σ cannot be directly obtained from Eq. (15). Therefore, we propose the use of an iterative procedure to obtain the conditional first-order approximate BLUEs, given $J = j$. Given the current estimates of μ and σ as $\mu^{(h)}$ and $\sigma^{(h)}$, respectively, the $(h+1)$th iteration of the procedure can be described as

1. Compute $T^{*(h)} = F^*\left(\frac{T-\mu^{(h)}}{\sigma^{(h)}}\right)$.
2. Compute the conditional expectations and covariances of $U_{i:m:n}$, $i = 1, 2, \ldots, m$, by fixing $T^* = T^{*(h)}$ and then obtain $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ from Eqs. (16)–(18).
3. The updated estimates of μ and σ can be obtained from Eq. (15) by replacing $\boldsymbol{\gamma}$ and Γ with $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$, respectively:

$$\begin{pmatrix} \tilde{\mu}^{(h+1)} \\ \tilde{\sigma}^{(h+1)} \end{pmatrix} = (\mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y})^{-1}(\mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{X}), \tag{19}$$

   where $\mathbf{Y} = [\mathbf{1}, \boldsymbol{\alpha}]$.
4. Repeat steps 1–3 until convergence occurs.

Here, a numerical example is used to illustrate the computation of the conditional approximate BLUEs. A progressively censored sample with $n = 10$, $m = 5$, censoring scheme (1, 0, 2, 1, 1) from the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, and time constraint $T = 0.3$ is generated. The simulated data is $(X_{1:5:10}, X_{2:5:10}, \ldots, X_{5:5:10})$ = (-0.55, -0.86, -0.03, 0.51, 0.54) with $J = 3$. The iterative procedure is said to have converged when

$$\max(|\tilde{\mu}^{(h+1)} - \tilde{\mu}^{(h)}|, |\tilde{\sigma}^{(h+1)} - \tilde{\sigma}^{(h)}|) < 5 \times 10^{-7}.$$

With the initial values $\mu^{(0)} = 0$ and $\sigma^{(0)} = 1$, the proposed iterative procedure takes 29 iterations to converge to the final conditional approximate BLUEs $\tilde{\mu}^{(29)} = 0.2190709$ and $\tilde{\sigma}^{(29)} = 0.7035756$ with the variance-covariance matrix

$$\begin{bmatrix} Var(\tilde{\mu}) & Cov(\tilde{\mu}^{(29)}, \tilde{\sigma}^{(29)}) \\ Cov(\tilde{\mu}^{(29)}, \tilde{\sigma}^{(29)}) & Var(\tilde{\sigma}^{(29)}) \end{bmatrix} \approx \begin{bmatrix} 0.014 & -0.005 \\ -0.005 & 0.043 \end{bmatrix}$$

5 Summary

In this paper, simple computational formulae for the conditional single and product moments of the progressively censored order statistics with time constraint are presented. These results can be applied in conditional inference of the lifetime data obtained from different hybrid progressive censoring schemes. The conditional statistical estimation methods illustrated in Sect. 4 provide alternatives to the unconditional inference by taking the number of observed progressively censored order statistics

being observed before the time constraint into account. Comparison of the performances of conditional and unconditional inference will be a interesting research topic. Further research on conditional inference based on hybrid progressive censoring schemes and evaluations of their performances is currently in progress and we hope to report these findings in future papers.

Acknowledgments The authors are grateful to the anonymous reviewer for his/her constructive comments which led to this substantial improvement on an earlier version of the paper. This project was supported by The Chinese University of Hong Kong Faculty of Science Direct Grant (Project ID 4053085) and a Grant from the Simons Foundation (#280601 to Tony Ng).

References

1. Arnold, B.C., N. Balakrishnan, and H.N. Nagaraja. 1992. *A first course in order statistics*. New York: Wiley.
2. Bairamov, I., and S. Parsi. 2011. On flexible progressive censoring. *Journal of Computational and Applied Mathematics* 235: 4537–4544.
3. Balakrishnan, N. 2007. Progressive censoring methodology: An appraisal (with discussion). *Test* 16: 211–296.
4. Balakrishnan, N., and R. Aggarwala. 2000. *Progressive censoring: Theory methods and applications*. Boston: Birkhäuser.
5. Balakrishnan, N., and E. Cramer. 2014. *The art of progressive censoring: Applications to reliability and quality*. Boston: Birkhäuser.
6. Balakrishnan, N., N. Kannan, C.T. Lin, and H.K.T. Ng. 2003. Point and interval estimation for the normal distribution based on progressive Type-II censored samples. *IEEE Transactions on Reliability* 52: 90–95.
7. Balakrishnan, N., N. Kannan, C.T. Lin, and J.S. Wu. 2004. Inference for the extreme value distribution under progressive Type-II censoring. *Journal of Statistical Computation and Simulation* 74: 25–45.
8. Cramer, E., and G. Iliopoulos. 2010. Adaptive progressive Type-II censoring. *Test* 19: 342–358.
9. Iliopoulos, G., and N. Balakrishnan. 2009. Conditional independence of blocked ordered data. *Statistics and Probability Letters* 79: 1008–1015.
10. Kundu, D., and A. Joarder. 2006. Analysis of Type-II progressively hybrid censored data. *Computational Statistics and Data Analysis* 50: 2509–2528.
11. Lin, C.T., H.K.T. Ng, and P.S. Chan. 2009. Statistical inference of Type-II progressively hybrid censored data with weibull lifetimes. *Communications in Statistics - Theory and Methods* 38: 1710–1729.
12. Ng, H.K.T., D. Kundu, and P.S. Chan. 2009. Statistical analysis of exponential lifetimes under an adaptive Type-II progressive censoring scheme. *Naval Research Logistics* 56: 687–698.
13. Park, S., H.K.T. Ng, and P.S. Chan. 2015. On fisher information and design of a flexible progressive censored experiment. *Statistics and Probability Letters* 97: 142–149.
14. Ye, Z.S., P.S. Chan, M. Xie, and H.K.T. Ng. 2014. Statistical inference for the extreme value distribution under adaptive Type-II progressive censoring schemes. *Journal of Statistical Computation and Simulation* 84: 1099–1114.

Adaptive Progressive Censoring

Erhard Cramer and George Iliopoulos

Abstract The notion of adaptive progressive Type-II censoring has been introduced in Cramer and Iliopoulos (2010) to analyse data from a progressively Type-II censored life test with observation dependent removals of units. Such a scheme gives more flexibility to the experimenter since it allows him/her to choose the number of units to be removed at each failure time during the life test. In this paper, the idea is generalised to a more general setting of progressive censoring. Our generalised model allows for arbitrary inspection times and possible removals of units during the experiment. The inspection times and removals depend on what has been observed so far. In particular, this approach includes adaptive progressive Type-I and Type-II censoring with random or fixed inspection timepoints.

Keywords Adaptive process · Progressive censoring · Type-I censoring · Type-II censoring · Likelihood inference

1 Introduction

Many variations of the basic idea of progressive censoring have been discussed, dating back to Herd [11] and Cohen [8] (see also [3–5]). In progressively censored life tests, some units are removed during the conduction of the experiment, which means that we do not observe the failure time of any unit. We only know that it has survived up to some censoring time. The censoring procedure is prescribed by the so-called censoring plan $(R_1, \ldots, R_k)$ and the censoring times $T_1 < \cdots < T_k$, meaning that R_j units are withdrawn from the life test at time T_j (if possible). The

E. Cramer (✉)
Institute of Statistics, RWTH Aachen University, 52056 Aachen, Germany
e-mail: erhard.cramer@rwth-aachen.de

G. Iliopoulos
Department of Statistics and Insurance Science, School of Finance and Statistics, University of Piraeus, 80, Karaoli & Dimitriou Str., 18534 Piraeus, Greece
e-mail: geh@unipi.gr

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_5

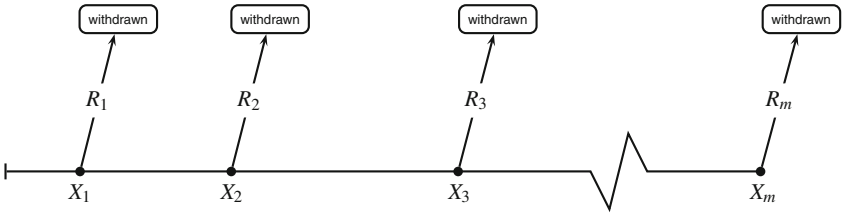


Fig. 1 Generation process of progressively Type-II censored order statistics

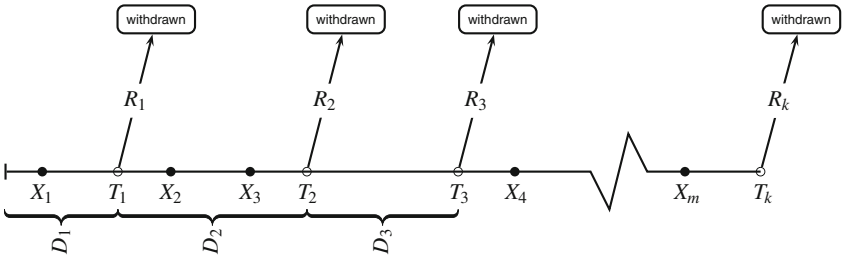


Fig. 2 Generation process of progressively Type-I censored order statistics

standard models of progressive censoring are progressive Type-I and progressive Type-II censoring as first presented in [8]. In these models, the censoring plan is fixed in advance but the censoring times are

- the observed failure times $X_1 \leq \cdots \leq X_m$ in the reduced sample in the case of Type-II censoring (see Fig. 1) and
- prefixed timepoints $T_1 < \cdots < T_k$ in the case of Type-I censoring (see Fig. 2).

Further modifications include hybrid censoring procedures, as introduced in [7] and [13], as well as progressive interval censoring (see [1]). For a review on these models, we refer to the recent monograph [5]. All these approaches have in common that the censoring scheme is fixed in advance. Some attempts have been made to get rid of this restriction. As a first step, progressive censoring with random removals allows some random choice of the scheme by introducing a probability distribution on the set of admissible censoring schemes (see [17]). In the Ng-Kundu-Chan model [16], a threshold parameter T is used to switch from the initially planned scheme to a modified scheme (for details, see Example 1). This idea has been taken up in [2, 12], where some extensions of this basic idea are presented. The general theory of adaptive progressive Type-II censoring is developed in [9].

Roughly speaking, an adaptive progressive censoring scheme is a progressive censoring procedure where the number of withdrawals and/or the time points at which these withdrawals occur can be modified during the experiment depending on what has been observed so far. According to this definition, standard progressive Type-I and Type-II censoring are self-adaptive by their very nature. In progressive Type-II censoring the *times of withdrawals* are random (see Fig. 1) while in progressive

Type-I censoring the *number of withdrawals* depends on what has been observed so far (see Fig. 2). For instance, it is possible that all items are either censored from the experiment or failed before some time $T_j < T_k$.

In this paper, we discuss progressive Type-I and Type-II censoring schemes for which the withdrawals and the times they occur depend on the observations. Moreover, we present a very general approach to adaptive progressive censoring called fully adaptive progressive censoring. This new scheme allows us to choose both the censoring times and the censoring plan. In particular, it comprises both standard progressive censoring procedures. Furthermore, it turns out that this approach gives much flexibility to the experimenter in conducting progressively censored experiments. For instance, as mentioned above, one may face the problem that a progressively Type-I censored experiment terminates without observing a failure time. In this regard, fully adaptive progressive censoring allows us to observe a desired minimum sample size. For illustration, we present two strategies for such life testing procedures which ensure a minimum sample size of the data set. In this regard, fully adaptive progressive censoring contributes to the experimental design of progressively censored life tests.

Finally, it should be mentioned that likelihood and Bayesian inference in all these models is the same as in the case of prefixed censoring times and censoring plans. In particular, this yields explicit representations of the maximum likelihood estimators for exponential lifetimes. But, except for some special cases, the distribution of the estimators will be different and, in most cases, more complicated.

In what follows, we avoid the standard notation $x_{i:m:n}$ for progressively censored order statistics, which will be denoted simply by x_i. The x's (and the corresponding random variable X's) correspond to observations ordered according to their indices:

$$x_1 < x_2 < \cdots .$$

2 Adaptive Progressive Type-II Censoring

The general setup of adaptive progressive Type-II censoring as proposed by [9] considers n items on test. By construction, m failure times will be observed and the censoring scheme is adaptively chosen by the experimenter depending on the history of the experiment. In fact, the censoring number R_j is chosen as a random number depending on the previous failure times $x_1, \ldots, x_j$ and the previous censoring numbers $r_1, \ldots, r_{j-1}$. This process is modelled by probability mass functions (PMF) $g_j(r_j|r_1, \ldots, r_{j-1}, x_1, \ldots, x_j)$. The resulting sample is given by $(X_1, R_1, \ldots, X_m, R_m)$. In detail, this procedure works as follows:

- Observe $X_1 = x_1$ and remove R_1 items where $R_1 \sim g_1(r_1|x_1)$.
- Observe $X_2 = x_2$ and remove R_2 items where $R_2 \sim g_2(r_2|r_1, x_1, x_2)$.

$\vdots$

- Observe $X_j = x_j$ and remove R_j items where $R_j \sim g_j(r_j|r_1, \ldots, r_{j-1}, x_1, \ldots, x_j)$.

$$\vdots$$

- Observe $X_m = x_m$ and remove all remaining $R_m = n - m - \sum_{j=1}^{m-1} R_j$ items.

The supports of distributions $g_1, \ldots, g_{m-1}$ are such that $\sum_{i=1}^{\ell} R_i \leq n - m$ for all ℓ.

Assume now that the n items' lifetimes are iid with probability density function (PDF) f_θ and cumulative distribution function (CDF) F_θ and let $\boldsymbol{x}_j = (x_1, \ldots, x_j)$, $\boldsymbol{r}_j = (r_1, \ldots, r_j)$ for $j = 1, \ldots, m$ (and $\boldsymbol{r}_0 \equiv \emptyset$). Then, the joint PDF of the data (both X's and R's) can be shown to be

$$\left\{ \prod_{j=1}^{m} g_j(r_j|\boldsymbol{r}_{j-1}, \boldsymbol{x}_j) \right\} \left\{ C(\boldsymbol{r}_m) \prod_{j=1}^{m} f_\theta(x_j)\{1 - F_\theta(x_j)\}^{r_j} \right\},$$

where $C(\boldsymbol{r}_m)$ is an appropriately chosen normalizing constant (see [9]). Hence, the likelihood of θ is the same as in the case when the observed progressive censoring scheme $(r_1, \ldots, r_{m-1}, r_m)$ had been fixed in advance (see [5], p. 146/7). The same fact is true for the observed Fisher information matrix,

$$I_{\text{obs}}(\theta) = -\nabla^2 \sum_{j=1}^{m} \Big[\log f_\theta(x_j) + r_j \log\{1 - F_\theta(x_j)\} \Big].$$

This implies that the MLE of θ is obtained as usual. However, its distribution is, in general, difficult to obtain. In cases where this distribution does not depend on the observed censoring plan $\boldsymbol{r}$, it can be explicitly determined (see [9]). For instance, it is shown in [9] for a two-parameter exponential distribution $\mathscr{E}(\mu, \sigma)$ with PDF $f_\theta(x) = \sigma^{-1} e^{-(x-\mu)/\sigma}$, $x > \mu$, $\theta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty)$, that

$$\hat{\mu} = X_1 \sim \mathscr{E}(\mu, \sigma/n), \quad \frac{m\hat{\sigma}}{\sigma} = \frac{1}{\sigma} \sum_{i=1}^{m} (R_i + 1)(X_i - X_1) \sim \chi^2_{2(m-1)}$$

as for progressive Type-II censoring (see [9], [5], p. 146/7). Moreover, the estimators $\hat{\mu}, \hat{\sigma}$ are independent.

As mentioned above, some particular models of adaptive progressive Type-II censoring have been discussed in the literature. For illustration, we briefly describe the censoring strategies in the following examples and explain how they fit into the above approach. At this point, we would like to mention that the model of progressive Type-II censoring with random removals also fits into the previous approach. However, since it does not include the observed failures in the decision rule, we do not really consider it as adaptive. In these procedures, the number of removals R_j in step j is randomly chosen according to a PMF (e.g., a discrete uniform or binomial PMF) whose support is given by $\{0, \ldots, n - \sum_{i=1}^{j-1}(R_i + 1) - m\}$. Further details on these censoring procedures can be found in [5], Sect. 6.2.

Example 1 Ng et al. [16] proposed the first adaptive progressive Type-II censoring scheme called the Ng-Kundu-Chan model. Their procedure starts with a (fixed) progressive censoring scheme

$$\boldsymbol{r}^{\circ}_{m-1} = (r^{\circ}_1, \ldots, r^{\circ}_{m-1})$$

and a (fixed) timepoint $T > 0$. The experiment is conducted as described in the general setup with the one-point PMF

$$g_j(r_j|\boldsymbol{r}_{j-1}, \boldsymbol{x}_j) = \begin{cases} I(r_j = r^{\circ}_j), & \text{if } x_j < T, \\ 0, & \text{if } x_j \geqslant T, \end{cases} \qquad j = 1, \ldots, m-1$$

(and $r_m = n - m - \sum_{i=1}^{m-1} r_i$), where $I(\cdot)$ denotes the indicator function.

Example 2 Bairamov and Parsi [2] generalized the experiment of [16] by considering two (fixed) progressive censoring schemes

$$\boldsymbol{r}^{(1)}_{m-1} = (r^{(1)}_1, \ldots, r^{(1)}_{m-1}) \quad \text{and} \quad \boldsymbol{r}^{(2)}_{m-1} = (r^{(2)}_1, \ldots, r^{(2)}_{m-1})$$

with $r^{(2)}_j \leq r^{(1)}_j$ and (fixed) timepoints $0 < T_1 \leq \cdots \leq T_{m-1}$. Then, they used the one-point PMFs

$$g_j(r_j|\boldsymbol{r}_{j-1}, \boldsymbol{x}_j) = \begin{cases} I(r_j = r^{(1)}_j), & \text{if } x_j < T_j, \\ I(r_j = r^{(2)}_j), & \text{if } x_j \geqslant T_j, \end{cases} \qquad j = 1, \ldots, m-1$$

(and $r_m = n - m - \sum_{i=1}^{m-1} r_i$). Clearly, for $\boldsymbol{r}^{(1)}_{m-1} = \boldsymbol{r}^{\circ}_{m-1}$, $\boldsymbol{r}^{(2)}_{m-1} = (0, \ldots, 0)$, and $T_1 = \cdots = T_{m-1} \equiv T$, this reduces to the adaptive scheme of [16]. Kinaci [12] extended the model of [2] by considering $k+1$ (fixed) progressive censoring schemes,

$$\boldsymbol{r}^{(i)}_{m-1} = (r^{(i)}_1, \ldots, r^{(i)}_{m-1}), \quad i = 1, \ldots, k+1,$$

as well as a set of (fixed) timepoints T_{ij}, $i = 0, 1, \ldots, k, k+1$, $j = 1, \ldots, m-1$, $0 \equiv T_{0j} < T_{1j} < \cdots < T_{kj} < T_{k+1,j} \equiv \infty$. Then, for $j = 1, \ldots, m-1$,

$$g_j(r_j|\boldsymbol{r}_{j-1}, \boldsymbol{x}_j) = I(r_j = r^{(i)}_j), \quad \text{if } T_{i-1,j} < x_j \leq T_{ij},\ i = 1, \ldots, k+1$$

(and $r_m = n - m - \sum_{i=1}^{m-1} r_i$).

3 Adaptive Progressive Type-I Censoring

Nothing has been published so far discussing explicitly adaptive versions of progressive Type-I censoring. This is probably due to the fact that even standard progressive Type-I censoring is in general unattractive since the related distribution theory is intractable in most cases. Recall that in a progressive Type-I censoring scheme, a vector of timepoints $0 < T_1 < \cdots < T_k$ is fixed in advance along with the number of items $R_1^\circ, \ldots, R_{k-1}^\circ, R_k^\circ$ to be withdrawn at these particular timepoints (see Fig. 2). Notice that R_k° is always random since it depends on the number of observed failures.

Let $D_1, \ldots, D_k$ be the number of observed failures within the k time intervals $(T_{i-1}, T_i], i = 1, \ldots, k$, where $T_0 \equiv 0$. Notice that $D_1, \ldots, D_k$ are random counters which may be zero. Since there is a chance of having too many early observations it is clear that it is possible not to end up with the pre-planned censoring scheme. In fact, the number of items which are withdrawn from the experiment at time j equals $R_j = \min\left\{R_j^\circ, n - \sum_{i=1}^{j} D_i - \sum_{i=1}^{j-1} R_i\right\}, j = 1, \ldots, k$. The vector $(R_1, \ldots, R_{k-1}, R_k)$ is called *the effectively applied progressive censoring scheme*.

A general simple adaptive progressive Type-I censoring procedure may be described as follows. Fix k timepoints $T_1 < \cdots < T_k$ and put n items on test.

- Set $D_1 = \#\{X's \le T_1\}$ and $\boldsymbol{X}_1 = (X_1, \ldots, X_{D_1})$ ($\equiv \emptyset$ if $D_1 = 0$). At T_1, remove R_1 items, where $R_1 \sim g_1(r_1|\boldsymbol{x}_1)$.
- Set $D_2 = \#\{X's : T_1 < X \le T_2\}$ and $\boldsymbol{X}_2 = \boldsymbol{X}_1 \cup (X_{D_1+1}, \ldots, X_{D_1+D_2})$ ($\equiv \boldsymbol{X}_1$ if $D_2 = 0$). At T_2 remove R_2 items, where $R_2 \sim g_2(r_2|r_1, \boldsymbol{x}_2)$. Here, $a \cup b$ denotes the concatenation of two vectors a and b.

$\vdots$

- Set $D_k = \#\{X's : T_{k-1} < X \le T_k\}$ and $\boldsymbol{X}_k$ the vector of all $X's$ ($\equiv \emptyset$ if $D_1 = \cdots = D_k = 0$). At T_k, remove all R_k remaining items.

Notice that a formal definition is quite complicated (see [5], pp. 11–12, for standard progressive Type-I censoring). Therefore, we present a demonstrative description of the process only as illustrated in Fig. 2. A rigorous formulation can be given as in Procedure 1.1.7 of [5] by choosing the number of random removals in each step according to the given PMFs.

Let $D = \sum_{i=1}^{k} D_i$ be the number of observed failure times. Then, the joint distribution of the data ($X's$ and $R's$) is given by

$$\left\{\prod_{i=1}^{k-1} g_i(r_i|\boldsymbol{r}_{i-1}, \boldsymbol{x}_i)\right\}\left\{C(\boldsymbol{r}_m, \boldsymbol{d}_m)\left[\prod_{j=1}^{d} f_\theta(x_j)\right]\left[\prod_{i=1}^{k}\{1 - F_\theta(T_i)\}^{r_i}\right]\right\},$$

for $x_1 < \cdots < x_d$ and $r_k = n - \sum_{i=1}^{k} d_i - \sum_{i=1}^{k-1} r_i$. Since the likelihood of θ is proportional to the likelihood when the progressive censoring scheme has been fixed in advance, i.e.,

$$\left[\prod_{j=1}^{d} f_\theta(x_j)\right]\left[\prod_{i=1}^{k}\{1 - F_\theta(T_i)\}^{r_i}\right] \tag{1}$$

(see [5], p. 313), the MLE of the parameter is obtained as usual. However, its distribution cannot be easily determined. Notice that the fixed timepoints $T_1, \ldots, T_k$ cause problems even in the non-adaptive case, as can be seen in [5], p. 215, and [6].

Example 3 For an underlying one-parameter exponential distribution $\mathscr{E}(\theta)$, i.e., $f_\theta(x) = \theta^{-1} e^{-x/\theta}$, $x > 0$, the MLE is given by

$$\hat{\theta} = D^{-1}\left(\sum_{j=1}^{D} X_j + \sum_{i=1}^{k} R_i T_i\right),$$

provided $D \geqslant 1$. Its exact distribution in the non-adaptive case is given in [6] and is a mixture of truncated gamma distributions. Since this distribution depends on the censoring scheme $\boldsymbol{R}$, the distribution is difficult to obtain when the censoring numbers $R_1, R_2, \ldots$ are random except for progressive censoring with random removals, i.e., when $g(r_i|\boldsymbol{r}_{i-1}, \boldsymbol{x}_i) = g(r_i|\boldsymbol{r}_{i-1})$ for all i. In this case, the probability distribution can also be expressed as a mixture of truncated gamma distributions.

Example 4 For the two-parameter exponential distribution $\mathscr{E}(\mu, \sigma)$, we get, provided $D \geqslant 1$,

$$\hat{\mu} = X_1, \quad \hat{\sigma} = D^{-1}\left\{\sum_{j=1}^{D}(X_j - X_1) + \sum_{i=1}^{k} R_i (T_i - X_1)_+\right\},$$

where $(x)_+ = \max(x, 0)$. Notice that the result is the same as in standard progressive Type-I censoring (see [10]). However, here, the joint distribution of $\hat{\mu}$ and $\hat{\sigma}$ depends on $\boldsymbol{R}$ in a nonstandard way unless $g(r_i|\boldsymbol{r}_{i-1}, \boldsymbol{x}_i) = g(r_i|\boldsymbol{r}_{i-1})$ for all i (see Example 3).

Notice that, although the joint CDF of $(\hat{\mu}, \hat{\sigma})$ can be expressed in closed form in the non-adaptive case, exact inference does not work as usual since the distribution of $\hat{\sigma}$ is not stochastically monotone in σ in general. This was noticed in [15] in a different context involving $\mathscr{E}(\mu, \sigma)$, but it is true here, too.

4 Fully Adaptive Progressive Censoring

In a standard progressive Type-I censoring, we fix k timepoints

$$T_1 < \cdots < T_k$$

and the numbers of items to be removed

$$R_1^\circ, \ldots, R_{k-1}^\circ.$$

In fact, the T's are "checkpoints" at which the experimenter plans to remove some (or none) still surviving items. In simple adaptive progressive Type-I censoring, the censoring numbers R_i depend on what has been observed so far, i.e., on $\boldsymbol{R}_{i-1}$ and on $\boldsymbol{X}_i$. Thus, the experimenter can decide during the experiment what to remove. Fully adaptive progressive Type-I censoring goes one step further and allows us to choose the checkpoints depending on the previous observations.

4.1 Description of Fully Adaptive Progressive Censoring

Begin by determining some (possibly random) candidate checkpoint $\widetilde{T}_{10} > 0$. If there are no failures in the interval $(0, \widetilde{T}_{10}]$ set $T_1 = \widetilde{T}_{10}$ and $D_1 = 0$ else at the time of each failure update sequentially $\widetilde{T}_{10}$ to $\widetilde{T}_{11}, \widetilde{T}_{12}, \ldots$, based on some rule. That is, as soon as you observe x_j set

$$\widetilde{T}_{1j} \sim h_{1j}(\widetilde{T}_{1j} | x_1, \ldots, x_j, \widetilde{T}_{10}, \widetilde{T}_{11}, \ldots, \widetilde{T}_{1,j-1}), \quad \widetilde{T}_{1j} > x_j, \tag{2}$$

where $h_{1j}(\cdot | x_1, \ldots, x_j, \widetilde{T}_{10}, \widetilde{T}_{11}, \ldots, \widetilde{T}_{1,j-1})$ denotes a (known) PDF. Notice that $\widetilde{T}_{1j}$ need not be larger nor smaller than $\widetilde{T}_{1,j-1}$. Let now

$$D_1 = \arg\min_j \left\{ \widetilde{T}_{1j} : \text{there are no failures in the time interval } (x_j, \widetilde{T}_{1j}] \right\}$$

and $T_1 = \widetilde{T}_{1d_1}, \boldsymbol{x}_1 = (x_1, \ldots, x_{d_1})$. After having observed T_1, remove R_1 items where

$$R_1 \sim g_1(r_1 | T_1, \boldsymbol{x}_1).$$

Notice that the contribution to the joint density of what has been observed so far will be

$$h_{10}(\widetilde{T}_{10} | T_0 \equiv 0) \times \left\{ \prod_{j=1}^{d_1} h_{1j}(\widetilde{T}_{1j} | \widetilde{T}_{10}, \ldots, \widetilde{T}_{1,j-1}, x_1, \ldots, x_j) \right\} I(T_1 = \widetilde{T}_{1d_1})$$
$$\times\ g_1(r_1 | T_1, \boldsymbol{x}_1) \times \frac{n!}{(n-d_1)!} \left\{ \prod_{j=1}^{d_1} f_\theta(x_j) \right\} \{1 - F_\theta(T_1)\}^{r_1}.$$

Determine now some new candidate checkpoint $\widetilde{T}_{20} > T_1$. If there are no failures in $(T_1, \widetilde{T}_{20}]$ set $T_2 = \widetilde{T}_{20}$ and $D_2 = 0$ else at the time of each failure "update" sequentially $\widetilde{T}_{20}$ to $\widetilde{T}_{21}, \widetilde{T}_{22}, \ldots$. That is, as soon as you observe x_{d_1+j}, $j \geqslant d_1 + 1$, set

$$\widetilde{T}_{2j} \sim h_{2j}(\widetilde{T}_{2j}|\boldsymbol{x}_1, x_{d_1+1} \ldots, x_{d_1+j}, T_1, \widetilde{T}_{20}, \widetilde{T}_{21}, \ldots, \widetilde{T}_{2,j-1}),\ \widetilde{T}_{2j} > x_{d_1+j}$$

by similarity with (2). Let

$$D_2 = \arg\min_j \left\{\widetilde{T}_{2j} : \text{there are no failures in the time interval } (x_{d_1+j}, \widetilde{T}_{2j}]\right\}$$

and $T_2 = \widetilde{T}_{2d_2}$, $\boldsymbol{x}_2 = \boldsymbol{x}_1 \cup (x_{d_1+1}, \ldots, x_{d_1+d_2})$. After having determined T_2, remove R_2 items where

$$R_2 \sim g_2(r_2|T_1, T_2, \boldsymbol{x}_2, r_1).$$

Proceed similarly in order to obtain

$$(T_3, D_3, \boldsymbol{X}_3, R_3), \ldots, (T_k, D_k, \boldsymbol{X}_k, R_k).$$

Notice that k can be either fixed in advance or random as well. For instance, it can be the first checkpoint T_i that reaches or exceeds some predetermined timepoint T. The procedure terminates with probability one if it is reasonably planned. The joint density of the data has the form

$$\prod_{i=1}^{k}\left[\left\{\prod_{j=0}^{d_i} h_{ij}(\widetilde{T}_{ij}|\text{everything observed so far})\right\} I(T_i = \widetilde{T}_{id_i})\right]$$
$$\times \left\{\prod_{j=1}^{k} g_j(r_j|T_1, \ldots, T_j, \boldsymbol{x}_j, \boldsymbol{r}_{j-1})\right\} \times \left\{C(\boldsymbol{r}_k, \boldsymbol{d}_k)\left[\prod_{j=1}^{d} f_\theta(x_j)\right]\left[\prod_{i=1}^{k}\{1 - F_\theta(T_i)\}^{r_i}\right]\right\},$$

for $d = \sum_{i=1}^{k} d_i, x_1 < \cdots < x_d, 0 = T_0 < T_1 < \cdots < T_k, 0 \le r_\ell \le n - \sum_{i=1}^{\ell} d_i - \sum_{i=1}^{\ell-1} r_i$ and $r_k = n - \sum_{i=1}^{k} d_i - \sum_{i=1}^{k-1} r_i$. Notice that, once more, the likelihood of θ is proportional to the likelihood when both the censoring times $(T_1, \ldots, T_k)$ and the censoring numbers $(R_1, \ldots, R_{k-1})$ had been fixed in advance (see Eq. (1)).

Example 5 In order to illustrate the fully adaptive procedure, we present an artificial example. An illustration of the adaption process is depicted in Fig. 3.

Suppose that we put $n = 15$ items on test and we choose $\widetilde{T}_{10} = 0.5$ as a first inspection time.

- Let the first failure time be $x_1 = 0.40$. Then, for whatever reason, we 'shift' the originally planned inspection time $\widetilde{T}_{10} = 0.5$ to $\widetilde{T}_{11} = 0.75$ and proceed with the monitoring of the life test.
- The next failure is observed at $x_2 = 0.60$. Then, we shift the checkpoint again but to the left: $\widetilde{T}_{12} = 0.65$.
- Since that time no further failures occur. Thus, we set $T_1 = 0.65$ to be the first censoring time. Using the inputs $d_1 = 2, \boldsymbol{x}_1 = (0.40, 0.60)$, a random generation of the censoring number R_1 is conducted according to the assumed PMF $g(\cdot|x_1)$. Suppose that its outcome is given by $r_1 = 5$ so that $r_1 = 5$ surviving items are removed at $T_1 = 0.65$. Then, the next inspection time is defined as $\widetilde{T}_{20} = 0.90$.

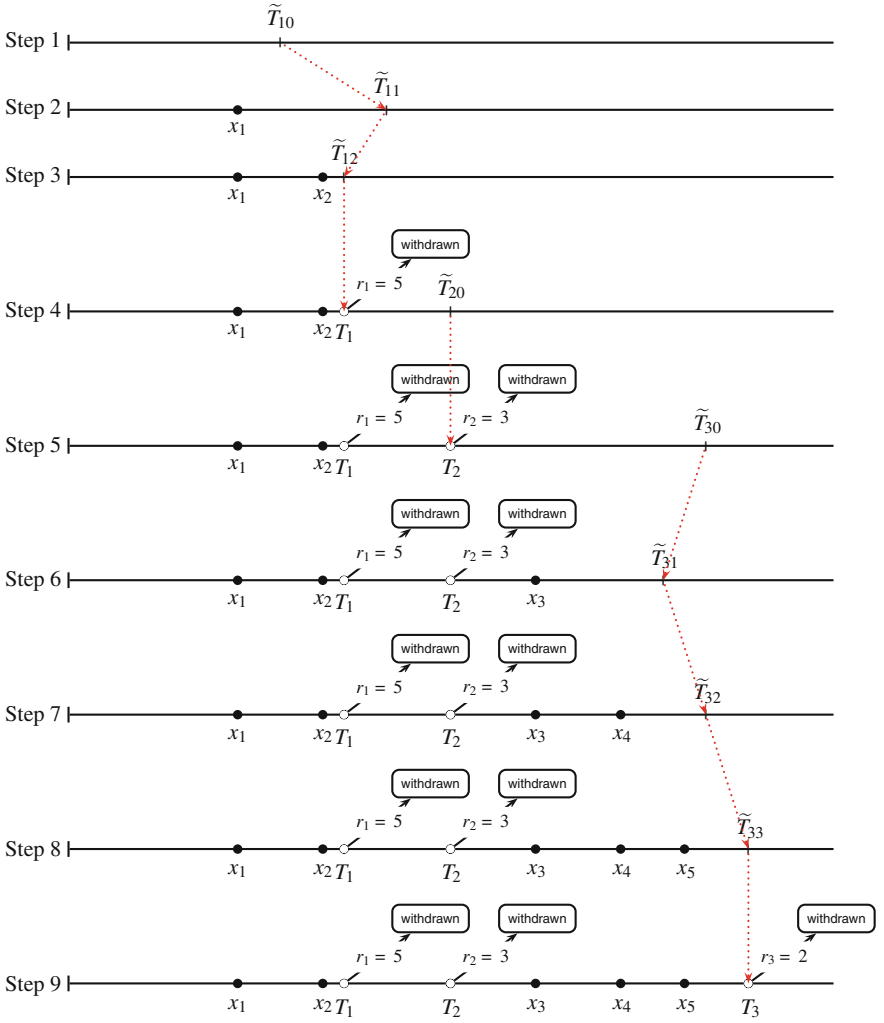


Fig. 3 A fully adaptive progressive censoring procedure as described in Example 5. The *red dotted lines* represent the shifts of the inspection times. They illustrate how the experimenter may change the inspection times depending on the observed failures (represented by •)

- In the interval (0.65, 0.90] no failures occur. Thus, we set $T_2 = 0.90$. With $d_2 = 0$, $\boldsymbol{x}_2 = (0.40, 0.60)$, the random procedure to generate R_2 is started again. Let $r_2 = 3$ and remove another $r_2 = 3$ surviving items. As above, we fix the next checkpoint $\widetilde{T}_{30} = 1.50$ and continue.

- At time $x_3 = 1.10$, another failure occurs and we 'move' the next checkpoint to $\widetilde{T}_{31} = 1.40$. After observing the fourth failure time $x_4 = 1.30$, we 'move' the checkpoint again to $\widetilde{T}_{32} = 1.50$. Then, we observe $x_5 = 1.45$ and we 'move' the inspection time to $\widetilde{T}_{33} = 1.60$. Since no further failures occur until 1.60, we set $T_3 = 1.60$. Then, with $d_3 = 3$, $\boldsymbol{x}_3 = (0.40, 0.60, 1.10, 1.30, 1.45)$, we again toss the dice. This yields $r_3 = 2$ so that all remaining surviving items are withdrawn and the experiment terminates.

Based on the observed data, the likelihood becomes proportional to

$$f_\theta(0.40)\, f_\theta(0.60)\, f_\theta(1.10)\, f_\theta(1.30)\, f_\theta(1.45) \\ \times \{1 - F_\theta(0.65)\}^5 \{1 - F_\theta(0.90)\}^3 \{1 - F_\theta(1.60)\}^2.$$

Remark 1 The above construction gives rise to the following remarks:

1. In the above procedure, it is possible to choose the inspection times as failure times, i.e., $\widetilde{T}_{ij} = x_\ell$ for some i, j, ℓ. This will lead us to a combination of progressive Type-I and Type-II censoring.
2. In particular, if $T_i = X_i, i = 1, \ldots, m$, is fixed in advance, then we get a progressively Type-II censored sample. Thus, progressive Type-II censoring is a special case of fully adaptive progressive Type-I censoring.
3. If we introduce some threshold $T > 0$ and put $\widetilde{T}_{i0} = \min\{x_i, T\}$ for all i, then we have an adaptive hybrid progressive censoring scheme.
4. Based on the above remarks, the fully adaptive progressive censoring scheme described earlier is very general since it has as special cases conventional and progressive Type-I and Type-II censoring schemes as well as their hybrid versions.
5. In the construction process of fully adaptive progressive censoring, it is assumed that the probability mass functions $g_j(r_j|T_1, \ldots, T_j, \boldsymbol{x}_j, \boldsymbol{r}_{j-1})$ as well as the densities $h_{ij}(\widetilde{T}_{ij}|\text{everything observed so far})$ do not depend on the parameter θ. They depend only on the past in the sense that they include information about the inspection times $T_1, \ldots, T_j$, the observed failure time $\boldsymbol{x}_j$, and the employed previous censoring numbers $\boldsymbol{r}_{j-1}$. For instance, since the parameter θ is unknown and g governs the generation of the current censoring numbers, g must not depend on θ. Otherwise, it will not be possible to select a censoring scheme. Of course, technically, we can allow g to depend on θ in which case the likelihood would be more complicated.
   However, one may think of sequential procedures which measure the information in the observations to define a stopping rule. For instance, we may consider estimates of the information about θ included in $T_1, \ldots, T_j, \boldsymbol{x}_j, \boldsymbol{r}_{j-1}$ or perform statistical tests as in sequential analysis (see, e.g., [14]). Depending on the outcome of this procedure, we may stop or continue the experiment. However, this will lead to the same approach where, for example, g is replaced by a different probability mass function $\widetilde{g}$.
6. Since, under fully adaptive progressive censoring, the likelihood of θ is the same as in the case of standard Type-I progressive censoring, Bayesian inference would work as usual, too. For instance, if $\pi(\theta)$ is some prior distribution for θ, then the

corresponding posterior distribution will be

$$\pi(\theta|\text{data}) \propto \pi(\theta)\Big[\prod_{j=1}^{d} f_\theta(x_j)\Big]\Big[\prod_{i=1}^{k}\{1 - F_\theta(T_i)\}^{r_i}\Big],$$

i.e., it will coincide with the posterior density we would have obtained based on a fixed Type-I progressive censoring scheme.

To get an impression of the flexibility of fully adaptive progressive censoring, we present two further examples not covered by the standard models.

Example 6 First, we introduce a progressively censored experiment with a random number of removals at random timepoints. Suppose that n items are put on a life test and that, as the experiment evolves, we have to withdraw some items at certain (but not predetermined) timepoints due to customers' demands. The demands $R_1, R_2, \ldots$ arrive at random times $T_1, T_2, \ldots$. The experiment terminates either at the last observed failure or at some timepoint T_k due to a demand of all remaining items. In this case the joint density of the X's, T's, R's can be expressed as

$$\omega(T_1, \ldots, T_k, r_1, \ldots, r_{k-1})\Big\{C(\boldsymbol{r}_k, \boldsymbol{d}_k)\Big[\prod_{j=1}^{d} f_\theta(x_j)\Big]\Big[\prod_{i=1}^{k}\{1 - F_\theta(T_i)\}^{r_i}\Big]\Big\},$$

where ω is a product of the conditional PDFs of the T's given R's and R's given T's.

Example 7 The following scenario discusses an approach to determine the checkpoints based on observed failures. An obvious issue in progressive Type-I censoring is the choice of the times $T_1 < \cdots < T_k$ of withdrawals. Of course, these times may be imposed by rules unrelated to the experiment. However, when the experimenter is allowed to choose the timepoints $T_1, T_2, \ldots$, he/she wants them to be meaningful. Moreover, in most cases, in order for the inferential procedures to be valid, a minimum number of observations, m say, must be obtained. Thus, a possible plan might be the following:

Run the experiment until a desired minimum of m failure times $X_1, \ldots, X_m$ has been observed. At X_m, remove R_1 items (so that $T_1 = X_m$) and determine the next checkpoint based on what has been observed so far. Such a construction ensures that enough data will be available for an intended statistical analysis. For instance, such an experimental design may help to avoid conditional inference or the undesired situations where all observations have been censored from the life test. In this regard, progressive Type-I censoring may be modified as follows:

Suppose a progressive Type-I censoring procedure with censoring times $T_1 < \cdots < T_k$ and censoring plan $(R_1^\circ, \ldots, R_{k-1}^\circ)$ is to be conducted. Then, in order to ensure at least m observations, the experimenter may apply the following decision rules:

1. Before starting the progressive Type-I procedure, the experimenter waits for the first m failure times $X_{1:n}, \ldots, X_{m:n}$. Then, the experiment is continued as planned

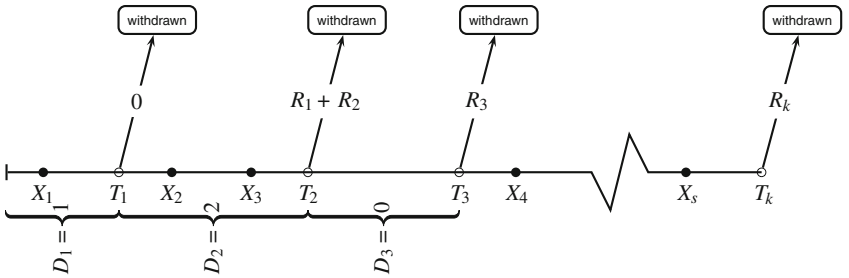


Fig. 4 Generation process of modified progressively Type-I censored order statistics when $D_1 = 1 < m = 2$. The process ensures that a minimum number of $m = 2$ observations will be available for the inferential procedures. Thus, the sample size s satisfies $s \geq m$, as desired

for the checkpoints exceeding $X_{m:n}$. If T_j is the first checkpoint exceeding $X_{m:n}$, the units not withdrawn at the previous timepoints $T_1, \ldots, T_{j-1}$ may be withdrawn at T_j (see Fig. 4 with $m = 2$, $D_1 = 1$).
Alternatively, the experimenter may decide to keep them in the experiment and to remove them at the termination time T_k (if they have not failed before).

2. The progressive Type-I procedure is carried out as planned but the experimenter ensures that m observations will be available by stopping the removal process immediately when the number of observations falls below a prefixed sample size m.

References

1. Aggarwala, R. 2001. Progressive interval censoring: some mathematical results with applications to inference. *Communications in Statistics - Theory and Methods* 30: 1921–1935.
2. Bairamov, I., and S. Parsi. 2011. On flexible progressive censoring. *Journal of Computational and Applied Mathematics* 235: 4537–4544.
3. Balakrishnan, N. 2007. Progressive censoring methodology: An appraisal (with discussions). *TEST* 16: 211–296.
4. Balakrishnan, N., and R. Aggarwala. 2000. *Progressive censoring: Theory, methods, and applications*. Boston: Birkhäuser.
5. Balakrishnan, N., and E. Cramer. 2014. *The art of progressive censoring. Applications to reliability and quality*. New York: Birkhäuser.
6. Balakrishnan, N., D. Han, and G. Iliopoulos. 2011. Exact inference for progressively Type-I censored exponential failure data. *Metrika* 73: 335–358.
7. Childs, A., B. Chandrasekar, and N. Balakrishnan. 2008. Exact likelihood inference for an exponential parameter under progressive hybrid censoring schemes. In *Statistical models and methods for biomedical and technical systems*, ed. F. Vonta, M. Nikulin, N. Limnios, and C. Huber-Carol, 323–334. Boston: Birkhäuser.
8. Cohen, A.C. 1963. Progressively censored samples in life testing. *Technometrics* 5: 327–329.
9. Cramer, E., and G. Iliopoulos. 2010. Adaptive progressive Type-II censoring. *TEST* 19: 342–358.

10. Cramer, E., and M. Tamm. 2014. On a correction of the scale MLE for a two-parameter exponential distribution under progressive type-I censoring. *Communications in Statistics - Theory and Methods* 43: 4401–4414.
11. Herd, R.G. 1956 Estimation of parameters of a population from a multi-censored sample. Ph.D. thesis, Iowa State College, Ames, Iowa
12. Kinaci, I. 2013. A generalization of flexible progressive censoring. *Pakistan Journal of Statistics* 29: 377–387.
13. Kundu, D., and A. Joarder. 2006. Analysis of Type-II progressively hybrid censored data. *Computational Statistics & Data Analysis* 50: 2509–2528.
14. Lai, T.L. 2001. Sequential analysis: Some classical problems and new challenges (with discussion). *Statist. Sinica* 11: 303–408.
15. Mitra, S., A. Ganguly, D. Samanta, and D. Kundu. 2013. On the simple step-stress model for two-parameter exponential distribution. *Statistical Methodology* 15: 95–114.
16. Ng, H.K.T., D. Kundu, and P.S. Chan. 2009. Statistical analysis of exponential lifetimes under an adaptive Type-II progressive censoring scheme. *Naval Research Logistics* 56: 687–698.
17. Yuen, H.K., and S.K. Tse. 1996. Parameters estimation for Weibull distributed lifetimes under progressive censoring with random removals. *Journal of Statistical Computation and Simulation* 55: 57–71.

Renyi Entropy of Progressively Censored Data

Z.A. Abo-Eleneen and B. Almohaimeed

Abstract In this paper, we discuss the calculation of Renyi entropy in a set of consecutive order statistics (OS) and a set of progressively Type-II censored OS. We propose a useful, but indirect, computational approach for computing the Renyi entropy of consecutive order statistics that simplifies the calculations. Some recurrence relations for the Renyi entropy of a set of consecutive order statistics are also derived to facilitate the Renyi entropy computation using the proposed decomposition. Moreover, an extension of the calculation of Renyi entropy for a set of progressively Type-II censored OS is established. Efficient methods are derived which simplify the computation of the Renyi entropy in both settings.

Keywords Renyi entropy · Progressive censoring · Order statistics · Recurrence relations · Markov chain

1 Introduction

The field of information theory has increased rapidly due to its applications in many areas, including statistical inference [21], signal processing, pattern recognition [17], biomedical engineering [14], statistical mechanics [16] and stochastic processes [13].

Renyi entropy of individual order statistics has been studied in [1], and residual Renyi entropy of order statistics and record values has been studied in [23]. The entropy of both single and consecutive order statistics have been studied in [12, 17, 18, 22]. The entropy of the progressively Type-II censored OS has been studied in

Z.A. Abo-Eleneen (✉)
Department of Mathematics, Faculty of Science, Qassim University, Al-Montazah, 1300, Al Qassim 51431, Saudi Arabia
e-mail: zaher_aboeleneen@yahoo.com

B. Almohaimeed
Department of Mathematics, Faculty of Science, Qassim University, 6644, Al Qassim 51452, Saudi Arabia
e-mail: bsmhiemied@qu.edu.sa

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_6

[3] and [5]. Recently Jomhoori and Yousefzadeh [15] discussed the estimation of residual Renyi entropy under progressive censoring.

Park [17] reported that order statistics can be applied to signal processing as order statistic filters, which include the median filter as a special case. For choosing an appropriate length of an order statistic filter, it will be useful to consider some measures of information in consecutive order statistics. In this paper, we provide the methods for calculating the Renyi entropy in consecutive order statistics. Calculations of the Renyi entropy of a set of consecutive order statistics is relatively more complicated than that of the Renyi entropy of the individual order statistics, which has been studied by Abbasnejad and Arghami [1], because the joint Renyi entropy of consecutive order statistics is an n-dimensional integral.

Let X be an absolutely continuous random variable with a cumulative distribution function (CDF) $F(x)$ and probability density function (PDF) $f(x)$. Shannon entropy, which is a key measure of information, is defined as:

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx. \tag{1}$$

A flexible extension of Shannon entropy was introduced by Renyi [19]. The Renyi entropy of order α, $H^{\alpha}(X)$, of the random variable X is defined by Cover and Thomas [9] as:

$$H^{\alpha}(X) = -\frac{1}{\alpha - 1} \log \int_{-\infty}^{\infty} f^{\alpha}(x) dx, \tag{2}$$

where $\lim_{\alpha \to 1} H^{\alpha}(X) = H(X)$ is the Shannon entropy. Suppose that we have an independent and identically distributed (i.i.d.) random sample of size n and arrange the sample in ascending order such that

$$X_{1:n} < X_{2:n} < \cdots < X_{n:n},$$

where $X_{r:n}$ is the rth order statistic.

Progressive censoring has received considerable attention in recent decades, particularly in reliability analysis. It is a more general censoring mechanism than the traditional Type-I and Type-II censoring. Following Balakrishnan and Aggarwala [6], a sample of progressively Type-II censored OS can be described as follows. Let n units be placed in test at time zero. Immediately following the first failure, R_1 surviving units are removed from the test at random. Then, immediately following the second failure, R_2 surviving units are removed from the test at random. This process continues until, at the time of the mth observed failure, the remaining $R_m = n - R_1 - R_2 - \cdots - R_{m-1} - m$, have all been removed from the experiment, so the life testing stops at the mth failure. The observed failure times $\mathbf{X} = (X_{1:m:n}, \ldots, X_{m:m:n})$ constitute progressive Type II censored OS. The joint Renyi entropy contained in $(X_{1:m:n}, \ldots, X_{i:m:n})$, i.e., the collection of the first i pro-

gressively Type II censored OS, is defined as

$$H^{\alpha}_{1\ldots i:m:n}(X) = \frac{-1}{\alpha-1}\log\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{x_{2:m:n}} f^{\alpha}_{1:m:n,\ldots,i:m:n}(x_{1:m:n},\ldots x_{i:m:n})dx_{1:m:n}\ldots dx_{i:m:n}, \tag{3}$$

where $f_{1:m:n,\ldots,i:m:n}(x_{1:m:n},\ldots,x_{i:m:n})$ is the density function of $(X_{1:m:n},\ldots,X_{i:m:n})$. To our knowledge, the exact values of the joint Renyi entropy $H^{\alpha}_{1\ldots i:m:n}$ have not been previously obtained. In the case of $H^{\alpha}_{1\ldots i:m:n}$, a difficulty arises due to the involvement of an integral over i random variables. We provide a method that simplifies the calculation of $H^{\alpha}_{1\ldots i:m:n}$.

In this paper, we focus on the calculation of the joint Renyi entropy, both in a set of order statistics and in a consecutive set of progressively Type-II censored OS. In Sect. 2, we consider the decomposition of the Renyi entropy of order statistics. In Sect. 3, we derive some recurrence relations for the first r order statistics and consider a dual principle for the Renyi entropy of order statistics. In Sect. 4, we extend the decomposition of the Renyi entropy of order statistics to Renyi entropy in progressively Type-II censored OS and derive a recurrence relation. In Sect. 5, we provide an efficient computational method that avoids the integrals in the calculation of the Renyi entropy in progressively Type-II censored OS. A conclusion is provided in Sect. 6.

2 Decomposition of the Joint Renyi Entropy

Since the Renyi entropy is a measure of uncertainty, it is natural that the total amount of Renyi entropy is decreased if the i.i.d. observations are ordered. The following identity shows the magnitude of this reduction:

Lemma 1

$$H^{\alpha}_{1\ldots:n:n} = nH^{\alpha}_{1:1} - \frac{\log(n!)^{\alpha}}{\alpha-1}. \tag{4}$$

Proof The proof can be obtained directly by using the Renyi entropy definition in (2) and the joint PDF of the ordered sample. □

As order statistics form a Markov chain [11], we have the following results for the conditional Renyi entropy and the mutual information.

Lemma 2

$$H^{\alpha}_{r+1\ldots:n|i\ldots r:n} = H^{\alpha}_{r+1\ldots n|r:n}, i = 1,\ldots,r, \tag{5}$$

$$H^{\alpha}_{r+1\ldots i:n} - H^{\alpha}_{r+1\ldots i:n|r:n} = H^{\alpha}_{r+1:n} - H^{\alpha}_{r+1:n|r:n}, i = r+1,\ldots,n. \tag{6}$$

Proof From the Markov chain property of order statistics, (5) follows directly. Furthermore, (6) can be obtained by using (5) and the symmetry of the mutual information [10],

$$\begin{aligned} H^{\alpha}_{r+1\ldots i:n} - H^{\alpha}_{r+1\ldots i:n|r:n} &= H^{\alpha}_{r+1:m:n} - H^{\alpha}_{r+1:n|r:n} \\ &= H^{\alpha}_{r:n} - H^{\alpha}_{r:n|r+1:n}. \end{aligned}$$

□

Next, we show the decomposition of the Renyi entropy of an ordered sample.

Lemma 3

$$H^{\alpha}_{1\ldots n:n} = H^{\alpha}_{1\ldots r:n} + H^{\alpha}_{r+1\ldots n:n|r:n}. \tag{7}$$

Proof The proof follows by the additive property of the Renyi entropy measure and Lemma 1. □

The following lemma shows that $H^{\alpha}_{i\ldots s:n}$ can be obtained jointly from $H^{\alpha}_{1\ldots s:n}$ and $H^{\alpha}_{1\ldots n:n}$.

Lemma 4

$$H^{\alpha}_{i\ldots r:n} = H^{\alpha}_{1\ldots r:n} + H^{\alpha}_{i\ldots n:n} - H^{\alpha}_{1\ldots n:n}. \tag{8}$$

Proof Since $H^{\alpha}_{r+1\ldots n:n|r:n} = H^{\alpha}_{r+1\ldots n:n|r\ldots n:n}$ we have in view of (7):

$$H^{\alpha}_{1\ldots n:n} = H^{\alpha}_{1\ldots r:n} + H^{\alpha}_{r+1\ldots n:n|r:n}.$$

Hence,

$$H^{\alpha}_{1\ldots n:n} = H^{\alpha}_{1\ldots r:n} + H^{\alpha}_{i\ldots n:n} - H^{\alpha}_{i\ldots r:n}.$$

□

We see from (4) and (7) that the Renyi entropy of r ordered data points $H^{\alpha}_{1\ldots r:n}$ can be obtained from $H^{\alpha}_{r+1\ldots n:n|r:n}$. So we consider $H^{\alpha}_{r+1\ldots n|r:n}$ to study $H^{\alpha}_{1\ldots r:n}$.

Let $(X_{1:n}, \ldots, X_{n:n})$ be an ordered sample. The Renyi entropy in the first i order statistics $(X_{1:n}, \ldots, X_{i:n})$ can be written as

$$H^{\alpha}_{1\ldots i:n} = -\frac{1}{\alpha - 1} \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{x_{2:n}} f^{\alpha}_{1..i:n}(x_{1:n}, \ldots, x_{i:n}) dx_{1:n} \ldots dx_{i:n}, \tag{9}$$

where $f_{1\ldots i:n}(x_{1:n}, \ldots, x_{i:n})$ is the joint PDF of the first i order statistics.

Using the Markov chain property of order statistics, we can obtain the following decomposition for the score function:

$$\log f_{1\ldots n:n} = \log f_{1\ldots i:n} + \log f_{i+1\ldots n:n|i:n},$$

where $f_{i+1\ldots n:n|i:n}$ is the PDF of $(X_{i+1\ldots n:n}, \ldots, X_{n:n})$ given $X_{i:n} = x_i$. The following decomposition follows from the strong additivity of the Renyi entropy,

$$H^{\alpha}_{1\ldots n:n} = H^{\alpha}_{1\ldots i:n} + H^{\alpha}_{i+1\ldots n:n|i:n}$$

where $H^{\alpha}_{i+1\ldots n:n|i:n}$ is the average of the conditional information in $(X_{i+1:n}, \ldots, X_{n:n})$ given $X_{i:n} = x_i$.

On the other hand, in view of the result of David and Nagaraja [11], $f_{i+1\ldots:n|i:n}$ is the joint density of the OS sample of size $(n-i)$, drawn from the parent distribution $f(x)$ and truncated from the left at x_i, with density $\frac{f(x)}{1-F(x_i)}$, $x > x_i$. Therefore, $H^{\alpha}_{i+1\ldots n:n|i:n}$ can be written as the double integral

$$H^{\alpha}_{i+1\ldots:n|i:n} = (n-i)\int_{-\infty}^{\infty} g(w) f_{i:n}(w) dw - \frac{\log((n-i)!)^{\alpha}}{\alpha - 1}, \tag{10}$$

where

$$g(w) = -\frac{1}{\alpha - 1}\log \int_{w}^{\infty} \{\frac{f(x)}{1-F(w)}\}^{\alpha} dx$$

and $f_{i:n}(x)$ is defined by

$$f_{i:n}(x) = \frac{n!}{(r-1)!(n-i)!} F(x)^{i-1} [1-F(x)]^{n-i+1} f(x),$$
$$-\infty < x < \infty,\ 1 \le i \le n.$$

3 Recurrence Relations

We have shown that the Renyi entropy of any consecutive order statistics can be obtained from the entropies of the right- and left-censored data. In this section, we consider $H^{\alpha}_{1\ldots i:n}$, while $H^{\alpha}_{r\ldots n:n}$ will be considered in the next one. Since the Renyi entropy of the complete sample $H^{\alpha}_{1\ldots n:n}$ is available, the Renyi entropy $H^{\alpha}_{1\ldots i:n}$ can be derived from (7) and (10) and $H^{\alpha}_{r\ldots n:n}$ can be derived from the duality of the Renyi entropy of order statistics.

Proposition 1

$$H^{\alpha}_{i+1\ldots n-1|i:n-1} = \frac{(n-i-1)}{n} H^{\alpha}_{i+1\ldots n|i:m:n} + \frac{i}{n} H^{\alpha}_{i+2\ldots n|i+1:n} + C_1(n,i,\alpha), \tag{11}$$

where

$$C_1(n,i,\alpha) = \frac{\alpha}{n(\alpha-1)}\{(n-i)\log(n-i) - \log(n-i)!\}. \tag{12}$$

Proof From (10) we have

$$H^{\alpha}_{i+1\ldots n:n-1|i:n-1} = (n-i-1)\int_{-\infty}^{\infty} g(w) f_{i:n-1}(w) dw - \frac{\log((n-i-1)!)^{\alpha}}{\alpha-1}, \quad (13)$$

on the other hand $f_{i:n-1}(w)$ can be written as Cole [8]

$$f_{i:n-1}(w) = \frac{n-i}{n} f_{i:n}(w) + \frac{i}{n} f_{i+1:n}(w). \quad (14)$$

Combining (13) and (14) proves this result. □

The following proposition shows that the Renyi entropy of the first i OS of sample size $n-1$ can be expressed as a linear combination of the first i and $i+1$ OS of sample size n.

Proposition 2

$$H^{\alpha}_{1\ldots i:n-1} = \frac{(n-i-1)}{n} H^{\alpha}_{1\ldots i:n} + \frac{i}{n} H^{\alpha}_{1\ldots i+1:n} + C_2(n,i,\alpha), \quad (15)$$

where

$$C_2(n,i,\alpha) = \frac{\alpha}{n(\alpha-1)}\Big\{(n-1)\log n - (n-i)\log(n-i) - \log(n-1)! + \log(n-i)!\Big\}.$$

Proof For a sample of size $n-1$, the general decomposition of the Renyi entropy of OS takes the form

$$H^{\alpha}_{1\ldots n-1:n-1} = H^{\alpha}_{1\ldots i:n-1} + H^{\alpha}_{i+1\ldots n-1:n-1|i:n-1}. \quad (16)$$

By applying Proposition 1 to (16), we get

$$H^{\alpha}_{1\ldots n-1:n-1} = H^{\alpha}_{1\ldots i:n-1} + \frac{(n-i-1)}{n} H^{\alpha}_{i+1\ldots n:n|i:n} + \frac{i}{n} H^{\alpha}_{i+2\ldots n:n|i+1:n} + C_1(n,i,\alpha), \quad (17)$$

where C_1 is defined by (12). By using (4) and (7), the expression (17) can be written as:

$$(n-1)H^{\alpha}_{1:1} - \frac{\alpha\log(n-1)!}{\alpha-1} = H^{\alpha}_{1\ldots i:m:n-1} + \frac{n-i-1}{n}\left\{nH^{\alpha}_{1:1} - \frac{\alpha\log(n)!}{\alpha-1} - H^{\alpha}_{1\ldots i:n}\right\} + \frac{i}{n}\left\{nH^{\alpha}_{1:1} - \frac{\alpha\log(n)!}{\alpha-1} - H^{\alpha}_{1\ldots i+1:n}\right\} + C_1(n,R_1\ldots,R_i).$$

After some simplifications Proposition 2 follows. □

Suppose that we have obtained the sequences $H^{\alpha}_{1:i}$ for $i=1,\ldots,n$, then we can use the following recurrence relation to obtain $H^{\alpha}_{1\ldots i:n}$, for $2 \le i \le n$.

Proposition 3

$$H^{\alpha}_{1\ldots i:n} = \sum_{r=n-i+1}^{n} C^{r-2}_{n-r-1} C^{n}_{r} (-1)^{r-n+i-1} H^{\alpha}_{1:r} + C_3(n, i, \alpha), \tag{18}$$

where $C^{n}_{r} = \frac{n!}{r!(n-r)!}$ *and*

$$C_3(n, i, \alpha) = \frac{\alpha}{\alpha - 1} \left\{ \log (n-r)! - \log(n)! + \sum_{r=n-i+1}^{n} C^{r-2}_{n-r-1} C^{n}_{r} (-1)^{r-n+i-1} \log i \right\}.$$

Proof If we use Renyi entropy decomposition in (10), $H^{\alpha}_{i+1\ldots n:n|i:n}$ can be written as

$$H^{\alpha}_{i+1\ldots:n|i:n} = (n-i) \int_{-\infty}^{\infty} g(w) f_{i:n}(w) dw - \frac{\alpha \log(n-i)!}{\alpha - 1}.$$

Using the recurrence relation between the densities of order statistics established by Srikantan [20],

$$f_{i:n} = \sum_{r=n-i+1}^{n} C^{r-2}_{n-r-1} C^{n}_{r} (-1)^{r-n+i-1} f_{1:r}. \tag{19}$$

Then, $H^{\alpha}_{i+1\ldots n:n|i:n}$ can be written as

$$\begin{aligned} H^{\alpha}_{i+1\ldots:n|i:n} &= (n-i) \sum_{r=n-i+1}^{n} C^{r-2}_{n-r-1} C^{n}_{r} (-1)^{r-n+i-1} \int_{-\infty}^{\infty} g(w) f_{1:r}(w) dw \\ &\quad - \frac{\alpha \log(n-i)!}{\alpha - 1} \\ &= (n-i) \sum_{r=n-i+1}^{n} C^{r-2}_{n-r-1} C^{n}_{r} (-1)^{r-n+i-1} \frac{\alpha \log(i-1)!}{\alpha - 1} + H^{\alpha}_{2\ldots:i|1:i} \\ &\quad - \frac{\alpha \log(n-i)!}{\alpha - 1}. \end{aligned}$$

Proposition 3 follows by using (4) and (7) and noting that

$$\sum_{r=n-i+1}^{n} C^{r-2}_{n-r-1} C^{n}_{r} (-1)^{r-n+i-1} r = n.$$

□

3.1 The Dual Principle

Balasubramanian and Balakrishnan [7] established the dual principle for the moments and distributions of order statistics. Park [17], in Lemma 4.1, derived the duality for the entropy of order statistics by considering the mirror image of $f(x)$ about $x = 0$. It is easy to see that Park [17], Lemma 4.1, also satisfies the duality of the Reyni entropy of order statistics. Hence, we can obtain a dual relation to Proposition 2 as

$$H^{\alpha}_{r\ldots n-1:n-1} = \frac{(n-r)}{(n)} H^{\alpha}_{r\ldots n:n} + \frac{r-1}{(n)} H^{\alpha}_{r+1\ldots n:n} + C_2(n,i,\alpha) \tag{20}$$

and a dual relation to Proposition 3 as

$$H^{\alpha}_{r\ldots n:n} = \sum_{i=r}^{n} C^{i-2}_{r-2} C^{n}_{i} (-1)^{i-r} H^{\alpha}_{i:i} + C_3(n,i,\alpha). \tag{21}$$

Thus Proposition 3 and (21) along with (8) show that the Renyi entropy of any set of order statistics can be expressed as a weighted sum of $H^{\alpha}_{1:r}, r = i-s+1,\ldots,n$ and $H^{\alpha}_{r:r}, r = i,\ldots,n$.

Remark 1 For the case when $\alpha \to 1$, the results in Sects. 2 and 3 simplify to corresponding results in Park [17] for the Shannon entropy of consecutive OS.

3.2 Examples

In order to calculate $H^{\alpha}_{1\ldots i:n}$ and $H^{\alpha}_{j\ldots n:n}$, it is enough to calculate $H^{\alpha}_{1:n}$ and $H^{\alpha}_{n:n}$. Note that $H^{\alpha}_{1:n}$ may be written in terms of the hazard rate:

$$H^{\alpha}_{1:n} = -\log n + \frac{1}{\alpha-1} \log \int_{-\infty}^{\infty} f_{1:n}(x) h^{\alpha-1}(x) (1-F(x))^{n(\alpha-1)} dx. \tag{22}$$

Furthermore, Abbasnejad and Arghami [1] used the probability integral transformation to find a representation for $H^{\alpha}_{i:n}$ in a single order statistic $X_{i:n}$. Below, we provide expressions for the exponential and Pareto distributions.

Example 1 For the exponential distribution $f(x) = \theta \exp(-\theta x),\ x > 0,\ \theta > 0$

$$H^{\alpha}_{1:n} = -\log n\theta + \frac{\log \alpha}{\alpha-1},$$

$$H^{\alpha}_{n:n} = -\frac{\alpha}{\alpha-1} \log \theta - \frac{1}{\alpha-1} \log \beta(\alpha, \alpha(n-1)+1).$$

Example 2 For the Pareto distribution $f(x) = \frac{\theta\beta^\theta}{x^\theta},\ x \geq \beta > 0,\ \theta > 0,$

$$H^\alpha_{1:n} = -\frac{\alpha}{\alpha-1}\log n\theta + \frac{1}{\alpha-1}\log \beta(\alpha(n\theta+1)-1), \alpha((n-1)+1).$$

4 Progressively Type-II Censored Order Statistics

The joint Renyi entropy contained in a sample of progressively Type-II censored order statistics, $(X_{1:m:n}, \ldots, X_{m:m:n})$, is defined as

$$H^\alpha_{1\ldots m:m:n}(X) = \frac{-1}{\alpha-1}\log \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f^\alpha_{1:m:n,\ldots,m:m:n}(x_{1:m:n}, \ldots, x_{m:m:n})dx_{1:m:n}\ldots dx_{m:m:n},$$

where the joint density function $f_{1:m:n,\ldots,\ m:m:n}(x_{1:m:n}, \ldots, x_{i:m:n})$ can be written as in [6],

$$f_{1:m:n,\ldots,m:m:n}(x_{1:m:n}, \ldots, x_{m:m:n}) = c\prod_{i=1}^{m} f(x_{i:m:n};\theta)[1-F(x_{i:m:n};\theta)]^{R_i},$$

where $c = n(n-R_1-1)(n-R_1-R_2-2)\cdots(n-R_1-R_2-R_3\ldots-R_{m-1}-m+1)$.

In Lemma 1, we have shown how much the Renyi entropy of an i.i.d. random sample of size n is reduced if it is ordered (see (4)). In view of (4) and noting that a progressive Type II censored OS sample can be seen as an ordered sample,

$$(X_{1:m:n}, \ldots, X_{m:m:n}),$$

with the removals $(R_1, R_2, \ldots, R_m)$, we have the following result for the Renyi entropy of the progressive Type II censored OS.

Lemma 5

$$H^\alpha_{1\ldots m:m:n} = nH_{1:1:1} - \frac{\alpha\log c}{(\alpha-1)}, \tag{23}$$

where $H^\alpha_{1:1:1} = H^\alpha_{1:1} = \frac{-1}{\alpha-1}\log\int_{-\infty}^{\infty} f^\alpha(x)dx.$

We can obtain the following result by taking steps similar to those outlined in Sect. 2.

Lemma 6

$$H^\alpha_{r+1\ldots m:m:n|i\ldots r:m:n} = H^\alpha_{r+1\ldots m:m:n|r:m:n}, i = 1, \ldots, r \tag{24}$$

$$H^\alpha_{r+1\ldots i:m:n} - H^\alpha_{r+1\ldots i:m:n|r:m:n} = H^\alpha_{r+1:m:n} - H^\alpha_{r+1:m:n|r:m:n}, i = r+1, \ldots m. \tag{25}$$

Next, we have the following decomposition of the Renyi entropy of progressively Type-II censored OS:

$$H^{\alpha}_{1\ldots m:m:n} = H^{\alpha}_{1\ldots r:m:n} + H^{\alpha}_{r+1\ldots m:m:n|r:m:n}. \tag{26}$$

From (26) and (23), the Renyi entropy of the first i progressively Type-II censored OS, $H^{\alpha}_{1\ldots i:m:n}$, can be obtained from $H^{\alpha}_{i+1\ldots m:m:n|i:m:n}$. We consider using $H^{\alpha}_{i+1\ldots m:m:n|i:m:n}$ to obtain $H^{\alpha}_{1\ldots i:m:n}$, where $H^{\alpha}_{1\ldots i:m:n}$ is defined by (3). We can obtain $H_{i+1\ldots m:m:n|i:m:n}$ as a double integral calculating $H_{i+1\ldots n:n|i:n}$ using a method similar to the one described in Sect. 2:

$$H^{\alpha}_{i+1\ldots m:m:n|i:m:n} = (n - \sum_{j=1}^{i} R_j - i) \int_{-\infty}^{\infty} g(w) f_{i:m:n}(w) dw - \frac{\log\left((n - \sum_{j=1}^{i} R_j - i)!\right)^{\alpha}}{1-\alpha}, \tag{27}$$

where

$$g(w) = -\frac{1}{\alpha - 1} \log \int_{w}^{\infty} \left\{ \frac{f(x)}{1 - F(w)} \right\}^{\alpha} dx$$

and $f_{i:m:n}(x)$ is defined by

$$f_{i:m:n}(x) = c_{i-1} \sum_{j=1}^{i} a_j(i)(1 - F(x))^{\gamma_j - 1} f(x), \quad -\infty < x < \infty, \ 1 \le i \le m \tag{28}$$

with

$$\gamma_i = n - i + 1 + \sum_{j=i}^{m} R_i, \qquad c_{i-1} = \prod_{j=1}^{i} \gamma_j \qquad 1 \le i \le m$$

and

$$a_j(i) = \prod_{r=1, r\neq j}^{i} \frac{1}{\gamma_r - \gamma_j}, \qquad 1 \le j \le i \le m.$$

Since we have the Renyi entropy $H^{\alpha}_{1\ldots m:m:n}$ of the complete sample, then the Renyi entropy $H^{\alpha}_{1\ldots i:m:n}$ can be now easily derived from (26) and (27).

4.1 Recurrence Relations

Recurrence relations between the CDF (PDF) of progressively Type-II censored OS have been studied by many authors to simplify the calculation of moments of progressively Type-II censored OS. Abo-Eleneen [2] obtained the following recurrence relation between the PDFs of the progressively Type-II censored OS:

$$(m+\sum_{j=1}^{m} R_j) f_{i:m:n-1} = (m-i+\sum_{j=r+1}^{m} R_j) f_{i:m:n} + (i+\sum_{j=1}^{i} R_j) f_{i+1:m:n}. \quad (29)$$

Using (29) and the decomposition of the Renyi entropy in (27), we have the following results for the Renyi entropy in the progressive Type II censoring scheme.

Proposition 4

$$\begin{aligned} H^{\alpha}_{i+1\ldots m:m:n-1|i:m:n-1} &= \frac{(n-\sum_{j=1}^{i} R_j - i)}{(m+\sum_{j=1}^{m} R_j)} H^{\alpha}_{i+1\ldots m:m:n|i:m:n} \\ &+ \frac{(\sum_{j=1}^{i} R_j + i)}{(m+\sum_{j=1}^{m} R_j)} H^{\alpha}_{i+2\ldots m:m:n|i+1:m:n} \\ &+ d_1(n, m, R_1, \ldots, R_i, \alpha), \end{aligned}$$

where $d_1(n, m, R_1, \ldots, R_i, \alpha) =$

$$\frac{\alpha}{n(\alpha-1)}\{(n-\sum_{j=1}^{i} R_j - i)\log(n-\sum_{j=1}^{i} R_j - i) - \log(n-\sum_{j=1}^{i} R_j - i)!\},$$

and $n = \sum_{j=1}^{m} R_j + m$.

Proof The result can be obtained by taking similar steps as in Proposition 1 and using the Renyi entropy decomposition in (27). □

The next proposition shows that the Renyi entropy of the first r progressively Type-II censored OS of sample size $n-1$ can be obtained as a linear combination of the first r and $r+1$ of the progressively Type II censored OS of sample size n.

Proposition 5

$$\begin{aligned} H^{\alpha}_{1\ldots i:m:n-1} &= \frac{(n-\sum_{j=1}^{i} R_j - i - 1)}{(m+\sum_{j=1}^{m} R_j)} H^{\alpha}_{1\ldots i:m:n} \\ &+ \frac{(\sum_{j=1}^{i} R_j + i)}{(m+\sum_{j=1}^{m} R_j)} H^{\alpha}_{1\ldots i+1:m:n} + d_2(n, m, R_1, \ldots, R_i, \alpha), \end{aligned}$$

where

$$d_2(n, m, R_1, \ldots, R_i, \alpha) = \frac{\alpha}{n(\alpha-1)}\{(n-1)\log n + \log(n - \sum_{j=1}^{i} R_j - i)! \\ - \log(n-1)! - (n - \sum_{j=1}^{i} R_j - i)\log(n - \sum_{j=1}^{i} R_j - i)\}.$$

Proof This result can be obtained by taking similar steps as in Proposition 2. □

Remark 2 For $R_1 = R_2 = \cdots = R_m = 0$ all results of Sect. 4 simplify to corresponding results for the Renyi entropy for consecutive OS.

5 Computational Method for Calculating $H^{\alpha}_{1\ldots i:m:n}$

In this section, we provide another approach to simplifying the calculation of the Renyi entropy in a collection of progressively Type-II censored OS. We avoid the integrals in the calculation of $H^{\alpha}_{1\ldots r:m:n}$ in which the computation of the Renyi entropy in a sample of progressively Type-II censored OS simplifies to a summation of Renyi entropy of the smallest OS of varying sample size, $H^{\alpha}_{1:n}$. Using (22), we have the following representation of Renyi entropy in the smallest progressive Type-II censored OS, $X_{1:m:n}$:

$$H^{\alpha}_{1:m:n} = -\log n + \frac{1}{\alpha-1}\log\int_{-\infty}^{\infty} f_{1:m:n}(x)h^{\alpha-1}(x)(1-F(x))^{n(\alpha-1)}dx. \quad (30)$$

Theorem 1 *Let $(X_{1:m:n}, \ldots, X_{m:m:n})$ be a sample of progressively Type-II censored OS with censoring scheme $(R_1, R_2, \ldots, R_m)$. The Renyi entropy in the first i progressively Type-II censored OS $(X_{1:m:n}, \ldots, X_{r:m:n})$ can be written as*

$$H^{\alpha}_{1\ldots r:m:n} = -\log c'(r) + \frac{1}{1-\alpha} \\ \times \sum_{s=1}^{r}\log c'(s)\sum_{i=1}^{s-1}\frac{c_{i,s-1}(R_1+1, \ldots, R_{s-1}+1)}{R'_i}\exp(1-\alpha)(\log R'_i + H^{\alpha}_{1:R'_i}),$$

where, $R'_i = (R^*_s + 1) + \sum_{j=s-i}^{s-1}(R_j + 1)$, $R^*_s = (n - s - R_1 - \cdots - R_{s-1} + 1)$, $c'(t) = n(n - R_1 - 1)\cdots(n - R_1 - \cdots - R_{t-1} - t + 1)$ and

$$c_{i,s}(R_1, \ldots, R_s) = \frac{(-1)^i}{\{\prod_{j=1}^{i}\sum_{k=s-i+1}^{s-i+j} R_k\}\{\prod_{j=1}^{s-i}\sum_{k=j}^{s-i} R_k\}}$$

in which empty products are defined as 1.

Proof By the Markov property of progressively Type-II censored OS, one can write

$$f_{1:m:n,\ldots,r:m:n}(x_{1:m:n},\ldots,x_{r:m:n}) = f_{1:m:n}(x_1) f_{2|1:m:n}(x_2|x_1)\ldots f_{r|r-1:m:n}(x_r|x_{r-1}),$$

where $f_{i+1|i:m:n}(x_{i+1}|x_i)$ is the conditional PDF of $X_{i+1:m:n}$ given $X_{i:m:n} = x_i$. Therefore, we have

$$H^{\alpha}_{1:m:n,\ldots,r:m:n} = H^{\alpha}_{1:m:n} + H_{2|1:m:n} + \cdots + H^{\alpha}_{r|r-1:m:n}, \tag{31}$$

where $H^{\alpha}_{i+1|i:m:n}$ is the expected Renyi entropy in $X_{i+1:m:n}$ given $X_{i:m:n} = x_i$, i.e.,

$$H^{\alpha}_{i+1|i:m:n} = E\{\frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} f^{\alpha}_{i+1|i:m:n}(x|x_{i:m:n})dx\}. \tag{32}$$

The conditional PDF $f_{i+1|i:m:n}(x_{i+1}|x_i)$ is given in [4] as

$$f_{i+1|i:m:n}(x_{i+1} \mid x_i) = (n - R_1 - \cdots - R_i - i)h(x_i)\left(\frac{1-F(x_i)}{1-F(x_{i+1})}\right)^{(n-R_1-\cdots-R_i-i)}, \tag{33}$$

where $h(x_i) = \frac{f(x_i)}{1-F(x_{i+1})}$. We now use (30) and note that, given $X_{i:m:n} = x_i$, $X_{i+1:m:n}$ has the same PDF as the first order statistic from a random sample of size $(n - R_1 - \cdots - R_i - i)$ with PDF $g(x) = \frac{f(x)}{1-F(x_i)}$, $x > x_i$. The expression in (32) can then be written as

$$\begin{aligned} H^{\alpha}_{i+1|i:m:n} &= -\log(n - R_1 - \cdots - R_i - i) \\ &\quad + \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} f_{i+1:m:n}(x)h^{\alpha-1}(x)(1-F(x))^{n(\alpha-1)}dx_i. \end{aligned} \tag{34}$$

Thus by using (31) and (34), $H^{\alpha}_{1\ldots r:m:n}$ can be expressed as a summation of a single integral as

$$H^{\alpha}_{1\ldots r:m:n} = -\log c'(r) + \frac{1}{1-\alpha}\sum_{i=1}^{r}\log\int_{-\infty}^{\infty} h^{\alpha-1}(x)(1-F(x))^{n(\alpha-1)} f_{i:m:n}(x)dx, \tag{35}$$

where $c'(r)$ is defined above and $f_{s:m:n}$ is defined by (28). We can also express (28) as

$$f_{s:m:n} = c'(s)\sum_{i=1}^{s-1}\frac{c_{i,s-1}(R_1+1,\ldots,R_{s-1}+1)}{R'_i} f_{1:R_i}(x_s) \tag{36}$$

where $f_{1:R_i}$ is the smallest order statistic in a sample of size R'_i. If we use (36) and (22) in (35) the result follows. □

Table 1 The Renyi entropy in a collection of order statistics from a sample of progressively Type-II censored order statistics from the exponential distribution

n	m	Censoring scheme	r	Progressively censored OS	Renyi entropy
5	3	(2,0,0)	1	$(X_{1:3:5})$	–0.799
5	3	(2,0,0)	2	$(X_{1:3:5}, X_{2:3:5})$	–0.681
5	3	(2,0,0)	3	$(X_{1:3:5}, X_{2:3:5}, X_{3:3:5})$	0.130
5	3	(0,0,2)	1	$(X_{1:3:5})$	–0.799
5	3	(0,0,2)	2	$(X_{1:3:5}, X_{2:3:5})$	–0.374
5	3	(0,0,2)	3	$(X_{1:3:5}, X_{2:3:5}, X_{3:3:5})$	–1.662
5	3	(1,1,0)	1	$(X_{1:3:5})$	–0.799
5	3	(1,1,0)	2	$(X_{1:3:5}, X_{2:3:5})$	–1.086
5	3	(1,1,0)	3	$(X_{1:3:5}, X_{2:3:5}, X_{3:3:5})$	–0.275
5	5	(0,0,0,0,0)	1	$(X_{1:5})$	–0.799
5	5	(0,0,0,0,0)	2	$(X_{1:5}, X_{2:5})$	–1.374
5	5	(0,0,0,0,0)	3	$(X_{1:5}, X_{2:5}, X_{3:5})$	–1.662
5	5	(0,0,0,0,0)	5	$(X_{1:5}, \ldots, X_{5:5})$	–0.733
10	5	(5,0,0,0,0)	1	$(X_{1:5:10})$	–1.492
10	5	(5,0,0,0,0)	2	$(X_{1:5:10}, X_{2:5:10})$	–2.067
10	5	(5,0,0,0,0)	3	$(X_{1:5:10}, X_{2:5:10}, X_{3:5:10})$	–2.355
10	5	(5,0,0,0,0)	4	$(X_{1:5:10}, \ldots, X_{4:0:10})$	–0.237
10	5	(5,0,0,0,0)	5	$(X_{1:5:10}, \ldots, X_{5:5:10})$	–1.426
10	5	(0,0,0,0,5)	1	$(X_{1:5:10})$	–1.492
10	5	(0,0,0,0,5)	2	$(X_{1:5:10}, X_{2:5:10})$	–2.878
10	5	(0,0,0,0,5)	3	$(X_{1:5:10}, X_{2:5:10}, X_{3:5:10})$	–4.146
10	5	(0,0,0,0,5)	4	$(X_{1:5:10}, \ldots, X_{4:5:10})$	–5.281
10	5	(0,0,0,0,5)	5	$(X_{1:5:10}, \ldots, X_{5:5:10})$	–6.262
10	5	(3,2,0,0,0)	1	$(X_{1:5:10})$	–1.492
10	5	(3,2,0,0,0)	2	$(X_{1:5:10}, X_{2:5:10})$	–2.473
10	5	(3,2,0,0,0)	3	$(X_{1:5:10}, X_{2:5:10}, X_{3:5:10})$	–2.760
10	5	(3,2,0,0,0)	4	$(X_{1:5:10}, \ldots, X_{4:5:10})$	–2.642
10	5	(3,2,0,0,0)	5	$(X_{1:5:10}, \ldots, X_{5:5:10})$	–1.831
10	10	(0,0,0,0,0)	1	$(X_{1:10})$	–1.492
10	10	(0,0,0,0,0)	2	$(X_{1:10}, X_{2:10})$	–2.878
10	10	(0,0,0,0,0)	3	$(X_{1:5:10}, X_{2:10}, X_{3:10})$	–4.146
10	10	(0,0,0,0,0)	4	$(X_{1:10}, \ldots, X_{4:10})$	–5.281
10	10	(0,0,0,0,0)	5	$(X_{1:10}, \ldots, X_{5:10})$	–6.262
10	10	(0,0,0,0,0)	4	$(X_{1:10}, \ldots, X_{6:10})$	–7.061
10	10	(0,0,0,0,0)	4	$(X_{1:10}, \ldots, X_{7:10})$	–7.636
10	10	(0,0,0,0,0)	4	$(X_{1:10}, \ldots, X_{8:10})$	–7.924
10	10	(0,0,0,0,0)	10	$(X_{1:10}, \ldots, X_{10:10})$	–6.995

We have written a program in the algebraic manipulation package MATHEMATICA for applying Theorem 1. For a pre-determined progressively Type-II censoring scheme $(n, m, R_1, R_2, \ldots R_m)$, the program returns the numerical values of the Renyi entropy. The electronic version of the computer program can be obtained from the authors.

Remark 3 For the case when $\alpha \to 1$, all the results in Sects. 4 and 5 can be simplified to the corresponding results for the Shannon entropy in a progressively Type-II censored OS obtained by Abo-Eleneen [3].

Example 3 For the standard exponential distribution $f(x) = \exp(-x)$, $x > 0$, we have

$$H^{\alpha}_{1:n} = -\log n + \frac{\log \alpha}{\alpha - 1}. \tag{37}$$

We can use Theorem 1 and (37) to calculate the $H^{\alpha}_{1\ldots r:m:n}$ presented in Table 1.

The table provides the values of the Renyi entropy, $H^{\alpha}_{1\ldots r:m:n}$, for $\alpha = 1.5$, $n = 5, 10$ and $m = 3, 5$ for different censoring schemes and $r = 1, \ldots, m$. The entries were computed using Theorem 1 and (37) in MATHEMATICA. For $r < m$, Table 1 gives the values of the Renyi entropy in a collection of r OS from a progressively Type-II censored sample. For $r = m$, the table lists the values of Renyi entropy in a complete sample of progressively Type-II censored OS. Furthermore, the table includes the cases $r_1 = r_2 = \cdots = r_{m-1} = 0$, $r_m = n - m$, which correspond to the Type II censored sample and $r_1 = r_2 = \cdots = r_m = 0$, $n = m$, which correspond to the OS of a complete sample.

6 Conclusion

We have discussed some properties of the Renyi entropy of consecutive OS and progressively Type-II censored OS. First, we considered the decomposition of the Renyi entropy in both settings to derive some useful recurrence relations and computational methods. We showed that the Renyi entropy of both consecutive OS and progressively Type-II censored OS can be simplified to a summation of Renyi entropy of the smallest OS of varying sample size. This representation is useful for computational purposes.

Acknowledgments It is a pleasure to contribute to this volume honoring Professor Nagaraja, a prolific researcher in the area of order statistics, record values and health research methods, and we wish him many more years of active academic life. The authors would like to express deep thanks to the referees for their helpful comments and suggestions which led to a considerable improvement in the presentation of this paper. The authors are thankful to Qassim University which provided financial support for this work under Grant SR-D-2819.

References

1. Abbasnejad, M., and N.R. Arghami. 2011. Renyi entropy properties of order statistics. *Communications in Statistics - Theory and Methods* 40: 40–52.
2. Abo-Eleneen, Z.A. 2008. Fisher Information in progressive Type II Censored Samples. *Communications in Statistics - Theory and Methods* 37: 1–10.
3. Abo-Eleneen, Z.A. 2011. The entropy of progressively censored samples. *Entropy* 13: 437–449.
4. Balakrishnan, N. 2007. Progressive censoring methodology: An appraisal (with discussions). *TEST* 16: 211–259.
5. Balakrishnan, N., A. Habibi Rad, and N.R. Arghami. 2007. Testing exponentiality based on Kullback-Leibler information with progressively Type-II censored data IEEE Trans. *Reliab*. 56: 301–307.
6. Balakrishnan, N., and R. Aggarwala. 2000. *Progressive Censoring: Theory, Methods, and Applications*. Boston: Birkhauser.
7. Balasubramanian, K., and N. Balakrishnan. 1993. Dual principle of order statistics. *JRSS B* 55: 687–691.
8. Cole, R.H. 1951. Relations between moments of order statistics. *The Annals of Mathematical Statistics* 22: 308–310.
9. Cover, T.M., and J.A. Thomas. 2005. *Elements of Information Theory*. New Jersey: Wiley.
10. Csiszár, I., Körner, J., 1981. *Information theory. Probability and Mathematical Statistics*. London, UK: Academic Press.
11. David, H.A., and H.N. Nagaraja. 2003. *Order statistics*, 3rd ed. New York: Wiley.
12. Ebrahimi, N., E.S. Soofi, and H. Zahedi. 2004. Information properties of order statistics and spacings. *IEEE Transactions on Information Theory* 50: 177–183.
13. Golshani, L., E. Pasha, and G. Yari. 2009. Some properties of Renyi entropy and Renyi entropy rate. *Information Sciences* 179: 2426–2433.
14. Javorka, M., Z. Trunkvalterova, L. Tonhaizerova, K. Javorka, and M. Baumert. 2008. Short-term heart rate complexity is reduced in patients with type 1 diabetes mellitus. *Clinical Neurophysiology* 119: 1071–1081.
15. Jomhoori, S., and F. Yousefzadeh. 2014. On Estimating the Residual Renyi Entropy under Progressive Censoring. *Communications in Statistics - Theory and Methods* 43: 2395–2405.
16. Kirchanov, V.S. 2008. Using the Renyi entropy to describe quantum dissipative systems in statistical mechanics. *Theoretical and Mathematical Physics* 156: 1347–1355.
17. Park, S. 1995. The entropy of consecutive order statistics. *IEEE Transactions on Information Theory* 41: 2003–2007.
18. Park, S. 2005. Testing exponentiality based on the Kullback-Leibler information with the type II censored data. *IEEE Transactions on Reliability* 54: 22–26.
19. Renyi, A., 1961. On measures of entropy and information. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: I, pp. 547–561. University of California Press, Berkeley (1960)
20. Srikantan, K.S. 1962. Recurrence relations between the pdfs of order statistics, and some applications. *Annals of Mathematical Statistics* 33: 169–177.
21. Vasicek, O. 1976. A test for normality based on sample entropy. *Journal of the Royal Statistical Society: Series B* 38: 54–59.
22. Wong, K.M., and S. Chen. 1990. The entropy of ordered sequences and order statistics. *IEEE Transactions Information Theory* 36: 276–284.
23. Zarezadeh, S., and M. Asadi. 2010. Results on residual Renyi entropy of order statistics and record values. *Information Sciences* 180: 4195–4206.

Part II
Stochastic Modeling and Estimation

Estimation in a Model of Sequential Order Statistics with Ordered Hazard Rates

Marco Burkschat, Udo Kamps and Maria Kateri

Abstract As a generalization of order statistics from independent and identically distributed random variables, sequential order statistics (SOSs) may be applied as a model for ordered data, when assuming changes of underlying distributions immediately after the occurrences of ordered observations. For example, in the case of a model for k-out-of-n-systems, where the $n-k+1$ failures of components in a system of n components occur successively, a change of the respective underlying distribution after a failure is motivated by an increased load put on the remaining components. The corresponding cumulative distribution functions are assumed to have possibly different ordered hazard rates, which are further multiplied by factors, in order to build the hazard rates of the SOSs. These factors are the parameters of interest. Estimation of the parameters is considered by means of maximum likelihood under order restriction, by means of link functions, and in a Bayesian set-up with an order statistics prior.

Keywords Sequential order statistics · Increasing hazard rate · Proportional hazard rate · Maximum likelihood estimation · Link function · Bayes estimation

1 Introduction

As an extension of common order statistics based on independent and identically distributed (i.i.d.) random variables, sequential order statistics (SOSs) have been introduced to model data in ascending order when the situation may change each time immediately after an observation. Thus, different underlying distributions are

M. Burkschat · U. Kamps (✉) · M. Kateri
Institute of Statistics, RWTH Aachen University, Aachen, Germany
e-mail: udo.kamps@rwth-aachen.de

M. Burkschat
e-mail: marco.burkschat@rwth-aachen.de

M. Kateri
e-mail: maria.kateri@rwth-aachen.de

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_7

incorporated in the model. As an application, such a model is useful to describe k-out-of-n systems, where, after some failure, there is an increased load put on the remaining components (cf., e.g., [1, 3, 10, 11, 15, 18]).

In terms of hazard rates $\lambda_j = \frac{f_j}{1-F_j}$ with cumulative distribution functions $F_1, F_2, \ldots$, and respective probability density functions $f_1, f_2, \ldots, j = 1, 2, \ldots$, we suppose n initial items in some k-out-of-n system to have lifetime distribution F_1. After the j-th observation, the hazard rate of all remaining $n - j$ items changes from λ_j to λ_{j+1}, $j = 1, \ldots, n-1$.

In the analysis of SOSs it is usually assumed that the hazard rates $\lambda_1, \ldots, \lambda_n$ are proportional, i.e.,

$$\lambda_j = \vartheta_j h\,, \quad j = 1, \ldots, n, \tag{1}$$

where h is some baseline hazard rate. Non-proportional hazard rates settings enable higher flexibility in modeling. SOSs based on different cumulative distribution functions $F_1, F_2, \ldots$ allow for structural changes of hazard rates, which, e.g., may all be increasing, but of different functional form (see [9, 18]). In what follows, we consider the situation

$$\lambda_j = \vartheta_j h_j\,, \quad j = 1, \ldots, n, \tag{2}$$

where the hazard rates $h_1, \ldots, h_n$ may be different, but are assumed to be pre-fixed, and the positive quantities $\vartheta_1, \ldots, \vartheta_n$ are the parameters of interest.

When modelling an increasing load put on the remaining components, it is reasonable to require

$$\lambda_j(t) \le \lambda_{j+1}(t)\,, \quad t > 0,\ j = 1, \ldots, n-1.$$

As examples for ordered hazard rates in families of distributions, we consider three choices with increasing failure rates, and with cumulative distribution functions of the type

$$F(t) = 1 - \exp\{-\vartheta H(t)\}, \tag{3}$$

where H denotes the cumulative hazard rate and $h = H'$. Hence, respective hazard rates are given by

$$\lambda(t) = \vartheta h(t).$$

Gompertz and exponentiated power function distributions have this particular form.

Gompertz distributions with cumulative distribution function

$$F(t) = 1 - \exp\left\{-\vartheta\left(e^{\delta t} - 1\right)\right\}, \quad t \in (0, \infty)\,,\ \delta > 0\,,\ \vartheta > 0,$$

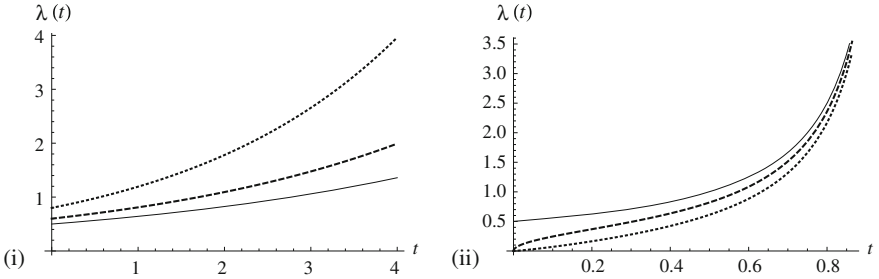


Fig. 1 Hazard rates of **i** Gompertz distributions with common $\vartheta = 2$ and $\delta_1 = 0.25$, $\delta_2 = 0.3$ and $\delta_3 = 0.4$ (*solid*, *dashed* and *dotted line*, respectively) and **ii** exponentiated power distributions with common $\vartheta = 0.4$ and $\delta_1 = 1$, $\delta_2 = 1.5$ and $\delta_3 = 2.2$ (*solid*, *dashed* and *dotted line*, respectively)

possess increasing hazard rates

$$\lambda(t) = \vartheta \delta e^{\delta t}, \quad t \in (0, \infty),$$

which are monotonically increasing in δ.

Exponentiated power function distributions are given by the cumulative distribution functions

$$F(t) = 1 - \left(1 - t^{\delta}\right)^{\vartheta}, \quad t \in (0, 1), \ \delta > 0, \ \vartheta > 0,$$

and, for $\delta \geq 1$, have increasing hazard rates

$$\lambda(t) = \vartheta \ \frac{\delta t^{\delta - 1}}{1 - t^{\delta}}, \quad t \in (0, 1),$$

which are monotonically decreasing in δ.

In Fig. 1, hazard rates of Gompertz and exponentiated power functions are shown.

Another interesting class of distributions are the linear hazard rate distributions given by

$$F(t) = 1 - \exp\left\{-\vartheta (at + bt^2)\right\}, \quad t \in (0, \infty), \ a, b > 0, \ \vartheta > 0,$$

with increasing hazard rates

$$\lambda(t) = \vartheta (a + 2bt), \ t \in (0, \infty),$$

which are monotonically increasing in b.

As is obvious from the hazard rate representations and from Fig. 1, the hazard rates can become close if their respective δ-parameters are close. In any case, the hazard rates for the linear hazard rate distributions class can become arbitrarily close. Hence, it is of particular interest to assume and to estimate the model parameters

$\vartheta_1, \vartheta_2, \ldots$ in increasing order, to ensure that ordered functions $h_1 \leq h_2 \leq \ldots$ will lead to ordered hazard rates $\lambda_1 \leq \lambda_2 \leq \ldots$.

The model of SOSs for hazard rates of the form (2) is presented in Sect. 2 and expressed in a form that exhibits its exponential family structure. The remainder of the paper deals with inference for the model parameters in case of s independent samples, each consisting of the first r SOSs. Maximum likelihood estimators (MLEs) and order restricted MLEs are considered in Sect. 3. More parsimonious models are derived by assuming that the model parameters are specified by some link function that simultaneously ensures their ordering as well. Inference under this set-up is dealt with in Sect. 4. Bayesian estimation for a specific family of priors that lead to explicit Bayes estimators for the model parameters is examined in Sect. 5 (also ordered). Finally, Sect. 6 deals with ordered Bayes estimation for the special case of a two-component system, for which explicit representations of the posterior distributions are possible under weaker conditions on the underlying prior distribution considered in Sect. 5.

2 Sequential Order Statistics

In the general version of SOSs $X^*_{1,n} \leq \cdots \leq X^*_{n,n}$, they are based on absolutely continuous cumulative distribution functions $F_1, \ldots, F_n$ with $F_1^{-1}(1) \leq \cdots \leq F_n^{-1}(1)$ and corresponding probability density functions $f_1, \ldots, f_n$. For $1 \leq r \leq n$, the joint probability density function of $X^*_{1,n}, \ldots, X^*_{r,n}$ is given by

$$f^{X^*_{1,n},\ldots,X^*_{r,n}}(x_1, \ldots, x_r) = \frac{n!}{(n-r)!} \prod_{j=1}^{r} \left(\frac{1 - F_j(x_j)}{1 - F_j(x_{j-1})} \right)^{n-j} \frac{f_j(x_j)}{1 - F_j(x_{j-1})} \tag{4}$$

on the region $-\infty = x_0 < x_1 \leq \cdots \leq x_r < \infty$ [18].

When modeling an $(n - r + 1)$-out-of-n system, as mentioned in the introduction, the r-th SOS $X^*_{r,n}$ describes the life-length of the system.

It is shown in [16] that SOSs with a probability density function as in (4) can be recursively generated via

$$X^*_{j,n} = F_j^{-1}\left(1 - V_j\left(1 - F_j\left(X^*_{j-1,n}\right)\right)\right), \quad 2 \leq j \leq n,$$

where $V_2, \ldots, V_n$ are independent power distributed random variables with $V_j \sim pow(n - j + 1)$, $2 \leq j \leq n$.

In terms of the setting (2) with cumulative hazard rates $H_2, \ldots, H_n$, the relation reads

$$X^*_{j,n} = H_j^{-1}\left(Z_j + H_j\left(X^*_{j-1,n}\right)\right), \quad 2 \leq j \leq n,$$

where $Z_2, \ldots, Z_n$ are independent, exponentially distributed random variables with $EZ_j = \left((n-j+1)\vartheta_j\right)^{-1}$, $2 \le j \le n$.

In most of the previous works on SOSs, the cumulative distribution functions $F_1, F_2, \ldots$ are chosen according to $F_j(t) = 1 - (1 - F(t))^{\vartheta_j}$, $\vartheta_j > 0$, $1 \le j \le n$ with some baseline cumulative distribution function F with probability density function f and hazard rate $h(t) = \frac{f(t)}{1-F(t)}$. Hence, the hazard rates of $F_1, F_2, \ldots$ are given by (1), which is a proportional hazards setting.

Remark 1 In the particular setting $F_j(t) = 1 - (1 - F(t))^{\vartheta_j}$, $\vartheta_j > 0$, $1 \le j \le r$, the joint density (4) can be rewritten as

$$f^{X^*_{1,n},\ldots,X^*_{r,n}}(x_1, \ldots, x_r) = \frac{n!}{(n-r)!} \left(\prod_{j=1}^{r} \vartheta_j \right) \left(\prod_{j=1}^{r-1} (1 - F(x_j))^{\gamma_j - \gamma_{j-1} - 1} f(x_j) \right) \times (1 - F(x_r))^{\gamma_r - 1} f(x_r), \quad x_1 \le \cdots \le x_r, \tag{5}$$

with $\gamma_j = (n-j+1)\vartheta_j$, $1 \le j \le r$.

If F is chosen to be a shifted exponential distribution with probability density function

$$f(x) = \lambda e^{-\lambda(x-c)}, \quad x > c, \tag{6}$$

then (5) with $r = n$ reads

$$n!\lambda^n \left(\prod_{j=1}^{n} \vartheta_j \right) \exp\left\{ -\lambda \sum_{j=1}^{n} (n-j+1)\vartheta_j \left(x_j - x_{j-1}\right) \right\}, \quad c = x_0 < x_1 \le \cdots \le x_n.$$

The particular choice $\vartheta_1 = \cdots = \vartheta_n = \vartheta > 0$, say, leads to the joint density of common order statistics in the i.i.d. setting (see [17]) based on the cumulative distribution function $1 - (1 - F(t))^{\vartheta}$.

Here, we choose $F_1, F_2, \ldots$ to be of the form

$$F_j(t) = 1 - \exp\left\{ -\vartheta_j H_j(t) \right\} \tag{7}$$

with cumulative hazard rates $H_1, H_2, \ldots$ and respective derivatives $h_1, h_2, \ldots$ leading to a hazard rate λ_j of F_j given by (2), i.e.

$$\lambda_j(t) = \vartheta_j h_j(t), \quad 1 \le j \le n\,.$$

In the following, it is assumed that the cumulative distribution functions $F_1, F_2, \ldots$ in (7) have the same support which is given by an interval of the real line.

In an increasing failure rate (IRF) setting, it would be interesting to ensure the IFR property of $X^*_{r,n}$ by means of properties of $F_1, \ldots, F_r$ (for the IFR property, see [4]). Results in this direction are shown in [13, 20], although there is no explicit

representation of the marginal density function of $X^*_{r,n}$. In the particular case of common order statistics $X_{1,n} \leq \cdots \leq X_{n,n}$ based on F, the sufficient conditions in [13, 20] amount to assuming F to have the IFR property, which is a well known result. In an outstanding article, Nagaraja (see [19]) states related general results: if the rth order statistic in a sample of size n possesses the IFR (or increasing failure rate average (IFRA), new better than used (NBU), decreasing mean residual lifetime (DMRL)) property, so do the neighboring order statistics $X_{r+1,n}$, $X_{r,n-1}$ and $X_{r+1,n+1}$.

Choosing $F_1, F_2, \ldots$ as in (7), the joint probability density function (4) of SOSs can be rewritten to obtain

$$f^{X^*_{1,n},\ldots,X^*_{r,n}}(x_1,\ldots,x_r) = \left(\prod_{j=1}^{r}\vartheta_j\right)\exp\left\{\sum_{j=1}^{r}\vartheta_j T_j(\mathbf{x})\right\}\left(\frac{n!}{(n-r)!}\prod_{j=1}^{r}h_j(x_j)\right), \tag{8}$$

where

$$\begin{aligned}
&\mathbf{x} = (x_1,\ldots,x_r) \quad \text{with } F_1^{-1}(0) < x_1 \leq \cdots \leq x_r < F_1^{-1}(1),\\
&T_1(\mathbf{x}) = -nH_1(x_1),\\
&T_j(\mathbf{x}) = -(n-j+1)\left(H_j(x_j) - H_j(x_{j-1})\right), \quad 2 \leq j \leq r.
\end{aligned}$$

It is seen (see [5, 6]) that the densities in (8) form an r-parametric exponential family in $\boldsymbol{T} = (T_1,\ldots,T_r)$ and $\boldsymbol{\vartheta} = (\vartheta_1,\ldots,\vartheta_r)$. Moreover, the statistics $T_1,\ldots,T_r$ are independent, $-T_j$ is exponentially distributed with mean $1/\vartheta_j$, $1 \leq j \leq r$, and $\boldsymbol{T}$ is minimal sufficient and complete (see also [9]).

The notation $\boldsymbol{X} = (X^*_{1,n},\ldots,X^*_{r,n})$ denotes that we have a sample $\boldsymbol{X}^{(1)},\ldots,\boldsymbol{X}^{(s)}$ of independent copies with density function (8), each. Then, the joint probability density function of $\boldsymbol{X}^{(1)},\ldots,\boldsymbol{X}^{(s)}$ is obtained as

$$f^{(s)}_{\boldsymbol{\vartheta}}(\tilde{\boldsymbol{x}}^{(s)}) = \left(\frac{n!}{(n-r)!}\right)^s\left(\prod_{j=1}^{r}\vartheta_j^s\right)\exp\left\{\boldsymbol{\vartheta}'\boldsymbol{T}^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right)\right\}\prod_{i=1}^{s}\prod_{j=1}^{r}h_j(x_{ij}), \tag{9}$$

where the vector $\boldsymbol{T}^{(s)} = (T_1^{(s)},\ldots,T_r^{(s)})'$ of statistics is given by

$$T_j^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right) = \sum_{i=1}^{s}T_j\left(\boldsymbol{x}^{(i)}\right), \tag{10}$$

$\tilde{\boldsymbol{x}}^{(s)} = (\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(s)})$ and $\boldsymbol{x}^{(i)} = (x_{i1},\ldots,x_{ir})$ with $F_1^{-1}(0) < x_{i1} \leq \cdots \leq x_{ir} < F_1^{-1}(1)$ the realization of $\boldsymbol{X}^{(i)}$, $1 \leq i \leq s$.

Hence, the densities in (9) form an r-parametric multivariate exponential family in $\boldsymbol{T}^{(s)}$ and $\boldsymbol{\vartheta}$, and $\boldsymbol{T}^{(s)}$ turns out to be minimal sufficient and complete. It is seen that $-T_j^{(s)} \sim \Gamma(s, 1/\vartheta_j)$, where the probability density function of this gamma distribution is

$$\frac{\vartheta_j^s}{(s-1)!}\, t^{s-1}\, e^{-\vartheta_j t}, \quad t > 0.$$

The structure of an exponential family directly leads to various results in statistical inference with SOSs as well as to structural findings (cf. [5–7, 9, 21]).

3 Maximum Likelihood Estimation

In the model of SOSs specified in Sect. 2 with s independent samples, consisting of the first r SOSs, each, [8] considers maximum likelihood estimators (MLEs) for the model parameters $\vartheta_1, \ldots, \vartheta_r$. We summarize some findings for both the unrestricted situation and estimation under the simple order restriction.

Theorem 1 *In the set-up of Sect. 2 we obtain:*

(i) The unique MLE $\hat{\boldsymbol{\vartheta}}^{(s)} = \left(\hat{\vartheta}_1^{(s)}, \ldots, \hat{\vartheta}_r^{(s)}\right)$ *of* $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_r)$ *is determined by* $\hat{\vartheta}_j^{(s)} = -s/T_j^{(s)}$, $1 \le j \le r$. *These quantities are jointly independent and inverse gamma distributed.*

(ii) For $s > 2$, *the uniformly minimum variance unbiased estimator of* ϑ_j *is* $\frac{s-1}{s}\hat{\vartheta}_j$, $1 \le j \le r$.

(iii) The sequence $\left(\sqrt{s}\left(\hat{\boldsymbol{\vartheta}}^{(s)} - \boldsymbol{\vartheta}\right)\right)_{s\in\mathbb{N}}$ *converges in distribution to a multivariate normal distribution with zero mean and diagonal covariance matrix diag* $(\vartheta_1^2, \ldots, \vartheta_r^2)$.

As described in Sect. 1, under the assumption of ordered hazard rates $\lambda_1(t) \le \lambda_2(t) \le \ldots$ for all t, it is reasonable to estimate the model parameters in ascending order.

Theorem 2 *In the set-up of Sect. 2 we find:*

The unique MLE $\tilde{\boldsymbol{\vartheta}}^{(s)} = \left(\tilde{\vartheta}_1^{(s)}, \ldots, \tilde{\vartheta}_r^{(s)}\right)$ *of* $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_r)$ *under the restriction* $\vartheta_1 \le \cdots \le \vartheta_r$ *is determined by*

$$\tilde{\vartheta}_j^{(s)} = \min_{1\le q\le r} \max_{1\le p\le j} \frac{q-p+1}{\sum_{k=p}^{q} 1/\hat{\vartheta}_k^{(s)}}, \quad 1 \le j \le r\,.$$

For further details such as strong consistency of the MLEs, we refer to [8].

4 Estimation by Means of Link Functions

Instead of considering r arbitrary positive (maybe increasingly ordered) parameters $\vartheta_1, \ldots, \vartheta_r$ as in the previous sections, the number of unknown parameters in the model could be reduced by assuming a functional relationship for these parameters.

Recently, two kinds of link functions have been considered in the setting of SOSs in more detail, namely proportional and linear link functions (see [2]). In the first case, a relation

$$\vartheta_j = \tau g_j, \quad j = 1, \ldots, r,$$

with known values $g_j > 0$ and an unknown proportionality factor $\tau > 0$ is assumed. By arguing analogously to [2], for example, the following results can be obtained for the model under consideration.

Theorem 3 *(i) The unique MLE of τ is given by*

$$\hat{\tau}^{(s)} = -rs \left(\sum_{j=1}^{r} g_j T_j^{(s)} \right)^{-1},$$

with $T_j^{(s)}$ as defined in (10). This estimator follows an inverse gamma distribution.

(ii) For $rs > 2$, the uniformly minimum variance unbiased estimator of τ is $\frac{rs-1}{rs}\hat{\tau}^{(s)}$.

(iii) The sequence $\left(\sqrt{rs}\left(\hat{\tau}^{(s)} - \tau\right)\right)_{s\in\mathbb{N}}$ converges in distribution to a normal distribution with zero mean and variance τ^2.

In the case of a linear link function, we assume the relation

$$\vartheta_j = \tau_1 + \tau_2 g_j, \quad j = 1, \ldots, r,$$

with known values $g_j \in \mathbb{R}$ and unknown parameters $\tau_1, \tau_2 \in \mathbb{R}$ such that $\vartheta_j > 0$ for every $j = 1, \ldots, r$. In [8] (see also [2]), the following results for maximum likelihood estimation can be found.

Theorem 4 *Let $r \geq 2$ and $g_1, \ldots, g_r$ be not all equal.*

(i) The unique MLE $\hat{\boldsymbol{\tau}}^{(s)} = \left(\hat{\tau}_1^{(s)}, \hat{\tau}_2^{(s)}\right)$ of (τ_1, τ_2) is given as the only solution of the equations

$$\tau_1 = -\frac{\tau_2 \tilde{T}_2^{(s)} + rs}{\tilde{T}_1^{(s)}}, \qquad \sum_{j=1}^{r} \frac{s}{\left(\tilde{T}_2^{(s)} - g_j \tilde{T}_1^{(s)}\right)\tau_2 + rs} = 1$$

for $\tau_1, \tau_2 \in \mathbb{R}$ such that $\tau_1 + \tau_2 g_j > 0$ for every $j = 1, \ldots, r$, where

$$\tilde{T}_1^{(s)}(\tilde{\boldsymbol{x}}^{(s)}) = \sum_{j=1}^{r} T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)}), \quad \tilde{T}_2^{(s)}(\tilde{\boldsymbol{x}}^{(s)}) = \sum_{j=1}^{r} g_j\, T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)}).$$

(ii) The sequence $\left(\sqrt{s}\left(\hat{\boldsymbol{\tau}}^{(s)} - \boldsymbol{\tau}\right)\right)_{s\in\mathbb{N}}$ *converges in distribution to a bivariate normal distribution with zero mean and covariance matrix given by the inverse of the matrix*

$$\begin{pmatrix} \sum_{j=1}^{r} \frac{1}{(\tau_1+\tau_2 g_j)^2} & \sum_{j=1}^{r} \frac{g_j}{(\tau_1+\tau_2 g_j)^2} \\ \sum_{j=1}^{r} \frac{g_j}{(\tau_1+\tau_2 g_j)^2} & \sum_{j=1}^{r} \frac{g_j^2}{(\tau_1+\tau_2 g_j)^2} \end{pmatrix}.$$

As an alternative to the preceding MLEs, particular plug-in estimators with explicit representations are proposed in [2], in a different set-up. For instance, if $r, s \geq 2$ and $g_1, ..., g_r$ are pairwise different, then

$$\tilde{\tau}_1 = \frac{s-1}{rs}\sum_{j=1}^{r} \hat{\vartheta}_j^{(s)} - \frac{\tilde{\tau}_2}{r}\sum_{j=1}^{r} g_j, \quad \tilde{\tau}_2 = \frac{s-1}{s(r-1)}\sum_{j=1}^{r-1} \frac{\hat{\vartheta}_{j+1}^{(s)} - \hat{\vartheta}_j^{(s)}}{g_{j+1}-g_j}$$

are plug-in estimators of τ_1 and τ_2 in our model (9), respectively. Note that both estimators are unbiased. Moreover, we obtain by construction

$$\frac{1}{r}\sum_{j=1}^{r}(\tilde{\tau}_1 + \tilde{\tau}_2 g_j) = \tilde{\tau}_1 + \frac{\tilde{\tau}_2}{r}\sum_{j=1}^{r} g_j = \frac{s-1}{rs}\sum_{j=1}^{r} \hat{\vartheta}_j^{(s)} > 0$$

almost surely. However, because the supports of the estimators $\hat{\vartheta}_1^{(s)}, \ldots, \hat{\vartheta}_r^{(s)}$ are given by the non-negative half axis of the real line (see Sect. 3, Theorem 3), $\tilde{\tau}_1 + \tilde{\tau}_2 g_j \geq 0$ does not necessarily hold almost surely for $j = 1, \ldots, r$, in contrast to the MLEs. For instance, if $g_j = j, j = 1, \ldots, r$, then

$$\tilde{\tau}_1 + \tilde{\tau}_2 g_1 = \frac{s-1}{s}\left[\frac{1}{r}\sum_{j=1}^{r} \hat{\vartheta}_j^{(s)} - \frac{1}{2}(\hat{\vartheta}_r^{(s)} - \hat{\vartheta}_1^{(s)})\right]$$

is negative with positive probability for $r \geq 3$.

Under additional assumptions on the linear link function, modified versions $\tilde{\tau}_{1,sor}$ and $\tilde{\tau}_{2,sor}$ of the preceding plug-in estimators can be constructed along the lines in [2] by using the MLEs $\tilde{\vartheta}_1^{(s)}, \ldots, \tilde{\vartheta}_r^{(s)}$ under the simple order restriction $\vartheta_1 \leq \cdots \leq \vartheta_r$ (cf. Theorem 4). These estimators satisfy $\tilde{\tau}_{1,sor} + \tilde{\tau}_{2,sor} g_j \geq 0$ almost surely for every $j = 1, \ldots, r$.

5 Bayes Estimation

Let $B_1, \ldots, B_r$ denote the parameter variables with outcomes $\vartheta_1, \ldots, \vartheta_r$. Then, (9) is the joint conditional density of $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(s)}$ given $B_1 = \vartheta_1, \ldots, B_r = \vartheta_r$.

In the Bayesian framework, in order to obtain explicit posterior distributions and closed-form expressions for the Bayes estimators, we consider conjugate priors for the model parameters. Thus, ϑ_j are realizations of B_j $(j = 1, \ldots, r)$, which are assumed to follow a gamma distribution or an extended truncated Erlang distribution (ETED) as introduced in [11].

5.1 Gamma Prior

Let B_j be gamma distributed, i.e., $B_j \sim \Gamma(a, b)$, $a, b > 0$, with probability density function

$$f^{B_j}(t) \propto t^{a-1} e^{-bt}, \quad t > 0.$$

Then, by (9), the posterior density of B_j given the data $\boldsymbol{X}^{(1)} = \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{X}^{(s)} = \boldsymbol{x}^{(s)}$ is given by

$$f^{B_j|\boldsymbol{X}^{(1)}=\boldsymbol{x}^{(1)},\ldots,\boldsymbol{X}^{(s)}=\boldsymbol{x}^{(s)}}\left(\vartheta_j\right) \propto \exp\left\{\vartheta_j \left(T_j^{(s)}(\boldsymbol{x}) - b\right)\right\} \vartheta_j^{s+a-1}, \quad \vartheta_j > 0,$$

which is a $\Gamma(s + a, b - T_j^{(s)}(\boldsymbol{x}))$-density.

The posterior mean is thus

$$\begin{aligned} E\left(B_j|\boldsymbol{X}^{(1)} = \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{X}^{(s)} = \boldsymbol{x}^{(s)}\right) &= \frac{s+a}{b - T_j^{(s)}(\boldsymbol{x})} = \frac{s+a}{b + s/\hat{\vartheta}_j^{(s)}} \\ &= \left(\frac{s}{s+a}\,\frac{1}{\hat{\vartheta}_j^{(s)}} + \frac{a}{s+a}\,\frac{1}{a/b}\right), \end{aligned}$$

which is represented as a weighted harmonic mean of the MLE $\hat{\vartheta}_j^{(s)}$ of ϑ_j and the expected value EB_j of the prior distribution. The improper prior distribution with $a = b = 0$ leads to equality of MLE and posterior mean. It is worth mentioning that the above posterior distribution is always gamma whatever initial distribution functions $F_1, F_2, \ldots$ are chosen.

The resulting structure for Bayes estimation coincides with that of Bayes estimation of proportionality factors (i.e., model parameters) in the common SOSs set-up (cf. [11]). It is easily seen that simultaneous Bayes estimation of $\vartheta_1, \ldots \vartheta_r$ under independent gamma priors leads to independent gamma posteriors.

5.2 ETED Prior

Explicit posterior densities may also be obtained by choosing ETED (c, a, b) priors [11], where the densities are given by

$$g_{c,a,b}(t) \propto t^{a-1}\, e^{-bt}\,, \quad t > c > 0,$$

with parameters $a \in \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ and $b > 0$.

For positive integers a, these are truncated Erlang (gamma) densities. In the case $a \in \{0, -1, -2, \ldots\}$, the normalizing constant is given by $\psi_{a,b}^{-1}(c)$ with

$$\psi_{a,b}(c) = \frac{b^{-a}}{(-a)!} Ei(-bc) + e^{-bc} \sum_{k=1}^{-a} \frac{(-1)^{k-a}\, b^{k-1}}{c^{1-a-k}(-a)\cdot\cdots\cdot(1-a-k)}$$

where Ei denotes the exponential integral defined by

$$Ei(-bt) = -\psi_{0,b}(t) = -\int_t^\infty \frac{e^{-bx}}{x}\, dx\,, \quad t > 0.$$

Some calculation shows (cf. [11]) that the Bayes estimator for ϑ_j, under the squared error loss, is given by

$$E\left(B_j | \boldsymbol{X}^{(1)} = \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{X}^{(s)} = \boldsymbol{x}^{(s)}\right)$$

$$= \begin{cases} \left[1 + \frac{((b-T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)}))c)^{s+a}}{(s+a)!} \left(\sum_{l=0}^{s+a-1} \frac{((b-T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)}))c)^l}{l!}\right)^{-1}\right] \frac{s+a}{b-T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)})}, & s+a \geq 1 \\ \left(\left(T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)}) - b\right) Ei\left(\left(T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)}) - b\right)c\right)\right)^{-1} \exp\left(T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)}) - b\right), & s+a = 0 \\ \psi_{s+a+1,b-T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)})}(c) / \psi_{s+a,b-T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)})}(c), & s+a \leq -1 \end{cases}.$$

As in the gamma-case, it is obvious that simultaneous Bayes estimation of $\vartheta_1, \ldots, \vartheta_r$ under independent ETED priors results in independent ETED posteriors.

5.3 Ordered Bayes Estimators

As for maximum likelihood estimation, we are also interested in obtaining ordered Bayes estimators.

Following the idea in [11], the prior distribution of $(\vartheta_1, \ldots, \vartheta_r)$ is chosen to be the joint distribution of common order statistics based on an ETED distribution. Of course, any underlying distribution of the order statistics could be considered as a prior. The ETED approach is exposed here, because, in a particular case, it leads to an explicit posterior density function.

Let $B_1, \ldots, B_r$ be i.i.d. random variables according to ETED (c, a, b) for some $a \leq 0$ and $b > 0$, and let $B_{1,r} \leq \cdots \leq B_{r,r}$ denote their respective order statistics. For a specific choice of the parameter a, namely for $a = 1 - s$, where s is the number of independent samples of SOSs, we obtain the following result.

Theorem 5 *Given the sampling situation of Sect. 2, and letting ETED* (c, a, b) *with* $a = 1 - s$ *be the underlying distribution of* $B_{1,r} \leq \cdots \leq B_{r,r}$*, then their posterior distribution coincides with the distribution of r SOSs with model parameters* $\mu_j = b - \frac{1}{r-j+1} \sum_{l=j}^{r} T_j^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right)$, $1 \leq j \leq r$*, from a shifted exponential distribution with parameters* $\lambda = 1$ *and c (see (6) in Remark 1).*

Proof Since, for the prior density, we have

$$f^{B_{1,r},\ldots,B_{r,r}}(b_1, \ldots, b_r) \propto \prod_{j=1}^{r} b_j^{a-1}\, e^{-bb_j},$$

where $c = b_0 < b_1 \leq \cdots \leq b_r$, the posterior density of $B_{1,r}, \ldots, B_{r,r}$ given $\boldsymbol{X}^{(1)} = \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{X}^{(s)} = \boldsymbol{x}^{(s)}$ is determined by

$$\begin{aligned}
&f^{B_{1,r},\ldots,B_{r,r}|\boldsymbol{X}^{(1)}=\boldsymbol{x}^{(1)},\ldots,\boldsymbol{X}^{(s)}=\boldsymbol{x}^{(s)}}(b_1, \ldots, b_r) \\
&\quad \propto \left(\prod_{j=1}^{r} b_j^{s}\right) \exp\left\{\sum_{j=1}^{r} b_j\, T_j^{(s)}(\tilde{\boldsymbol{x}}^{(s)})\right\} \prod_{j=1}^{r} b_j^{a-1}\, e^{-bb_j} \\
&\quad = \prod_{j=1}^{r} \exp\left\{b_j \left(T_j^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right) - b\right)\right\} \\
&\quad \propto \prod_{j=1}^{r} \exp\left\{(b_j - c) \left(T_j^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right) - b\right)\right\} \\
&\quad = \prod_{j=1}^{r} \exp\left\{(b_j - b_{j-1}) \sum_{l=j}^{r} \left(T_j^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right) - b\right)\right\}
\end{aligned}$$

since $\displaystyle\sum_{j=1}^{r} (b_j - c)\, t_j = \sum_{j=1}^{r} \left(\sum_{l=1}^{j} (b_l - b_{l-1})\right) t_j = \sum_{l=1}^{r} (b_l - b_{l-1}) \sum_{j=l}^{r} t_j.$

From Remark 1 it is seen that this posterior density coincides with the density of r SOSs from a sample of size r with model parameters $\mu_j = b - \frac{1}{r-j+1} \sum_{l=j}^{r} T_j^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right)$.

The posterior distribution of $B_{1,r}, \ldots, B_{r,r}$ coincides with the distribution of SOSs from a shifted exponential distribution whatever distributions $F_1, \ldots, F_r$ are chosen.

From [15], the posterior marginal means, and thus the Bayes estimators of ϑ_j under the squared error loss and for the particular setting of Theorem 5, can be derived as

$$E\left(B_{j,r}|\boldsymbol{X}^{(1)} = \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{X}^{(s)} = \boldsymbol{x}^{(s)}\right) = c + \sum_{l=1}^{j} \frac{1}{(r-l+1)\mu_l}$$
$$= c - \sum_{l=1}^{j} \left(\sum_{j=l}^{r} \left(T_j^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right) - b \right) \right)^{-1}.$$

The ordering of these estimates is guaranteed by construction.

Furthermore, the joint posterior density in Theorem 5 is multivariate log-concave (cf. [14]) and all univariate marginal distributions are unimodal. These properties are useful for obtaining Bayesian credible sets.

Note that the choice $a = 1 - s$ is not data-dependent and does not lead to empirical Bayes inference, since s is fixed by the experimental design and preknown. However, for an alternative choice of a or, furthermore, for alternative prior distributions not in the ETED family, more flexible prior assumptions can be considered with the cost of non-closed form expressions for the estimators.

6 Ordered Bayes Estimators for Systems with Two Components

In the setting of the preceding section, explicit representations of the posterior distributions can be obtained under weaker conditions on the underlying ETED distribution, if the case of systems with two components, i.e., $r = 2$, is considered. In [12] a two-sample SOSs model with arbitrary numbers of ordered quantities within the samples is examined.

Let $B_{1,2} \leq B_{2,2}$ be the order statistics from i.i.d. random variables B_1 and B_2 which are distributed according to ETED (c, a, b) with $c > 0$, $a \in \mathbb{Z}$, $b > 0$ (see Sect. 5.2). Moreover, if $a < 0$, then assume $s + a > 0$. Introducing the notations

$$g(\vartheta_1, \vartheta_2) = (\vartheta_1 \vartheta_2)^{s+a-1} e^{-\gamma_1 \vartheta_1 - \gamma_2 \vartheta_2},$$

with $\gamma_j = b - T_j^{(s)}\left(\tilde{\boldsymbol{x}}^{(s)}\right), j = 1, 2$, and

$$G_{k,d}(c) = 1 - e^{-dc} \sum_{i=0}^{k-1} \frac{(dc)^i}{i!}\,,\ k \in \mathbb{N}, d > 0,$$

where $G_{k,d}$ is the distribution function of $\Gamma(k, d)$, the density of the posterior distribution of $B_{1,2}$, $B_{2,2}$ can be expressed as

$$f^{B_{1,2},B_{2,2}|\boldsymbol{X}^{(1)}=\boldsymbol{x}^{(1)},\ldots,\boldsymbol{X}^{(s)}=\boldsymbol{x}^{(s)}}(\vartheta_1, \vartheta_2) = \frac{g(\vartheta_1, \vartheta_2)}{\tilde{K}(\tilde{\boldsymbol{x}}^{(s)})}, \quad c < \vartheta_1 < \vartheta_2,$$

with

$$\begin{aligned}\tilde{K}(\tilde{\boldsymbol{x}}^{(s)}) = {} & \tfrac{(s+a-1)!}{\gamma_1^{s+a}} \Big[\Big(1 - G_{s+a,\gamma_1}(c)\Big) \tfrac{(s+a-1)!}{\gamma_2^{s+a}} \Big(1 - G_{s+a,\gamma_2}(c)\Big) \\ & - \sum_{i=0}^{s+a-1} \tfrac{\gamma_1^i}{i!} \tfrac{(s+a+i-1)!}{(\gamma_1+\gamma_2)^{s+a+i}} \Big(1 - G_{s+a+i,\gamma_1+\gamma_2}(c)\Big) \Big].\end{aligned}$$

The corresponding marginal densities are then given by

$$\begin{aligned}& f^{B_{1,2}|\boldsymbol{X}^{(1)}=\boldsymbol{x}^{(1)},\ldots,\boldsymbol{X}^{(s)}=\boldsymbol{x}^{(s)}}(\vartheta_1) \\ & \quad = \frac{1}{\tilde{K}(\tilde{\boldsymbol{x}}^{(s)})} \frac{(s+a-1)!}{\gamma_2^{s+a}} \vartheta_1^{s+a-1} e^{-(\gamma_1+\gamma_2)\vartheta_1} \times \sum_{i=0}^{s+a-1} \frac{(\gamma_2\vartheta_1)^i}{i!}, \\ & f^{B_{2,2}|\boldsymbol{X}^{(1)}=\boldsymbol{x}^{(1)},\ldots,\boldsymbol{X}^{(s)}=\boldsymbol{x}^{(s)}}(\vartheta_2) \\ & \quad = \frac{1}{\tilde{K}(\tilde{\boldsymbol{x}}^{(s)})} \frac{(s+a-1)!}{\gamma_1^{s+a}} \vartheta_2^{s+a-1} e^{-\gamma_2\vartheta_2} \times \big(G_{s+a,\gamma_1}(\vartheta_2) - G_{s+a,\gamma_1}(c)\big)\end{aligned}$$

for $\vartheta_1, \vartheta_2 > c$. The Bayes posterior means can be also explicitly stated:

$$\begin{aligned}E(B_{1,2}|\boldsymbol{X}^{(1)} = \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{X}^{(s)} = \boldsymbol{x}^{(s)}) = {} & \frac{1}{\tilde{K}(\tilde{\boldsymbol{x}}^{(s)})} \frac{(s+a-1)!}{\gamma_2^{s+a}(\gamma_1+\gamma_2)^{s+a+1}} \\ & \times \sum_{i=0}^{s+a-1} \frac{(s+a+i)!}{i!} \left(\frac{\gamma_2}{\gamma_1+\gamma_2}\right)^i \Big(1 - G_{s+a+i+1,\gamma_1+\gamma_2}(c)\Big)\end{aligned}$$

and

$$\begin{aligned}& E(B_{2,2}|\boldsymbol{X}^{(1)} = \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{X}^{(s)} = \boldsymbol{x}^{(s)}) \\ & = \frac{1}{\tilde{K}(\tilde{\boldsymbol{x}}^{(s)})} \frac{(s+a-1)!}{\gamma_1^{s+a}} \Bigg[\frac{(s+a)!}{\gamma_2^{s+a+1}} \big(1 - G_{s+a,\gamma_1}(c)\big)\big(1 - G_{s+a+1,\gamma_2}(c)\big) \\ & \quad - \frac{1}{(\gamma_1+\gamma_2)^{s+a+1}} \sum_{i=0}^{s+a-1} \frac{(s+a+i)!}{i!} \Big(\frac{\gamma_1}{\gamma_1+\gamma_2}\Big)^i \Big(1 - G_{s+a+i+1,\gamma_1+\gamma_2}(c)\Big) \Bigg].\end{aligned}$$

If $a > 0$, then the preceding results can be extended to $c = 0$. In this case the prior is given by the distribution of two order statistics based on a common gamma distribution. The resulting expressions for densities and means can be obtained by replacing values of the form $G_{k,d}(c)$ by zero in the above representations.

References

1. Balakrishnan, N., E. Beutner, and U. Kamps. 2008. Order restricted inference for sequential k-out-of-n systems. *Journal of Multivariate Analysis* 99: 1489–1502.
2. Balakrishnan, N., E. Beutner, and U. Kamps. 2011. Modeling parameters of a load-sharing system through link functions in sequential order statistics models and associated inference. *IEEE Transactions on Reliability* 60: 605–611.
3. Balakrishnan, N., U. Kamps, and M. Kateri. 2012. A sequential order statistics approach to step-stress testing. *Annals of the Institute of Statistical Mathematics* 64: 303–318.
4. Barlow, R.E., and F. Proschan. 1981. *Statistical theory of reliability and life testing*. Silver Spring, Maryland: To Begin With.
5. Bedbur, S. 2010. UMPU tests based on sequential order statistics. *Journal of Statistical Planning and Inference* 140(9): 2520–2530.
6. Bedbur, S., E. Beutner, and U. Kamps. 2012. Generalized order statistics: An exponential family in model parameters. *Statistics* 46: 159–166.
7. Bedbur, S., E. Beutner, and U. Kamps. 2014. Multivariate testing and model-checking for generalized order statistics with applications. *Statistics* 48(6): 1297–1310.
8. Bedbur, S., M. Burkschat, and U. Kamps, U. 2015. Inference in a model of successive failures with shape-adjusted hazard rates. *Annals of the Institute of Statistical Mathematics*, to appear.
9. Bedbur, S., U. Kamps, and M. Kateri. 2015. Meta-analysis of general step-stress experiments under repeated Type-II censoring. *Applied Mathematical Modelling* 39: 2261–2275.
10. Beutner, E., and U. Kamps. 2009. Order restricted statistical inference for scale parameters based on sequential order statistics. *Journal of Statistical Planning and Inference* 139: 2963–2969.
11. Burkschat, M., U. Kamps, and M. Kateri. 2010. Sequential order statistics with an order statistics prior. *Journal of Multivariate Analysis* 101: 1826–1836.
12. Burkschat, M., U. Kamps, and M. Kateri. 2013. Estimating scale parameters under an order statistics prior. *Statistics & Risk Modeling* 30: 205–219.
13. Burkschat, M., and J. Navarro. 2011. Aging properties of sequential order statistics. *Probability in the Engineering and Informational Sciences* 25: 449–467.
14. Chen, H., H. Xie, and T. Hu. 2009. Log-concavity of generalized order statistics. *Statistics & Probability Letters* 79: 396–399.
15. Cramer, E., and U. Kamps. 2001. Sequential k-out-of-n systems. In *Handbook of Statistics*, vol. 20, ed. N. Balakrishnan, and C.R. Rao, 301–372. Advances in Reliability, Amsterdam: Elsevier.
16. Cramer, E., and U. Kamps. 2003. Marginal distributions of sequential and generalized order statistics. *Metrika* 58: 293–310.
17. David, H.A., and H.N. Nagaraja. 2003. *Order Statistics*, 3rd ed. Hoboken: Wiley.
18. Kamps, U. 1995. A concept of generalized order statistics. *Journal of Statistical Planning and Inference* 48: 1–23.
19. Nagaraja, H.N. 1990. Some reliability properties of order statistics. *Communications in Statistics - Theory and Methods* 19: 307–316.
20. Torrado, N., R.E. Lillo, and M.P. Wiper. 2012. Sequential order statistics: Ageing and stochastic orderings. *Methodology and Computing in Applied Probability* 14: 579–596.
21. Vuong, Q.N., S. Bedbur, and U. Kamps. 2013. Distances between models of generalized order statistics. *Journal of Multivariate Analysis* 118: 24–36.

Properties of the Vacancy Statistic in the Discrete Circle Covering Problem

Gadi Barlevy and H.N. Nagaraja

Abstract Holst [10] introduced a discrete spacings model that is related to the Bose-Einstein distribution and obtained the distribution of the number of vacant positions in an associated circle covering problem. We correct his expression for its probability mass function, obtain the first two moments, and describe their limiting properties. We then examine the properties of the vacancy statistic when the number of covering arcs in the associated circle covering problem is random. We also discuss applications of our results to a study of contagion in networks.

Keywords Occupancy problems · Spacings · Bose-Einstein distribution · Sampling without replacement · Sampling with replacement

1 Introduction

This paper examines the discrete version of the circle covering problem first introduced by Stevens [15] in which m arcs of length $a(< 1)$ are randomly placed on a circle with unit circumference, and the question of interest is the fraction of the circle left uncovered by any arc. The discrete version of this problem can be described as follows. Consider r (≥ 2) boxes arranged in a ring numbered $0, 1, \ldots, r-1$. Draw $m-1$ boxes by simple random sampling without replacement from the boxes numbered $1, 2, \ldots, r-1$, where $2 \leq m \leq r$. Let $1 \leq R_1 < \cdots < R_{m-1} \leq r-1$ be the drawn numbers, and set $R_0 = 0$ and $R_m = r$. Define

$$S_k = R_k - R_{k-1},$$

G. Barlevy (✉)
Economic Research Department, Federal Reserve Bank of Chicago,
230 South LaSalle, Chicago 60604, USA
e-mail: gbarlevy@frbchi.org

H.N. Nagaraja
College of Public Health Division of Biostatistics, The Ohio State University,
1841 Neil Avenue, Columbus 43210-1351, USA
e-mail: nagaraja.1@osu.edu

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_8

for $k = 1, 2, \ldots, m$, i.e., S_k are spacings. Next, for an integer b where $1 \leq b \leq r$, define

$$V = \sum_{k=1}^{m} (S_k - b)_+ , \tag{1}$$

where $(x)_+ = \max\{x, 0\}$. This setup can be interpreted as follows. We can think of $\{R_0, \ldots, R_{m-1}\}$ as m distinct starting points whose location among the r boxes is chosen at random. From each starting point, we designate the next b boxes including the starting point as covered. Any remaining uncovered boxes are designated as vacant. The random variable (rv) V represents the total number of vacant boxes. For reference, it will be convenient to also define

$$N = \sum_{k=1}^{m} I\,(S_k > b)\,, \tag{2}$$

where $I(C) = 1$ if condition C is true and 0 otherwise. Thus, N represents the number of distinct blocks of vacant boxes.

We are interested in characteristics of V, specifically its distribution, some of its moments, and its behavior when m is random. Holst [10] derived the marginal and joint distributions of $\{S_k\}_{k=1}^{m}$ and showed they are exchangeable. He also explored the connection between these random variables (rv's) and the Bose-Einstein distribution. Feller [7, Sect. II.5(a)] provides a nice introduction to the Bose-Einstein urn model.

As anticipated by our comments above, in the limit as $r \to \infty$ while $b/r \to a$ for some constant $a < 1$ this problem converges to the circle covering problem in which m points are chosen uniformly from the circumference of a circle, and each of the m points forms the end point of an arc of length a. The latter problem has been extensively analyzed; see for example, Siegel [13]. The limit of V/r in our discrete setup corresponds to the fraction of the circumference that is uncovered. However, the finite version of the problem has been less studied, even though, as we discuss later, this version arises in certain applications.

We derive in Sect. 2 an explicit expression for the probability mass function (pmf) of V including an exploration of the range of its values; in this process we correct an error in the expression for the pmf given in Holst [10]. We derive the first two moments of V in Sect. 3 using several properties of the joint distribution of S_k derived by Holst [10]. We discuss an extension of the model in Sect. 4 in which the number of starting points m is random, allowing us among other things to discuss an alternative version of our model in which the boxes are chosen with replacement. This section is motivated by a query by the referee regarding the distribution of V when we sample with replacement. We establish limiting properties of V in Sect. 5 and link our results to those of Siegel [13]. In Sect. 6, we discuss an application concerning financial contagion that coincides with the discrete circle covering problem under certain conditions. This application suggests generalizations of the circle covering problem that have not been explored in previous work.

2 Exact Distribution of V

2.1 The Range

The value of V must be non-negative, and the lowest value it can assume is $r - mb$. Further, the largest possible value of V occurs when the chosen boxes are consecutive (implying $N \leq 1$) and V takes on the value $r - m - b + 1$. Thus, the support of V is the set $\{(r - mb)_+, \ldots, (r - m - b + 1)_+\}$, and so V is degenerate at 0 whenever $r < m + b$. Further, when $r - mb \geq 0$, the total number of points in the support of V is $(m - 1)(b - 1) + 1$ independently of r. Hence when $b = 1$, we have a single support point at $r - m > 0$. We now examine the form of the pmf $P(V = x)$ for various x values when $r \geq m + b$ and $b > 1$.

2.2 Probability Mass Function of V

Holst [10, Theorem 2.2] argues that $P(V = x)$ is given by

$$\sum_{y=1}^{m} \binom{m}{y} \sum_{t=0}^{m-y} (-1)^t \binom{m-y}{t} \binom{x-1}{y-1} \binom{r-(y+t)\,b-x-1}{m-y-1} \Big/ \binom{r-1}{m-1}. \quad (3)$$

Note that expression (3) may include improper binomial coefficients $\binom{n}{k}$ where either $n < 0$ or $k \notin \{0, \ldots, n\}$. Such terms are traditionally set to 0. We now argue that this convention may yield an incorrect expression for $P(V = x)$ for $x = 0$ and for $x = r - mb$, and offer correct expressions for $P(V = x)$ for these cases.

Observe first that for $x = 0$, the right-hand side of (3) would, under the usual convention, equal 0 due to the presence of the $\binom{x-1}{y-1}$ term. But this is at odds with the fact that $P(V = 0) = 1$ whenever $r < m + b$.

To derive $P(V = 0)$, we use the observation noted by Holst that

$$P(V = 0) = P\left(\sum_{j=1}^{m} I(S_j > b) = 0\right) = P(N = 0). \quad (4)$$

Holst [10] derives an expression for the right-hand side of (4) in part (a) of his Theorem 2.2. Using his result, we can deduce that for $x = 0$, (3) must be replaced by

$$P(V = 0) = \sum_{j=0}^{m} (-1)^j \binom{m}{j} \binom{r-jb-1}{m-1} \Big/ \binom{r-1}{m-1}. \quad (5)$$

We next turn to the case where $r \geq m + b$ and $x > 0$, and examine the range of values for y and t for which the associated terms on the right-hand side of (3) are all positive, i.e., when $0 \leq t \leq \min\{m - y, (r - m - x - (b - 1)y)/b\}$, and $1 \leq y \leq \min\{m-1, x, (r-m-x)/(b-1)\}$ for $b > 1$ and $1 \leq y \leq \min\{m-1, x\}$ for $b = 1$.

We shall now argue that for this range, (3) holds except when $x = r - mb \geq m$. To see this, we begin with the observation by Holst in proving his Theorem 2.2 that if $I_k = I(S_k > b)$, then $P(V = x)$ must equal

$$\sum_{y=1}^{m} \binom{m}{y} \sum_{t=0}^{m-y} (-1)^t \binom{m-y}{t} P\left(\sum_{k=1}^{y} (S_k - b)_+ = x, I_1 = \cdots = I_{y+t} = 1 \right). \quad (6)$$

Expression (6) can in turn be rewritten as

$$\sum_{y=1}^{m-1} \binom{m}{y} \sum_{t=0}^{m-y-1} (-1)^t \binom{m-y}{t} P\left(\sum_{k=1}^{y} (S_k - b)_+ = x, I_1 = \cdots = I_{y+t} = 1 \right) + \sum_{y=1}^{m} \binom{m}{y} (-1)^{m-y} P\left(\sum_{k=1}^{y} (S_k - b)_+ = x, I_1 = \cdots = I_m = 1 \right). \quad (7)$$

Holst then computes the following probabilities based respectively on parts (E) and (D) of his Theorem 2.1:

$$P\left(I_1 = \cdots = I_{y+t} = 1\right) = \frac{\binom{r-(y+t)b-1}{m-1}}{\binom{r-1}{m-1}}$$

and

$$P\left(\sum_{k=1}^{y} (S_k - b)_+ = x \,\middle|\, I_1 = \cdots = I_{y+t} = 1 \right) = \frac{\binom{x-1}{y-1}\binom{r-(y+t)b-x-1}{m-y-1}}{\binom{r-(y+t)b-1}{m-1}} \quad (8)$$

and thus concludes that

$$P\left(\sum_{k=1}^{y} (S_k - b)_+ = x, I_1 = \cdots = I_{y+t} = 1 \right) = \frac{\binom{x-1}{y-1}\binom{r-(y+t)b-x-1}{m-y-1}}{\binom{r-1}{m-1}}. \quad (9)$$

The expression for the conditional probability in (8) is valid and nonzero whenever $y \leq m - 1$, and $y + t < m$.

Next we consider the last sum on the right in (7). Since the event $\{I_1 = \cdots = I_m = 1\}$ implies $S_i > b$ for $i = 1, \ldots, m$, the sum $\sum_{k=1}^{y} (S_k - b)_+$ is strictly increasing in y for $y \le m$ and $\sum_{k=1}^{y} (S_k - b)_+ < \sum_{k=1}^{m} (S_k - b)_+ \equiv r - mb$. This means

$$\sum_{y=1}^{m} \binom{m}{y} (-1)^{m-y} P\left(\sum_{k=1}^{y} (S_k - b)_+ = x, I_1 = \cdots = I_m = 1 \right)$$

is equal to 0 when $x > r - mb$ and is equal to the last term in the sum,

$$P\left(\sum_{k=1}^{m} (S_k - b)_+ = r - mb, I_1 = \cdots = I_m = 1 \right),$$

when $x = r - mb$. Since each term in $\sum_{k=1}^{m} (S_k - b)_+$ is at least one, the sum should be at least m. In other words, the only nonzero term in the last sum on the right-hand side of (7) is

$$\begin{aligned} &P\left(\sum_{k=1}^{m} (S_k - b)_+ = r - mb, I_1 = \cdots = I_m = 1 \right) \\ &\quad = P\left(I_1 = \cdots = I_m = 1\right) = P\left(S_1 > b, \ldots, S_m > b\right) = P(S_1 > mb) \\ &\quad = \binom{r - mb - 1}{m - 1} \Big/ \binom{r - 1}{m - 1} \end{aligned} \tag{10}$$

provided $r - mb \ge m$. Upon collecting all of our findings in (4), (6), (8), and (9), we have the following modification of Theorem 2.2 of Holst [10].

Theorem 1 *The support of the rv V representing the length of the vacant region is given by $\{(r - mb)_+, \ldots, (r - m - b + 1)_+\}$. When $r < m + b$, V is degenerate at 0. When $r > m$ and $b = 1$, V is degenerate at $(r - m)$. When $r \ge m + b$ and $r - mb \le 0$, $P(V = 0)$ is given by (5). In all other cases, $P\,(V = x)$ is given by*

$$\begin{aligned} &\left\{ \sum_{y=1}^{m-1} \binom{m}{y} \sum_{t=0}^{m-y-1} (-1)^t \binom{m - y}{t} \binom{x - 1}{y - 1} \binom{r - (y + t)\, b - x - 1}{m - y - 1} \right. \\ &\quad \left. + I(x = r - mb \ge m) \binom{r - mb - 1}{m - 1} \right\} \Big/ \binom{r - 1}{m - 1}. \end{aligned} \tag{11}$$

Remark 1 The actual range of values for y and t for which the associated terms are positive is more restricted than given by the limits in the double sum in (11) in a way that depends on x. For example, when $x = r - mb$, the lowest value V can assume is positive, the terms are positive for all $1 \le y \le m - 1$ and $0 \le t \le \min\{m - y - 1, m - y - (m - y)/b\}$. In contrast, when $x = r - m - b + 1$, the highest value V can assume, $(y, t) = (1, 0)$ is the only combination that produces a positive term. In that case, (11) yields

Table 1 The pmf $P(V = x)$ for $r = 10, m = 5$ for various values of b

b	x					
	0	1	2	3	4	5
1	0	0	0	0	0	1
2	0.008	0.159	0.476	0.317	0.040	
3	0.405	0.397	0.159	0.040		
4	0.802	0.159	0.040			
5	0.960	0.040				
6	1					

$$P(V = r - m - b + 1) = m \Big/ \binom{r-1}{m-1},$$

a quantity free of b.

Table 1 provides the pmf of V for $r = 10$ and $m = 5$. It shows how the probability mass shifts towards values close to 0 as b increases.

3 Moments of V

Instead of using the pmf for V to compute the first two moments of V, we take advantage of an exchangeability argument to derive them from those of S_k. We will use the following representations for the first two moments of nonnegative integer valued rv's X and Y.

$$E(X) = \sum_{i=0}^{\infty} P(X > i); \tag{12}$$

$$E(X^2) = 2\sum_{i=0}^{\infty} i\,P(X > i) + E(X); \tag{13}$$

$$E(XY) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} P(X > i, Y > j). \tag{14}$$

The first two are well-known. Equations (12) and (13) are given in, for example, David and Nagaraja [5, p. 43], and go back to Feller's [7] classical work. Expression (14) is similar to known results for the continuous case; see, for example, the formula for the covariance in Barlow and Proschan [3, p. 31], and the idea goes back to [9]; also see Wellner [16].

Here we give a short proof of (14) when X and Y are nonnegative integer valued rv's.

$$E(XY) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} ij\,P(X=i, Y=j)$$
$$= \sum_{i=0}^{\infty} i \sum_{j=0}^{\infty} P(X=i, Y>j)$$
$$= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} P(X>i, Y>j)$$

upon using the idea of the form on the right-hand side of (12) twice.

The moment expressions simplify further by the use of the following well-known identity: For positive integers $c \le a$,

$$\sum_{k=c}^{a} \binom{k-1}{c-1} = \binom{a}{c}. \tag{15}$$

Harris et al. [8, p.141] derive this identity using an induction argument (see their Eq. (2.10)). We give a simple probabilistic proof.

Proof Multiply both sides by $(1/2)^a$. Then the right-hand side, $\binom{a}{c}(1/2)^a$, represents the probability that in a tosses of a fair coin there are exactly c heads. Now if we have c heads, this event can be written as the union of disjoint events $E_c, \ldots, E_a$ where E_k is the event that we have exactly c heads and the cth head appears at the kth toss. By a negative binomial type argument we know that this probability is

$$\binom{k-1}{c-1}(1/2)^c(1/2)^{a-c} = \binom{k-1}{c-1}(1/2)^a.$$

Now sum this over k from c to a. □

Theorem 2 *Let $W_i = (S_i - b)_+$, for $i = 1, 2$. The W_i are exchangeable and for $r \ge m + b$*

$$E(W_1) = \binom{r-b}{m} \Big/ \binom{r-1}{m-1}, \tag{16}$$

and

$$E(W_1^2) = (2(r-b)+1)E(W_1) - 2m\frac{\binom{r-b+1}{m+1}}{\binom{r-1}{m-1}}. \tag{17}$$

For $r \ge m + 2b$,

$$E(W_1W_2) = \binom{r-2b+1}{m+1} \Big/ \binom{r-1}{m-1}, \tag{18}$$

and $E(W_1W_2) = 0$ if $r < m + 2b$.

Proof Exchangeability follows from Theorem 2.1 of Holst [10]. Now

$$\begin{aligned} E(W_1) &= \sum_{i=0}^{r-m-b} P(W_1 > i) \ \text{[from (12)]} \\ &= \sum_{j=b}^{r-m} P(S_1 > j) \\ &= \sum_{j=b}^{r-m} \binom{r-j-1}{m-1} \Big/ \binom{r-1}{m-1} \ \text{[from Theorem 2.1(B), Holst [10]]}. \quad (19) \end{aligned}$$

From (15), the numerator on the right-hand side of (19) reduces to $\binom{r-b}{m}$.

To establish (17), we use the expression for the second moment in (13). Consider

$$\begin{aligned} \sum_{i=0}^{\infty} i\,P(W_1 > i) &= \sum_{j=b}^{r-m} (j-b) P(S_1 > j) \\ &= \sum_{j=b}^{r-m} \{r-(r-j)\} P(S_1 > j) - bE(W_1) \\ &= r \sum_{j=b}^{r-m} P(S_1 > j) - \sum_{j=b}^{r-m} (r-j) P(S_1 > j) - bE(W_1) \\ &= (r-b)E(W_1) - \sum_{j=b}^{r-m} (r-j) \binom{r-j-1}{m-1} \Big/ \binom{r-1}{m-1} \quad (20) \end{aligned}$$

from the expression for $P(S_1 > i)$ in Theorem 2.1, Part (B) of Holst [10]. The numerator in the second term in (20) above can be expressed as

$$m \sum_{i=b}^{r-m} \binom{r-i}{m} = m \sum_{j=m+1}^{r-b+1} \binom{j-1}{m} = m \binom{r-b+1}{m+1}, \quad (21)$$

where the last equality follows from (15). Upon using (13) with $W_1 = X$ and applying (20) and (21), we obtain (17).

Using (14) with $W_1 = X$ and $W_2 = Y$, and applying Theorem 2.1 Part (E), and Part (B) of Holst [10] in succession, we obtain

$$
\begin{aligned}
E(W_1W_2) &= \sum_{i\geq b}\sum_{j\geq b} P(S_1 > i, S_2 > j) \\
&= \sum_{i=b}^{r-m-b}\sum_{j=b}^{r-m-i} P(S_1 > i, S_2 > j) \\
&= \sum_{i=b}^{r-m-b}\sum_{j=b}^{r-m-i} P(S_1 > i + j) \\
&= \sum_{i=b}^{r-m-b}\sum_{j=b}^{r-m-i} \binom{r-i-j-1}{m-1} \Big/ \binom{r-1}{m-1}.
\end{aligned}
$$

Now, with $k = r - i - j$,

$$
\sum_{j=b}^{r-m-i} \binom{r-i-j-1}{m-1} = \sum_{k=m}^{r-b-i} \binom{k-1}{m-1} = \binom{r-b-i}{m}
$$

from (15). Hence

$$
\sum_{i=b}^{r-m-b}\sum_{j=b}^{r-m-i} \binom{r-i-j-1}{m-1} = \sum_{i=b}^{r-m-b} \binom{r-b-i}{m}.
$$

With $k = r - b - i + 1$, the above sum can be expressed as

$$
\sum_{k=m+1}^{r-2b+1} \binom{k-1}{m} = \binom{r-2b+1}{m+1}.
$$

Hence the claim in (18) holds. Clearly, when $r < m + 2b$, W_1 and W_2 cannot both be positive simultaneously and hence $E(W_1W_2) = 0$. □

From (1) and the exchangeability of the W_i we see that

$$
\begin{aligned}
E(V) &= mE(W_1) \\
\mathrm{Var}(V) &= m\mathrm{Var}(W_1) + m(m-1)\mathrm{Cov}(W_1, W_2) \\
&= m[E(W_1^2) - \{E(W_1)\}^2] + m(m-1)[E(W_1W_2) - \{E(W_1)\}^2] \\
&= mE(W_1^2) + m(m-1)E(W_1W_2) - m^2\{E(W_1)\}^2, \qquad (22)
\end{aligned}
$$

where the expectations on the right-hand side of (22) are given by Theorem 2. Thus we have the following result.

Theorem 3 *If $r \geq m+b$, the first two moments of the rv V representing the number of vacant boxes are given by*

$$E(V) = m\frac{\binom{r-b}{m}}{\binom{r-1}{m-1}}, \tag{23}$$

$$Var(V) = \frac{m(2(r-b)+1)\binom{r-b}{m} - 2m^2\binom{r-b+1}{m+1} + m(m-1)\binom{r-2b+1}{m+1}}{\binom{r-1}{m-1}} - m^2\left\{\frac{\binom{r-b}{m}}{\binom{r-1}{m-1}}\right\}^2, \tag{24}$$

where the coefficient of $m(m-1)$ in (24) is taken to be 0 whenever $r < m+2b$.

Notes

1. After deriving the expression for $E(V)$, we discovered it was previously reported in Ivchenko [11, p. 108]. However, he does not derive a formula for the variance of V.
2. Ivchenko [11, p. 108] also derives an expression for $E(N)$. Using similar exchangeability arguments, we can derive the same expression as well as an expression for the variance of N. In particular, from Theorem 2.1 of Holst [10], we see that

$$E(I_1) = E(I_1^2) = P(S_1 > b) = \binom{r-b-1}{m-1} \Big/ \binom{r-1}{m-1} \equiv p_1,$$

and

$$E(I_1 I_2) = P(S_1 > b, S_2 > b) = \binom{r-2b-1}{m-1} \Big/ \binom{r-1}{m-1} \equiv p_2.$$

Thus from (2) we obtain

$$E(N) = mE(I_1) = m\binom{r-b-1}{m-1} \Big/ \binom{r-1}{m-1}$$

and

$$\begin{aligned} \mathrm{Var}(N) &= m\mathrm{Var}(I_1) + m(m-1)\mathrm{Cov}(I_1, I_2) \\ &= mp_1(1-p_1) + m(m-1)(p_2 - p_1^2) \\ &= m(m-1)p_2 + mp_1 - (mp_1)^2, \end{aligned} \tag{25}$$

where the p_i's are given above.

3. For $m \leq r - b$,

$$E(V) = m\binom{r-b}{m}\Big/\binom{r-1}{m-1} = r\binom{r-b}{m}\Big/\binom{r}{m} \tag{26}$$

$$= \frac{(r-b)!\,(r-m)!}{(r-b-m)!\,(r-1)!}. \tag{27}$$

   As seen from (27), $E(V)$ is symmetric in b and m, even though the pmf for V is not symmetric in these parameters. The symmetry also does not hold for the second moment.
4. As mentioned in Theorem 1, if $r < m + b$, $P(V = 0) = 1$. Thus if $b \geq r - m + 1$ or $m + b > r$, all the S_i are b or less. Thus, $E(V) = \text{Var}(V) = 0$ whenever $r < m + b$. When $b = 1$ and $r > m$, V is degenerate at $r - m$ and in that case $E(V) = r - m$ and $\text{Var}(V) = 0$.

4 The Case with Random *m*

Up to now we have assumed the number of starting points m is fixed and restricted to values in $\{2, \ldots, r\}$. We now consider an extension in which the number of starting points is an rv M with support $\{0, \ldots, r\}$ and pmf $P(M = m) = p_M(m)$. We let V_M denote the number of vacant boxes in $\{0, \ldots, r - 1\}$ in this case to highlight that the number of starting points is allowed to be random here.

Formally, we construct V_M as follows. We first draw a value for M. If $M = 0$, we designate all boxes as vacant and set $V_M = r$. If $M = m > 0$, we draw m starting points without replacement from the boxes labeled $0, \ldots, r - 1$. Let $\{R_0, \ldots, R_{m-1}\}$ denote the identities of these starting points. For each $k \in \{0, \ldots, m-1\}$ we designate the boxes

$$\{R_k, (R_k + 1) \mod r, \ldots, (R_k + b - 1) \mod r\}$$

as covered. Any box that is not covered is labeled vacant. Define J_i, for each $i = 0, \ldots, r - 1$, as equal to 1 if box i is vacant and 0 if covered. Then the number of vacant boxes is

$$V_M = \sum_{i=0}^{r-1} J_i. \tag{28}$$

The discrete circle covering problem we started with is thus a special case of this formulation in which M is degenerate with full mass at a single value m. We now describe two different approaches for deriving the mean and variance of V_M in the general case where M has a nondegenerate distribution.

When $2 \leq m \leq r$, the rv V_M given $M = m$ has the same distribution as V in the circle covering problem with m starting points. When $M = 0$ and $M = 1$, the rv

V_M has a degenerate distribution with full mass at r and $r-b$, respectively. We can therefore use a simple conditioning argument to obtain the following:

Theorem 4 *The pmf, mean, and variance of V_M are respectively given by*

$$P(V_M = x) = \sum_{m=0}^{r} P(V_m = x)\, p_M(m), \tag{29}$$

$$\begin{aligned} E(V_M) &= \sum_{m=0}^{r} E(V_m)\, p_M(m) \\ &= r \sum_{m=0}^{r-b} \frac{\binom{r-b}{m}}{\binom{r}{m}}\, p_M(m), \end{aligned} \tag{30}$$

and

$$Var(V_M) = Var(E(V_M|M)) + E(Var(V_M|M)), \tag{31}$$

where, for $2 \le m \le r$, $P(V_m = x)$ is given in Theorem 1 and $E(V_m)$ and $Var(V_m)$ are given in Theorem 3. Further, V_0 is degenerate at r and V_1 is degenerate at $(r-b)_+$.

In (30) we have used the second form of $E(V_m)$ in (26) and the fact that $E(V_1) = (r-b)_+ = r\binom{r-b}{1}/\binom{r}{1}$.

An alternative approach to computing moments for V_M makes use of (28). As long as the sampling mechanism is completely symmetric with respect to the r available boxes, the J_i's are identically distributed (but not exchangeable) and

$$\begin{aligned} E(V_M) &= r\,P(J_0 = 1), \\ Var(V_M) &= r\,P(J_0 = 1)P(J_0 = 0) + r \sum_{k=1}^{r-1} Cov(J_0, J_k); \end{aligned} \tag{32}$$

and $Cov(J_0, J_k) = P(J_0 = 1, J_k = 1) - [P(J_0 = 1)]^2$. As we discuss below, this approach will often be more useful for computing moments than summing conditional moments.

4.1 Moments for Particular Cases

We now describe three specific examples in which m is random. The first two are motivated by cases that have been studied in the application to financial contagion we discuss later in Sect. 6, although the first one turns out to be of more general interest, as we discuss below.

Binomial M. Let M assume a Binomial (r, p) distribution, i.e., $p_M(m) = \binom{r}{m} p^m (1-p)^{r-m}$, $0 \le m \le r$. This is equivalent to assuming that each box $i \in \{0, \ldots, r-1\}$ is chosen as a starting point with probability p independently of whether any other box is chosen as a starting point. The expression for the mean in (30) simplifies substantially;

$$E(V_M) = r \sum_{m=0}^{r-b} \binom{r-b}{m} p^m (1-p)^{r-m} = r(1-p)^b. \tag{33}$$

We could in principle compute $Var(V_M)$ using (31). However, it is easier to compute this statistic using the indicator variables J_i.

To evaluate the moments for V_M using these indicator variables, note that box 0 is vacant if and only if none of the boxes labeled $0, r-1, \ldots, r-b+1$ are starting points, i.e. $\{R_0, \ldots, R_{M-1}\} \cap \{0, r-1, \ldots, r-b+1\} = \varnothing$. Recall that when M is binomial, each box i is chosen as a starting point with probability p independently of whether any other box is a starting point. Hence, $P(J_0 = 1) = q^b$, where $q = 1-p$, and $E(V_M) = rq^b$ in line with (33).

We now use (32) to compute $Var(V_M)$. When $b = 1$ the J_i are independent, and $Var(V_M) = rpq$. When $2 \le b \le (r+1)/2$, we have

$$P(J_0 = 1, J_k = 1) = \begin{cases} q^{b+k}, & \text{if } 1 \le k \le b-1 \\ q^{b+r-k}, & \text{if } r-b+1 \le k \le r-1 \\ q^{2b}, & \text{if } b \le k \le r-b; \end{cases}$$

$$Cov(J_0, J_k) = \begin{cases} q^{b+k} - q^{2b}, & \text{if } 1 \le k \le b-1 \\ q^{b+r-k} - q^{2b}, & \text{if } r-b+1 \le k \le r-1 \\ 0, & \text{if } b \le k \le r-b, \end{cases}$$

and thus we obtain

$$\begin{aligned} \sum_{k=1}^{r-1} Cov(J_0, J_k) &= \sum_{k=1}^{b-1} q^{b+k} + \sum_{k=r-b+1}^{r-1} q^{b+r-k} - 2(b-1)q^{2b} \\ &= 2q^b \frac{q - q^b}{p} - 2(b-1)q^{2b}. \end{aligned} \tag{34}$$

It follows that when $1 < b \le (r+1)/2$,

$$Var(V_M) = rq^b \left\{ 1 + \frac{2q}{p} - q^b \left(\frac{2}{p} + 2b - 1 \right) \right\}. \tag{35}$$

This form also holds when $b = 1$. When $(r+1)/2 < b < r-1$, we have

$$P(J_0 = 1, J_k = 1) = \begin{cases} q^{b+k}, & \text{if } 1 \le k \le r-b-1 \\ q^r, & \text{if } r-b \le k \le b \\ q^{b+r-k}, & \text{if } b+1 \le k \le r-1; \end{cases}$$

$$Cov(J_0, J_k) = \begin{cases} q^{b+k} - q^{2b}, & \text{if } 1 \le k \le r-b-1 \\ q^r - q^{2b}, & \text{if } r-b \le k \le b \\ q^{b+r-k} - q^{2b}, & \text{if } b+1 \le k \le r-1. \end{cases}$$

Hence

$$\begin{aligned} \sum_{k=1}^{r-1} Cov(J_0, J_k) &= \sum_{k=1}^{r-b-1} q^{b+k} + \sum_{k=b+1}^{r-1} q^{b+r-k} + (2b-r+1)q^r - (r-1)q^{2b} \\ &= 2q^b \frac{q - q^{r-b}}{p} + (2b-r+1)q^r - (r-1)q^{2b}. \end{aligned} \tag{36}$$

It now follows from (32) that when $(r+1)/2 < b < r-1$,

$$Var(V_M) = rq^b \left\{ 1 + \frac{2q}{p} - q^{r-b}\left(\frac{2}{p} + r - 2b - 1\right) - rq^b \right\}. \tag{37}$$

The form in (37) holds for $b = r-1$ as well, in which case it reduces to

$$Var(V_M) = rq^{r-1}\left\{1 + q(r-1) - rq^{r-1}\right\}.$$

For $b \ge r$, V_M has a two point distribution with support $\{0, r\}$ and $E(V_M) = rq^r$, and $Var(V_M) = r^2 q^r(1-q^r)$. Thus, for $1 \le b \le r$, $E(V_M)$ has the same form for all b, but the variance expression has three distinct forms depending on b.

Mixture of Binomials. As above let M be Binomial(r,p) and assume p is also an rv with support $(0, 1)$, i.e., M is a mixture of binomials. From (33), the conditional mean of $E(V_M|p) = r(1-p)^b$ and hence $E(V_M) = rE(1-p)^b$. If p has a Beta(α, β) distribution $(\alpha, \beta > 0)$, this expectation simplifies to

$$E(V_M) = r\frac{(\alpha+\beta)\cdots(\alpha+\beta+b-1)}{\beta\cdots(\beta+b-1)},$$

since b is a positive integer. Using the expressions for the conditional variance of V_M for a given p and the relation

$$Var(V_M) = Var(E(V_M|p)) + E(Var(V_M|p)),$$

we can also obtain an expression for $Var(V_M)$ if p has a Beta(α, β) distribution with $\alpha > 1$. For $\alpha \geq 1$, $Var(V_M)$ does not exist.

Sampling With Replacement. Suppose we choose starting points by taking n (≥ 1) draws from $\{0, \ldots, r-1\}$ with replacement. The resulting number of starting points M will be random with support $\{1, \ldots, n\}$, and its pmf is given by (see, e.g., Feller [7, p. 102])

$$\begin{aligned} p_M(m) &= \binom{r}{m} \frac{1}{r^n} \sum_{j=0}^{m} (-1)^j \binom{m}{j} (m-j)^n \\ &= \frac{r(r-1)\cdots(r-m+1)}{r^n} S(n, m), \end{aligned} \tag{38}$$

where the $S(n, m)$ are Stirling numbers of the second kind (see, e.g., Johnson and Kotz [12, pp. 110] for the Stirling number connection).

Alternatively, we can use (32) to obtain the first two moments of V_M for this case. Note that $P(J_0 = 1) = \{(r-b)/r\}^n$; thus

$$E(V_M) = r\left(1 - \frac{b}{r}\right)^n, 1 \leq b \leq r. \tag{39}$$

Since for $2 \leq b \leq r/2$,

$$P(J_0 = 1, J_k = 1) = \begin{cases} \left(\frac{r-(b+k)}{r}\right)^n, & \text{if } 1 \leq k \leq b-1 \\ \left(\frac{k-b}{r}\right)^n, & \text{if } r-b+1 \leq k \leq r-1 \\ \left(\frac{r-2b}{r}\right)^n, & \text{if } b \leq k \leq r-b, \end{cases} \tag{40}$$

we obtain

$$Cov(J_0, J_k) = \begin{cases} \left(\frac{r-(b+k)}{r}\right)^n - \left(\frac{r-b}{r}\right)^{2n}, & \text{if } 1 \leq k \leq b-1 \\ \left(\frac{k-b}{r}\right)^n - \left(\frac{r-b}{r}\right)^{2n}, & \text{if } r-b+1 \leq k \leq r-1 \\ \left(\frac{r-2b}{r}\right)^n - \left(\frac{r-b}{r}\right)^{2n}, & \text{if } b \leq k \leq r-b. \end{cases}$$

Thus,

$$\begin{aligned} &\sum_{k=1}^{r-1} Cov(J_0, J_k) \\ &\quad = 2\sum_{k=1}^{b-1} \left(\frac{r-b-k}{r}\right)^n + (r-2b+1)\left(\frac{r-2b}{r}\right)^n - (r-1)\left(\frac{r-b}{r}\right)^{2n}, \end{aligned}$$

and for $2 \le b \le r/2$, (32) yields the following expression for $Var(V_M)/r$:

$$\left(1-\frac{b}{r}\right)^n + 2\sum_{k=1}^{b-1}\left(1-\frac{b+k}{r}\right)^n + (r-2b+1)\left(1-\frac{2b}{r}\right)^n - r\left(1-\frac{b}{r}\right)^{2n}. \quad (41)$$

When $b = 1$, the second term above is to be interpreted as 0. For $r/2 \le b \le r-2$,

$$P(J_0 = 1, J_k = 1) = \begin{cases} \left(\frac{r-(b+k)}{r}\right)^n, & \text{if } 1 \le k \le r-b-1 \\ \left(\frac{k-b}{r}\right)^n, & \text{if } b+1 \le k \le r-1 \\ 0, & \text{if } r-b \le k \le b, \end{cases}$$

and consequently we obtain

$$\frac{Var(V_M)}{r} = \left(1-\frac{b}{r}\right)^n + 2\sum_{k=1}^{r-b-1}\left(1-\frac{b+k}{r}\right)^n - r\left(1-\frac{b}{r}\right)^{2n}.$$

When $b = r-1$, the summation above is to be interpreted as 0. In fact, in that case, V_M is a Bernoulli rv with success probability $(1/r)^{n-1}$. For $b \ge r$, V_M is degenerate at 0.

Without-Replacement Sample. In this setup, we have

$$P(J_0 = 1) = \binom{r-b}{m} \Big/ \binom{r}{m}$$

leading to the second form for $E(V)$ given in (26). Using a representation for $P(J_0 = 1, J_k = 1)$ that parallels (40), an expression for the variance can be written using (32). We will not pursue it as we already have a compact expression available in (24).

5 Limiting Properties of V

5.1 Limiting Distributions

We now return to the case where the number of starting points m is nonrandom. Holst [10, Theorem 3.2] argues that as $r, b \to \infty$ with $b/r \to a$ for some $0 < a < 1$, $V/r \xrightarrow{d} V_a$ where V_a has the same distribution as the length of non-covered segments when m arcs of length a are dropped at random on a circle with unit circumference. Siegel [13, Theorem 3] has shown that the distribution of V_a can be expressed as the mixture of a degenerate and a continuous rv. Specifically, he shows that $P(V_a(m) = (1-ma)_+) = p_a(m)$ where

$$p_a(m) = \sum_{i=0}^{m-1} (-1)^i \binom{m}{i} (1 - ia)_+^{m-1}, ma > 1 \tag{42}$$

$$= (1 - ma)^{m-1}, ma \le 1, \tag{43}$$

and with probability $1 - p_a(m)$, $V_a(m)$ behaves like a continuous rv $W_a(m)$ having the pdf $f(w; a, m)$ given by

$$f(w; a, m)$$

$$= \frac{m}{1 - p_a(m)} \sum_{i=1}^{m} \sum_{j=1}^{m-1} (-1)^{i+j} \binom{m-1}{i-1} \binom{m-1}{j} \binom{i-1}{j-1} w^{j-1} (1 - ia - w)_+^{m-j-1},$$

$$(1 - ma)_+ < w < 1 - a, \tag{44}$$

with the convention that $(1 - ia - w)_+^0$ is interpreted as 1 if $1 - ia - w \ge 0$, and as 0, otherwise. We now show the following.

Lemma 1 *If $r, b \to \infty$ such that $b/r \to a, 0 < a < 1$, then*

$$P\{V = (r - mb)_+\} \to p_a(m) \equiv P\{V_a = (1 - ma)_+\},$$

given by (42) when $ma > 1$, and by (43) when $ma < 1$. When $ma = 1$, both (42) and (43) reduce to 0.

Proof When $ma > 1$, $r - mb$ is eventually negative, our interest then is in the limiting form of $P(V = 0)$ given in (5). Consider the jth term there, excluding the factor $(-1)^m \binom{m}{j}$:

$$\frac{\binom{r-jb-1}{m-1}}{\binom{r-1}{m-1}} = \frac{(r - jb - 1) \cdots (r - jb - m + 1)}{(r - 1) \cdots (r - m + 1)},$$

if $r - jb = r(1 - j(b/r)) > m - 1$; and it is 0 if $r(1 - j(b/r)) \le m - 1$. So if $b/r \to a$ with $1 - ma < 0$, the above ratio converges to $(1 - ja)_+^{m-1}$. Thus the limit is given by (42).

Whenever $ma < 1$, since $r - mb = r(1 - m(b/r))$, $r - mb$ eventually exceeds any fixed m. In that case the term (10) converges to $(1 - ma)^{m-1}$. The remaining finite number of terms in the numerator on the right in (11) are finite and each is of $o(r^{m-1})$ whereas the denominator is $O(r^{m-1})$. Thus, the only nonzero term in the limit is that of (10) and it coincides with (43).

If $ma = 1$, (43) is obviously 0 and now we show that (42) also converges to 0 as $a \to (1/m)^+$. For this we consider the continuous uniform spacing problem where one chooses at random $m - 1$ points $U_1, \ldots, U_{m-1}$ from the interval $(0, 1)$. With spacings defined as $Y_i = U_{i:m-1} - U_{i-1:m-1}, i = 1, \ldots, m$, where $U_{0:m-1} = 0$ and $U_{m:m-1} = 1$, it is known that the distribution function of the continuous rv $Y_{(m)}$

representing the maximal spacing can be expressed as (see, e.g., David and Nagaraja [5], p. 135)

$$P(Y_{(m)} \le a) = 1 - P(Y_{(m)} > a) = \sum_{i=0}^{m} (-1)^i \binom{m}{i} (1 - ia)_+^{m-1}, \qquad (45)$$

for all a in $(0, 1)$. Since by construction the maximal spacing $Y_{(m)}$ exceeds $1/m$ with probability 1, the right-hand sum in (45) is 0 whenever $a \le 1/m$ or $ma \le 1$ and thus approaches 0 as $a \to (1/m)^+$. The difference between this sum and the sum in (42), $(-1)^m(1 - am)_+^{m-1}$, is 0 whenever $a \ge (1/m)$. Thus we conclude that as $a \to (1/m)^+$ the expression in (42) converges to 0. □

Notes

5. When $b = ar$, with $ma < 1$, we have seen that the expression in (10) converges to $(1-ma)^{m-1}$, while the other terms contributing to $P(V = r-mb)$ converge to 0, indicating the dominant nature of this term missing in Holst's Theorem 2.2 [10]. That is, the term missing from Holst's expression is precisely what converges to the degenerate component of the rv in the continuous case.
6. Holst's Theorem 3.2 gives expressions for $P(V_a = 0)$ and the pdf of the continuous part. Lemma 1 reveals that his expressions are imprecise and fail to properly account for the range of V.
7. Siegel's [13] version of (42) [his expression (3.23)] has a summation that includes an additional term with $i = m$. In view of the assumption that $ma > 1$, the corresponding term is 0, and hence they coincide. Further, in view of the above lemma, we can conclude that when $ma = 1$ both (42) and (43) hold.
8. Consider the case where M is random, specifically where M is generated by taking n draws with replacement from $\{1, \ldots, r\}$. As $r \to \infty$ while n is held fixed, the first factor in the expression for the pmf of M in (38) converges to 0 whenever $m < n$ and to 1 when $m = n$. Since $S(n, n) = 1$, it follows that $M \xrightarrow{P} n$, the sample size, and the limiting distribution of V_M is the same as the limiting distribution of V when m is replaced by n. More generally, given any process M that converges in probability to a degenerate distribution as $r \to \infty$, the distribution for V_M will converge to the same limiting distribution as V with m corresponding to the value that M collapses to.

For $m = 5$ and selected r values, Figs. 1 and 2 respectively provide the normalized conditional pmf of V given the event $\{V > (r - mb)_+\}$, for $a = 0.1$ and 0.25. The case of $r = \infty$ corresponds to the conditional pdf $f(w; a, m)$ of the continuous case, given in (44). Both these figures suggest that by the time r reaches 500, we are close to the limiting result, indicating that when the sampling fraction is under 1 %, $f(w; a, m)$ provides a close approximation to the conditional pmf of V.

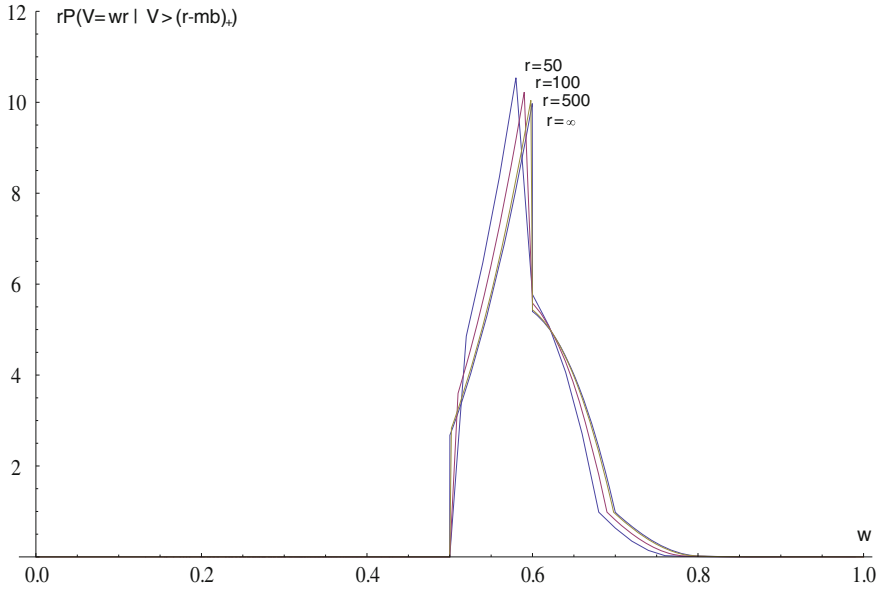


Fig. 1 $rP\{V = rw|V > (r - bm)_+\}$ for $m = 5$, $a = 0.1$, $b = ar$, for selected r; $f(w; a, m)$, given in (44), corresponds to $r = \infty$

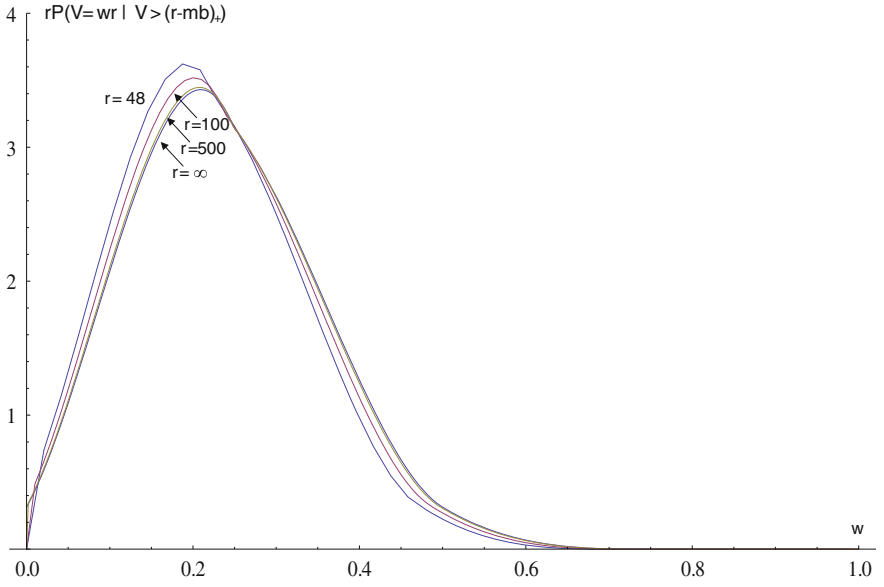


Fig. 2 $rP\{V = rw|V > (r - bm)_+\}$ for $m = 5$, $a = 0.25$, $b = ar$, for selected r; $f(w; a, m)$, given in (44), corresponds to $r = \infty$

5.2 Limiting Moments

Limits for Large r and b

Since V/r is uniformly bounded, convergence in distribution implies that $E(V/r)^k \to E(V_a^k)$ when $r \to \infty$ and $b/r \to a$ and $ma \neq 1$. Siegel has shown in his Theorem 2 that,

$$E(V_a^k) = \binom{k+m-1}{m}^{-1} \sum_{i=1}^{k} \binom{k}{i}\binom{m-1}{i-1}(1-ia)_+^{m+k-1}, k \geq 1. \tag{46}$$

Hence we can obtain approximations to any moment of V when m is small and r and b are large using the moments of V_a.

Table 2 provides some key facts about the features of the distributions of V and the limiting rv V_a for $m = 5$, $r = 20, 50$ and b values up to 20 corresponding to a good range of a values. It shows that as a increases $p_a(m)$ decreases for $a \leq 1/m$, it is 0 when $ma = 1$ and then $p_a(m)$ increases. Note that whenever a reaches $1/m$ from below, the lower limit of the support of V_a moves towards 0 and whenever $ma > 1$, the lower limit remains at 0. This limiting pattern is closely followed by V when $r = 50$, but not that closely when $r = 20$. The moments converge fairly quickly to the limiting values. The mean is better approximated by the limit for small b (= ar) whereas for the standard deviation, large b values tend to be slightly more efficient.

Limits for Large r and m

When b is held fixed and $r, m \to \infty$ such that $m/r \to p = 1-q, 0 < p < 1$, Holst [10, Theorem 4.2] has shown that V is asymptotically normal and has given the first two moments of the limit distribution. We now derive asymptotic approximations for the expressions for $E(V)$ and $Var(V)$ in (23) and (24) to study their limiting properties. When b is held fixed, we have

$$C(r,m;b) \equiv \frac{(r-b)!(r-m)!}{r!(r-m-b)!} = \frac{(r-m)(r-m-1)\cdots(r-m-b+1)}{r(r-1)\cdots(r-b+1)} \approx q^b.$$

Hence

$$\frac{\binom{r-b}{m}}{\binom{r-1}{m-1}} = \frac{r}{m}C(r,m;b) \approx \frac{r}{m}q^b;$$

$$\frac{\binom{r-b+1}{m+1}}{\binom{r-1}{m-1}} = \frac{r(r-b+1)}{m(m+1)}C(r,m;b) \approx \frac{r(r-b+1)}{m(m+1)}q^b;$$

Table 2 Properties of V and V_a for $m = 5$, $r = 20, 50$, and selected b values; $a = b/r$

b	$P(V = (r - mb)_+)$	$p_a(m)$	$E(V)$	$rE(V_a)$	$SD(V)$	$rSD(V_a)$
$r = 20$						
1	1	0.3164	15	15.48	0	0.49
2	0.008	0.0625	11.05	11.14	0.80	1.14
3	0.405	0.0040	7.98	8.87	1.42	1.71
4	0.802	0	5.63	6.55	1.84	2.11
5	0.960	0.0040	3.87	4.75	2.04	2.31
$r = 50$						
1	1	0.6561	45	45.196	0	0.341
2	0.6407	0.4096	40.408	40.769	0.586	0.899
3	0.3882	0.2401	36.199	36.695	1.213	1.536
4	0.2189	0.1296	32.348	32.954	1.875	2.196
5	0.1121	0.0625	28.832	29.525	2.532	2.842
6	0.0502	0.0256	25.628	26.387	3.156	3.452
7	0.0183	0.0081	22.716	23.521	3.726	4.009
8	0.0047	0.0016	20.075	20.911	4.229	4.503
9	0.0006	0.0001	17.685	18.537	4.658	4.925
10	0.0000	0	15.528	16.384	5.007	5.270
11	0.0006	0.0001	13.587	14.436	5.273	5.538
12	0.0047	0.0016	11.845	12.678	5.457	5.728
13	0.0179	0.0081	10.287	11.095	5.560	5.842
14	0.0452	0.0246	8.897	9.675	5.586	5.881
15	0.0885	0.0545	7.661	8.404	5.541	5.852
16	0.1467	0.0989	6.566	7.27	5.430	5.759
17	0.2158	0.1561	5.601	6.262	5.261	5.609
18	0.2912	0.2226	4.752	5.369	5.041	5.408
19	0.3689	0.2944	4.010	4.581	4.780	5.164
20	0.4455	0.3680	3.363	3.888	4.486	4.886

$$\frac{\binom{r-2b+1}{m+1}}{\binom{r-1}{m-1}} = \frac{r(r-2b+1)}{m(m+1)} C(r, m; 2b) \approx \frac{r(r-2b+1)}{m(m+1)} q^{2b}.$$

Upon plugging these approximations into the expressions given in Theorem 3, we obtain

$$E(V) \approx rq^b, \tag{47}$$

and $Var(V)/r$ is

$$\approx q^b \left\{ 2(r-b) + 1 - 2\frac{m}{m+1}(r-b+1) \right\} + q^{2b} \left\{ \frac{m-1}{m+1}(r-2b+1) - r \right\}$$
$$= \frac{q^b}{m+1}\{2(r-b)+1-m\} - \frac{q^{2b}}{m+1}\{2r+(2b-1)(m-1)\}$$
$$\approx q^b \frac{1+q}{p} - q^{2b}\left\{ \frac{2}{p} + 2b - 1 \right\}. \tag{48}$$

We note that the approximations of the mean and variance of V above, where m grows deterministically with r, match with the exact moments of V_M in the case where M is Binomial(r, p) with p equal to the limiting value of m/r. In particular, (47) matches with (33) and (48) matches with (35), as we would expect given b is fixed and r tends to ∞. The expressions we obtain for the variances differ from the variance of the limiting normal distribution reported in Theorem 4.2 of Holst [10]. While the convergence in distribution and convergence of moments are not directly related, it appears his expression is incorrect.

Remark 2 To appreciate why the asymptotic approximations for the mean and variance of V coincide with the exact moments of V_M when M has a binomial distribution, observe that when $m/r \to p$, if we take any pair of boxes i and j, the probability that each is drawn as a starting point converges to p while the probability that both are drawn converges to p^2, i.e., the two events are asymptotically independent. But recall that this independence is what distinguishes the case where M is binomial when r is finite. Consistent with this, J_i converges in probability to a Bernoulli rv with success probability $(1-p)^b$, which is the exact distribution of J_i when M is binomial. In other words, when the number of starting points m is deterministic but grows proportionately with r, the number of vacant boxes in a block of boxes of fixed size behaves asymptotically in the same way as the number of vacant boxes in that block for the same r where M is distributed Binomial(r, p).

Sampling with Replacement: Limits for Large r and n

Finally, consider the case where M is random and generated by drawing n times with replacement from all boxes. As we discussed above, this case converges to the discrete circle covering problem with n boxes. Consistent with this, suppose $r, n \to \infty$ such that $n/r \to \theta$, $0 < \theta < 1$. Using (39) and (41) we obtain

$$E(V_M) \approx re^{-b\theta} = rq^b \tag{49}$$
$$\frac{\mathrm{Var}(V_M)}{r} \approx e^{-b\theta} + 2\sum_{k=1}^{b-1} e^{-(b+k)\theta} + (r-2b+1)e^{-2b\theta} - re^{-2b\theta}$$
$$= q^b \frac{1+q}{p} - q^{2b}\left\{ \frac{2}{p} + 2b - 1 \right\} \tag{50}$$

upon simplification with $-\log(q) = \theta$, and $p = 1 - q$. This matches (33) and (35). Under this scheme, the probability that a particular box is never among the n boxes drawn is $(1 - r^{-1})^n$, and thus the expected number of distinct boxes chosen is

$$E(M) = r\left\{1 - (1 - \frac{1}{r})^n\right\} \approx r(1 - e^{-\theta}) = r(1 - q) = rp.$$

In other words $E(M) \approx rp$.

6 Application to Financial Contagion

We conclude by showing how the discrete circle covering problem we analyzed is related to the literature on financial contagion. One of the workhorse models of financial contagion is due to Eisenberg and Noe [6]. Their model posits that banks are connected via a directed network based on the obligations banks owe one another. If some banks incur losses, they will be unable to meet their required payments to other banks, inflicting losses on other banks. Thus, shocks that affect certain nodes in a network can propagate to other nodes. Eisenberg and Noe derive the vector of clearing payments given the network structure and the identity of the nodes that incur direct losses. One can use this vector to deduce which banks will be adversely affected when certain banks are hit.

Subsequent work has extended the Eisenberg and Noe model by assuming that the number of banks that experience losses as well as their identity is random. For tractability, this literature has focused on simple network structures. For example, Caballero and Simsek [4], Acemoglu et al. [1] and Alvarez and Barlevy [2] all consider contagion in circular networks in which the network of obligations across banks is isomorphic to a circular graph. While these networks bear little resemblance to the pattern of obligations across banks in practice, these settings still provide useful intuition about what determines contagion and how banks might behave when they are uncertain about the extent of contagion. We now show how contagion in circular networks is related to the discrete circle covering problem. We further argue that the connection between the two suggests generalizations of the circle covering problem that to our knowledge have not been noted previously.

Suppose there are r banks, that each bank owes an amount λ to one other bank, and that each bank is in turn owed λ by another bank. Banks can be viewed as connected to one another via a directed network in which a bank points to another bank if the former owes something to the latter. Formally, index banks in the network by $i \in \{0, \ldots, r-1\}$. A circular network is one where bank i owes λ to bank $i + 1$ mod r. We henceforth drop the reference to mod r. As in Alvarez and Barlevy [2], we impose the following assumptions: (1) Each bank owns μ worth of assets that it can sell to repay its outstanding obligations if it needs to; (2) Among the r banks, a random number M with pmf $P(M = m) = p_M(m)$ will be "bad", meaning they incur a loss of size ϕ that must be subtracted from their initial asset holdings μ; (3)

Given $M = m$, each of the $\binom{r}{m}$ groups of size m is equally likely to be those which are bad; and (4) $\mu < \phi < \frac{r}{m}\mu$. The last assumption implies bad banks incur losses that exceed what they can afford to pay by liquidating their assets, but total losses across all m bad banks are still less than the combined value of all assets held among all r banks in the network. Banks are required to pay their full obligation λ if possible, and must sell their asset holdings if they fall short. Although the distribution of the number of bad banks M is unrestricted, Alvarez and Barlevy [2] draw particular attention to the cases where M has a degenerate distribution (i.e., $p_M(m) = 1$ for some m), a binomial distribution, and a mixture of binomials as instructive special cases.[1]

Let x_i denote the amount bank i pays bank $i + 1$ and assume that the bad banks are labeled R_k, $k = 0, \ldots, m-1$, and these are the ones who have incurred a direct external loss of ϕ and the others have no external losses. Given that banks must pay their obligations if they can, the payments $\{x_i\}_{i=0}^{r-1}$ satisfy the following system of equations

$$x_i = \begin{cases} \min\{(x_{i-1} + \mu - \phi)_+, \lambda\}, & i \in \{R_0, \ldots, R_{m-1}\} \\ \min\{x_{i-1} + \mu, \lambda\}, & i \notin \{R_0, \ldots, R_{m-1}\} \end{cases}$$

with $x_{-1} \equiv x_{r-1}$. For $\phi < \frac{r}{m}\mu$, there exists a unique solution $\{x_i\}_{i=0}^{r-1}$ to this system. Bank i is said to be insolvent if $x_i < \lambda$, i.e., if it cannot meet its obligation, and solvent otherwise. Each of the m bad banks are insolvent, since even if they received the full amount λ from the bank that is obligated to them, the fact that $\mu < \phi$ implies $x_{i-1} + \mu - \phi < \lambda$ and so they would be unable to pay in full even after liquidating their assets. Beyond these m bad banks, banks that do not directly suffer losses may still end up insolvent because they are exposed to bad banks either directly—meaning the bank that owes them λ is bad—or indirectly—meaning the bank that owes them λ is good but is exposed to a bad bank. A central question in this framework is to determine the number of banks that are insolvent, i.e., to gauge the extent of contagion when only M banks suffer direct losses.

The results turn out to depend on the parameter λ. Suppose $\lambda \leq \phi - \mu$, and $\lambda = b\mu$, for some integer b. Then $(b+1)\mu \leq \phi$, and $x_{R_k} = 0$, and $x_{R_k+j} = j\mu$, $j = 1, \ldots, \min\{b-1, S_{k+1} - 1\}$, for $k = 0, \ldots, m-1$. Hence, the number of insolvent banks starting from each bad bank is a fixed number b, unless one of those b banks is itself bad. It should be clear that the number of bad banks corresponds to the number of starting points M, while the number of solvent banks corresponds to the number of vacant boxes V with b equal to $\frac{\lambda}{\mu}$. Thus our results provide the small as well as large sample properties of the number of solvent banks in a circular network with a random number of bad banks M when $\lambda \leq \phi - \mu$.

[1] These cases are of interest because the unconditional probability that a bank is bad is higher than, equal to, and lower than the probability that a bank is bad conditional on news that another bank is good when M is degenerate, binomial, and a mixture of binomials, respectively. These distinctions turn out to matter when there is some possibility that news about some banks might be revealed.

The situation where $\lambda > \phi - \mu$ provides a new generalization of the discrete circle covering problem. In this case, b_k, the number of insolvent banks induced by the kth bad bank, becomes a rv that depends on the location of the other bad banks. However, unlike in Siegel and Holst [14], who discuss the continuous case of the circle covering problem assuming the length of the arc b starting at any point is an i.i.d. rv, here the number of insolvent banks starting at each R_k will depend on the distribution of the spacings between the bad banks. To elaborate, $x_{R_k} = (x_{R_k-1} - (\phi - \mu))_+$ is no longer identically 0, and this will affect the number of banks immediately following bank R_k that are insolvent. That is, the number of banks that must be covered starting from the kth bad bank is a rv that depends on the entire collection of spacings $\{S_k\}_{k=1}^{M}$. Alvarez and Barlevy [2] show that when M has a degenerate distribution so that $p_M(m) = 1$ for some m, if $\lambda > m(\phi - \mu)$ then the number of solvent banks V (vacant boxes) is degenerate and equals $r - mb$ where $b = \lambda/\mu$. In the intermediate case where $\phi - \mu < \lambda < m(\phi - \mu)$ the distribution of V is non-degenerate. We leave the investigation of the closed-form expression for this distribution where b_k is a function of $\{S_k\}_{k=1}^{m}$ for future work. More generally, results for Bose-Einstein statistics may prove useful for analyzing contagion in networks that are not circular but still symmetric.

References

1. Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi. 2015. Systemic risk and stability in financial networks. *American Economic Review* 105: 564–608.
2. Alvarez, F., G. Barlevy. 2015. Mandatory disclosure and financial contagion. *University of Chicago Working Paper*.
3. Barlow, R.E., and F. Proschan. 1981. *Statistical theory of reliability and life testing*. Silver Springs, MD: To Begin With.
4. Caballero, R.J., and A. Simsek. 2013. Fire sales in a model of complexity. *The Journal of Finance* 68: 2549–2587.
5. David, H.A., and H.N. Nagaraja. 2003. *Order statistics*, 3rd ed. Hoboken: Wiley.
6. Eisenberg, L., and T.H. Noe. 2001. Systemic risk in financial systems. *Management Science* 47: 236–249.
7. Feller, W. 1968. *An introduction to probability theory and its applications*, vol. 1, 3rd edn. New York: Wiley.
8. Harris, J., J.L. Hirst., and M. Mossinghoff. 2008. *Combinatorics and graph theory*. Springer.
9. Hoeffding, W. 1940. Masstabinvariante Korrelationstheorie. Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universitat Berline, 5, Heft 3, 179–233.
10. Holst, L. 1985. On discrete spacings and the Bose–Einstein distribution. In: Lanke, J., G. Lindgren (Eds.), *Contributions to probability and statistics in honour of Gunnar Blom*, pp. 169–177. Lund.
11. Ivchenko, G.I. 1994. On the random covering of a circle: a discrete model (in Russian). *Diskretnaya Matematika* 6(3): 94–109.
12. Johnson, N.L., and S. Kotz. 1977. *Urn models and their application*. New York: Wiley.
13. Siegel, A.F. 1978. Random arcs on the circle. *Journal of Applied Probability* 15: 774–789.
14. Siegel, A.F., and L. Holst. 1982. Covering the circle with random arcs of random sizes. *Journal of Applied Probability* 19: 373–381.

15. Stevens, W.L. 1939. Solution to a geometric problem in probability. *Annals of Eugenics* 9: 315–320.
16. Wellner, J.A. 1994. Covariance formulas via marginal martingales. *Statistica Neerlandica* 48: 201–207.

A Note on Marginal Count Distributions for Diversity Estimation

John Bunge

Abstract Our problem is to estimate the total number of classes in a population, both observed and unobserved. This is often called the species problem, where the classes are (biological) species, but the same methods apply to "single source" capture-recapture, where only the number of captures for each individual is available (as opposed to the complete capture history). The data is summarized by the frequency counts, i.e., the number of classes observed exactly once, twice, three times, and so on, in the sample. Almost every known statistical procedure uses a mixed Poisson distribution to model the frequency counts, which assumes that the class sizes were independently generated from some latent or underlying mixing distribution, and that the classes independently contributed members to the sample. To depart from these assumptions we require different marginal distributions for the frequency counts. Here we consider distributions having probability generating functions based on generalized hypergeometric functions, first proposed by Kemp in 1968. We show that many of these are not mixed Poisson, and are useful and valuable in the species problem.

Keywords Species problem · Kemp distribution · Generalized hypergeometric function

1 Introduction

Suppose that a population of discrete units is partitioned into C classes. The classes may be regarded as species in a biological application, and the units are the organisms or representatives of the species. There are many other applications: the classes may be types of some kind in information science, such as symbols in an alphabet, and the units may be particular instantiations of the symbols. The same scenario applies

J. Bunge (✉)
Department of Statistical Science, Cornell University, 294 Ives Hall, Ithaca, NY 14853, USA
e-mail: jab18@cornell.edu; j.bunge@cornell.edu

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_9

to "single source" capture-recapture; here the "classes" are the individuals we are interested in counting, and the units are the multiple observations of each (observed) individual. In this case we only have the number of observations for each individual rather than its complete capture history as in standard capture-recapture. In these applications the problem is to estimate the total number of classes C, including both those observed in the sample and those that eluded the sampling mechanism.

Because the identities of the classes are not all known, the data can only be summarized by frequency counts, i.e., $f_1 =$ the number of classes observed exactly once in the sample (the singletons); $f_2 =$ the number of classes observed exactly twice, and so on. Most methods for estimating C proceed by first making an inference about the distribution that (purportedly) generated these counts [1]. Since the only available data consists of the f_j's, the distribution in question is the marginal distribution for the counts. Note that the number of *unobserved* classes is the unobservable random variable f_0. Thus the distribution that generates all counts, observed plus unobserved, has support $\{0, 1, 2, \ldots\}$, while the distribution generating the observable data has support $\{1, 2, \ldots\}$. Most methods begin with a model for $f_0, f_1, f_2, \ldots$ and then assume that the given distribution is zero-truncated; the latter version is then fitted to the data $f_1, f_2, \ldots$. Here we focus on models rather than statistical procedures so we will discuss distributions on $\{0, 1, 2, \ldots\}$, bearing in mind the possibility and potential complication of subsequent zero-truncation. On the latter point, note that if a distribution on $\{0, 1, 2, \ldots\}$ is given by $p_0, p_1, p_2, \ldots$ (where $\sum_{j\geq 0} p_j = 1$), then the zero-truncated version of the distribution is concentrated on $\{1, 2, \ldots\}$ and the corresponding (renormalized) probability masses are $p_1/(1 - p_0), p_2/(1 - p_0), \ldots$. The presence of the denominator $1 - p_0$ can cause considerable numerical difficulties with estimation (based on $f_1, f_2, \ldots$), especially when $p_0, p_1, p_2, \ldots$ is negative binomial, for example. One advantage of methods based on the distributions discussed herein is that it is possible to evade this problem altogether (e.g., see [12]).

The classical approach to estimating C, dating back to the 1943 paper of Fisher et al. [3], is based on the mixed Poisson model. Here we assume that the species' sampling intensities or abundances $\Lambda_1, \ldots, \Lambda_C$ are independent and identically distributed (i.i.d.) draws from a "stochastic abundance distribution" on $(0, \infty)$, say F. The ith species independently contributes a Poisson-distributed number of members to the sample, with Λ_i as its Poisson mean. Marginally, the counts $f_0, f_1, f_2, \ldots$ then summarize the values of C F-mixed Poisson random variables. Various problems have been noted with this model in the literature, including: (i) in the parametric setting, model selection for F and consequent multiple hypothesis testing; (ii) in the nonparametric setting, unbounded bias in the estimate of C and nonidentifiability of F; and (iii) in general, data-analytic issues such as the necessity of truncating large frequencies to obtain an acceptable fit. Thus we are motivated to look for marginal distributions for counts that are not mixed Poisson. Such alternative distributions may arise from a variety of mechanisms, but they necessarily admit alterations to, or generalizations of, parts of the aforementioned mixed Poisson mechanism.

In this note we consider a class of power series distributions, first studied by A. Kemp in 1968 [2, 4, 5], which have probability generating functions based on generalized hypergeometric functions. We use a result of Puri and Goldie

[8] to show that large sub-classes of these distributions are not mixed Poisson. These "Kemp-type" distributions are important in the diversity estimation problem for a variety of reasons, including but not limited to the following. First, they include the negative binomial and Poisson, which are two distributions that have often been used in the past in this problem; but the class extends far beyond these. Consequently we have a way of embedding classical models in a larger family, which allows us to test the adequacy of the former, usually to their disadvantage. In particular the Kemp-type distributions admit much heavier tails than the negative binomial (or Poisson). Second, the Kemp-type distributions are parameterized in a way that permits all models based on them to be nested (see the next section), which leads to an elegant method of numerical fitting and a systematic approach to model selection (as discussed in [12]). Third, they typically allow use of more of the count data, requiring less truncation of the high frequency counts than classical models. We conclude by briefly noting certain statistical methods that will facilitate the use of Kemp-type models in the species problem.

2 Kemp-Type Distributions

Dacey [2] describes Kemp's [4] class of distributions via the probability generating function (p.g.f.)

$$C\,{}_pF_q[(a);(c);\lambda z] \tag{1}$$

where ${}_pF_q$ is the generalized hypergeometric function with p numeratorial and q denominatorial parameters ($p, q \geq 0$), and $C^{-1} = {}_pF_q[(a);(c);\lambda]$; z is the argument of the p.g.f. The parameters of the distribution are $\lambda > 0$, $(a) = (a_1, \ldots, a_p)$ and $(c) = (c_1, \ldots, c_q)$. Considerable effort is expended in [2, 4] to specify the relationships between the parameters required to produce a valid p.g.f., and we do not reproduce these requirements here in detail, but we discuss some aspects below. For information about generalized hypergeometric functions we refer to Spanier and Oldham's *Atlas of Functions* [11] and references therein.

The p.g.f.'s defined by (1) display a wide variety of behaviors depending on the values of p, q, (a) and (c). Here we show that certain of these distributions, at least for low values of p and q, are not mixed Poisson. For this we need the following result of Puri and Goldie [8].

Theorem 1 *A p.g.f. G is a p.g.f. corresponding to a Poisson mixture if and only if $G(\cdot)$ is defined, has continuous derivatives of all orders, and satisfies*

$$G(1) = 1, \quad 0 \leq G(s) \leq 1, \quad \textit{and} \quad 0 \leq G^{(n)}(s) < \infty, \quad n = 1, 2, \ldots,$$

for all real $s \in (-\infty, 1)$.

We proceed by showing that the p.g.f.'s in (1) assume negative values for certain negative values of the argument. Remarkably, this is true for the functions themselves and we will not need to consider derivatives here.

Dacey's Table 1 [2] gives information about the distributions in (1) for $p+q=0,1,2,3$. We first note that "terminating" distributions, i.e., having bounded support, cannot be mixed Poisson (this is also clear from Theorem 1), so we consider only nonterminating distributions, i.e., with unbounded support $\{0,1,2,\ldots\}$. We look at these in Dacey's order, for $p+q\le 2$.

1. $p=q=0$. This is the Poisson distribution.
2. $p=1$, $q=0$. This is the negative binomial distribution $\equiv$ gamma-mixed Poisson.
3. $p=0$, $q=1$. The p.g.f. is

$$C_0F_1[c;\lambda z]=C\sum_{j\ge 0}\frac{(\lambda z)^j}{(c)_j\,j!},$$

where $c>0$, $(c)_j=\Gamma(c+j)/\Gamma(j)$, and $\lambda>0$. Such distributions do not seem to be "named." We have

$${}_0F_1\left[1+\nu;\frac{-x^2}{4}\right]=\Gamma(1+\nu)\left(\frac{2}{x}\right)^{\nu}J_\nu(x),$$

$\nu\neq -1,-2,-3,\ldots$, where J_ν is the Bessel function. Since the latter oscillates around zero on the positive half-line, these distributions are not mixed Poisson.
4. $p=q=1$. Here the p.g.f. is

$$C_1F_1[a;c;\lambda z]=C\sum_{j\ge 0}\frac{(a)_j}{(c)_j}\frac{(\lambda z)^j}{j!},$$

where $a,c,\lambda>0$. Again these distributions are not named in general, although ${}_1F_1$ is called the Kummer function. The latter function is equal to one when $z=0$. Furthermore Spanier and Oldham [11] give a complete list of the numbers of zeroes of ${}_1F_1$ on the negative half-line, for all values of a and c. There are no zeroes when $0<a<c$. We also have

$$\frac{d^n}{dx^n}\left({}_1F_1[a;c;x]\right)=\frac{(a)_n}{(c)_n}\left({}_1F_1[n+a;n+c;x]\right).$$

It follows that for $0<a<c$ the function and all of its derivatives remain positive on the negative half-line, and hence for these parameter values the distributions are mixed Poisson. On the other hand, for $0<c<a$, there is at least one zero on the negative half-line so the function must assume both negative and positive values there. This can be seen because ${}_1F_1$ satisfies the differential equation

$$x\frac{d^2 f}{dx^2} + (c - x)\frac{df}{dx} - af = 0,$$

so if both ${}_1F_1$ and its first derivative are zero at some strictly negative x, its second derivative must also be zero there. Thus for $0 < c < a$ the distributions are not mixed Poisson.

5. Higher order generalized hypergeometric functions: $p = 0, q = 2$; or $p + q \geq 3$. For these cases analytical results on zero-crossings of the relevant functions do not appear to be readily available. However, numerical examples can easily be constructed (in every case we have investigated) that admit negative values on the negative half-line, so that the corresponding distributions are not mixed Poisson. For example: $p = 1, q = 2, a_1 = 3/2, c_1 = 3/2, c_2 = 3$. But in general this is a topic for further research.

3 Statistical Estimation of Population Diversity

We now note some aspects of the problem of estimating C. In general, one begins by assuming a model for both the observed and unobserved frequency counts, $f_0, f_1, f_2, \ldots$. Denote the assumed marginal distribution for these counts by $\{p_j, j = 0, 1, 2, \ldots\}$ where $p_j = \mathbb{P}(j) =$ the probability mass assigned to j; our interest here is in the Kemp-type distributions. The observable data consists of $f_1, f_2, \ldots$ so it is logical to model these counts using the zero-truncated distribution $\{p_j/(1 - p_0), j = 1, 2, \ldots\}$. The question then is how to fit the latter distribution to the data and how to use the resulting information to estimate C. Let θ denote the vector of parameters of $\{p_j\}$ or of $\{p_j/(1 - p_0)\}$.

There are (at least) three potential methods. Under the classical maximum likelihood (ML) approach we fit $\{p_j/(1 - p_0)\}$ to $\{f_1, f_2, \ldots\}$ by ML, which yields an MLE $\hat{\theta}$ of θ. The estimator of C is then either an empirical Horvitz-Thompson estimator, also known as the conditional MLE, $c/(1 - \hat{p}_0) = c/(1 - p_0(\hat{\theta}))$, where c is the observed number of species $c = f_1 + f_2 + \cdots$, or the unconditional MLE which results from globally maximizing the likelihood over both θ and C. It is known that these are asymptotically equivalent [10]. The difficulty in implementing this approach for the Kemp-type distributions is that we do not have an explicit likelihood in most cases, rendering ML infeasible.

The second approach was recently pioneered by Rocchetti et al. [9] and extended by Willis and Bunge [12]. Here we model the ratios of successive frequency counts f_{j+1}/f_j by ratios of successive probabilities p_{j+1}/p_j. The model is fitted by nonlinear regression (which is complicated by heteroscedasticity, dependence, and numerical difficulties). It is then possible to produce a prediction $\hat{f}_0$ of the unobserved count f_0; the resulting estimate of C is $\hat{f}_0 + f_1 + f_2 + \cdots = \hat{f}_0 + c$. This method is ideally suited to the Kemp-type distributions because for these the ratios of probabilities take the convenient form of a rational function of j:

$$\frac{p_{j+1}}{p_j} = \frac{(a_1 + j)(a_2 + j) \cdots (a_p + j)\lambda}{(c_1 + j)(c_2 + j) \cdots (c_q + j)(j + 1)}, \tag{2}$$

$j = 0, 1, \ldots$. Thus this approach admits estimation of C across a broad spectrum of non-mixed Poisson distributions $\{p_j\}$. It is worked out in detail in [12]. In particular, model selection, which consists mainly of choosing the numerator and denominator orders p and q, is dealt with in [12] via an algorithm that essentially steps through models of increasing complexity (higher p and q). The method selects the lowest-complexity (most parsimonious) model that (i) converges numerically; (ii) has no singularities (zeroes of the denominator) in the relevant domain; and (iii) yields a positive prediction of f_0 (this is not guaranteed in this or in many other diversity estimation procedures). It is found that p and q are at most 2 in all cases studied.

While the nonlinear regression method of Willis and Bunge [12] does not require the likelihood function and provides good fits and reasonable estimates and standard errors (and good simulation results), asymptotic mathematical analysis of the procedure is rendered difficult by the fact that the procedure uses the standard Gaussian-error nonlinear regression model for the behavior of f_{j+1}/f_j relative to p_{j+1}/p_j. This is at best an approximation to the true error structure. It would be preferable to base a model on the assumed operative probability distribution rather than on a continuous approximation or analogue—but still without using the likelihood function.

Recently there have been developments in fitting distributions directly from an empirical version of the probability generating function, e.g., [7]. Since the p.g.f. of the zero-truncated distribution is

$$\sum_{j \geq 1} z^j \frac{p_j}{1 - p_0} = \frac{G(z) - p_0}{1 - p_0}, \tag{3}$$

where G is the p.g.f. of the original, untruncated distribution, it should be straightforward to adapt these methods to the zero-truncated case. In particular, the empirical p.g.f. here is $(\sum_{i \geq 1} i f_i)^{-1} \sum_{j \geq 1} f_j z^j$, which is an estimate of (3) for $z \in [0, 1]$. If (3) is a parametric distribution, as it is here, then the parameters can be estimated using well-known methods as described in [6]. A hybrid method for estimating C would then proceed by estimating the parameter (vector) θ via the empirical (zero-truncated) p.g.f.; given the resulting estimate (say) $\tilde{\theta}$ of θ, an estimator of C would then be a new empirical Horvitz-Thompson estimator, $c/(1 - \tilde{p}_0) = c/(1 - p_0(\tilde{\theta}))$. Again, we can abandon the mixed-Poisson assumption if $\{p_j\}$ is Kemp-type (for example). We are currently investigating this approach.

Acknowledgments The author is grateful to Pankaj Choudhury for his invitation to the 60th birthday conference for H.N. Nagaraja, for which this note formed part of a presentation. He thanks an anonymous referee for detailed comments which significantly improved the paper. Most of all he is grateful to H.N. Nagaraja himself, for the latter was the author's Ph.D. adviser and taught him everything that set him on his subsequent path in statistics.

References

1. Bunge, J., A. Willis, and F. Walsh. 2014. Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application*. doi:10.1146/annurev-statistics-022513-115654.
2. Dacey, M.F. 1972. A family of discrete probability distributions defined by the generalized hypergeometric series. *Sankhyā Series B* 34: 243–250.
3. Fisher, R.A., S. Corbet, and C.B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* 12: 42–58.
4. Kemp, A.W. 1968. A wide class of discrete distributions and the associated differential equations. *Sankhyā Series A* 30: 401–410.
5. Kemp, A.W. 2010. Families of power series distributions, with particular reference to the Lerch family. *Journal of Statistical Planning and Inference* 140: 2255–2259.
6. Nakamura, M., and V. Pérez-Abreu. 1993. Empirical probability generating function: An overview. *Insurance: Mathematics and Economics* 12: 287–295.
7. Ng, C.M., S.-H. Ong, and H.M. Srivastava. 2013. Parameter estimation by Hellinger type distance for multivariate distributions based upon probability generating functions. *Applied Mathematical Modelling* 37: 7374–7385.
8. Puri, P.S., and C.M. Goldie. 1979. Poisson mixtures and quasi-infinite divisibility of distributions. *Journal of Applied Probability* 16: 138–153.
9. Rocchetti, I., J. Bunge, and D. Böhning. 2011. Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics* 5: 1512–1533.
10. Sanathanan, L. 1977. Estimating the size of a truncated sample. *Journal of the American Statistical Association* 72: 669–672.
11. Spanier, J., and K.B. Oldham. 1987. *An atlas of functions*. Washington: Hemisphere.
12. Willis, A., and J. Bunge. 2014. Estimating diversity via frequency ratios. Submitted. arXiv:1408.3333.

Approximate Bayesian Estimation for Multivariate Count Time Series Models

Volodymyr Serhiyenko, Nalini Ravishanker and Rajkumar Venkatesan

Abstract In many areas of application, there is increasing interest in modeling multivariate time series of counts on several subjects as a function of subject-specific and time-dependent covariates. We propose a level correlated model (LCM) to account for the association among the components of the response vector, as well as possible overdispersion. The flexible LCM framework allows us to combine different marginal count distributions and to build a hierarchical model for the vector time series of counts. We employ the Integrated Nested Laplace Approximation (INLA) for fast approximate Bayesian modeling using the R package INLA (r-inla.org). We illustrate it by modeling the monthly prescription counts by physicians of a focal drug from a multinational pharmaceutical firm along with monthly counts of other drugs with a sizable market share for the same therapeutic category.

Keywords Bayesian framework · Discrete-valued time series · INLA · Marketing · Multivariate Poisson · ZIP model

1 Introduction

In many applications, including marketing, we observe counts of some event of interest at different times and for different subjects. Increasing attention is being given to the problem of accurate modeling of such time series of univariate or multivariate counts for N subjects over T time periods as functions of relevant subject-specific and/or time-varying covariates, incorporating dependence over time and association between the components of the response vector. While there is considerable literature on count data regression [13], models for count time series are less common. Zeger

V. Serhiyenko · N. Ravishanker (✉)
University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269, USA
e-mail: nalini.ravishanker@uconn.edu

V. Serhiyenko
e-mail: volodymyr.serhiyenko@gmail.com

R. Venkatesan
University of Virginia, Darden School of Business, Charlottesville, VA 22903, USA
e-mail: venkatesanr@darden.virgina.edu

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_10

[29] described a regression type approach using the quasi-likelihood approach. The Generalized Linear AutoRegressive Moving Average (GLARMA) model was discussed in Davis et al. [5], while the Dynamic Generalized Linear Model (DGLM) framework was described in West and Harrison [28], Gamerman [6], and Landim and Gamerman [15], among others. In this article, we describe modeling of univariate and multivariate time series of counts in the context of a marketing application that uses data from the pharmaceutical industry. We propose dynamic models for modeling prescription counts. These models may be viewed as extensions of the Gaussian Dynamic Linear Models (DLMs).

Kalman [9], and Kalman and Bucy [10] popularized a recursive algorithm for optimal estimation (filtering, smoothing and forecasting) of the state vector, and forecasting of the observation vector for Gaussian DLMs, also referred to as Gaussian State Space Models (SSMs). Gaussian Hierarchical Dynamic Linear Models (HDLMs) include a set of one or more dimension reducing structural equations along with the observation equation and state (or system) equation of the DLM [7, 15]. For count time series, DLMs have been generalized to Dynamic Generalized Linear Models (DGLMs) or exponential family SSMs, which assume that the sampling distribution is a member of the exponential family of distributions, such as the Poisson or negative binomial distributions [6, 28]. DGLMs may be viewed as dynamic versions of the Generalized Linear Models (GLMs) discussed in McCullagh and Nelder [17]. For non-Gaussian or nonlinear models, Bayesian inference is usually facilitated by sampling based approaches such as the Metropolis-Hastings algorithm combined with the Gibbs sampler; for details, see Carlin et al. [2], Chen et al. [4].

Hu [8] described hierarchical dynamic models for univariate and multivariate count times series, while Ravishanker et al. [21] discussed similar multivariate dynamic models for ecological data that varies over time and by location. These papers used a modified Gibbs sampling framework for Bayesian inference under a multivariate Poisson or a mixture of multivariate Poisson sampling distribution as discussed in Karlis and Meligkotsidou [11, 12]. The computational time increases considerably with the sample size and vector dimension. Ravishanker et al. [22] describe the use of models for count time series on several subjects in the context of a marketing example, and include a discussion of a hierarchical dynamic model using univariate Poisson and zero-inflated Poisson (ZIP) sampling distributions, and a fully Bayesian inferential approach. The computational time is once again a consideration in fitting these models.

In this article we discuss a level correlated ZIP model for multivariate count time series from a marketing application and carry out the estimation using R-INLA. For each drug in the same therapeutic category, this model enables us to estimate both the expected number of new prescriptions for each drug as well as the probability of retention of a physician. Through approximate Bayesian inference, INLA enables relatively fast computation. The format of the paper follows. Section 2 gives a description of the marketing application and a description of the data. Section 3.1 reviews the univariate ZIP model which is estimated in a fully Bayesian framework. Section 3.2 gives details on multivariate level correlated ZIP fitting with an R-INLA implementation. Section 4 provides a discussion and summary of the results.

2 Problem and Data Description

We describe statistical analyses pertaining to marketing data from a large multinational pharmaceutical firm. Analysis of the drivers of new prescriptions written by physicians is of interest to marketing researchers. Most existing research focuses on physician level sales for a single drug within a therapeutic category and do not consider the association between the sales of a drug and its competitors over time, see Venkatesan et al. [26] for a detailed discussion. Further, there is interest in knowing the effect of the firm's detailing efforts (visits by the firm's sales representatives to physicians) on the sales of its own drug and the sales of competitors. Mizik and Jacobson [18] discussed existing research showing that detailing, sampling (giving samples of drugs to physicians), and previous behavior influence new prescriptions from physicians. Montoya et al. [19] state that after accounting for dynamics in physician prescription writing behavior, detailing seems to be most effective in acquiring new physicians, whereas sampling is most effective in obtaining recurring prescriptions from existing physicians. As do most other research studies in this context, we treat physicians as customers of the pharmaceutical firm.

The behavioral data collected monthly by the firm over a period of 3 years consists of the number of new prescriptions from a physician (sales) and the number of sales calls directed toward the physicians (detailing). As in Venkatesan et al. [26], our focus is on one of the newer drugs launched by the firm in a large therapeutic drug category (one of the ten largest therapeutic categories in the United States). The database consists of a monthly prescription history of the drug for 45 continuous months within the last decade from a sample of physicians from the American Medical Association (AMA) database. The time window of our data starts after 1 year since the introduction of the focal drug. For our analysis, we have chosen three drugs in the same therapeutic category with the highest market shares. The focal drug is a drug made by the firm of interest with a market share of 13 %. The leader drug has a market share of 47 % and a challenger drug has a 15 % market share. We denote the focal drug by the abbreviation "F", the leader drug is denoted by "L" and the challenger drug by "C". Let $\boldsymbol{Y}_{it} = (Y_{F,it}, Y_{L,it}, Y_{C,it})'$ be a 3-dimensional vector of count responses of the number of new prescriptions for each drug written by the ith physician at equally spaced times t, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Our analysis is based on a random sample of $n = 100$ physicians, and Fig. 1 shows the time series of prescription counts for the focal, leader, and challenger drugs for a randomly chosen physician. Due to confidentiality concerns, we are unable to reveal any other information about the drug category or the pharmaceutical firm. We are interested in modeling patterns in the number of prescriptions written by the physician on the focal drug, as well as on the leader and challenger drugs. The sales calls directed towards the customer by the firm constitute *customer relationship management* (CRM) actions.

Exploratory data analysis shows that while the firm obtains on average three new prescriptions per month from a single physician, and salespeople call on a physician on average about twice a month, there is large variation in both the monthly level of sales per physician and the number of sales calls directed toward the physician each

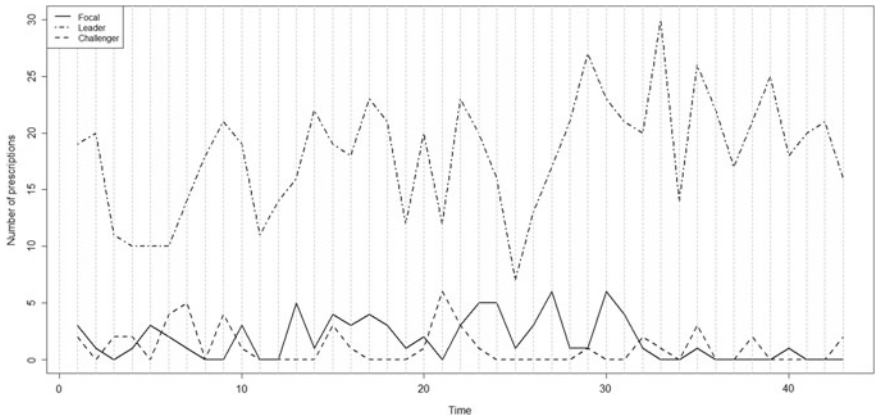


Fig. 1 Prescription counts of three drugs for a randomly selected physician

month. The correlation matrix indicates that sales calls toward the firm correlate positively with sales of the focal drug. The focal drug represents a significantly different chemical formulation, and further targets a different function of the human body to cure the disease condition than the drugs available at the time of introduction in the therapeutic category. It is therefore reasonable to expect that physicians will learn about the efficacy of the drug over time, resulting in a variation (either increase or decrease) in sales over time. This expectation is supported by multiple exploratory analyses of sales histories. We observe that the average level of sales (across all physicians) ranges from about 1 in the first month to 4 in the last month. An ANOVA test, reported in Venkatesan et al. [26], rejected the null hypothesis that the mean level of sales was the same across the months.

The variation in sales over time has motivated us to develop a dynamic model framework where the coefficients in the customer level sales response model could vary across customers and over time. During these 45 months, the pharmaceutical firm also collected attitudinal data, viz., monthly information on customer attitudes regarding all the drugs in the therapeutic category and their corresponding salespeople. Ravishanker et al. [22] discussed a hierarchical dynamic zero-inflated Poisson framework for sales of the focal drug only, which combines sparse survey based customer attitude data that is not available at regular intervals, with customer level transaction and marketing histories that are available at regular time intervals; see also Venkatesan et al. [26]. We omit discussion of attitudes in this paper for modeling multivariate count time series and only focus on behavioral data.

An important step of the marketing research is to jointly model the number of prescriptions of different drugs written by the physicians over time, taking into account possible associations between them. Almost all the current research focuses on physician level sales for a single drug within a category, and do not consider the association between the sales over time of a drug and its competitions within the category. The effect of a firm's detailing on the sales of its own drug and competitors is also of

interest. The computational requirements of the fully Bayesian approach described in Hu [8] prompted us to explore an approximate Bayesian framework. In Sect. 3.2, we describe a model which we call the level correlated model, which provides a useful framework for studying the evolution of sales of a set of competing drugs within a category. We use this model for multivariate counts in order to decompose the association in sales among competing drugs between marketing activities of a drug in the category and coincidence induced by general industry trends.

3 Bayesian Inference

In Sect. 3.1, we briefly review univariate hierarchical dynamic ZIP modeling for the focal drug counts. Details may be found in Hu [8] and Ravishanker et al. [22], while details of a static model are discussed in Venkatesan et al. [26]. Approximate Bayesian inference for a model for multivariate counts using R-INLA is then discussed in Sect. 3.2.

3.1 Univariate ZIP Models

Let $Y_{i,t}$ denote observed new prescription counts of the focal drug from physician i in month t, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Let $D_{i,t}$ denote the level of detailing (sales calls) directed at the ith physician in month t. As mentioned earlier, Hu [8], Venkatesan et al. [27] and Ravishanker et al. [22] use behavioral and attitudinal data, while we restrict our discussion here to just behavioral data. Suppose $Y_{i,t}$ follows a zero-inflated Poisson model [14]. Under this model, it is assumed that the ith physician at time t can belong to one of these two latent (unobserved) states: an inactive state, or an active state. The states have the interpretation that zero new prescriptions will be observed with high probability from physicians in the inactive state. When the physician is in the active state, the number of new prescriptions can assume values $k = 0, 1, 2, \ldots$. Due to market forces, marketing actions from the focal firm, and other influences, a physician is likely to move from the active to the inactive state from time to time, and vice versa. We may interpret a physician in the active state as being retained by the focal firm and a physician in the inactive state to be dormant. We also assume that a physician never quits his/her relationship with the focal firm, and that there is always a finite probability that he/she will return to prescribing the drugs of the focal firm. Venkatesan et al. [26] described a univariate ZIP model [14] for focal drug counts $Y_{i,t}$ as a function of detailing and previous history of prescription counts. The regression coefficients were assumed to be static (not time-varying) random variables, and the model was estimated in the fully Bayesian framework using Markov chain Monte Carlo (MCMC) algorithms.

In order to extend the model to include dynamic or time-varying behavior of the regression coefficients, a hierarchical dynamic ZIP model for the focal drug counts

was discussed in Ravishanker et al. [22]. Let $\lambda_{i,t}$ denote the mean of $Y_{i,t}$, and let $\Pi_{i,t}$ denote the probability of zero. The ZIP model for the focal drug counts is

$$\begin{aligned} P(Y_{i,t} = 0|\lambda_{i,t}, \Pi_{i,t}) &= \Pi_{i,t} + (1 - \Pi_{i,t}) \exp(-\lambda_{i,t}) \\ P(Y_{i,t} = k|\lambda_{i,t}, \Pi_{i,t}) &= (1 - \Pi_{i,t}) \exp(-\lambda_{i,t})\lambda_{i,t}^{k}/k!, \quad k = 1, 2, \ldots \end{aligned} \tag{1}$$

It is well known that the distribution for $Y_{i,t}$ can be written as a mixture distribution, i.e., $Y_{i,t} = V_{i,t}(1 - B_{i,t})$, where $B_{i,t} \sim$ Bernoulli$(\Pi_{i,t})$, $V_{i,t} \sim$ Poisson$(\lambda_{i,t})$, and $B_{i,t}$ and $V_{i,t}$ are independent. In the dynamic model, both $\lambda_{i,t}$ and $\Pi_{i,t}$ are latent (unobserved) physician specific dynamic parameters. Ravishanker et al. [22] modeled $\log(\lambda_{i,t})$ and $\text{logit}(\Pi_{i,t})$ as functions of log detailing $\log(D_{i,t})$ and a measure of physicians' behavioral loyalty $R_{i,t}$, which is a weighted average of the time since last prescription (recency), number of months with positive sales (frequency), and cumulative level of sales (monetary value). For each physician i and time t, these three variables were calculated as moving averages over 3 months prior to time t, with respective weights determined empirically as 0.6, 0.3, and 0.1. A brief summary of this model formulation is given below.

Let $\boldsymbol{\beta}_{i,t}^{\lambda}$ and $\boldsymbol{\beta}_{i,t}^{\Pi}$ respectively denote the physician and time specific coefficients. Including an intercept, a coefficient for $\log(D_{i,t})$ and a coefficient for $R_{i,t}$, each of these is a three-dimensional parameter vector for each t and each i. Then, $\boldsymbol{\beta}_{i,t} = (\boldsymbol{\beta}_{i,t}^{\lambda\prime}, \boldsymbol{\beta}_{i,t}^{\Pi,\prime})'$ is a $p = 6$-dimensional vector. The hierarchical or structural equation modeled $\boldsymbol{\beta}_{i,t}$ as a function of a p-dimensional dynamic state vector $\boldsymbol{\gamma}_t$. Physician level variables (such as demographics or specialty, if available) may be included with static coefficients in this equation. The errors $\mathbf{v}_{i,t}$ are assumed to be $N_p(\mathbf{0}, \mathbf{V}_i)$ vectors. The state equation described the dynamic evolution of the state vector $\boldsymbol{\gamma}_t$:

$$\boldsymbol{\gamma}_t = \mathbf{G}\boldsymbol{\gamma}_{t-1} + \mathbf{w}_t, \tag{2}$$

where $\mathbf{G}$ is an identity matrix if a random walk evolution is assumed, and $\mathbf{w}_t \sim N_p(\mathbf{0}, \mathbf{W})$. Usual conjugate prior distributions such as multivariate normal and inverse Wishart distributions were assumed for the model parameters, and the Gibbs sampling algorithm was employed to estimate the posterior distribution of the model parameters. While static coefficients were routinely drawn from known distributions, the Forward-Filtering-Backward-Sampling (FFBS) algorithm [3] enabled sampling $\boldsymbol{\gamma}_t$, and the Metropolis-Hastings algorithm was used to generate samples from other parameters. Modeling details as well as detailed results and comparisons between several dynamic models are given in Hu [8], while details and results for the static ZIP models are discussed in Venkatesan et al. [27]. Using posterior results and summaries, physicians were classified into quintiles based on the actual customer lifetime value (CLV) as well as the CLV predicted from the hierarchical dynamic ZIP model, enabling effective marketing actions by the firm. ZIP models for multivariate times series of counts using approximate Bayesian inference using R-INLA are discussed in Sect. 3.2, and provide a computationally feasible approach for modeling such data.

3.2 Approximate Bayesian Inference for Level Correlated Models

Accurate modeling of multivariate time series of counts for N subjects over T time periods as functions of relevant subject-specific and/or time-varying covariates, incorporating dependence over time and association between the components of the response vector, is discussed in Hu [8] and Ravishanker et al. [21] using multivariate Poisson and mixtures of multivariate Poisson distributions as sampling distributions and employing a fully Bayesian framework. The evaluation of the corresponding likelihood function and the Gibbs sampling framework can be time consuming, especially as N, T and the vector dimension increase. As a fast alternative, we explore approximate Bayesian inference as discussed below.

Regression models which use a multivariate extension of the Poisson lognormal mixture distribution have become popular in different areas of research [1, 16, 20]. There are many situations where researchers wish to model dependence in the response vector for data that are possibly overdispersed. A Bayesian framework using Markov Chain Monte Carlo (MCMC) has been developed for model estimation [16], although full Bayesian sampling can be computationally expensive and time consuming, especially for big data sets. In this paper, we propose an approach that combines different marginal count distributions in multivariate time series data and use the Integrated Nested Laplace Approximation (INLA) method to carry out approximate Bayesian inference. The observation equation of the Level Correlated Model (LCM) for multivariate time series in the hierarchical dynamic form is

$$Y_{F,it}|\pi_F, \lambda_{F,it} \sim \text{ZIP}(\pi_F, \lambda_{F,it}) \tag{3}$$

$$Y_{L,it}|\lambda_{L,it} \sim \text{Poisson}(\lambda_{L,it}) \tag{4}$$

$$Y_{C,it}|\pi_C, \lambda_{C,it} \sim \text{ZIP}(\pi_C, \lambda_{C,it}) \tag{5}$$

with the natural logarithmic link function for the parameter $\lambda_{j,it}$ for $j = F, L, C$ (corresponding to the focal, leader, and challenger drugs). We based our choice of ZIP models for $j = F, C$ and a Poisson model for $j = L$ on exploratory univariate models that we fit to the data. Although we do not show the details here, it is possible to verify these specifications as part of the model selection procedure. We model $\lambda_{j,it}$ as follows:

$$\log(\lambda_{j,it}) = \eta_i + \gamma_{j,t} + \mathbf{z}'_{j,it}\boldsymbol{\beta}_j + \alpha_{j,it}, \tag{6}$$

where $i = 1, \ldots, n$, $t = 1, \ldots, T$ and $j = F, L, C$. In (6), the random effect η_i is a physician specific effect for the number of prescriptions, $\gamma_{j,t}$ represents a drug specific time effect, the random effect $\alpha_{j,it}$ is a drug and time specific level correlated component, the vector $\mathbf{z}'_{j,it}$ denotes a P_j-dimensional vector of covariates with a vector of one's as a first column, and $\boldsymbol{\beta}_j$ is a P_j-dimensional vector of coefficients corresponding to the predictors. In general, $\boldsymbol{\beta}_j$ can be physician specific as well as time-varying, and consists of intercepts for each component of the response vector

$\boldsymbol{Y_{it}} = (Y_{F,it}, Y_{L,it}, Y_{C,it})'$. The natural logarithm of the number of prescriptions written during the previous time point for each component of $\boldsymbol{Y_{it}}$ is a predictor in the model. For prediction of the focal drug, another predictor is the natural logarithm of detailing, i.e., the number of the sales calls to the physician from the firm's representative regarding the focal drug. A small correction term is added to avoid taking logarithms of zeros.

Let $\boldsymbol{\alpha}_{it} = (\alpha_{F,it}, \alpha_{L,it}, \alpha_{C,it})'$. In (6), the dependence between different types of counts in (3)–(5) is introduced at the physician level through $\boldsymbol{\alpha}_{it} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a variance-covariance matrix for the level correlated random effect term. We assume that the physician specific random effect follows a normal distribution, i.e., $\eta_i \sim \text{Normal}(0, 1/\tau)$. We also assume that the components of the drug specific time effect vector $\boldsymbol{\gamma}_t = (\gamma_{F,t}, \gamma_{L,t}, \gamma_{C,t})'$ are independent and evolve according to a random walk process in the state equation in the HDLM. The state equation is:

$$\gamma_{j,t} = \gamma_{j,t-1} + w_{j,t}, \tag{7}$$

where $i = 1, \ldots, n$, $t = 1, \ldots, T$, $j = F, L, C$ and the error term is defined as $w_{j,t} \sim \text{Normal}(0, 1/V_j)$.

We implement the model estimation through an approximate sampling based Bayesian framework, assuming usual prior specifications for the parameters. We assume a normal prior for β_j's in (6), a Wishart prior for $\boldsymbol{\Sigma}$ in the distribution of $\boldsymbol{\alpha}_{it}$ and a log gamma prior for $\log(\tau)$ and $\log(V_j)$ in the distribution of η_i and $w_{j,t}$, respectively, in (2). We also assume a normal prior for $\text{logit}(\pi_F)$ and $\text{logit}(\pi_C)$ in (3) and (5), respectively. Let $\boldsymbol{\theta}$ denotes all the hyperparameters associated with a model. We use the recently proposed INLA approach [25] which provides a mechanism for Bayesian inference based on accurate approximations to the posterior distributions of the parameters. Since INLA does not rely on MCMC, the approximate approach greatly reduces computational time. More details on the approach as well as the R-INLA package are available on the website www.r-inla.org. We give a brief overview of the INLA approach.

INLA performs approximate Bayesian inference for structured additive regression models with latent Gaussian field specification (or, latent Gaussian models), such as the Bayesian additive model with normal priors. Let $\boldsymbol{\xi}$ denote the vector of all components of the latent Gaussian model in (3)–(7). We are interested in deriving the marginal posterior distribution for each ξ_k, which denotes the kth component of the vector $\boldsymbol{\xi}$. The marginal posterior distribution can be written in the following form:

$$\pi(\xi_i|\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \pi(\xi_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{8}$$

where ξ_i denotes each component of the latent Gaussian field given by (3)–(7), $\boldsymbol{\theta}$ denotes all the hyperparameters associated with a model and $\boldsymbol{y}$ is the observed data vector. Using the hierarchical structure of the joint distribution, we can rewrite $\pi(\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{y}) = \pi(\boldsymbol{\xi}|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{y})$. Then, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ can be approximated by the

Laplace approximation of a marginal posterior distribution.

$$\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{\pi(\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{y})}{\widetilde{\pi}_G(\boldsymbol{\xi}|\boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*(\boldsymbol{\theta})}, \tag{9}$$

where $\boldsymbol{\xi}^*$ denotes the mode of the full conditional $\pi(\boldsymbol{\xi}|\boldsymbol{\theta}, \boldsymbol{y})$. In (9), $\widetilde{\pi}_G(\boldsymbol{\xi}|\boldsymbol{\theta}, \boldsymbol{y})$ denotes the Gaussian approximation to $\pi(\boldsymbol{\xi}|\boldsymbol{\theta}, \boldsymbol{y})$ [23]. To integrate out $\boldsymbol{\theta}$, we need to find a good set of evaluation points θ_k for numerical integration in (8). This is done by exploring the properties of (9), via an iterative algorithm with appropriate choice of weights Δ_k, which are assigned to each θ_k [25].

Another part that needs to be approximated is $\pi(\xi_i|\boldsymbol{\theta}, \boldsymbol{y})$. According to Rue et al. [25] and Rue and Martino [24], there are three alternatives: a Laplace approximation, a simplified Laplace approximation, and a Gaussian approximation (the simplest one). The non-normal distribution under this alternative is approximated with a Gaussian density by matching the mode and the curvature at the mode [23]. Overall, the method gives reasonable results, but the approximation can be improved by applying the Laplace or simplified Laplace approximation to $\pi(\xi_i|\boldsymbol{\theta}, \boldsymbol{y})$. To summarize, an approximation of the posterior marginal density (8) can be obtained:

$$\widetilde{\pi}(\xi_i|\boldsymbol{y}) = \sum_k \widetilde{\pi}(\xi_i|\theta_k, \boldsymbol{y})\widetilde{\pi}(\theta_k|\boldsymbol{y})\Delta_k. \tag{10}$$

4 Discussion of Results

We use the default hyperparameter specifications in the *inla* function from the R-INLA package. Results from the model fit are given in Table 1.

Table 1 Posterior estimates for the fixed effects

Parameters	Model 1	Model 2	Model 3	Model 4
$\beta_{F,0}$	−0.35(0.044)	−0.23(0.046)	−0.46(0.053)	−0.50(0.051)
$\beta_{F,1}$	0.73(0.19)	0.71(0.019)	0.75(0.021)	0.78(0.021)
$\beta_{F,2}$	0.16(0.026)	0.14(0.027)	0.16(0.028)	0.16(0.029)
$\beta_{L,0}$	1.20(0.046)	1.22(0.044)	1.29(0.047)	1.26(0.048)
$\beta_{L,1}$	0.44(0.015)	0.44(0.014)	0.41(0.015)	0.42(0.015)
$\beta_{C,0}$	−0.16(0.038)	−0.05(0.040)	−0.32(0.044)	−0.29(0.044)
$\beta_{C,1}$	0.77(0.017)	0.73(0.018)	0.82(0.020)	0.82(0.019)
DIC	53722	54028	53382	53461
PMAE	2.36	2.48	2.30	2.40

For model comparison purposes, we consider four slightly different models. Model 1 is given by Eqs. (3)–(7). Model 2 is the same as Model 1, but instead of assuming different time effects $\gamma_{j,t}$ in (6), we assume the same time effect γ_t across all components in (6). For Model 3, we assume marginal Poisson distributions for all three components in (3)–(5), while all other equations stay the same as in Model 1. Model 4 assumes the same underlying Poisson distribution as in Model 3 with the same time effect as in Model 2. The parameters $\beta_{F,0}$, $\beta_{L,0}$ and $\beta_{C,0}$ respectively denote the intercepts in the models for the focal, leader and challenger drugs. The parameters $\beta_{F,1}$, $\beta_{L,1}$ and $\beta_{C,1}$ respectively denote the coefficients corresponding to the logarithm of the prescription counts in the previous time period in models for the focal, leader and challenger drugs, while $\beta_{F,2}$ is the coefficient corresponding to the logarithm of detailing in the model for the focal drug. For each fitted model, the table shows the posterior means with posterior standard deviations in parentheses. The last two rows also show the DIC and the Predictive Mean Absolute Error (PMAE) for these models and enable model comparison. To construct the PMAE, the last time point across all physicians was used as hold-out predictive performance evaluation. Model 3 with the lowest DIC and PMAE values slightly outperforms the other models. Note that the posterior means for the fixed effect parameters are similar across all four models, and that the posterior mean for $\beta_{F,2}$ is positive. This suggests that an increase in sales calls from the firm results in an increase in the expected number of the prescriptions for the focal drug.

The posterior mean for the zero probability parameters π_F and π_C are 0.03 and 0.04 in Model 1. This suggests that the overall probability of a physician going inactive for the focal drug is lower than that for the challenger drug. The most useful outcome from multivariate modeling is the correlation coefficient between all three drugs, whose posterior summary is extracted from the posterior distribution of $\mathbf{\Sigma}$, corresponding to $\alpha_{j,it}$. In all models, the posterior mean of the correlation coefficient between the leader drug and the challenger drug ranges between -0.20 and -0.33. The correlation coefficients between the focal drug and the leader drug and between the focal drug and the challenger drug are very close to zero, suggesting that after controlling for the fixed effects predictors, the physician effect, and the time effect, an increase in the number of the prescriptions for the leader drug results in a decrease in the number of the prescriptions for the challenger drug, while there is no significant effect on the focal drug. The next step in the analysis is to investigate the temporal behavior for all three drugs.

In Fig. 2, we plot the posterior mean of $\gamma_{j,t}$ from Model 3 for $j = F, L, C$. After the steep decrease in the beginning, the leader drug shows an increase for the rest of the observational period. By contrast, after an increase during the first 9 months for the challenger drug, there is a decreasing trend until the end of the observational period. For the focal drug, the relatively stable period in the beginning is followed by a decrease. Nevertheless, we notice that there is a flattening out for the focal drug towards the end of the observational period.

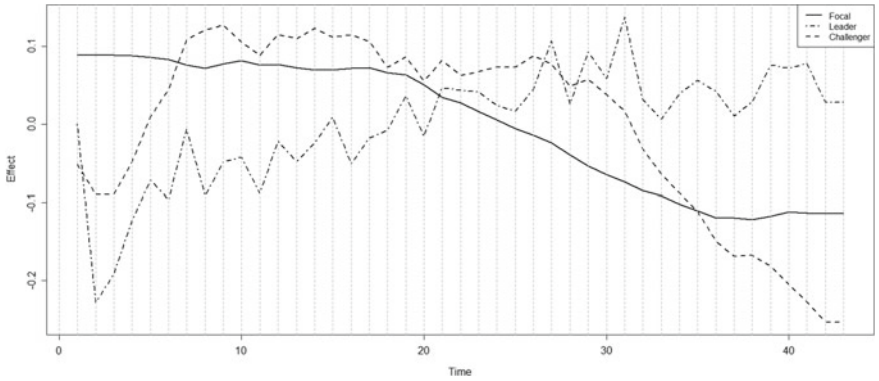


Fig. 2 Time effect for the focal, leader, and challenger drugs

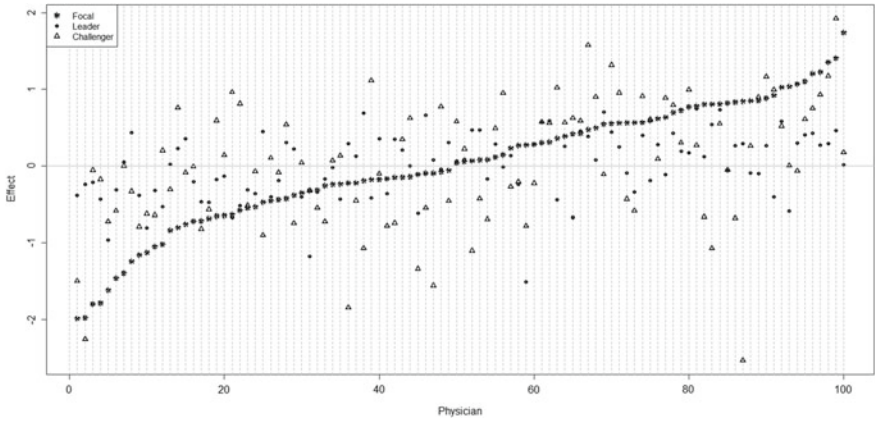


Fig. 3 Physician effect for the focal, leader, and challenger drugs

One of the research goals is to estimate the probability that a physician goes inactive for the focal drug. We estimate the probability using a univariate ZIP model for the focal drug, and compare the probability to those obtained from univariate ZIP models for the leader and challenger drugs. To construct Fig. 3, we compute ordered posterior mean values for the physician effect for the focal drug, and plot those together with the physician effect from the other two drugs. The plot clearly shows that there is a group of a dozen physicians whose mean prescription level is lower than that of the others, and perhaps future marketing efforts could be directed at such physicians in an effort to improve sales of the focal drug.

Acknowledgments We thank the reviewers for their very helpful comments.

References

1. Aitchison, J., and C. Ho. 1989. The multivariate Poisson-log normal distribution. *Biometrika* 76(4): 643–653.
2. Carlin, B., N. Polson, and D. Stoffer. 1992. A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association* 87: 493–500.
3. Carter, C., and R. Kohn. 1994. On Gibbs sampling for state-space models. *Biometrika* 81: 541–553.
4. Chen, M.-H., Q.-M. Shao, and J. Ibrahim. 2000. *Monte Carlo methods in Bayesian computation*. New York: Springer.
5. Davis, R., W. Dunsmuir, and S. Streett. 2003. Observation-driven models for Poisson counts. *Biometrika* 90(4): 777–790.
6. Gamerman, D. 1998. Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika* 85(1): 215–227.
7. Gamerman, D., and H. Migon. 1993. Dynamic hierarchical models. *Journal of the Royal Statistical Society Series (B)* 55(3): 629–642.
8. Hu, S. 2012. Dynamic modeling of discrete-valued time series with applications. Ph.D. Thesis, University of Connecticut.
9. Kalman, R. 1960. A new approach to linear filtering and prediction theory. *Transactions of the ASME Series D, Journal of Basic Engineering* 82: 35–45.
10. Kalman, R., and R. Bucy. 1961. New results in filtering and prediction theory. *Transactions of the ASME Series D, Journal of Basic Engineering* 83: 95–108.
11. Karlis, D., and L. Meligkotsidou. 2005. Multivariate Poisson regression with covariance structure. *Statistics and Computing* 15: 255–265.
12. Karlis, D., and L. Meligkotsidou. 2007. Finite mixtures of multivariate Poisson regression with application. *Journal of Statistical Planning and Inference* 137: 1942–1960.
13. Kedem, B., and K. Fokianos. 2002. *Regression models for time series analysis*. Hoboken: Wiley.
14. Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34(1): 1–13.
15. Landim, F., and D. Gamerman. 2006. Dynamic hierarchical models: an extension to matrix-variate observations. *Computational Statistics and Data Analysis* 35: 11–42.
16. Ma, J., K. Kockelman, and P. Damien. 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention* 40(3): 964–975.
17. McCullagh, P., and J. Nelder. 1989. *Generalized linear models*, 2nd ed. London: Chapman and Hall.
18. Mizik, N., and R. Jacobson. 2004. Are physicians easy marks? Quantifying the effects of detailing and sampling on new prescriptions. *Management Science* 50(12): 1704–1715.
19. Montoya, R., O. Netzer, and K. Jedidi. 2010. Dynamic allocation of pharmaceutical detailing and sampling for long-term profitability. *Marketing Science* 29(5): 909–924.
20. Park, E., and D. Lord. 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record: Journal of the Transportation Research Board* 2009(1): 1–6.
21. Ravishanker, N., V. Serhiyenko, and M. Willig. 2014. Hierarchical dynamic models for multivariate times series of counts. *Statistics and Its Interface* 7: 559–570.
22. Ravishanker, N., R. Venkatesan, and S. Hu 2015. Dynamic models for time series of counts with a marketing application. In *Handbook of discrete-valued time series*, ed. R.A. Davis, S.H. Holan, R.L. Lund, and N. Ravishanker, 1–10.
23. Rue, H., and L. Held. 2005. *Gaussian Markov random fields theory and applications*. New York: Chapman & Hall/CRC.
24. Rue, H., and S. Martino. 2007. Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference* 137: 3177–3192.

25. Rue, H., S. Martino, and N. Chopin. 2009. Approximate Bayesian inference for latent Gaussian models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society Series B* 71: 319–392.
26. Venkatesan, R., W. Reinartz, and N. Ravishanker. 2012. The role of attitudinal information in CLV-based customer management. *MSI Working Paper Series*, 12: 107.
27. Venkatesan, R., W. Reinartz, and N. Ravishanker, 2014. ZIP models for CLV based customer management using attitudinal and behavioral data. Technical Report, Department of Statistics, University of Connecticut.
28. West, M., and P. Harrison. 1989. *Bayesian forecasting and dynamic models*. New York: Springer.
29. Zeger, S. 1998. A regression model for time series of counts. *Biometrika* 75: 621–629.

Modeling and Analysis of Method Comparison Data with Skewness and Heavy Tails

Dishari Sengupta, Pankaj K. Choudhary and Phillip Cassey

Abstract The analysis of method comparison data is mainly concerned with evaluating agreement between methods of measuring a continuous variable. The methodology commonly assumes normally distributed data, which are usually modeled using a standard linear mixed model that assumes normality for both random effects and errors. In practice, however, the data often exhibit skewness and have tails heavier than those of a normal distribution, possibly due to outlying observations. When such data are analyzed using the standard mixed model, the non-normality may become apparent from model diagnostics. This article develops a methodology for agreement evaluation by modeling data using a recent robust mixed model that assumes a skew-t distribution for random effects and an independent t-distribution for errors. As the standard model is a special case of the robust model, the new methodology offers a unified framework for analyzing data with skewness and heavy tails as well as normally distributed data. The methodology is presented for both unreplicated and replicated data. A real example is used for illustration.

Keywords Concordance correlation · Heavy tailed distribution · Mixed effects model · Robust model · Skew-t distribution · Total deviation index

D. Sengupta
RainMan Consulting Pvt Ltd, Bangalore 560075, Karnataka, India
e-mail: dishari.stat@gmail.com

P. K. Choudhary (✉)
Department of Mathematical Sciences, University of Texas at Dallas,
Richardson, TX 75083-0688, USA
e-mail: pankaj@utdallas.edu

P. Cassey
School of Biological Sciences, University of Adelaide, North Terrace, SA 5005, Australia
e-mail: phill.cassey@adelaide.edu.au

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_11

1 Introduction

Method comparison studies compare two or more methods of measuring a continuous response variable with the primary aim of evaluating agreement between them. The comparison rests on the premise that if the methods agree well, they may be used interchangeably or we may prefer the cheapest or the least invasive method. Such studies are conducted in many disciplines, including metrology [43], ecology [23], and biomedical fields such as medical imaging, biomedical engineering, physiology and clinical chemistry [8]. Reviews of literature on this topic can be found in Barnhart et al. [5] and in the books by Carstensen [12] and Lin et al. [34].

In method comparison studies, measurements are taken by each method on every subject. Sometimes the measurements may be replicated. The data from the same subject are dependent whereas those from different subjects are assumed to be independent. The analysis of these data may be thought of as a two-step procedure. The first step involves modeling of the data. For this, the framework of linear *mixed models* [37] is an especially attractive choice because it allows capturing the within-subject dependence through random subject effects and their interactions. Indeed, a number of authors have used linear mixed models for method comparison data, including [9, 11, 13, 14, 40]. The models are generally fit using likelihood-based methods. The second step involves evaluation of agreement by performing inference on *measures of agreement* derived from the model fit in the first step. Although a number of agreement measures are available [5], the *concordance correlation coefficient* (CCC; [31]) and the *total deviation index* (TDI; [16, 32]) have received the most attention in the statistical literature. See [15] for a description of the two-step methodology.

In its standard formulation, a linear mixed model assumes that random effects and errors follow independent normal distributions, implying normality for the observed data. However, method comparison data often exhibit skewness and heavy tails, i.e., tails heavier than those of a normal distribution, possibly due to the presence of outlying observations. When a standard mixed model is fit to such data, the non-normality may manifest itself during model checking through non-normality of either predicted random effects or residuals or both. The crab claws data in Choudhary et al. [18] is a real example of such data; it is used for illustration later in this article. It may be possible to transform the data for better adherence to the normality assumption. But transformation may render the difference in measurements from two methods difficult to interpret—an important issue in data analysis. Therefore, a transformation other than the logarithmic is generally not recommended in method comparison studies [8]. But the logarithmic transformation may not succeed in normalizing the data. Besides, it is often desirable to analyze the data on the original scale. These considerations call for approaches that deal with non-normality rather than avoid it by employing a transformation.

One can analyze non-normal method comparison data by simply ignoring the model violation and applying the methodology developed for normal data. But depending upon the seriousness of the violation, the estimates of agreement mea-

sures and their standard errors may be quite inaccurate. This has been demonstrated by Carrasco et al. [10] for estimation of CCC in the case of skewed data. Another approach is to use a procedure that does not require the normality assumption. These procedures include the nonparametric procedures developed by King and Chinchilli [25, 26], King et al. [27, 28] and Choudhary [15]; and semiparametric procedures based on generalized estimating equations proposed by Barnhart and Williamson [4], Barnhart et al. [6, 7] and Lin et al. [33]. Yet another parametric approach—also the focus of this article—is to employ a *robust* mixed model wherein the normality of random effects and errors is replaced by more general distributions that have the normal as a special case. Such distributions include mixture of normals, t, skew-normal [2], and skew-t distributions [3]. The literature on this kind of robust model is vast and is under active development. One may start with [1, 21, 29, 38, 42, 44] for an initiation into the topic.

The robust mixed model of interest in this article is the general skew-t (GST) mixed model developed in Choudhary et al. [18]. It assumes a skew-t distribution for the random effects and an independent t-distribution for the errors. This way the model not only incorporates both skewness and heavy-tailedness in the data but also lets their distributions differ in heaviness of tails. The latter feature, not shared by competing models, affords additional flexibility in modeling because it permits, e.g., the random effects to be a normal and the errors to be a t, and vice versa. This brings us to the specific goal of this article: To extend the two-step methodology for method comparison data analysis by modeling data using a GST mixed model instead of the standard mixed model. Because the latter model is a special case of the former, the proposed extension offers a unified parametric framework for analyzing data with normal distributions as well as with skewness and heavy tails.

The rest of the article is organized as follows. Section 2 presents GST mixed models for basic method comparison data. Section 3 adapts the usual agreement evaluation methodology to work under the GST models. The methodology is illustrated in Sect. 4. Section 5 contains some concluding remarks. The software `R` [39] is used for all the analysis in this article.

2 Modeling Basic Method Comparison Data

Consider a method comparison study comparing J (≥ 2) measurement methods. We focus on two scenarios. One is when the measurements are not replicated. In this case, there is just one measurement by each method on every subject. The data are of the form Y_{ij}, $j = 1, \ldots, J$, $i = 1, \ldots, n$, where Y_{ij} is the measurement by the jth method on the ith subject. The other is when the measurements are replicated. In this case, the data are of the form $Y_{ijk}, k = 1, \ldots, m_{ij}, j = 1, \ldots, J, i = 1, \ldots, n$, where Y_{ijk} is the kth replicate measurement from the jth method on the ith subject. It is assumed that the replications are independently and identically distributed measurements of the same underlying true value. In either case, the number of observations on the ith subject is M_i and $N = \sum_{i=1}^{n} M_i$ is the total number of observations in

the data. Obviously, $M_i = J$ for unreplicated data and $M_i = \sum_{j=1}^{J} m_{ij}$ for replicated data. Although the unreplicated data are a special case of replicated data with $m_{ij} = 1$, we treat the two scenarios separately as they warrant somewhat different model formulations.

The vectors and matrices in this article are denoted by boldface letters. The vectors are column vectors unless stated otherwise. Let $\mathbf{I}_J$ denote a $J \times J$ identity matrix; $\mathbf{a}'$ denote the transpose of $\mathbf{a}$; $|\boldsymbol{\Sigma}|$ denote the determinant of $\boldsymbol{\Sigma}$; $\boldsymbol{\Sigma}^{1/2}$ denote a symmetric square root of a symmetric, positive definite matrix $\boldsymbol{\Sigma}$ so that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$; and $\boldsymbol{\Sigma}^{-1/2}$ denote the inverse of $\boldsymbol{\Sigma}^{1/2}$. We also use $\Phi(\cdot)$ for the cumulative distribution function of a univariate standard normal distribution, and $f(\mathbf{y}|\boldsymbol{\theta})$ for the probability density function of a random vector $\mathbf{Y}$ with parameter vector $\boldsymbol{\theta}$.

2.1 A General Formulation of GST Mixed Models

We first describe the GST mixed models in general terms before presenting them for method comparison data. To lay some groundwork for this, let $\boldsymbol{\mu} \in \mathbb{R}^J$ be a vector of location parameters; $\boldsymbol{\Sigma}$ be a $J \times J$ positive definite scale matrix; $\boldsymbol{\lambda} \in \mathbb{R}^J$ be a vector of skewness parameters; and $\nu\,(> 0)$ be degrees of freedom. We use $\mathcal{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathcal{SN}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, $t_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ and $\mathcal{ST}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ respectively to denote J-dimensional normal, skew-normal, t and skew-t distributions. The last three distributions are defined in Appendix A.1. Azzalini and Capitanio [3] and Genton [19] may be consulted for their additional properties. The normal is a special case of the skew-normal and the t is a special case of the skew-t when the skewness parameter $\boldsymbol{\lambda} = \mathbf{0}$. Similarly, the normal becomes a special case of the t and the skew-normal becomes a special case of the skew-t in the limit when the degrees of freedom $\nu \to \infty$. The location vector $\boldsymbol{\mu}$ and the scale matrix $\boldsymbol{\Sigma}$ are actually the mean vector and the covariance matrix in case of the normal, but this may not be the case in general for other distributions.

Let $\mathbf{Y}_i$ denote the vector of M_i observations on subject $i = 1, \ldots, n$. A linear mixed model for the data $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ can be written as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}$ is the p-vector of fixed effects and $\mathbf{X}_i$ is the associated $M_i \times p$ design matrix; $\mathbf{b}_i$ is the q-vector of random effects and $\mathbf{Z}_i$ is the associated $M_i \times q$ design matrix; and $\mathbf{e}_i$ is the M_i-vector of within-subject random errors. The design matrices are assumed to have full column ranks.

The standard version of this model makes the normality assumption,

$$\mathbf{b}_i \sim \text{independent}\, \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi}), \mathbf{e}_i \sim \text{independent}\, \mathcal{N}_{M_i}(\mathbf{0}, \boldsymbol{\Sigma}_i), \tag{2}$$

and $\mathbf{b}_i$ and $\mathbf{e}_i$ are mutually independent. The GST mixed model in Choudhary et al. [18] replaces the normality in (2) with more general distributions,

$$\mathbf{b}_i \sim \text{independent}\mathcal{ST}_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \nu_b), \mathbf{e}_i \sim \text{independent}t_{M_i}(\mathbf{0}, \boldsymbol{\Sigma}_i, \nu_e), \quad (3)$$

while $\mathbf{b}_i$ and $\mathbf{e}_i$ remain mutually independent. The assumptions in (3) reduce to (2) in the limit when the skewness parameter $\boldsymbol{\lambda} = \mathbf{0}$ and the degrees of freedom parameters $\nu_b, \nu_e \to \infty$. Appendix A.2 gives a hierarchical representation of this model that is especially useful for studying its distributional properties. The scale matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}_i$ in both (2) and (3) are generally parameterized in terms of a small number of parameters that do not depend on i. To get the mean vector and covariance matrix of $\mathbf{Y}_i$ under the GST mixed model, $(\boldsymbol{\Psi}, \boldsymbol{\lambda})$ is reparameterized as $(\boldsymbol{\Gamma}, \boldsymbol{\gamma})$, where

$$\boldsymbol{\delta} = \boldsymbol{\lambda}/(1 + \boldsymbol{\lambda}'\boldsymbol{\lambda})^{1/2}, \boldsymbol{\gamma} = \boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}, \boldsymbol{\Gamma} = \boldsymbol{\Psi}^{1/2}(\mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}')\boldsymbol{\Psi}^{1/2} = \boldsymbol{\Psi} - \boldsymbol{\gamma}\boldsymbol{\gamma}'. \quad (4)$$

Then, we have

$$E[\mathbf{Y}_i] = \mathbf{X}_i\boldsymbol{\beta} + \sqrt{\frac{\nu_b}{\pi}}\frac{\text{gam}((\nu_b - 1)/2)}{\text{gam}(\nu_b/2)}\mathbf{Z}_i\boldsymbol{\gamma},$$
$$\text{var}[\mathbf{Y}_i] = \frac{\nu_b}{\nu_b - 2}\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i' + \frac{\nu_e}{\nu_e - 2}\boldsymbol{\Sigma}_i - \frac{\nu_b}{\pi}\left(\frac{\text{gam}((\nu_b - 1)/2)}{\text{gam}(\nu_b/2)}\right)^2\mathbf{Z}_i\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Z}_i', \quad (5)$$

where $E[\mathbf{Y}_i]$ assumes $\nu_b, \nu_e > 1$, $\text{var}[\mathbf{Y}_i]$ assumes $\nu_b, \nu_e > 2$, and $\text{gam}(\cdot)$ denotes the gamma function.

The likelihood function under the GST model is not available in a closed form. It can, however, be computed via one-dimensional numerical integration. Choudhary et al. [18] fit the model by a variant of the expectation-maximization (EM) algorithm [35]—the expectation-conditional maximization (ECM) algorithm [36]. All subsequent inference relies on the fact that, when the number of subjects n is large, under certain regularity conditions, the maximum likelihood (ML) estimator $\hat{\boldsymbol{\theta}}$ of the model parameter vector $\boldsymbol{\theta}$ approximately follows a normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\mathfrak{I}^{-1}$, where $\mathfrak{I}$ is the observed information matrix associated with the fitted model [30]. This matrix is obtained by numerically differentiating the log-likelihood function which itself is computed via numerical integration. The `numDeriv` package of [20] and the `statmod` package of [41] in `R` can be used for this task.

2.2 GST Mixed Models for Method Comparison Data

First, consider the case of unreplicated data. These data can be modeled as

$$Y_{ij} = \beta_j + b_i + e_{ij}, j = 1, \ldots, J, i = 1, \ldots, n, \quad (6)$$

where β_j is the fixed intercept of the jth method, b_i is the random effect of the ith subject, and e_{ij} is the within-subject random error. To write this model in the matrix

notation of (1), let $\mathbf{1}_J$ be a J-vector of ones, take $(p, q) = (J, 1)$, and define

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{iJ} \end{bmatrix},\ \mathbf{X}_i = \mathbf{I}_J,\ \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_J \end{bmatrix},\ \mathbf{Z}_i = \mathbf{1}_J, s\ \mathbf{b}_i = b_i,\ \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ \vdots \\ e_{iJ} \end{bmatrix}.$$

Here the q-dimensional random effects vector $\mathbf{b}_i$ is actually a scalar quantity b_i. We assume that b_i and $\mathbf{e}_i$ follow (3). Thus, $b_i \sim \mathcal{ST}_1(0, \Psi, \lambda, \nu_b)$ and $\mathbf{e}_i \sim t_J(\mathbf{0}, \boldsymbol{\Sigma}, \nu_e)$, with $\boldsymbol{\Sigma}$ as a $J \times J$ diagonal matrix consisting of the diagonal elements $(\sigma_1^2, \ldots, \sigma_J^2)$. It may be noted that this model is often not useful for $J = 2$ despite being identifiable because the unreplicated data may not have enough information for reliable estimation of all model parameters, especially the scale parameters. The same situation also arises in the case of the standard mixed model.

Next, consider the case of replicated data. These data can be modeled as

$$Y_{ijk} = \beta_j + b_{ij} + e_{ijk}, k = 1, \ldots, m_{ij}, j = 1, \ldots, J, i = 1, \ldots, n, \qquad (7)$$

where β_j is the fixed intercept of the jth method, b_{ij} is the random effect of the ith subject on the jth method, and e_{ijk} is the within-subject random error. One can also think of the b_{ij} as subject $\times$ method interactions. This model can be written in the matrix notation of (1) by letting $\mathbf{0}_J$ denote a J-vector of zeros, taking $(p, q) = (J, J)$ and $\boldsymbol{\beta}$ as in unreplicated data, and defining $\mathbf{Y}_{ij} = (Y_{i11}, \ldots, Y_{i1m_{ij}})'$, $\mathbf{e}_{ij} = (e_{i11}, \ldots, e_{i1m_{ij}})'$,

$$\mathbf{Y}_i = \begin{bmatrix} \mathbf{Y}_{i1} \\ \vdots \\ \mathbf{Y}_{iJ} \end{bmatrix},\ \mathbf{X}_i = \mathbf{Z}_i = \begin{bmatrix} \mathbf{1}_{m_{i1}} & \ldots & \mathbf{0}_{m_{i1}} \\ \vdots & \ddots & \vdots \\ \mathbf{0}_{m_{iJ}} & \ldots & \mathbf{1}_{m_{iJ}} \end{bmatrix},\ \mathbf{b}_i = \begin{bmatrix} b_{i1} \\ \vdots \\ b_{iJ} \end{bmatrix},\ \mathbf{e}_i = \begin{bmatrix} \mathbf{e}_{i1} \\ \vdots \\ \mathbf{e}_{iJ} \end{bmatrix}.$$

It is assumed that $\mathbf{b}_i$ and $\mathbf{e}_i$ follow (3) with $\boldsymbol{\Psi}$ as an unstructured $J \times J$ scale matrix and $\boldsymbol{\Sigma}_i$ as an $M_i \times M_i$ diagonal matrix with diagonal elements $(\sigma_1^2 \mathbf{1}'_{m_{i1}}, \ldots, \sigma_J^2 \mathbf{1}'_{m_{iJ}})$. This model can be used for $J = 2$ as well; the aforementioned problem that arises with unreplicated data usually does not arise if measurements are replicated.

The models (6) and (7) are similar except that the subject random effects are common to all methods in (6), whereas they vary with methods in (7). The mean vector and covariance matrix of $\mathbf{Y}_i$ for both data types are given by (5). Both (6) and (7) are models for basic method comparison data that we commonly encounter in applications. They have been used with the normality assumption (2) in Choudhary and Yin [17]. The models may need to be modified to incorporate additional structures that may be present in the data—see, e.g., the model (13) for the crab claws data in Sect. 4.

3 Evaluation of Agreement Under the Basic Models

Let the vector $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_J)'$ consist of one measurement from each of the J methods under comparison on a randomly selected subject from the population. The distribution of $\tilde{\mathbf{Y}}$ induced by the assumed data model is needed to derive expressions for agreement measures.

3.1 Distributional Properties of $\tilde{\mathbf{Y}}$

In the case of unreplicated data, the model (6) induces a companion model for $\tilde{\mathbf{Y}}$,

$$\tilde{Y}_j = \beta_j + \tilde{b} + \tilde{e}_j,\ j = 1, \ldots, J,$$

where $\tilde{b} \sim \mathcal{ST}_1(0, \Psi, \lambda, \nu_b)$ and $\tilde{\mathbf{e}} = (\tilde{e}_1, \ldots, \tilde{e}_J)' \sim t_J(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}, \nu_e)$ are identically distributed as b_i and $\mathbf{e}_i$ in (6). Here $\tilde{\boldsymbol{\Sigma}} = \text{diag}\{\sigma_1^2, \ldots, \sigma_J^2\}$. One can think of the observed data $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ as independent draws from the distribution of $\tilde{\mathbf{Y}}$. To get $E[\tilde{\mathbf{Y}}]$ and $\text{var}[\tilde{\mathbf{Y}}]$ for $\nu_b, \nu_e > 2$, we simply substitute $(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Psi}, \boldsymbol{\lambda}) = (\mathbf{I}_J, \mathbf{1}_J, \tilde{\boldsymbol{\Sigma}}, \Psi, \lambda)$ in (5). The same substitution in (A.4) but without the subscript i gives a hierarchical representation for $\tilde{\mathbf{Y}}$ which can be used to determine its distribution.

The case for replicated data is completely analogous to the unreplicated data. The companion model for $\tilde{\mathbf{Y}}$ induced by the data model (7) is

$$\tilde{Y}_j = \beta_j + \tilde{b}_j + \tilde{e}_j,\ j = 1, \ldots, J,$$

where $\tilde{\mathbf{b}} = (\tilde{b}_1, \ldots, \tilde{b}_J)' \sim \mathcal{ST}_J(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \nu_b)$ is identically distributed as $\mathbf{b}_i$ and $\tilde{\mathbf{e}}$ is the same as before. Moreover, the moments for $\tilde{\mathbf{Y}}$ and its hierarchical representation are obtained by substituting $(\mathbf{I}_J, \mathbf{I}_J, \tilde{\boldsymbol{\Sigma}})$ for $(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\Sigma}_i)$ in (5) and for $(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Sigma})$ in (A.4).

The resulting hierarchical representations of $\tilde{\mathbf{Y}}$ in both cases have the same general form,

$$\tilde{\mathbf{Y}} \mid U, V \sim \mathcal{SN}_J(\boldsymbol{\beta}, \tilde{\boldsymbol{\Pi}}_V/U, \tilde{\boldsymbol{\lambda}}_V),\ V \sim \mathcal{G}(\nu_b/2, \nu_b/2),\ U/V \sim \mathcal{G}(\nu_e/2, \nu_e/2), \tag{8}$$

where $\tilde{\boldsymbol{\Pi}}_V$ and $\tilde{\boldsymbol{\lambda}}_V$ are counterparts of $\boldsymbol{\Pi}_V$ and $\boldsymbol{\lambda}_V$ from (A.2) which are obtained by appropriate substitution as mentioned previously, and $\mathcal{G}(\alpha, \beta)$ denotes a gamma distribution with density (A.3).

Next, for $j \neq l = 1, \ldots, J$, let $\tilde{D}_{jl}$ be the difference $\tilde{Y}_j - \tilde{Y}_l$, and $\mathbf{a}_{jl}$ be a J-vector whose jth element is 1, lth element is -1 and the rest are zero. It follows from Proposition 1 in Appendix A.3 that

$$\tilde{D}_{jl} \mid U, V \sim \mathcal{SN}_1\big(\beta_j - \beta_l,\ \mathbf{a}_{jl}'\tilde{\boldsymbol{\Pi}}_V\mathbf{a}_{jl}/U,\ \mathbf{a}_{jl}'\tilde{\boldsymbol{\Pi}}_V^{1/2}\tilde{\boldsymbol{\delta}}_V/(\mathbf{a}_{jl}'\tilde{\boldsymbol{\Gamma}}_V\mathbf{a}_{jl})^{1/2}\big), \tag{9}$$

where $\tilde{\boldsymbol{\delta}}_V = \tilde{\boldsymbol{\lambda}}_V/(1+\tilde{\boldsymbol{\lambda}}_V'\tilde{\boldsymbol{\lambda}}_V)^{1/2}$ and $\tilde{\boldsymbol{\Gamma}}_V = \tilde{\boldsymbol{\Pi}}_V - \tilde{\boldsymbol{\Pi}}_V^{1/2}\tilde{\boldsymbol{\delta}}_V\tilde{\boldsymbol{\delta}}_V'\tilde{\boldsymbol{\Pi}}_V^{1/2}$. Let F_{jl} be the distribution function of this conditional distribution. We can now write the unconditional distribution function of $|\tilde{D}_{jl}|$ as

$$P(|\tilde{D}_{jl}| \le t) = \int_0^\infty \int_0^\infty \{F_{jl}(t) - F_{jl}(-t)\}\, f(u, v\,|(\nu_b, \nu_e))\, du\, dv, t > 0, \quad (10)$$

where $f(u, v\,|(\nu_b, \nu_e))$ is the joint density of (U, V) appearing in (8). This integral is not available in a closed-form but can be computed numerically.

3.2 Inference on Agreement Measures

Let φ be a scalar measure of agreement between two methods. By definition, φ associated with the method pair $(j, l), l > j = 1, \ldots, J$, is a function of parameters of bivariate distribution of $(\tilde{Y}_j, \tilde{Y}_l)$, or more generally of $\tilde{\mathbf{Y}}$, whose distributional properties were discussed previously. Since φ is scalar, either a large or a small value for it indicates good agreement. Here we focus on only two agreement measures—CCC [31] and TDI [32]. Others can be handled in a similar manner.

The CCC between methods j and l is defined as

$$\text{CCC}_{jl} = 2\,\text{cov}[\tilde{Y}_j, \tilde{Y}_l]/\left\{(E[\tilde{Y}_j] - E[\tilde{Y}_l])^2 + \text{var}[\tilde{Y}_j] + \text{var}[\tilde{Y}_l]\right\}. \quad (11)$$

It lies between -1 and 1 and a large positive value for it implies good agreement. Next, for a given large probability $0 < p_0 < 1$, the TDI between methods j and l, say $\text{TDI}_{jl}(p_0)$, is defined as the p_0th quantile of $|\tilde{D}_{jl}|$. It is positive and indicates how large the differences between the methods j and l can be in $100p_0\,\%$ of the population. A small value for TDI implies good agreement. For both unreplicated and replicated data, CCC is computed for $\nu_b, \nu_e > 2$ by substituting in (11) the moments of $(\tilde{Y}_j, \tilde{Y}_l)$ obtained using (5). Similarly, the TDI is computed by numerically solving $P(|\tilde{D}_{jl}| \le t) = p_0$ for $t > 0$, where the probability on the left is given by (10).

When only two methods are compared, i.e., $J = 2$, one is mainly concerned with measuring between the methods. For this, one typically constructs a one-sided confidence bound for the agreement measure φ—a lower bound if large values for φ imply good agreement (e.g., CCC) or an upper bound if small values for φ imply good agreement (e.g., TDI). This bound is then used to determine whether the methods have satisfactory agreement. In the case of $J > 2$, one is additionally concerned with *multiple comparisons*—a comparison of values of φ for each pair of methods of interest—besides measuring agreement between the method pairs. For this, one constructs *simultaneous* one-sided confidence bounds for the pairs of values of φ of interest. These bounds are then used to determine which method pairs, if any, have satisfactory agreement, and also to order the method pairs on the basis of their extent of agreement.

To treat the two cases of J together, we index the method pairs of interest as $1, \ldots, K$. Thus, $K = 1$ if $J = 2$. More generally, $K = \binom{J}{2}$ if all pairwise comparisons are desired, whereas $K = J - 1$ if there is a reference method and only the comparisons with the reference are desired. Next, let $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_K)'$ be the vector of values of φ associated with the K method pairs. Obviously, $\boldsymbol{\varphi}$ is a function of the model parameter vector $\boldsymbol{\theta}$. Its ML estimator $\hat{\boldsymbol{\varphi}} = (\hat{\varphi}_1, \ldots, \hat{\varphi}_K)'$ is obtained by plugging in $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in $\boldsymbol{\varphi}$. From the large sample theory [30], $\hat{\boldsymbol{\varphi}}$ approximately follows a $\mathcal{N}_K(\boldsymbol{\varphi}, \mathbf{H}\mathfrak{I}^{-1}\mathbf{H}')$ distribution when the number of subjects n is large. Here $\mathbf{H} = \partial\boldsymbol{\varphi}/\partial\boldsymbol{\theta}$ is the Jacobian matrix evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\mathfrak{I}$ is the observed information matrix associated with the fitted model.

Following [17], the simultaneous confidence bounds for $\varphi_1, \ldots, \varphi_K$ can be computed as

$$\hat{\varphi}_l - c_{1-\alpha,K}\, v_{ll}^{1/2} \ \text{(lower bounds)}, \quad \hat{\varphi}_l - d_{\alpha,K}\, v_{ll}^{1/2} \ \text{(upper bounds)}, \tag{12}$$

where v_{ll} is the lth diagonal element of $\mathbf{H}\mathfrak{I}^{-1}\mathbf{H}'$, $l = 1, \ldots, K$, and $c_{1-\alpha,K}$ and $d_{\alpha,K}$ are critical points such that the limiting simultaneous coverage probability of each set of bounds is $(1 - \alpha)$ as $n \to \infty$. Essentially $c_{1-\alpha,K}$ is the $(1 - \alpha)$th quantile of $\max_{l=1}^{K} G_l$ and $d_{\alpha,K}$ is the αth quantile of $\min_{l=1}^{K} G_l$, where $(G_1, \ldots, G_K)$ follows the same multivariate normal distribution as the limiting joint distribution of $(\hat{\varphi}_l - \varphi_l)/v_{ll}^{1/2}$, $l = 1, \ldots, K$. These critical points can be computed using the method of [22] implemented in their `R` package `multcomp`. The critical points are standard normal quantiles when $K = 1$. Also, the finite sample accuracy of the bounds in (12) can be improved by first constructing them after applying a normalizing transformation to the agreement measure and then applying its inverse transformation to the results. The Fisher's z-transformation ($\tanh^{-1}$) of CCC and the log transformation of TDI are commonly used as normalizing transformations.

4 Illustration

Consider the crab claws data from [18]. These data consist of lengths (in mm) of 25 fiddler crab claws measured by three observers, each using two calipers. The measurements are replicated three times. Thus, each claw specimen has a total of $3 \times 6 = 18$ observations, $2 \times 3 = 6$ from each observer. This results in a total of $25 \times 18 = 450$ observations in the data. The primary goal of this study was to compare the three observers with respect to their extent of agreement between the two calipers. Figure 1 shows trellis plots of the data. Each row in a trellis plot displays all measurements on a specimen. The claw measurements range between 20 and 47 mm. The 18 measurements on each subject essentially overlap, suggesting not only that there is little variation within the replications, but also that there is little difference between either the calipers or the observers. These differences are easier to see in Fig. 2, which presents Bland-Altman plots of differences against averages. The three

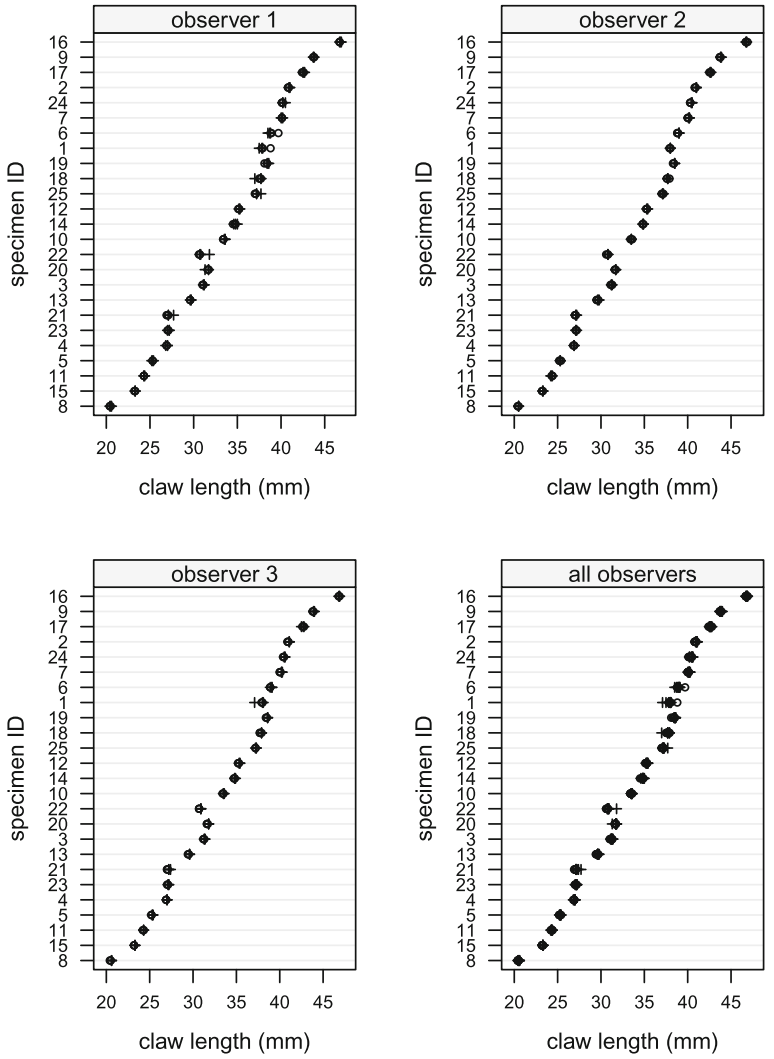


Fig. 1 Trellis plots of crab claws data. The circles (○) and the plus signs (+) represent measurements from calipers 1 and 2, respectively. Six measurements are displayed in each row of the first three plots. Eighteen measurements are displayed in each row of the last (*bottom-right*) plot

replicate measurements for each observer-caliper combination are averaged prior to creating these graphs. We see that the average measurements of caliper 2 tend to be slightly larger than caliper 1. Moreover, the differences in the average measurements seem to be largest for observer 1 and smallest for observer 2. Thus, while it appears that the agreement between the calipers is quite good for each observer, observer 2 seems to have the most agreement, followed by observers 3 and 1.

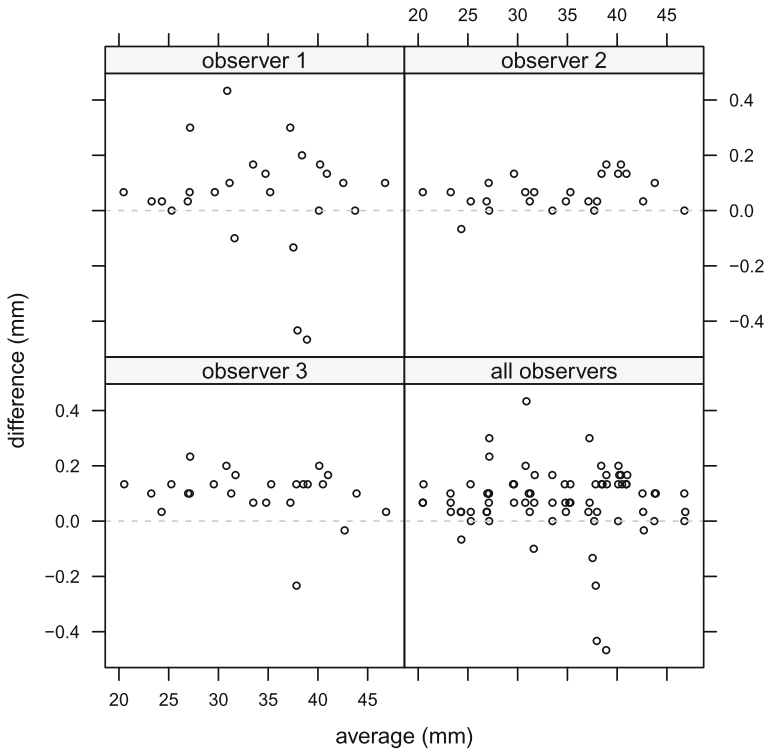


Fig. 2 Bland-Altman plots of differences (caliper 2 − caliper 1) against averages for crab claws data. The three replications for each observer-caliper combination are averaged prior to plotting. A horizontal line at zero is added in each plot

The modeling of these data calls for an extension of the basic model (7) for replicated data to incorporate effects of both observers and calipers. Choudhary et al. [18] adopt the model:

$$Y_{ijkl} = \beta_{jl} + b_{ij} + e_{ijkl},\ i = 1, \ldots, 25,\ j = 1, 2,\ k = 1, 2, 3,\ l = 1, 2, 3, \quad (13)$$

where Y_{ijkl} is the kth repeated measurement of the length of the ith claw, taken by the lth observer using the jth caliper; β_{jl} is the fixed intercept associated with the combination of caliper j and observer l; b_{ij} is the random interaction effect of claw i and caliper j; and e_{ijkl} is the random error. The repeated measurements for each combination of observer and caliper are independently and identically distributed. This model is written in the usual form (1) by taking

$$\begin{aligned}
&\mathbf{Y}_i = (Y_{i111}, Y_{i121}, Y_{i131}, Y_{i112}, Y_{i122}, Y_{i132}, \ldots, Y_{i213}, Y_{i223}, Y_{i233})',\\
&\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \ldots, \beta_{23})',\ \mathbf{b}_i = (b_{i1}, b_{i2})',\\
&\mathbf{e}_i = (e_{i111}, e_{i121}, e_{i131}, e_{i112}, e_{i122}, e_{i132}, \ldots, e_{i213}, e_{i223}, e_{i233})',
\end{aligned}$$

and conformably defining the matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$. The vectors $\mathbf{Y}_i$ and $\mathbf{e}_i$ have $M_i = 18$ elements, $\boldsymbol{\beta}$ has $p = 6$ elements, and $\mathbf{b}_i$ has $q = 2$ elements.

Initially, the model (13) is fit assuming the usual normality (2), where $\boldsymbol{\Psi}$ is a 2×2 unstructured matrix with ψ_1^2 and ψ_2^2 as diagonal elements and ψ_{12} as the off-diagonal element; and $\boldsymbol{\Sigma}$ is an 18×18 diagonal matrix

$$\boldsymbol{\Sigma} = \text{diag}\left\{\sigma_{11}^2, \sigma_{11}^2, \sigma_{11}^2, \sigma_{12}^2, \sigma_{12}^2, \sigma_{12}^2, \ldots, \sigma_{23}^2, \sigma_{23}^2, \sigma_{23}^2\right\},$$

parameterized in terms of six scale parameters σ_{jl}^2, $j = 1, 2$, $l = 1, 2, 3$. The assumption about the errors amounts to assuming that they are independent $\mathcal{N}_1(0, \sigma_{jl}^2)$ random variables. This model is fit using the `nlme` package of [37]. Figure 3 shows the resulting normal quantile-quantile (QQ) plots of the predicted random effects b_{i1} and b_{i2} and the standardized residuals. Also shown is a histogram of the residuals. These graphs suggest that the normality assumption is reasonable for the random effects, whereas a heavier tailed distribution than the normal is needed for the errors. The normality of random effects is further corroborated by the 0.47 p-value for the

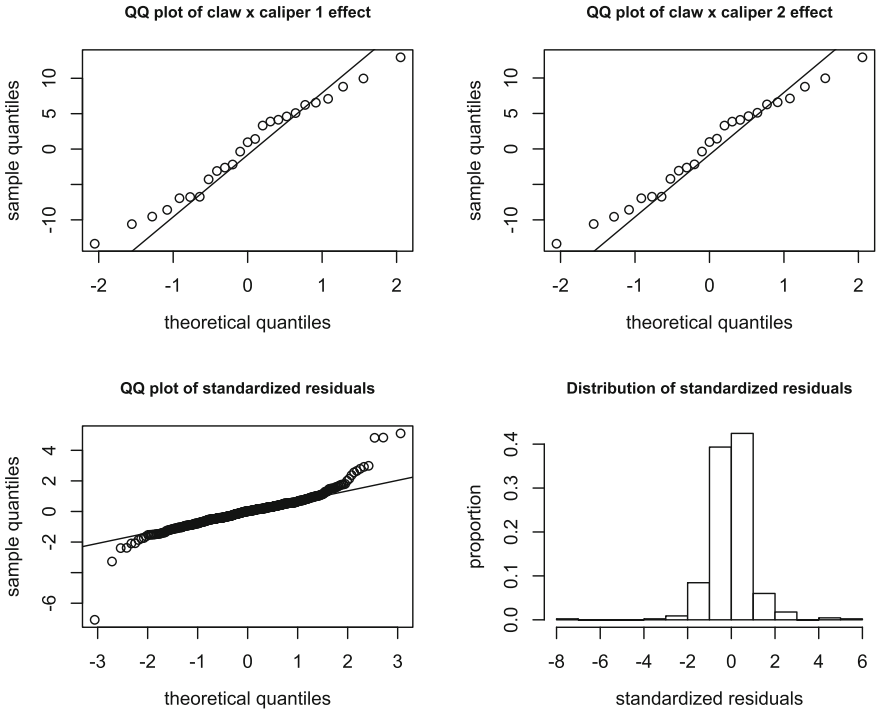


Fig. 3 Normal QQ plots of predicted claw × caliper random effects and standardized residuals, and a histogram of standardized residuals. A *line* passing through the first and third quartiles is added in each QQ plot

Shapiro-Wilk test of multivariate normality given by the `mvnormtest` package of [24] in R.

These diagnostics suggest modeling the data using a GST model where $\mathbf{b}_i$ is normally distributed as before, but the normality of $\mathbf{e}_i$ is replaced by a $t_{18}(\mathbf{0}, \boldsymbol{\Sigma}, \nu_e)$ distribution assumption. There are a total of 16 parameters in the GST model. It is fit using an ECM algorithm. Further details regarding estimation can be found in Choudhary et al. [18]. Here we simply note their conclusions that the model (13) with estimated error degrees of freedom $\hat{\nu}_e = 3.6$ fits reasonably well and the fit is substantially better than that of the standard mixed model. It is also preferred over a more general model that lets the random effects have a skew-t distribution with unknown parameters for skewness and degrees of freedom.

Our next task is to compare the extent of agreement that the three observers have between the two calipers. For this, we first need to adapt the approach of Sect. 3 to derive expressions for agreement measures based on the fitted GST model (13). Let $\tilde{Y}_{jl}$ be the length of a randomly selected claw specimen from the population as measured by the lth observer using the jth caliper, $j = 1, 2$, $l = 1, 2, 3$. The companion model for these $\tilde{Y}_{jl}$ induced by (13) is

$$\tilde{Y}_{jl} = \beta_{jl} + \tilde{b}_j + \tilde{e}_{jl},\ (\tilde{b}_1, \tilde{b}_2)' \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Psi}),\ (\tilde{e}_{11}, \tilde{e}_{12}, \ldots, \tilde{e}_{23})' \sim t_6(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}, \nu_e),$$

with $\tilde{\boldsymbol{\Sigma}} = \text{diag}\left\{\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{23}^2\right\}$. Next, the difference $\tilde{D}_l = \tilde{Y}_{1l} - \tilde{Y}_{2l}$ can be represented as

$$\tilde{D}_l \mid W_e \sim \mathcal{N}_1\big(E[\tilde{D}_l], \text{var}[\tilde{D}_l|W_e]\big),\ W_e \sim \mathcal{G}(\nu_e/2, \nu_e/2),$$

where $E[\tilde{D}_l] = \beta_{1l} - \beta_{2l}$ and $\text{var}[\tilde{D}_l|W_e] = \psi_1^2 + \psi_2^2 - 2\psi_{12} + (\sigma_{1l}^2 + \sigma_{2l}^2)/W_e$. Further, upon proceeding as in Sect. 3, we can see that for $\nu_e > 2$ the CCC between $(\tilde{Y}_{1l}, \tilde{Y}_{2l})$ for the lth observer is

$$\text{CCC}_l = \frac{2\psi_{12}}{(\beta_{1l} - \beta_{2l})^2 + \{\psi_1^2 + (\nu_e/(\nu_e - 2))\sigma_{1l}^2\} + \{\psi_2^2 + (\nu_e/(\nu_e - 2))\sigma_{2l}^2\}}.$$

Similarly, for a given large probability p_0, the TDI between $(\tilde{Y}_{1l}, \tilde{Y}_{2l})$ for the lth observer, say $\text{TDI}_l(p_0)$, is the solution of the equation

$$\begin{aligned} p_0 &= P(|\tilde{D}_l| \le t) \\ &= \int_0^\infty \Big[\Phi\{(t - E[\tilde{D}_l])/\text{sd}[\tilde{D}_l|w_e]\} - \Phi\{(-t - E[\tilde{D}_l])/\text{sd}[\tilde{D}_l|w_e]\}\Big] f(w_e\,|\nu_e)\, dw_e \end{aligned}$$

for $t > 0$. One can perform simultaneous inference on these measures as in Sect. 3.

The ML estimates for $z(\text{CCC}_l) = \tanh^{-1}(\text{CCC}_l)$ (the Fisher's z-transformation of CCC_l) and $\log(\text{TDI}_l)$ (with $p_0 = 0.90, 0.95$) for $l = 1, 2, 3$, their standard errors, and 95 % simultaneous confidence bounds—lower bounds for $z(\text{CCC})$ and upper bounds for $\log(\text{TDI})$— are presented in Table 1. The critical point for the CCC

Table 1 ML estimates, standard errors and one-sided 95 % simultaneous confidence bounds for $z(\text{CCC}_l)$, $l = 1, 2, 3$ and $\log(\text{TDI}_l)$, $l = 1, 2, 3$ with $p_0 = 0.90, 0.95$ under both GST and normal models

Measure	Estimate	Standard error	Confidence bound
GST model			
z(CCC)	(3.90, 4.57, 4.38)	(0.22, 0.21, 0.17)	(3.48, 4.18, 4.05)
log{TDI(0.90)}	(−0.87, −1.53, −1.34)	(0.12, 0.11, 0.09)	(−0.61, −1.30, −1.16)
log{TDI(0.95)}	(−0.61, −1.29, −1.14)	(0.15, 0.13, 0.11)	(−0.31, −1.02, −0.93)
Normal model			
z(CCC)	(3.90, 4.75, 4.34)	(0.15, 0.16, 0.16)	(3.60, 4.43, 4.03)
log{TDI(0.90)}	(−0.77, −1.62, −1.22)	(0.06, 0.08, 0.07)	(−0.65, −1.46, −1.06)
log{TDI(0.95)}	(−0.60, −1.45, −1.05)	(0.06, 0.08, 0.07)	(−0.47, −1.29, −0.91)

bounds is $c_{0.95,3} = 1.89$, and they are $d_{0.05,3} = -2.03$ and -1.98 for the TDI bounds with $p_0 = 0.90$ and 0.95, respectively. The critical points are computed using the `multcomp` package [22] in `R`.

The application of inverse transformation yields the simultaneous bounds for $(\text{CCC}_1, \text{CCC}_2, \text{CCC}_3)$ as (0.9981, 0.9995, 0.9994), and for $(\text{TDI}_1, \text{TDI}_2, \text{TDI}_3)$ as (0.54, 0.27, 0.31) with $p_0 = 0.90$ and as (0.73, 0.36, 0.39) with $p_0 = 0.95$. The CCC bounds are practically one and the TDI bounds are quite small compared to the magnitude of measurements that range between 20 and 47 mm. In particular, the bound of 0.54 implies that 90 % of differences in measurements from the two calipers by observer 1 are estimated to lie within ±0.54 mm. All three sets of bounds suggest excellent agreement between the two calipers for each observer. Furthermore, the ordering of the bounds imply that the calipers agree most in the case of observer 2, followed by observers 3 and 1. This conclusion is in line with what we expected from the exploratory data analysis.

Our next task is to assess the statistical significance of the differences in the extent of agreement across the observers. For this, we proceed in a manner similar to Sect. 3.2 and use the large-sample theory of ML estimators to construct 95 % simultaneous two-sided confidence intervals for pairwise differences in the observer-specific agreement measures. The confidence intervals for $z(\text{CCC}_1) - z(\text{CCC}_2)$, $z(\text{CCC}_1) - z(\text{CCC}_3)$ and $z(\text{CCC}_2) - z(\text{CCC}_3)$ are (−0.90, −0.44), (−0.74, −0.20) and (−0.04, 0.42), respectively. They indicate that the difference between observers 2 and 3 is not significant, but observer 1 differs significantly with them. The observer 1 has lower agreement between the calipers than the other observers. The difference, however, is not practically significant. The same conclusion is reached on the basis of confidence intervals for $\log(\text{TDI}_1) - \log(\text{TDI}_2)$, $\log(\text{TDI}_1) - \log(\text{TDI}_3)$ and $\log(\text{TDI}_2) - \log(\text{TDI}_3)$, which are (0.43, 0.88), (0.23, 0.70) and (−0.40, 0.02), respectively, in the case of $p_0 = 0.90$.

To see how the results under the assumed GST model compare with those under the standard model, we redo the analysis by replacing the t-distribution for errors with a normal distribution. These results are also presented in Table 1. There is no consistent pattern among the ML estimates, but all the standard errors are smaller under the normal model than the GST model. The latter finding appears reasonable as the GST model accounts for the heavy-tailedness in the errors (see Fig. 3). Interestingly, the bounds for each measure are such that, relative to the normal model, the implied agreement between the calipers under the GST model is less for observers 1 and 2 and more for observer 3. On the original scale, the bounds under the normal model are $(0.9985, 0.9997, 0.9994)$ for $(\text{CCC}_1, \text{CCC}_2, \text{CCC}_3)$, and $(0.52, 0.23, 0.35)$ and $(0.62, 0.27, 0.40)$ for $(\text{TDI}_1, \text{TDI}_2, \text{TDI}_3)$ with $p_0 = 0.90$ and 0.95, respectively. Upon comparing with their GST counterparts, we find that the bounds for CCC are virtually unchanged whereas those for TDI show a modest change for $p_0 = 0.90$. A somewhat more glaring change is evident for $p_0 = 0.95$, but only for observers 1 and 3. This finding may also be reasonable in that the impact of heavy tails is expected to be more stark for more extreme percentiles. On the whole, these findings for the crab data seem to suggest the following vis-à-vis the two models: (a) the standard errors appear better estimated under the GST model; (b) the CCC bounds under the two models appear quite similar because the agreement is very high; and (c) the difference between the two models becomes more evident in TDI bounds for values of p_0 that are closer to one.

5 Concluding Remarks

This article develops a methodology for analyzing method comparison data by modeling them within the framework of GST mixed models. The framework is flexible enough to incorporate data with skewness and heavy tails in addition to normally distributed data. This flexibility, however, comes at the cost of increased computational difficulty in fitting the model and standard error estimation. While the computations can be programmed using a statistical language such as `R`, it is important to do a sensitivity analysis to ensure that the results are reliable. At the minimum, this involves using different starting points for optimization algorithms and also different routines for optimization and numerical differentiation and integration. The results that we have presented appear to pass this sensitivity check. Comparing results with those based on the standard mixed model is also a good idea.

Acknowledgments The authors thank Golo Maurer, Rebecca Boulton and Leanne Reaney for assistance in collection of the crab claws data. They are also thankful to a reviewer for comments that greatly improved this article.

Appendix

A.1 Definitions

Let $\mathbf{Y}$ be a $J \times 1$ random vector with $\boldsymbol{\mu}$ as a $J \times 1$ location parameter vector and $\boldsymbol{\Sigma}$ as a $J \times J$ scale matrix. Define

$$\mathbf{Y}^* = \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}).$$

Also let $\phi_J(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the density function of a $\mathcal{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, and $\tau(\cdot, \nu)$ be the distribution function of a univariate t-distribution with ν degrees of freedom. The J-dimensional skew-normal, t and skew-t distributions are defined as follows.

Definition 1 $\mathbf{Y} \sim \mathcal{SN}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ if its density function is

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_J(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}'\mathbf{y}^*), \quad \mathbf{y} \in \mathbb{R}^J.$$

Definition 2 $\mathbf{Y} \sim t_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ if its density function is

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = (\nu\pi)^{-J/2} \frac{\text{gam}((\nu + J)/2)}{\text{gam}(\nu/2)} |\boldsymbol{\Sigma}|^{-1/2} \left(1 + \mathbf{y}^{*\prime}\mathbf{y}^*/\nu\right)^{-(\nu+J)/2}, \quad \mathbf{y} \in \mathbb{R}^J,$$

with $\text{gam}(\cdot)$ as the gamma function.

Definition 3 $\mathbf{Y} \sim \mathcal{ST}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ if its density function is

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) = 2\, f_t(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)\, \tau\left(\boldsymbol{\lambda}'\mathbf{y}^*\{(\nu + J)/(\nu + \mathbf{y}^{*\prime}\mathbf{y}^*)\}^{1/2} \mid \nu + J\right), \quad \mathbf{y} \in \mathbb{R}^J,$$

where $f_t(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ is the density function of a $t_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ distribution.

A.2 Hierarchical Representation for a GST Mixed Model

Let $\mathbf{Y}$ be an M-vector obtained by dropping the subscript i in $\mathbf{Y}_i$ defined by (1). From (3), the GST mixed model for $\mathbf{Y}$ can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}, \quad \mathbf{b} \sim \mathcal{ST}_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \nu_b), \quad \mathbf{e} \sim t_n(\mathbf{0}, \boldsymbol{\Sigma}, \nu_e), \tag{A.1}$$

where $\mathbf{b}$ and $\mathbf{e}$ are mutually independent. For a hierarchical representation of this model, define for $\nu > 0$,

$$\begin{aligned} \boldsymbol{\Pi}_\nu &= (\mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}' + \nu\boldsymbol{\Sigma}), \\ \boldsymbol{\lambda}_\nu &= \frac{\boldsymbol{\Pi}_\nu^{-1/2}\mathbf{Z}\boldsymbol{\Psi}^{1/2}\boldsymbol{\lambda}}{\left(1 + \boldsymbol{\lambda}'\boldsymbol{\Psi}^{-1/2}(\boldsymbol{\Psi}^{-1} + \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}/\nu)^{-1}\boldsymbol{\Psi}^{-1/2}\boldsymbol{\lambda}\right)^{1/2}}, \end{aligned} \tag{A.2}$$

and let $\mathcal{G}(\alpha, \beta)$ denote a gamma distribution with parameters $\alpha, \beta > 0$, and density

$$f(y|\alpha, \beta) = \frac{\beta^{\alpha}}{\text{gam}(\alpha)} y^{\alpha-1} \exp(-\beta y), \quad y > 0. \tag{A.3}$$

Now from [18], the model (A.1) can be represented as

$$\mathbf{Y}|U, V \sim \mathcal{SN}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Pi}_V/U, \boldsymbol{\lambda}_V),\ U \sim \mathcal{G}(\nu_b/2, \nu_b/2),\ U/V \sim \mathcal{G}(\nu_e/2, \nu_e/2). \tag{A.4}$$

A.3 Linear Combination of Skew-Normals

Proposition 1 *Let* $\mathbf{Y} \sim \mathcal{SN}_q(\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$ *and consider the quantities defined in* (4). *Let* $\mathbf{a} \in \mathbb{R}^q$ *with at least one non-zero element. Then*

$$\mathbf{a}'\mathbf{Y} \sim \mathcal{SN}_1\left(\mathbf{a}'\boldsymbol{\beta}, \mathbf{a}'\boldsymbol{\Psi}\mathbf{a}, \mathbf{a}'\boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}/(\mathbf{a}'\boldsymbol{\Gamma}\mathbf{a})^{1/2}\right).$$

Proof The proof relies on a stochastic representation of a skew-normal variate. Let $\mathbf{Y}^* \sim \mathcal{SN}_q(\mathbf{0}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$. Then, from [1],

$$\mathbf{Y}^* \stackrel{d}{=} \boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}|G_1^*| + \boldsymbol{\Psi}^{1/2}(\mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}')\mathbf{G}_2^*, \tag{A.5}$$

where $G_1^* \sim \mathcal{N}_1(0, 1)$, $\mathbf{G}_2^* \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}_q)$ independently of G_1^*, and the notation "$\stackrel{d}{=}$" means "equal in distribution." Using (A.5), we can write

$$\mathbf{a}'\mathbf{Y} \stackrel{d}{=} \mathbf{a}'\boldsymbol{\beta} + \mathbf{a}'\boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}|G_1^*| + (\mathbf{a}'\boldsymbol{\Gamma}\mathbf{a})^{1/2}G^*,$$

where $G^* \sim \mathcal{N}_1(0, 1)$ independently of G_1^*. Define

$$\lambda^* = \mathbf{a}'\boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}/(\mathbf{a}'\boldsymbol{\Gamma}\mathbf{a})^{1/2}, \quad \delta^* = \lambda^*/(1 + \lambda^{*2})^{1/2}.$$

From an application of (4), we have $(\mathbf{a}'\boldsymbol{\Psi}^{1/2}\boldsymbol{\delta})^2 + \mathbf{a}'\boldsymbol{\Gamma}\mathbf{a} = \mathbf{a}'\boldsymbol{\Psi}\mathbf{a}$, implying

$$\mathbf{a}'\boldsymbol{\Psi}^{1/2}\boldsymbol{\delta} = (\mathbf{a}'\boldsymbol{\Psi}\mathbf{a})^{1/2}\delta^*, \quad (\mathbf{a}'\boldsymbol{\Gamma}\mathbf{a})^{1/2} = (\mathbf{a}'\boldsymbol{\Psi}\mathbf{a})^{1/2}(1 - \delta^{*2})^{1/2}.$$

This allows us to write

$$\mathbf{a}'\mathbf{Y} \stackrel{d}{=} \mathbf{a}'\boldsymbol{\beta} + (\mathbf{a}'\boldsymbol{\Psi}\mathbf{a})^{1/2}\delta^*|G_1^*| + (\mathbf{a}'\boldsymbol{\Psi}\mathbf{a})^{1/2}(1 - \delta^{*2})^{1/2}G^*.$$

Now the result follows from the representation (A.5) for the univariate case.

References

1. Arellano-Valle, R.B., H. Bolfarine, and V.H. Lachos. 2005. Skew-normal linear mixed models. *Journal of Data Science* 3: 415–438.
2. Azzalini, A. 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12: 171–178.
3. Azzalini, A., and A. Capitanio. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B* 65: 367–389.
4. Barnhart, H.X., and J.M. Williamson. 2001. Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* 57: 931–940.
5. Barnhart, H.X., M.J. Haber, and L.I. Lin. 2007. An overview on assessing agreement with continuous measurement. *Journal of Biopharmaceutical Statistics* 17: 529–569.
6. Barnhart, H.X., M.J. Haber, and J. Song. 2002. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 58: 1020–1027.
7. Barnhart, H.X., J. Song, and M.J. Haber. 2005. Assessing intra, inter and total agreement with replicated readings. *Statistics in Medicine* 24: 1371–1384.
8. Bland, J.M., and D.G. Altman. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8: 135–160.
9. Carrasco, J.L., and L. Jover. 2003. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* 59: 849–858.
10. Carrasco, J.L., L. Jover, T.S. King, and V.M. Chinchilli. 2007. Comparison of concordance correlation coefficient estimating approaches with skewed data. *Journal of Biopharmaceutical Statistics* 17: 673–684.
11. Carrasco, J.L., T.S. King, and V.M. Chinchilli. 2009. The concordance correlation coefficient for repeated measures estimated by variance components. *Journal of Biopharmaceutical Statistics* 19: 90–105.
12. Carstensen, B. 2010. *Comparing Clinical Measurement Methods: A Practical Guide*. New York: Wiley.
13. Carstensen, B. Simpson, J. Gurrin, L.C. 2008. Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics* 4. doi:10.2202/1557-4679.1107
14. Choudhary, P.K. 2008. A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and Inference* 138: 1102–1115.
15. Choudhary, P.K. 2010. A unified approach for nonparametric evaluation of agreement in method comparison studies. *The International Journal of Biostatistics* 6. doi:10.2202/1557-4679.1235
16. Choudhary, P.K., and H.N. Nagaraja. 2007. Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* 137: 279–290.
17. Choudhary, P.K., and K. Yin. 2010. Bayesian and frequentist methodologies for analyzing method comparison studies with multiple methods. *Statistics in Biopharmaceutical Research* 2: 122–132.
18. Choudhary, P.K., D. Sengupta, and P. Cassey. 2014. A general skew-*t* mixed model that allows different degrees of freedom for random effects and error distributions. *Journal of Statistical Planning and Inference* 147: 235–247.
19. Genton, M.G. 2004. *Skew-Elliptical distributions and their applications - A journey beyond normality*. Boca Raton: Chapman & Hall/CRC Press.
20. Gilbert, P. and Varadhan, R. 2012. *numDeriv: Accurate Numerical Derivatives*. R package version 2012.9-1
21. Ho, H.J., and T.I. Lin. 2010. Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Biometrical Journal* 52: 449–469.
22. Hothorn, T., F. Bretz, and P. Westfall. 2008. Simultaneous inference in general parametric models. *Biometrical Journal* 50: 346–363.

23. Igic, B., M.E. Hauber, J.A. Galbraith, T. Grim, D.C. Dearborn, P.L.R. Brennan, C. Moskat, P.K. Choudhary, and P. Cassey. 2010. Comparison of micrometer—and scanning electron microscope-based measurements of avian eggshell thickness. *Journal of Field Ornithology* 81: 402–410.
24. Jarek, S. 2012. *mvnormtest: Normality test for multivariate variables*. R package version 0.1-9
25. King, T.S., and V.M. Chinchilli. 2001. A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* 20: 2131–2147.
26. King, T.S., and V.M. Chinchilli. 2001. Robust estimators of the concordance correlation coefficient. *Journal of Biopharmaceutical Statistics* 11: 83–105.
27. King, T.S., V.M. Chinchilli, and J.L. Carrasco. 2007. A repeated measures concordance correlation coefficient. *Statistics in Medicine* 26: 3095–3113.
28. King, T.S., V.M. Chinchilli, K.-L. Wang, and J.L. Carrasco. 2007. A class of repeated measures concordance correlation coefficients. *Journal of Biopharmaceutical Statistics* 17: 653–672.
29. Lachos, V.H., P. Ghosh, and R.B. Arellano-Valle. 2010. Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica* 20: 303–322.
30. Lehmann, E.L. 1998. *Elements of Large-Sample Theory*. New York: Springer.
31. Lin, L.I. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–268. Corrections: 2000, **56**, 324–325
32. Lin, L.I. 2000. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 19: 255–270.
33. Lin, L.I., A.S. Hedayat, and W. Wu. 2007. A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics* 17: 629–652.
34. Lin, L.I., A.S. Hedayat, and W. Wu. 2011. *Statistical Tools for Measuring Agreement*. New York: Springer.
35. McLachlan, G.J., and T. Krishnan. 2007. *The EM algorithm and extensions*, 2nd ed. New York: Wiley.
36. Meng, X.-L., and D.B. Rubin. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80: 267–278.
37. Pinheiro, J.C., and D.M. Bates. 2000. Mixed-Effects models in S and S-PLUS. New York: Springer.
38. Pinheiro, J.C., C. Liu, and Y.N. Wu. 2001. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution. *Journal of Computational and Graphical Statistics* 10: 249–276.
39. R Core Team. 2014. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
40. Roy, A. 2009. An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19: 150–173.
41. Smyth, G., Y. Hu, P. Dunn, B. Phipson, and Y. Chen. 2014. statmod: Statistical modeling. *R package version* 1.4.20.
42. Verbeke, G., and E. Lesaffre. 1996. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* 91: 217–221.
43. Wang, C.M., and H.K. Iyer. 2008. Fiducial approach for assessing agreement between two instruments. *Metrologia* 45: 415–421.
44. Zhang, D., and M. Davidian. 2001. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* 57: 795–802.

Part III
Statistical Methods for Health Research

Inference for a Poisson-Inverse Gaussian Model with an Application to Multiple Sclerosis Clinical Trials

Mallikarjuna Rettiganti and H.N. Nagaraja

Abstract Magnetic resonance imaging (MRI) based new brain lesion counts are widely used to monitor disease progression in relapsing remitting multiple sclerosis (RRMS) clinical trials. These data generally tend to be overdispersed with respect to a Poisson distribution. It has been shown that the Poisson-Inverse Gaussian (P-IG) distribution fits better than the negative binomial to MRI data in RRMS patients that have been selected for lesion activity during the baseline scan. In this paper we use the P-IG distribution to model MRI lesion count data from RRMS parallel group trials. We propose asymptotic and simulation based exact parametric tests for the treatment effect such as the likelihood ratio (LR), score and Wald tests. The exact tests maintain precise Type I error levels whereas the asymptotic tests fail to do so for small samples. The LR test remains empirically unbiased and results in 30–50 % reduction in sample sizes required when compared to the Wilcoxon rank sum (WRS) test. The Wald test has the highest power to detect a reduction in the number of lesion counts and provides a 40–57 % reduction in sample sizes when compared to the WRS test.

Keywords Poisson inverse gaussian distribution · Parallel group trials · Multiple sclerosis · Wald test · Sample size

1 Introduction

Multiple Sclerosis (MS) is an autoimmune disease which attacks the central nervous system and has the potential to cause severe crippling disabilities. Relapsing remitting

M. Rettiganti (✉)
ACHRI, University of Arkanasas for Medical Sciences, 1 Children's Way, Little Rock, AR 72202, USA
e-mail: mrrettiganti@uams.edu

H.N. Nagaraja
College of Public Health, The Ohio State University, Columbus 43210, USA
e-mail: nagaraja.1@osu.edu

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_12

multiple sclerosis (RRMS) is the first stage of this disease in which the patient experiences distinct phases of relapses and remissions. Disease progression in MS has been generally measured by clinical end points such as the relapse rate and the Expanded Disability Status Scale. However, cumulative magnetic resonance imaging (MRI) lesion counts from gadolinium injected T1-weighted scans of the brain are widely used as primary and secondary end points in Phase I and II MS trials and as secondary endpoints in Phase III trials. Unlike the clinical measures, these counts are not subjective and are highly sensitive to disease changes. The total number of new enhancing lesions observed during successive monthly scans is often used as the outcome variable. For the first month's scan, the new enhancing lesions are counted with respect to a reference scan.

Nonparametric methods have been used in the past to compare treatment groups in MS clinical trials that use MRI count data as an end point. Nauta et al. [6] used bootstrap resampling and nonparametric tests to estimate power and sample sizes for various trial designs involving RRMS patients. In recent years, parametric modeling of the count data has been pursued. Sormani et al. [13, 15] first proposed the negative binomial (NB) distribution as a model and showed that it gave a good fit for MRI data in untreated RRMS patients. Although these papers modeled the data and validated the NB model, sample sizes to detect treatment effect were computed using the nonparametric Wilcoxon rank sum (WRS) test. This approach was later improved by [1] who proposed and compared several parametric tests for the treatment effect in a two parallel group (PG) trial. They considered the likelihood ratio test (LRT), Rao's score test (RST) and the Wald test (WT) and used chi-squared approximations to the test statistics which resulted in inflated Type I error levels for small sample sizes. Rettiganti and Nagaraja [11] overcame this limitation by proposing *exact* parametric tests for PG trials based on the NB model that maintain precise Type I error levels even for very small sample sizes. They also proposed exact parametric tests for RRMS baseline versus treatment trials which showed marked improvement in power when compared to nonparametric tests.

Although the NB model has served well for many overdispersed count data problems, there are other mixed-Poisson distributions that can be more appropriate. Sormani et al. [14] showed that the Poisson-Inverse Gaussian (P-IG) distribution provides a better fit than the NB distribution to MRI count data from 115 RRMS patients who had at least one lesion count observed during the reference scan. However, parametric tests for the treatment effect based on the P-IG model have not been studied and we attempt to address that issue in this paper. In Sect. 2, we introduce the P-IG distribution and use it to model MRI count data in RRMS PG trials. Maximum likelihood estimates (MLEs) of the model parameters are obtained in Sect. 3. In Sect. 4, we propose likelihood based parametric tests for the treatment effect such as LRT, RST and WTs. We compare the performance of these tests with a detailed power analysis and obtain sample size estimates for RRMS PG trials in Sect. 5. In Sect. 6, we conclude with a summary of our results and provide some discussion.

We denote a chi-squared random variable (rv) with 1 degree of freedom as χ_1^2 and its upper νth percentile $\chi_1^2(1-\nu)$. For simplicity, we use c_0 to denote $\chi_1^2(0.95)$ (=3.8415). A Poisson rv with mean η is denoted as Poisson(η). The likelihood

of the observed data is denoted as L and $\log(L)$ is denoted as ℓ. For a scalar or vector parameter θ, Θ denotes the unrestricted parameter space while $\hat{\theta}$ denotes its maximum likelihood estimate (MLE) under Θ. Similarly, Θ_0 and $\tilde{\theta}$ denote the restricted parameter space and its associated MLE respectively. The vector of first order partial derivatives of ℓ, $\partial\ell/\partial\theta$, is called the score vector. All tests carried out in this paper are two-sided and have 5 % nominal significance level. All computations were done using the statistical software package [8].

2 The Poisson-Inverse Gaussian (P-IG) Model

2.1 The Basic Model

Let a rv Z have Inverse Gaussian (IG) distribution with parameters μ and λ and density

$$f_Z(z|\mu,\lambda) = \left(\frac{\lambda}{2\pi z^3}\right)^{\frac{1}{2}} \exp\left\{\frac{-\lambda(z-\mu)^2}{2\mu^2 z}\right\}, \quad z>0,\ \mu,\lambda>0. \tag{1}$$

We then write $Z \sim \text{IG}(\mu,\lambda)$. When $Y|Z=z$ has Poisson(z), the marginal pmf of Y is given by

$$\begin{aligned} P(Y=y) = p_y &= \int_0^\infty \frac{e^{-z}z^y}{y!} f_Z(z|\mu,\lambda)dz \\ &= \frac{\tau^y}{y!}\left(\frac{2\omega}{\pi}\right)^{\frac{1}{2}} \exp\left(\frac{\lambda}{\mu}\right) K_{y-\frac{1}{2}}(\omega), \qquad y=0,1,\ldots, \end{aligned} \tag{2}$$

with $\tau = (1/\mu^2 + 2/\lambda)^{-1/2}$, $\omega = \lambda/\tau$. Here, $\nu = y - 1/2$ is non-integer and $K_\nu(\cdot)$ is the modified Bessel function of the third kind, defined as

$$K_\nu(z) = \frac{\pi}{2}\cdot\frac{I_{-\nu}(z) - I_\nu(z)}{\sin\nu\pi}, \tag{3}$$

where $I(\cdot)$ is the modified Bessel function of the first kind given by

$$I_\nu(z) = \sum_{m=0}^{\infty} \frac{(\frac{z}{2})^{\nu+2m}}{m!\Gamma(m+\nu+1)}. \tag{4}$$

It can be seen from (2) that

$$p_0 = \exp\left(\frac{\lambda}{\mu} - \frac{\lambda}{\tau}\right) \text{ and } p_1 = \tau p_0. \tag{5}$$

For $y \geq 2$, the P-IG probabilities satisfy the recurrence relation

$$p_y = \tau^2 \left\{ \frac{p_{y-2}}{y(y-1)} + \frac{2y-3}{\lambda y} p_{y-1} \right\}. \tag{6}$$

If a rv Y has the pmf given by (2) we write $Y \sim$ P-IG(μ, λ). Both P-IG and NB pmfs are unimodal and right skewed but the former has a longer tail. Further, $E(Y) = \mu$ and $Var(Y) = \mu + \mu^3/\lambda$. The MLE of μ is the sample mean [16].

To compute the P-IG probabilities using the closed form expression given in (2) we need to evaluate the Bessel function of the third kind $K(\cdot)$. This can be computed easily for small values of y, but for large values, (3) and/or (2) may return infinite values in which case the probabilities cannot be computed directly. This problem can be avoided by using the recursive relation in (6).

2.2 A P-IG Model for PG Trials

Suppose there are n_1 subjects in the placebo group (group 1) and n_2 subjects in the treatment group (group 2). Let Y_i denote the total number of new enhancing lesions seen in a subject in the ith group and Z_i denote the associated random subject effect, $i = 1, 2$. We assume that $Y_i | Z_i = z \sim$ Poisson(z) and that $Z_1 \sim$ IG(μ, λ) and $Z_2 \sim$ IG$(\gamma\mu, \lambda)$ with common scale parameter λ. Here $1 - \gamma$ is the measure of the treatment effect, which can be viewed as the percentage reduction in the mean lesion counts seen in the treatment group. Then, Y_1 and Y_2 are independent, $Y_1 \sim$ P-IG(μ, λ) and $Y_2 \sim$ P-IG$(\gamma\mu, \lambda)$, and from (2) we obtain

$$\begin{aligned} P(Y_1 = y_1) &= \frac{\tau_1^{y_1}}{y_1!} \left(\frac{2\omega_1}{\pi}\right)^{\frac{1}{2}} \exp\left\{\frac{\lambda}{\mu}\right\} K_{y_1 - \frac{1}{2}}(\omega_1), \\ P(Y_2 = y_2) &= \frac{\tau_2^{y_2}}{y_2!} \left(\frac{2\omega_2}{\pi}\right)^{\frac{1}{2}} \exp\left\{\frac{\lambda}{\gamma\mu}\right\} K_{y_2 - \frac{1}{2}}(\omega_2), \end{aligned} \tag{7}$$

where

$$\tau_1 = \left(\frac{1}{\mu^2} + \frac{2}{\lambda}\right)^{-\frac{1}{2}}, \quad \tau_2 = \left(\frac{1}{\gamma^2\mu^2} + \frac{2}{\lambda}\right)^{-\frac{1}{2}}, \text{ and } \quad \omega_i = \frac{\lambda}{\tau_i}, \ i = 1, 2.$$

For observed counts $\mathbf{y}_1 = (y_{11}, y_{12}, \ldots, y_{1n_1})$ and $\mathbf{y}_2 = (y_{21}, y_{22}, \ldots, y_{2n_2})$, it follows from (7) that the likelihood function for a parallel group P-IG model is given by

$$L(\gamma,\mu,\lambda|\mathbf{y}_1,\mathbf{y}_2)=\prod_{i=1}^{n_1}P(Y_1=y_{1i})\times\prod_{j=1}^{n_2}P(Y_2=y_{2j})$$
$$=\frac{(\tau_1)^{n_1\bar{y}_1}}{\prod y_{1i}!}\left(\frac{2\lambda}{\pi\tau_1}\right)^{\frac{n_1}{2}}\exp\left\{\frac{n_1\lambda}{\mu}\right\}\prod_{i=1}^{n_1}K_{y_{1i}-\frac{1}{2}}(\omega_1)$$
$$\times\frac{(\tau_2)^{n_2\bar{y}_2}}{\prod y_{2j}!}\left(\frac{2\lambda}{\pi\tau_2}\right)^{\frac{n_2}{2}}\exp\left\{\frac{n_2\lambda}{\gamma\mu}\right\}\prod_{j=1}^{n_2}K_{y_{2j}-\frac{1}{2}}(\omega_2). \quad (8)$$

3 Parameter Estimation

The parameter of interest γ, and nuisance parameters μ and λ can be estimated using the method of maximum likelihood. The log-likelihood function is

$$\ell(\gamma,\mu,\lambda|\mathbf{y}_1,\mathbf{y}_2)=n_1\bar{y}_1\log(\tau_1)+\frac{n_1}{2}\log\left(\frac{2\lambda}{\pi\tau_1}\right)+\frac{n_1\lambda}{\mu}+\sum_{i=1}^{n_1}\log K_{y_{1i}-\frac{1}{2}}(\omega_1)$$
$$+\,n_2\bar{y}_2\log(\tau_2)+\frac{n_2}{2}\log\left(\frac{2\lambda}{\pi\tau_2}\right)+\frac{n_2\lambda}{\gamma\mu}+\sum_{j=1}^{n_2}\log K_{y_{2j}-\frac{1}{2}}(\omega_2)$$
$$+\sum_{i=1}^{n_1}\log y_{1i}!+\sum_{j=1}^{n_2}\log y_{2j}!. \quad (9)$$

Intermediate steps containing the first and second order partial derivatives needed to obtain the MLEs and the Fisher information matrix (FIM) are given in Appendix 1 and 2. The MLEs of γ, μ and λ can be obtained by equating the score vector equations (17)–(19) to zero and solving simultaneously for the three parameters. Doing so for the first two, we readily obtain

$$\sum_{i=1}^{n_1}R_{y_{1i}-\frac{1}{2}}(\omega_1)=\frac{n_1\mu}{\tau_1}\quad\text{and}\quad\sum_{j=1}^{n_2}R_{y_{2j}-\frac{1}{2}}(\omega_2)=\frac{n_2\gamma\mu}{\tau_2},$$

where $R_\nu(z)=K_{\nu+1}(z)/K_\nu(z)$. Using these results in the third and using the fact that

$$\left(\frac{\lambda-\tau_1^2}{\tau_1^2}\right)=\frac{\lambda+\mu^2}{\mu^2}\quad\text{and}\quad\left(\frac{\lambda-\tau_2^2}{\tau_2^2}\right)=\frac{\lambda+\gamma^2\mu^2}{\gamma^2\mu^2},$$

we obtain

$$n_1(\bar{y}_1-\mu)+n_2(\bar{y}_2-\gamma\mu)=0. \quad (10)$$

One solution for Eq. (10) is $\hat{\mu} = \bar{y}_1$ and $\hat{\gamma}\hat{\mu} = \bar{y}_2$, suggesting that it may be a local maximum. We examined the surface of the log-likelihood as a function of γ and μ for selected λ values (figures not shown). The log-likelihood is a smooth concave function for the values of the parameters considered with a unique maximum attained at $\hat{\mu} = \bar{y}_1$ and $\hat{\gamma} = \bar{y}_2/\bar{y}_1$, where λ maximized the profile log-likelihood $\ell(\bar{y}_2/\bar{y}_1, \bar{y}_1, \lambda)$. Thus the local maximum obtained as a solution to the score vector equations (17)–(19) can be argued to be the MLEs of the parameters, which leads us to the following conclusions.

3.1 Unrestricted MLEs

The MLEs of μ and γ are $\hat{\mu} = \bar{y}_1$ and $\hat{\gamma} = \bar{y}_2/\bar{y}_1$, and the MLE of λ, $\hat{\lambda}$, can be obtained as a solution to the equation

$$\left(\frac{\hat{\lambda} - \hat{\tau}_1^2}{\hat{\lambda}\hat{\tau}_1}\right)\sum_{i=1}^{n_1} R_{y_{1i}-\frac{1}{2}}(\hat{\omega}_1) + \left(\frac{\hat{\lambda} - \hat{\tau}_2^2}{\hat{\lambda}\hat{\tau}_2}\right)\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\hat{\omega}_2) = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{\hat{\lambda}} + \frac{n_1\hat{\gamma} + n_2}{\hat{\gamma}\hat{\mu}},$$

where $\hat{\omega}_i = \hat{\lambda}/\hat{\tau}_i,\ i = 1, 2$ and

$$\hat{\tau}_1 = \left[\frac{1}{\hat{\mu}^2} + \frac{2}{\hat{\lambda}}\right]^{-\frac{1}{2}} \quad \text{and} \quad \hat{\tau}_2 = \left[\frac{1}{\hat{\gamma}^2\hat{\mu}^2} + \frac{2}{\hat{\lambda}}\right]^{-\frac{1}{2}}.$$

Alternatively $\hat{\lambda}$ can also be obtained by numerically maximizing the profile log-likelihood function $\ell(\hat{\gamma}, \hat{\mu}, \lambda)$ with respect to λ.

3.2 MLEs Under the Restricted Hypothesis

When $\gamma = \gamma_0$ is assumed known, the MLEs of μ and λ can be obtained by setting the score equations (18) and (19) to 0 with $\gamma = \gamma_0$ and simultaneously solving for the two parameters. For a general γ_0 the MLEs are not available in closed form and numerical methods must be employed. However, for $\gamma_0 = 1$ (the null hypothesis of no treatment effect), we have $\tau_1 = \tau_2$ and $\omega_1 = \omega_2$, using which we can conclude that the MLE of μ is $\tilde{\mu} = (n_1\bar{y}_1 + n_2\bar{y}_2)/(n_1 + n_2)$, the grand mean. The MLE of λ, $\tilde{\lambda}$ solves the equation

$$\left(\frac{\lambda - \tilde{\tau}_1^2}{\lambda\tilde{\tau}_1}\right)\sum_{i=1}^{n_1} R_{y_{1i}+\frac{1}{2}}(\tilde{\omega}_1) + \left(\frac{\tilde{\lambda} - \tilde{\tau}_2^2}{\tilde{\lambda}\tilde{\tau}_2}\right)\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\tilde{\omega}_2) = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{\tilde{\lambda}} + \frac{n_1 + n_2}{\tilde{\mu}},$$

where $\tilde{\omega}_i = \tilde{\lambda}/\tilde{\tau}_i,\ i = 1, 2$ and

$$\tilde{\tau}_1 = \left(\frac{1}{\tilde{\mu}^2} + \frac{2}{\tilde{\lambda}}\right)^{-\frac{1}{2}} \quad \text{and} \quad \tilde{\tau}_2 = \left(\frac{1}{\tilde{\mu}^2} + \frac{2}{\tilde{\lambda}}\right)^{-\frac{1}{2}}.$$

The MLE of λ can also be obtained by numerically maximizing the profile log-likelihood function $\ell(1, \tilde{\mu}, \lambda)$ with respect to λ.

The asymptotic variance of $\hat{\gamma}$ for the P-IG model is the first element of the inverse of the Fisher information matrix (FIM). That is,

$$\sigma^2_{\hat{\gamma}}(\boldsymbol{\theta}) = I^{-1}_{1.2} = [I_{11} - I_{12} I^{-1}_{22} I_{21}]^{-1} \tag{11}$$

where the elements of the FIM are as in Appendix 2.

4 Hypothesis Testing

In this section we propose parametric tests such as the LRT, RST and WT to test for the treatment effect for a general $H_0 : \gamma = \gamma_0$ versus $H_1 : \gamma \neq \gamma_0$. For RRMS clinical trials, a test for no treatment effect would test $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$. In the results that follow, $\hat{\gamma}$, $\hat{\mu}$, $\hat{\lambda}$ denote the MLEs under Θ given in Sect. 3.1 and $\tilde{\mu}$, $\tilde{\lambda}$ denote the MLEs of the parameters under Θ_0 given in Sect. 3.2. We now display the test statistics for these tests.

4.1 Test Statistics

Likelihood Ratio Test (LRT [10]): The LRT statistic to test $H_0 : \gamma = \gamma_0$ versus $H_1 : \gamma \neq \gamma_0$ is

$$LRT = -2(\ell(\gamma_0, \tilde{\mu}, \tilde{\alpha}) - \ell(\hat{\gamma}, \hat{\mu}, \hat{\alpha})). \tag{12}$$

Rao Score Test (RST [9, 10]): The RST statistic to test $H_0 : \gamma = \gamma_0$ versus $H_1 : \gamma \neq \gamma_0$ is

$$\begin{aligned} RST &= \left\{ \frac{\partial \ell(\gamma, \mu, \alpha)}{\partial \gamma} \right\}^2_{\theta=\tilde{\theta}} \times \sigma^2_{\hat{\gamma}}(\tilde{\boldsymbol{\theta}}) \\ &= \left\{ \frac{\tilde{\lambda}\tilde{\tau}_2}{\gamma_0^3 \tilde{\mu}^2} \sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\tilde{\omega}_2) - \frac{n_2 \tilde{\lambda}}{\gamma_0^2 \tilde{\mu}} \right\}^2 \times \sigma^2_{\hat{\gamma}}(\tilde{\boldsymbol{\theta}}), \end{aligned} \tag{13}$$

where $\sigma^2_{\hat{\gamma}}(\tilde{\boldsymbol{\theta}})$ is the asymptotic variance of $\hat{\gamma}$ given in (11) evaluated at the MLEs under H_0.

Wald Test (WT [10, 17]): The WT statistic for testing $H_0 : \gamma = \gamma_0$ versus $H_1 : \gamma \neq \gamma_0$ is given by

$$WT(\gamma) = \left\{ \frac{\hat{\gamma} - \gamma_0}{\sigma_{\hat{\gamma}}} \right\}^2, \tag{14}$$

where $\sigma^2_{\hat{\gamma}}(\hat{\theta})$ is the asymptotic variance of $\hat{\gamma}$ given in Eq. (11) evaluated at the unrestricted MLE $\hat{\theta}$.

Estimation of $\sigma^2_{\hat{\gamma}}$: To compute the RST, $\sigma^2_{\hat{\gamma}}$ needs to be evaluated at the restricted MLEs. Since the FIM is not available in closed form, the observed information evaluated at the MLEs under H_0 is used in its place. However, this approach can generate negative variance estimates which leads to negative score test statistics and an inconsistent test. Thus an RST using the observed information may not produce a valid chi-square test. Freedman [3] gives a detailed discussion of these anomalies. Morgan et al. [5] give an example involving a zero-inflated Poisson distribution where the score test statistic using the observed information is negative. This is not an issue for the WT because $\sigma^2_{\hat{\gamma}}$ is evaluated at the unrestricted MLEs and the observed information will be positive definite, ensuring consistency of the test.

We also evaluated WTs for other differentiable functions $g(\gamma)$ such as $g(\gamma) = \log(\gamma)$, $\sqrt{\gamma}$, and γ^2. The WT for $\log(\gamma)$ had properties similar to the LRT, while the WT for γ^2 had properties similar to that of the WT for γ. Those results are not presented here.

4.2 Exact Percentile

Each of the above statistics is asymptotically distributed as a χ^2_1 rv under H_0 and an approximate level ν test rejects H_0 if the test statistic is $> \chi^2_1(1-\nu)(\equiv c_0)$. The validity of this asymptotic approximation for small sample sizes was evaluated using simulation. For a given sample size, the exact percentile estimate for any test statistic is computed by first simulating $M = 100{,}000$ data sets under the null hypothesis H_0 (no treatment effect) and computing the 95th percentile of the resulting distribution of test statistics. Figure 1 shows the simulation based exact 95th percentile estimates for the three asymptotic tests as a function of common sample size $n = n_1 = n_2$. The exact percentiles for the LRT and WT converge to c_0 as n increases.

We did a simulation study to evaluate further the effect of changing sample size n, μ, and λ on the simulated levels and exact percentiles. The initial parameter estimates used for the simulation were the ones given by [14]. They fit the P-IG model (using an alternate parametrization, μ and $\beta = \mu^2/\lambda$) and obtained MLE estimates $\hat{\mu} = 16.8$ and $\hat{\beta} = 43$ for a 6 month follow-up period. This translates to $\hat{\mu} = 16.8$ and $\hat{\lambda} = 6.56$ for the parametrization used in (2). For our simulation study, we considered these estimates as the actual parameter values for the placebo group. Four different sample sizes were considered: $n = 10, 20, 50, 100$ with μ ranging from 1 to 20 and $\lambda = 0.5, 1, 25, 10$. The simulated exact 95th percentiles of the LRT

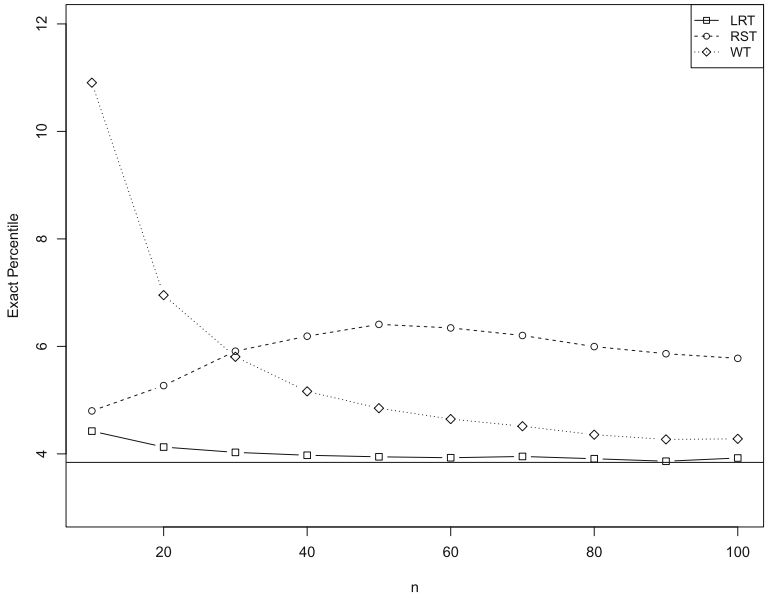


Fig. 1 Simulation based 95th percentile estimate for the null distribution of LRT, RST and WT statistics as a function of common sample size n; $\gamma = 1$, $\mu = 16.8$, $\lambda = 6.56$. The *solid horizontal line* refers to c_0

and the WT statistics are presented in Figs. 2 and 3 respectively. Increasing μ has no effect on the exact percentiles for the LRT. However, when μ increases the exact percentiles for WT move away from c_0. For both the LRT and the WT, increasing λ brings the exact percentiles closer to c_0.

The Type I error rates for the asymptotic LRT and WT are given in Figs. 4 and 5. The error rates for the asymptotic WT seem to be higher than those for the LRT. For the asymptotic LRT, the error rates are very close to the nominal level for per-group sample sizes 50 and above for all values of μ and λ. For the asymptotic WT, these rates are very high for sample sizes 20 or lower per group and for very large sample sizes (n = 50, 100), they can be very high for small λ. Thus, we suggest the use of the exact percentile based WT unless the sample size is 100 or above per group and λ is 10 or higher. Further, μ does not seem to have an effect on the empirical Type I error levels for the asymptotic LRT, but increasing μ seems to slightly increase the levels for the asymptotic WT. Exact percentile estimates and the Type I error rates for RST are not shown as it was an inconsistent test; (see Fig. 7) and the comments in Sect. 5.1.

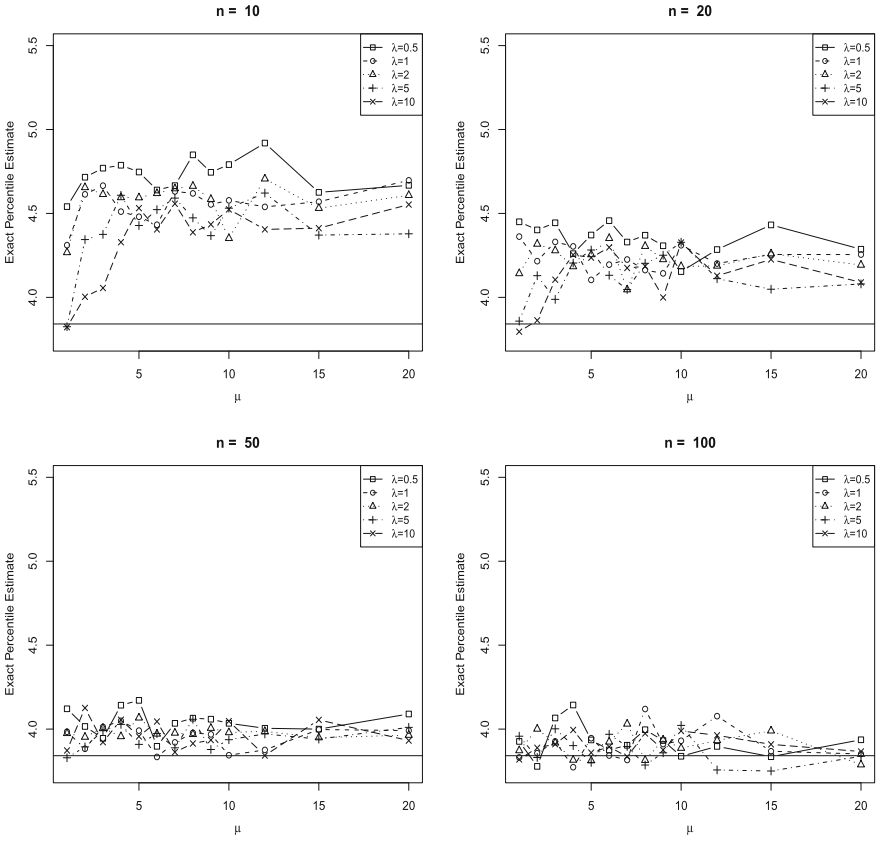


Fig. 2 Simulation based 95th percentile value for the null distribution of LRT statistic for different values of μ and λ and $n = 10, 20, 50, 100$. The *solid horizontal line* refers to c_0

5 Power and Sample Size

We now use selected parametric tests derived in Sect. 2.2 and obtain power and sample size estimates for RRMS PG trials. We compare these sample sizes with the ones obtained using the WRS test. The initial parameter estimates used for the comparative simulation study here are the same as the ones used to obtain Fig. 1.

5.1 Power Analysis

The power for a level ν test is estimated using simulation as follows. Assuming sample sizes n_1 and n_2 for the two groups, exact percentiles for the corresponding test statistic is calculated using the method described in Sect. 4.2. Then for a particular

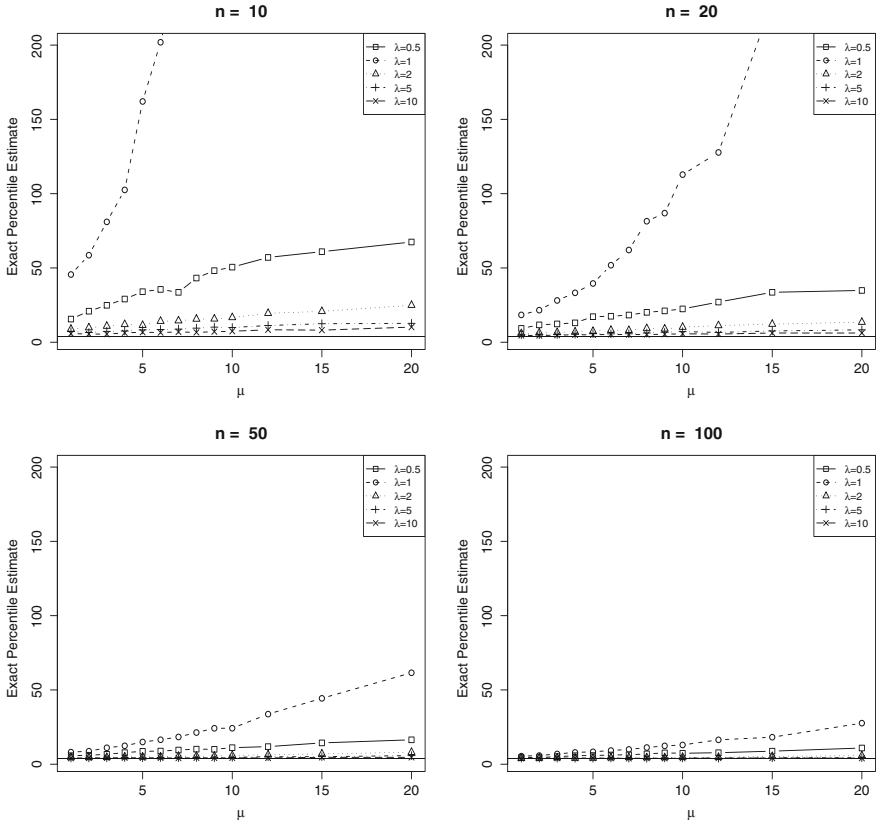


Fig. 3 Simulation based 95th percentile value for the null distribution of WT statistic for different values of μ and λ and $n = 10, 20, 50, 100$. The *solid horizontal line* refers to c_0

trial, two independent sets of observations, $y_{11}, y_{12}, \ldots, y_{1n_1}$ and $y_{21}, y_{22}, \ldots, y_{2n_2}$ are randomly sampled from P-IG(μ, λ) and P-IG($\gamma\mu, \lambda$), respectively. The model parameters are estimated under both Θ_0 and Θ and the test statistic is computed as described in Sect. 4. The null hypothesis is rejected if this test statistic is greater than the simulated percentile. This procedure is repeated 1000 times and the power is estimated as the proportion of trials for which the null hypothesis is rejected. (The R programs used to estimate the power will be provided upon request.)

The power curves as a function of γ for the three asymptotic tests are shown in Fig. 6. The LRT is empirically unbiased and also maintains Type I error rates very well (close to the nominal level 0.05 under $H_0 : \gamma = 1$) for a sample size of 50 per group. The simulated error rate for WT is still slightly higher than the nominal 5 %. Figure 7 shows the power curves for the exact tests. All the tests maintain the Type I error rates very precisely. The exact LRT is an unbiased test and WT, though not unbiased, has the highest power for $\gamma < 1$.

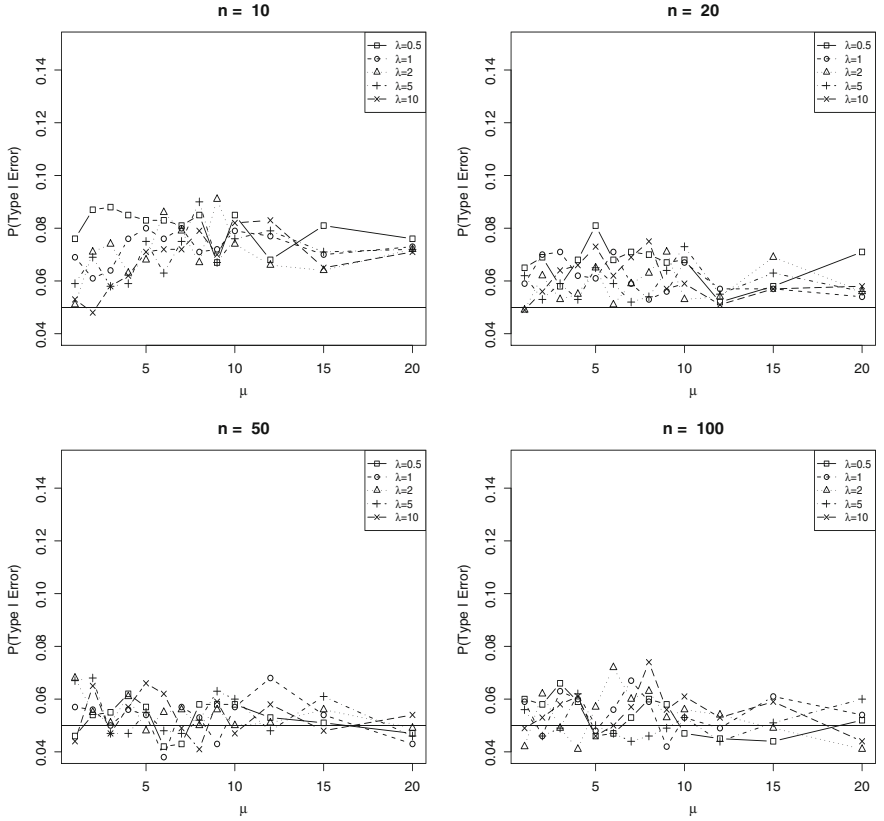


Fig. 4 Type I error rates for LRT with critical value c_0 for different values of μ and λ and $n = 10, 20, 50, 100$ subjects per group. The *solid horizontal line* refers to the nominal level $\nu = 0.05$

In conclusion, if an unbiased test is preferred, the asymptotic LRT can be used for larger sample sizes (>50); otherwise the exact percentile based LRT is recommended. If biased tests are allowed, since $\gamma < 1$ is the region of interest for RRMS clinical trials, the exact WT is recommended. We do not recommend the asymptotic WT even for very large sample sizes.

The RST statistic was computed using the observed information matrix evaluated at the MLEs under the null hypothesis. This sometimes led it to not being positive definite, resulting in a negative RST statistic and an inconsistent test. The power of this RST may also be non-monotonic. Moving away from H_0 does not necessarily result in an increase in power (see Figs. 6 and 7). Thus, the RST is not recommended.

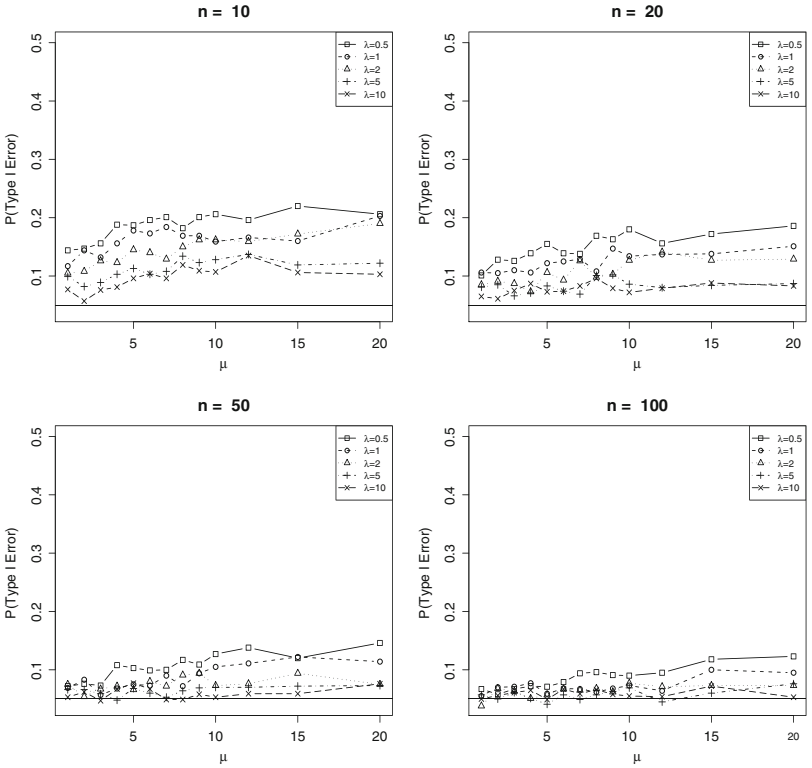


Fig. 5 Type I error rates for WT(γ) with critical value c_0 for different values of μ and α and $n = 10, 20, 50, 100$ subjects per group. The *solid horizontal line* refers to the nominal level $\nu = 0.05$

5.2 *Sample Sizes for Clinical Trials*

In this section, we present sample size estimates based on the LRT and WT for PG trials assuming the P-IG model. We present sample sizes using both the asymptotic and the exact LRT and only for the exact WT. The exact percentile and the power estimates are calculated using the methods described in Sect. 4. For comparison purposes, we also present the sample sizes obtained using the nonparametric WRS test.

Sample sizes (Table 1) for each group are given for 80 and 90 % power, follow-up period of 6 months, and treatment effect $1 - \gamma$ ranging from 0.50 to 0.80. The LRT sample sizes are approximately 30–52 % smaller than the WRS sample sizes. For example, for a 50 % treatment effect and 80 % power, WRS estimates a sample size of 134 per group, whereas the LRT (asymptotic or exact) estimates only 66 per group, a reduction of 51 %. The exact LRT sample sizes are higher than the asymptotic LRT sample sizes by at most 1. This difference is seen only for higher values of treatment effects which yield smaller sample sizes. Sample sizes using the exact WT

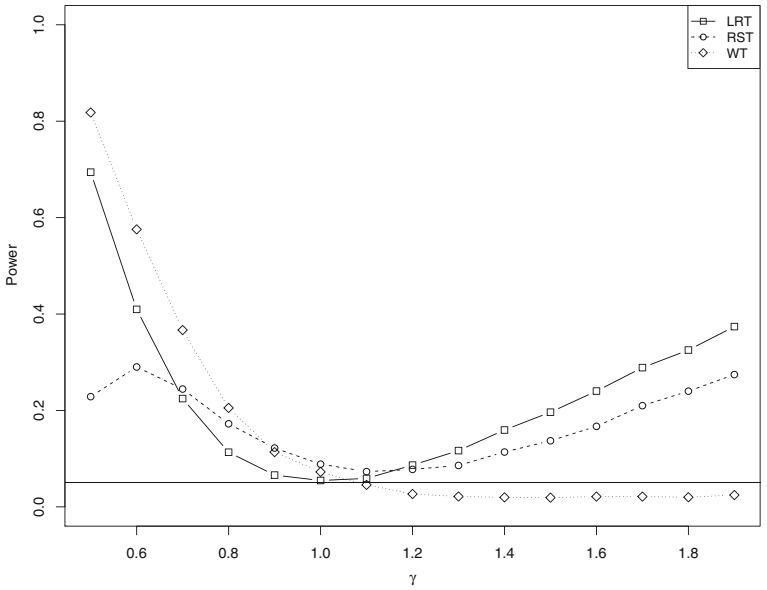


Fig. 6 Power of asymptotic 5 % level LRT, RST, and WT for treatment effect, assuming initial parameter estimates $\mu = 16.8$, $\lambda = 6.56$, sample sizes $n_1 = n_2 = 50$. The *solid horizontal line* refers to the nominal level $\nu = 0.05$

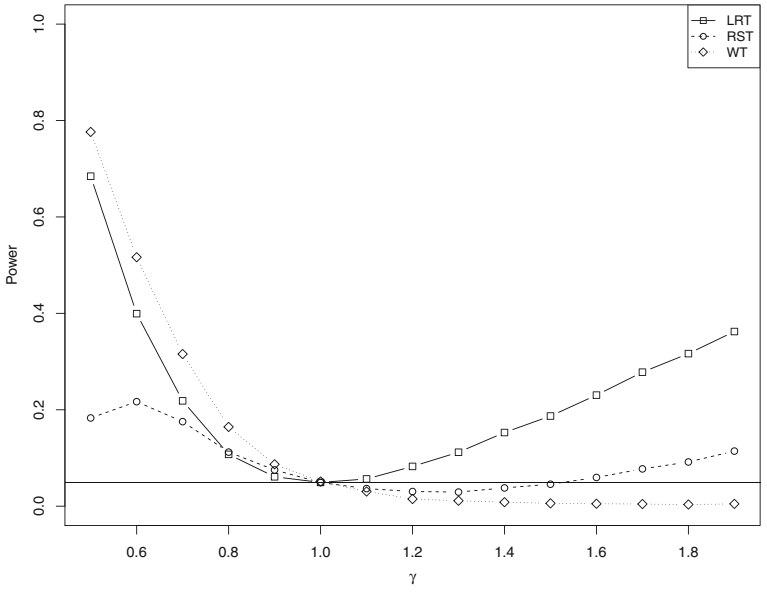


Fig. 7 Power of exact 5 % level LRT, RST, and WT for treatment effect, assuming initial parameter estimates $\mu = 16.8$, $\lambda = 6.56$, sample sizes $n_1 = n_2 = 50$. The *solid horizontal line* refers to the nominal level $\nu = 0.05$

Table 1 Sample sizes per group to achieve 80 and 90 % power for $100(1-\gamma)$ % treatment effect, follow-up period of 6 months, and initial estimates $(\mu, \lambda) = (16.8, 6.56)$; level $\nu = 0.05$

	80 % power; 5 % level					
	Test				Critical value	
		LRT		WT(γ)	LRT	WT(γ)
$1-\gamma$	WRS	Asymptotic	Exact	Exact	Simul.	Simul.
0.50	134	66	66	58	3.8402	3.9483
0.60	169	36	36	30	3.8846	4.0164
0.70	134	19	20	17	4.0961	4.3876
0.80	117	11	12	10	4.2762	4.7157
	90 % power; 5 % level					
	Test				Critical value	
		LRT		WT(γ)	LRT	WT(γ)
$1-\gamma$	WRS	Asymptotic	Exact	Exact	Simul.	Simul.
0.50	179	87	87	78	3.9009	4.0004
0.60	188	48	49	40	3.8530	4.0088
0.70	145	26	26	24	4.0377	4.0382
0.80	123	13	13	12	4.2129	4.4614

are 7–18 % smaller than the sample sizes using LRT and around 40–57 % smaller than the WRS sample sizes. Sample sizes to achieve 90 % power are 20–40 % higher in general than the sample sizes required for 80 % power.

6 Discussion

In this paper we propose likelihood based parametric tests such as LRT, RST, and WT assuming the P-IG model for PG trials. We compare their performance using the asymptotic and exact percentiles and obtain sample size estimates for selected tests. We show the reduction in sample sizes required to detect a significant treatment effect when parametric tests are used as opposed to a nonparametric test.

The recommendations in this paper are based on the properties of the test and the research hypothesis in question. Though not unbiased, the exact WT has the highest power when $\gamma < 1$ and is thus best suited for RRMS clinical trials. When an unbiased test is desired, irrespective of whether one is interested in the region $\gamma < 1$ or $\gamma > 1$, the LRT is preferred. Asymptotic approximation for the LRT can be used for sample sizes over 50 but for smaller sample sizes the exact test needs to be used to ensure the Type I error levels are close to nominal values. When the P-IG model is assumed and true, using the parametric tests proposed in this paper as opposed to nonparametric tests provide a way of significantly reducing sample sizes (by about 30–50 %) and associated costs in RRMS clinical trials. These parametric tests give

similar reduction in sample sizes when compared to the reduction observed by [11] in their study of PG trials based on the NB model.

The parameter estimates used in this study are representative of patients who were followed for a total of six months and the sample size estimates presented here are only representative of clinical trials where the follow-up period is six months. The parameter estimates, especially the mean parameter μ of the P-IG distribution will likely be different for other follow-up periods and this could impact sample sizes. Also, we have assumed that the parameter λ of the P-IG distribution is the same in both the treatment and control groups. It is quite difficult to evaluate this assumption without real data on patients from the treatment group and more research is required to study whether the methods presented in this paper are robust to violations to this assumption.

Further, in estimating the sample size we have assumed that the nuisance parameters (μ, λ) are known beforehand. Quite often in practice information regarding the nuisance parameters is limited and ignoring this uncertainty in computing sample sizes could lead to inappropriately sized clinical trials. One way around this is to consider a blinded sample size reestimation approach to estimate the nuisance parameters from an interim sample, and then use these estimates in simulations to compute the sample size.

Several other areas of application of the P-IG model have been discussed in the literature. Holla [4] first derived the P-IG distribution and discussed its applications to accident statistics. Willmot [18] showed that the model provides an extremely good fit to automobile claim frequency data and also showed that the P-IG model fits better than the NB model in most cases. Ord and Whitmore [7] discuss the P-IG distribution as a model for species abundance. Sankaran [12] illustrates the applicability of the P-IG model to larvae counts on corn bean plants. In all these cases the methods developed in this paper can be used.

Acknowledgments The authors would like to sincerely thank Dr. Marie Davidian and the anonymous referee whose comments have significantly strengthened this manuscript.

Appendix 1

The following lemma presents some useful results needed to obtain the score vector and the matrix of the second order derivatives.

Lemma 1

$$\frac{\partial \tau_1}{\partial \gamma} = 0; \quad \frac{\partial \tau_1}{\partial \mu} = \frac{\tau_1^3}{\mu^3}; \quad \frac{\partial \tau_1}{\partial \lambda} = \frac{\tau_1^3}{\lambda^2};$$
$$\frac{\partial \tau_2}{\partial \gamma} = \frac{\tau_2^3}{\gamma^3 \mu^2}; \quad \frac{\partial \tau_2}{\partial \mu} = \frac{\tau_2^3}{\gamma^2 \mu^3}; \quad \frac{\partial \tau_2}{\partial \lambda} = \frac{\tau_2^3}{\lambda^2}.$$

Using the above results we obtain

$$\frac{\partial \omega_1}{\partial \gamma} = 0; \qquad \frac{\partial \omega_1}{\partial \mu} = -\frac{\lambda \tau_1}{\mu^3}; \quad \frac{\partial \omega_1}{\partial \lambda} = \frac{\lambda - \tau_1^2}{\lambda \tau_1};$$
$$\frac{\partial \omega_2}{\partial \gamma} = -\frac{\lambda \tau_2}{\gamma^3 \mu^2}; \quad \frac{\partial \omega_2}{\partial \mu} = -\frac{\lambda \tau_2}{\gamma^2 \mu^3}; \quad \frac{\partial \omega_2}{\partial \lambda} = \frac{\lambda - \tau_2^2}{\lambda \tau_2}.$$

The following lemma provides results for modified Bessel functions that simplify the derivation of the score vector and the second order derivatives.

Lemma 2 (Modified Bessel function of the third kind (See Sect. 9.6, [2])).
The following relations hold for the modified Bessel function of the third kind $K_\nu(z)$*:*

$$\begin{aligned} K_{-\nu}(z) &= K_\nu(z) \\ K_{-\frac{1}{2}}(z) &= K_{\frac{1}{2}}(z) = \sqrt{\frac{\pi}{2z}} e^{-z} \\ K_{\nu+1}(z) &= K_{\nu-1}(z) + \frac{2\nu}{z} K_\nu(z) \\ \frac{\partial}{\partial z} K_\nu(z) &= K'_\nu(z) = -K_{\nu+1}(z) + \frac{\nu}{z} K_\nu(z). \end{aligned} \tag{15}$$

The ratio of modified Bessel functions $R_\nu(z) = \frac{K_{\nu+1}(z)}{K_\nu(z)}$*, satisfies the following relations:*

$$\begin{aligned} R_{-\frac{1}{2}}(z) &= 1; \\ R_\nu(z) &= \frac{2\nu}{z} + \frac{1}{R_{\nu-1}(z)}, \quad \nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots; \\ \frac{\partial}{\partial z} R_\nu(z) &= R'_\nu(z) = R^2_\nu(z) - \frac{2(\nu + 1/2)}{z} R_\nu(z) - 1. \end{aligned} \tag{16}$$

Appendix 2

First and Second Order Derivatives

The score vector components for the log-likelihood function given in (9) are

$$\frac{\partial \ell(\gamma,\mu,\lambda)}{\partial \gamma} = \frac{\lambda\tau_2}{\gamma^3\mu^2}\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2) - \frac{n_2\lambda}{\gamma^2\mu}, \tag{17}$$

$$\frac{\partial \ell(\gamma,\mu,\lambda)}{\partial \mu} = \frac{\lambda\tau_1}{\mu^3}\sum_{i=1}^{n_1} R_{y_{1i}-\frac{1}{2}}(\omega_1) + \frac{\lambda\tau_2}{\gamma^2\mu^3}\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2) - \frac{\lambda(n_1\gamma+n_2)}{\gamma\mu^2}, \tag{18}$$

$$\frac{\partial \ell(\gamma,\mu,\lambda)}{\partial \lambda} = \frac{n_1\bar{y}_1+n_2\bar{y}_2}{\lambda} + \frac{n_1\gamma+n_2}{\gamma\mu} - \left(\frac{\lambda-\tau_1^2}{\lambda\tau_1}\right)\sum_{i=1}^{n_1} R_{y_{1i}-\frac{1}{2}}(\omega_1) - \left(\frac{\lambda-\tau_2^2}{\lambda\tau_2}\right)\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2). \tag{19}$$

The second order derivatives of the log-likelihood function in (9) are

$$\frac{\partial^2 \ell(\gamma,\mu,\lambda)}{\partial \gamma^2} = -\frac{\lambda^2\tau_2^2}{\gamma^6\mu^4}\sum_{j=1}^{n_2} R'_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{\lambda\tau_2}{\gamma^6\mu^4}(\tau_2^2-3\gamma^2\mu^2)\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{2n_2\lambda}{\gamma^3\mu},$$

$$\frac{\partial^2 \ell(\gamma,\mu,\lambda)}{\partial \gamma\partial \mu} = -\frac{\lambda^2\tau_2^2}{\gamma^5\mu^5}\sum_{j=1}^{n_2} R'_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{\lambda\tau_2^3}{\gamma^5\mu^5}(\tau_2^2-2\gamma^2\mu^2)\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{n_2\lambda}{\gamma^2\mu^2},$$

$$\frac{\partial^2 \ell(\gamma,\mu,\lambda)}{\partial \gamma\partial \lambda} = \left(\frac{\lambda-\tau_2^2}{\gamma^3\mu^2}\right)\sum_{j=1}^{n_2} R'_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{\tau_2}{\gamma^3\mu^2\lambda}(\tau_2^2+\lambda)\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2) - \frac{n_2}{\gamma^2\mu}, \tag{20}$$

$$\frac{\partial^2 \ell(\gamma,\mu,\lambda)}{\partial \mu^2} = -\frac{\lambda^2\tau_1^2}{\mu^6}\sum_{i=1}^{n_1} R'_{y_{1i}-\frac{1}{2}}(\omega_1) + \frac{\lambda\tau_1}{\mu^6}(\tau_1^2-3\mu^2)\sum_{i=1}^{n_1} R_{y_{1i}-\frac{1}{2}}(\omega_1) - \frac{\lambda^2\tau_2^2}{\gamma^4\mu^6}\sum_{j=1}^{n_2} R'_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{\lambda\tau_2}{\gamma^4\mu^6}(\tau_2^2-3\gamma^2\mu^2)\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{2\lambda(n_1\gamma+n_2)}{\gamma\mu^3},$$

$$\frac{\partial^2 \ell(\gamma,\mu,\lambda)}{\partial \mu\partial \lambda} = \left(\frac{\lambda-\tau_1^2}{\mu^3}\right)\sum_{i=1}^{n_1} R'_{y_{1i}-\frac{1}{2}}(\omega_1) + \frac{\tau_1}{\lambda\mu^3}(\tau_1^2+\lambda)\sum_{i=1}^{n_1} R_{y_{1i}-\frac{1}{2}}(\omega_1) + \left(\frac{\lambda-\tau_2^2}{\gamma^2\mu^3}\right)\sum_{j=1}^{n_2} R'_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{\tau_2}{\lambda\gamma^2\mu^3}(\tau_2^2+\lambda)\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2) - \frac{n_1\gamma+n_2}{\gamma\mu^2},$$

$$\frac{\partial^2 \ell(\gamma,\mu,\lambda)}{\partial \lambda^2} = -\frac{n_1\bar{y}_1+n_2\bar{y}_2}{\lambda^2} - \left(\frac{\lambda-\tau_1^2}{\lambda\tau_1}\right)^2\sum_{i=1}^{n_1} R'_{y_{1i}-\frac{1}{2}}(\omega_1) + \frac{\tau_1^3}{\lambda^3}\sum_{i=1}^{n_1} R_{y_{1i}-\frac{1}{2}}(\omega_1) - \left(\frac{\lambda-\tau_2^2}{\lambda\tau_2}\right)^2\sum_{j=1}^{n_2} R'_{y_{2j}-\frac{1}{2}}(\omega_2) + \frac{\tau_2^3}{\lambda^3}\sum_{j=1}^{n_2} R_{y_{2j}-\frac{1}{2}}(\omega_2).$$

Fisher Information Matrix

Since $E(\bar{Y}_1) = \mu$ and $E(\bar{Y}_2) = \gamma\mu$, and $Y_{1i},\ i = 1, \ldots, n_1$ and $Y_{2j},\ j = 1, \ldots, n_2$ are respectively identically distributed, the elements $\left(I(\boldsymbol{\theta})\right)_{i,j} = -E\left\{\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i \partial\boldsymbol{\theta}_j}\right\}$ of the FIM $\mathbf{I}(\boldsymbol{\theta})$ with the parameter vector $\boldsymbol{\theta} = (\gamma, \mu, \lambda)$ can be expressed as follows:

$$
\begin{aligned}
I_{11}(\boldsymbol{\theta}) &= \frac{n_2\lambda^2\tau_2^2}{\gamma^6\mu^4} E\left\{R'_{Y_2-\frac{1}{2}}(\theta_2)\right\} - \frac{n_2\lambda\tau_2}{\gamma^6\mu^4}(\tau_2^2 - 3\gamma^2\mu^2) E\left\{R_{Y_2-\frac{1}{2}}(\theta_2)\right\} - \frac{2n_2\lambda}{\gamma^3\mu},\\
I_{12}(\boldsymbol{\theta}) &= \frac{n_2\lambda^2\tau_2^2}{\gamma^5\mu^5} E\left\{R'_{Y_2-\frac{1}{2}}(\theta_2)\right\} - \frac{n_2\lambda\tau_2^3}{\gamma^5\mu^5}(\tau_2^2 - 2\gamma^2\mu^2) E\left\{R_{Y_2-\frac{1}{2}}(\theta_2)\right\} - \frac{n_2\lambda}{\gamma^2\mu^2},\\
I_{13}(\boldsymbol{\theta}) &= -n_2\left(\frac{\lambda - \tau_2^2}{\gamma^3\mu^2}\right) E\left\{R'_{Y_2-\frac{1}{2}}(\theta_2)\right\} - \frac{n_2\tau_2}{\gamma^3\mu^2\lambda}(\tau_2^2 + \lambda) E\left\{R_{Y_2-\frac{1}{2}}(\theta_2)\right\} + \frac{n_2}{\gamma^2\mu},\\
I_{22}(\boldsymbol{\theta}) &= \frac{n_1\lambda^2\tau_1^2}{\mu^6} E\left\{R'_{Y_1-\frac{1}{2}}(\theta_1)\right\} - \frac{n_1\lambda\tau_1}{\mu^6}(\tau_1^2 - 3\mu^2) E\left\{R_{Y_1-\frac{1}{2}}(\theta_1)\right\} - \frac{2\lambda(n_1\gamma + n_2)}{\gamma\mu^3}\\
&\quad + \frac{n_2\lambda^2\tau_2^2}{\gamma^4\mu^6} E\left\{R'_{Y_2-\frac{1}{2}}(\theta_2)\right\} - \frac{n_2\lambda\tau_2}{\gamma^4\mu^6}(\tau_2^2 - 3\gamma^2\mu^2) E\left\{R_{Y_2-\frac{1}{2}}(\theta_2)\right\},\\
I_{23}(\boldsymbol{\theta}) &= -n_1\left(\frac{\lambda - \tau_1^2}{\mu^3}\right) E\left\{R'_{Y_1-\frac{1}{2}}(\theta_1)\right\} - \frac{n_1\tau_1}{\lambda\mu^3}(\tau_1^2 + \lambda) E\left\{R_{Y_1-\frac{1}{2}}(\theta_1)\right\} + \frac{n_1\gamma + n_2}{\gamma\mu^2}\\
&\quad - n_2\left(\frac{\lambda - \tau_2^2}{\gamma^2\mu^3}\right) E\left\{R'_{Y_2-\frac{1}{2}}(\theta_2)\right\} - \frac{n_2\tau_2}{\lambda\gamma^2\mu^3}(\tau_2^2 + \lambda) E\left\{R_{Y_2-\frac{1}{2}}(\theta_2)\right\},\\
I_{33}(\boldsymbol{\theta}) &= \frac{n_1\mu + n_2\gamma\mu}{\lambda^2} + n_1\left(\frac{\lambda - \tau_1^2}{\lambda\tau_1}\right)^2 E\left\{R'_{Y_1-\frac{1}{2}}(\theta_1)\right\} - \frac{n_1\tau_1^3}{\lambda^3} E\left\{R_{Y_1-\frac{1}{2}}(\theta_1)\right\}\\
&\quad + n_2\left(\frac{\lambda - \tau_2^2}{\lambda\tau_2}\right)^2 E\left\{R'_{Y_2-\frac{1}{2}}(\theta_2)\right\} - \frac{n_2\tau_2^3}{\lambda^3} E\left\{R_{Y_2-\frac{1}{2}}(\theta_2)\right\}.
\end{aligned}
$$

Further, $I_{ij} = I_{ji}$ for $i \neq j = 1, 2, 3$. The above expressions involve evaluating the expectation of $R(.)$, which is a ratio of two modified Bessel functions of the third kind, which cannot be computed in closed form. Instead, the observed information evaluated at the MLEs are used.

References

1. Aban, I., G.R. Cutter, and N. Mavinga. 2009. Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. *Computational Statistics and Data Analysis* 53: 820–833.
2. Abramowitz, M., and I.A. Stegun (eds.). 1970. *Handbook of mathematical functions*. New York: Dover Publications Inc.
3. Freedman, D.A. 2007. How can the score test be inconsistent? *The American Statistician* 61(4): 291–295.
4. Holla, M.S. 1971. Canonical expansion of the compounded correlated bivariate Poisson distribution. *The American Statistician* 23: 32–33.

5. Morgan, B.J.T., K.J. Palmer, and M.S. Ridout. 2007. Score test oddities: Negative score test statistic. *The American Statistician* 61(4): 285–288.
6. Nauta, J.J.P., A.J. Thompson, F. Barkhof, and D.H. Miller. 1994. Magnetic resonance imaging in monitoring the treatment of multiple sclerosis patients: Statistical power of parallel-groups and crossover designs. *Journal of Neurological Sciences* 122: 6–14.
7. Ord, J.K., and K.A. Whitmore. 1986. The Poisson-inverse Gaussian distribution as a model for species abundance. *Communications in Statistics - Theory and Methods* 15(3): 853–871.
8. R Development Core Team 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
9. Rao, C.R. 1948. Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation. *Proceedings of the Cambridge Philosophical Society* 44: 50–57.
10. Rao, C.R. 2005. *Advances in ranking and selection, multiple comparisons, and reliability*, Chapter Score Test: Historical Review and Recent Developments, pp. 3–20. Boston: Birkhäuser
11. Rettiganti, M.R., and H.N. Nagaraja. 2012. Power analysis for negative binomial models with application to multiple sclerosis clinical trials. *Journal of Biopharmaceutical Statistics* 22(2): 237–259.
12. Sankaran, M. 1968. Mixtures by the inverse Gaussian distribution. *Sankhya B* 30: 455–458.
13. Sormani, M.P., P. Bruzzi, D.H. Miller, C. Gasperini, F. Barkhof, and M. Filippi. 1999. Modelling MRI enhancing lesion counts in multiple sclerosis using a negative binomial model: Implications for clinical trials. *Journal of the Neurological Sciences* 163: 74–80.
14. Sormani, M.P., P. Bruzzi, M. Rovaris, F. Barkhof, G. Comi, D.H. Miller, G.R. Cutter, and M. Filippi. 2001a. Modelling new enhancing MRI lesion counts in multiple sclerosis. *Multiple Sclerosis* 7: 298–304.
15. Sormani, M.P., D.H. Miller, G. Comi, F. Barkhof, M. Rovaris, P. Bruzzi, and M. Filippi. 2001b. Clinical trials of multiple sclerosis monitored with enhanced MRI: New sample size calculations based on large data sets. *Journal of Neurology Neurosurgery and Psychiatry* 70: 494–499.
16. Stein, G.Z., W. Zucchini, and J.M. Juritz. 1987. Parameter estimation for the Sichel distribution and its multivariate distribution. *Journal of the American Statistical Association* 82(399): 938–944.
17. Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 54: 426–482.
18. Willmot, G. 1987. The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal* 87: 113–127.

Purely Sequential and Two-Stage Bounded-Length Confidence Intervals for the Bernoulli Parameter with Illustrations from Health Studies and Ecology

Nitis Mukhopadhyay and Swarnali Banerjee

Abstract Infestation affects supplies of food and nutrition as well as the environment, thus making a deep impact in the ecological balance of the health of humans, animals, plant populations, and other natural resources. It is well known, for example, that estimation of (i) the probability of presence of infestation, (ii) the chance of getting a disease, and (iii) the chance of a relapse are very important in entomology and health studies. They frequently involve binary data modelled by a Bernoulli(p) distribution where p is an unknown parameter, $0 < p < 1$. In this paper, we begin by summarizing selected existing methodologies of confidence interval estimation and illustrate how they may fail to estimate p efficiently. Consequently, we introduce new confidence interval methods for estimating p. Having fixed $0 < \alpha < 1$ and $d(> 1)$, we develop approximately $100(1 - \alpha)$ % confidence intervals (L_N, U_N) for p such that $0 < L_N < U_N < 1$ and $U_N - L_N \leq d$ w.p.1. Here, N is a properly designed and determined stopping variable obtained via both two-stage and purely sequential sampling strategies. The proposed two-stage and purely sequential bounded-length confidence interval methodologies are shown to enjoy both asymptotic first-order efficiency and asymptotic consistency properties. Then, we present summary performances of the new methodologies by analyzing data generated from simulations. We have also implemented the proposed methodologies for three real data sets of size small to moderate to large.

Keywords Environmental statistics · First-order properties · Infestation · Multi-stage sampling · Second-order properties · Statistical ecology

N. Mukhopadhyay (✉)
Department of Statistics, University of Connecticut, Storrs, CT, USA
e-mail: nitis.mukhopadhyay@uconn.edu

S. Banerjee
Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA, USA
e-mail: sbanerje@odu.edu

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_13

1 Introduction

Let $X_1, \ldots, X_n, \ldots$ be a sequence of *independent and identically distributed* (i.i.d.) Bernoulli(p) random variables, where p is an unknown parameter, $0 < p < 1$. Of interest is purely sequential or two-stage confidence interval estimation of p.

In general, for i.i.d. random variables with an unknown mean μ, $-\infty < \mu < \infty$, and unknown variance, Chow and Robbins [5] developed purely sequential estimation methods for μ by a fixed-width confidence interval centered at the sample mean. That methodology was distribution-free. Khan [9] developed analogous fixed-width confidence intervals for a parameter centered at its *maximum likelihood estimator* (MLE). To be specific, let $T_n \equiv T_n(X_1, \ldots, X_n)$ be a point estimator of p such that $T_n \xrightarrow{P} p$ as $n \to \infty$, that is, T_n is consistent for p.

One sets a preassigned width ($= 2c$) of the confidence interval with nearly $1 - \alpha$ confidence coefficient where $0 < \alpha < 1$ is also fixed in advance. Chow and Robbins [5] began with $T_n = \overline{X}_n$, the sample mean, and considered the associated confidence interval G_n:

$$G_n = \{p : |T_n - p| \le c\}, \text{ fixed } c > 0. \tag{1.1}$$

Khan's [9] MLE-based fixed-width confidence interval was identical with (1.1). Obviously $\widehat{p}_{n,\mathrm{MLE}}$, the MLE of p, is $\overline{X}_n$.

Alternatively, Ehrenfeld and Littauer [7, p. 339], and Zacks [16] had incorporated the *proportional closeness* criterion. Zacks [16] began with $T_n = \overline{X}_n$ and proposed the associated confidence interval H_n:

$$H_n = \{p : |T_n - p| \le \delta p\}, \text{ fixed } 0 < \delta < 1, \tag{1.2}$$

where $0 < \delta < 1$, a measure of proportional closeness, is fixed in advance. Nadas [12] adapted this proportional closeness criterion in the context of the population mean estimation problem of Chow and Robbins [5]. Willson and Folks [15] also adopted this criterion for estimating the mean in a negative binomial population with applications in ecology.

Recently, Mukhopadhyay and Banerjee [10] introduced *fixed-accuracy* confidence interval methodologies to estimate the mean parameter in a negative binomial population. This methodology was later generalized for estimating an unknown positive parameter in Banerjee and Mukhopadhyay [1]. With $T_n = \overline{X}_n$, a corresponding fixed-accuracy confidence interval for p would look like I_n:

$$I_n = \{p : p \in [d^{-1}T_n, dT_n]\}, \text{ fixed } d > 1. \tag{1.3}$$

In the context of (1.1)–(1.3), however, we suggest replacing $T_n = \overline{X}_n$ with

$$T_n = \overline{X}_n + n^{-\gamma}, \gamma > \tfrac{1}{2}, \tag{1.4}$$

since $\overline{X}_n$ can be zero with a positive probability whatever n may be. The term $n^{-\gamma}$ with $\gamma > \frac{1}{2}$ would ensure that $\overline{X}_n + n^{-\gamma}$ remains consistent for p, but $\overline{X}_n + n^{-\gamma}$ will satisfy the customary *central limit theorem* (CLT).

For a more complete review, one may additionally refer to Robbins and Siegmund [13], Cho [3, 4], Zacks and Mukhopadhyay [17], and also consider the literature cited in those references. Again, going back to (1.1)–(1.4), an associated confidence interval for p will simplify to:

$$(L_n, U_n)$$

where L_n (U_n) is the lower (upper) confidence limit.

For the three types of confidence intervals summarized in (1.1)–(1.3), the corresponding L_n, U_n and the expressions of the required but unknown optimal fixed-sample-sizes are summarized as follows:

(**a**) Interval $G_n : L_n = T_n - c$ and $U_n = T_n + c$ with an optimal fixed-sample -size : $n_c^0 = z_{\alpha/2}^2 p(1-p)/c^2$, fixed $c > 0$;

(**b**) Interval $H_n : L_n = (1+\delta)^{-1} T_n$ and $U_n = (1-\delta)^{-1} T_n$ with an optimal fixed-sample-size : $n_\delta^0 = z_{\alpha/2}^2 (p^{-1} - 1)/\delta^2$, fixed $0 < \delta < 1$;

(**c**) Interval $I_n : L_n = d^{-1} T_n$ and $U_n = d T_n$ with an optimal fixed-sample -size : $n_d^0 = z_{\alpha/2}^2 (p^{-1} - 1)/(\ln d)^2$, fixed $d > 1$,

(1.5)

where $z_{\alpha/2}$ is the upper $100(\alpha/2)\%$ point of a standard normal distribution.

Consequently, the corresponding purely sequential stopping rules respectively estimating n_c^0, n_δ^0, n_d^0 from (1.5) and the confidence interval estimation methodologies from the existing literature look like:

(**a**) $N \equiv N_c = \inf \left\{ n \geq n_0 : n \geq z_{\alpha/2}^2 \left(\widehat{p}_{n,\mathrm{MLE}}(1 - \widehat{p}_{n,\mathrm{MLE}}) + n^{-1} \right) / c^2 \right\}$:
$G_{N_c} = \{p : |T_{N_c} - p| \leq c\}$ with $L_{N_c} = T_{N_c} - c, U_{N_c} = T_{N_c} + c$;

(**b**) $N \equiv N_\delta = \inf \left\{ n \geq n_0 : n \widehat{p}_{n,\mathrm{MLE}}(1 - \widehat{p}_{n,\mathrm{MLE}})^{-1} \geq z_{\alpha/2}^2 / \delta^2 \right\}$:
$H_{N_\delta} = \{p : |T_{N_\delta} - p| \leq \delta p\}$ with $L_{N_\delta} = (1+\delta)^{-1} T_{N_\delta}, U_{N_\delta} = (1-\delta)^{-1} T_{N_\delta}$;

(**c**) $N \equiv N_d = \inf \left\{ n \geq n_0 : n \widehat{p}_{n,\,\mathrm{MLE}}(1 - \widehat{p}_{n,\mathrm{MLE}})^{-1} \geq z_{\alpha/2}^2 / (\ln d)^2 \right\}$:
$I_{N_d} = [d^{-1} T_{N_d}, d T_{N_d}]$ with $L_{N_d} = d^{-1} T_{N_d}, U_{N_d} = d T_{N_d}$.

(1.6)

The Paper's Layout and a Recommendation

Now, we explain this paper's layout. Section 2 highlights some drawbacks of selected existing methodologies laid down in (1.5)–(1.6) with the help of data analysis. We have used both simulated data and a classical set of real data on potato beetle infestation for this purpose.

In order to get rid of the said drawbacks, we proceed to introduce a new bounded-length purely sequential confidence interval estimation methodology for p in Sect. 3.

Next, we develop a bounded-length two-stage confidence interval estimation methodology in Sect. 4. In Sects. 3 and 4, we prove desirable theoretical results such as asymptotic first-order efficiency and asymptotic consistency properties associated with the respective methodologies.

In Sect. 5, we conduct extensive data analysis. First, we summarize findings from simulated exercises (Sect. 5.1). Section 5.2 includes three illustrations with real data analysis. We have implemented the proposed purely sequential and two-stage methodologies by handling three real data sets of size small to moderate to large covering interesting areas of applications: Estimating the chance of (i) relapse in bone marrow transplant patients, (ii) presence or absence of diabetes for Pima Indians, and (iii) presence or absence of potato beetle infestation from entomology.

2 Data Analysis to Point Out Some Drawbacks of Selected Existing Methodologies

Here, we highlight some important drawbacks of the purely sequential confidence interval estimation methodologies:

$$(N_c, G_{N_c}), (N_\delta, H_{N_\delta}) \text{ and } (N_d, I_{N_d})$$

for p as laid out in (1.6) parts (a), (b), and (c) respectively. First, we take resort to simulated data. Then, we validate our concerns with the help of real data from potato beetle infestation.

2.1 Drawbacks: Validation with Simulated Data

A close look at (1.6) reveals that it is possible that L_N may be negative and/or U_N may exceed 1 with positive probability. We provide simulation results to emphasize this point. Table 1 shows simulation results corresponding to the purely sequential procedure (1.6) part (a). Data are simulated from a Bernoulli distribution with $p = 0.90$ and each row shown corresponds to 10000 replications.

We considered $c = 0.10, 0.05$ in (1.5) part (a) and fixed $\gamma = 0.7$ in the definition of T_N from (1.4). Recall that any γ value exceeding $\frac{1}{2}$ will suffice. We found comparable performances with a wide variety of other choices of γ. For brevity, we include results from data analysis when $\gamma = 0.7$.

The second column n_c^0 shows the optimal fixed sample size as in (1.5) part (a). Then, we have n_0 that denotes the pilot sample size. This is assumed to be fixed and in this section we have assumed $n_0 = 10$ for all our results. We explored other choices of n_0 but found no difference in conclusions. Column 4 shows $\overline{n}$, the average from 10000 runs of sequential sampling ((1.6) part (a)) with its standard error $s(\overline{n})$.

Table 1 Simulation results from 10000 replications for the purely sequential methodology (1.6) part (a) in a Bernoulli($p = 0.90$) population when $\alpha = 0.05$ and $\gamma = 0.7$ in (1.4)

δ	n_c^0(1.5) part (b)	n_0	$\overline{n}(s(\overline{n}))$	$\overline{n}/n_d^*$	$\overline{n} - n_d^*$	$\overline{w}(s(\overline{w}))$
0.10	34.574	10	40.776 (0.334)	1.179	6.202	0.524 (0.005)
0.05	138.298	10	140.904 (0.334)	1.019	2.606	0.695 (0.005)

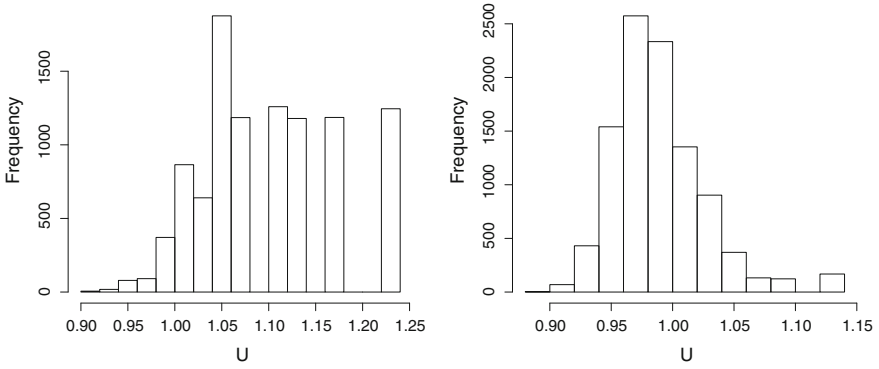


Fig. 1 Plots on the *left* and the *right* correspond respectively to the rows $c = 0.10$ and $c = 0.05$ in Table 1. The histograms plot values of upper confidence limit U_{N_c} using (1.6) part (a) under simulations with 10000 replications from Bernoulli($p = 0.9$) corresponding to $\gamma = 0.7, n_0 = 10$ and $\alpha = 0.05$

The next two columns 5–6 show the ratio and difference of the sample sizes $\overline{n}$ and n_c^0, which give an idea of the efficiency (first- and second-order, respectively) of the sampling rule. The ratio $\overline{n}/n_c^0$ is expected to be near 1 and this is a measure of the first-order efficiency. One may refer to Theorem 3.1 part (ii).

Let w denote an indicator variable taking the value 1 (or 0) if a constructed confidence interval (L_N, U_N) obtained for a run includes (or does not include) the true p. Then, $\overline{w}$ is the average w from 10000 replications with its standard errors $s(\overline{w})$ shown in column 7. This $\overline{w}$ gives us an idea about the achieved coverage probability. Since we fixed $\alpha = 0.05$, we expect $\overline{w}$ to be close to 0.95. Chow and Robbins [5] proved that the fixed-width confidence interval for the population mean parameter μ is asymptotically first-order efficient and asymptotically consistent.

In each of the two cases summarized in Table 1, we show Fig. 1 which plots all 10000 observed values of the upper confidence limit U_N. The histograms in Fig. 1 clearly show that $P(U_N > 1) > 0$. By fixing a small value of the Bernoulli parameter p (with a small value of c), we may similarly validate the possibility that L_N may be negative with a positive probability. We have omitted these results for brevity.

Similarly, Table 2 shows simulation results for the sequential procedure (1.6) part (b). Data are simulated from a Bernoulli distribution with $p = 0.90$ and each row corresponds to 10000 replications under $\delta = 0.10, 0.05$. The second column shows

Table 2 Simulation results from 10000 replications for the purely sequential methodology (1.6) part (b) in a Bernoulli($p = 0.90$) population when $\alpha = 0.05$ and $\gamma = 0.7$ in (1.4)

δ	n_δ^0(1.5) part (b)	n_0	$\overline{n}(s(\overline{n}))$	$\overline{n}/n_d^*$	$\overline{n} - n_d^*$	$\overline{w}(s(\overline{w}))$
0.10	42.684	10	32.964 (0.206)	0.772	−9.720	0.372 (0.005)
0.05	170.738	10	116.844 (0.818)	0.684	−53.894	0.474 (0.005)

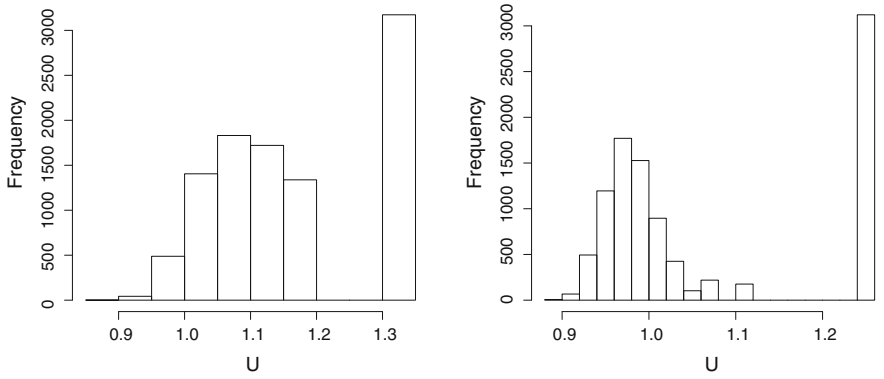


Fig. 2 Plots on the *left* and the *right* correspond respectively to the rows $\delta = 0.10$ and $\delta = 0.05$ in Table 2. The histograms plot values of upper confidence limit U_{N_δ} using (1.6) part (b) under simulations with 10000 replications from Bernoulli($p = 0.9$) corresponding to $\gamma = 0.7$, $n_0 = 10$ and $\alpha = 0.05$

Table 3 Simulation results from 10000 replications for the purely sequential methodology (1.6) part (c) in a Bernoulli($p = 0.90$) population when $\alpha = 0.05$ and $\gamma = 0.7$ in (1.4)

d	n_δ^0(1.5) part (b)	n_0	$\overline{n}(s(\overline{n}))$	$\overline{n}/n_d^*$	$\overline{n} - n_d^*$	$\overline{w}(s(\overline{w}))$
1:10	46:988	10	35:913 (0:228)	0:765	11:075	0:407 (0:005)
1:05	179:310	10	121:056 (0:855)	0:685	–58:254	0:485 (0:005)

n_δ^0 from (1.5) part (b). Figure 2 plots all 10000 values of the upper confidence limit U_N. The histograms from Fig. 2 clearly show that $P(U_N > 1) > 0$.

Table 3 summarizes simulation results for the sequential procedure (1.6) part (c). Figure 3 plots all 10000 values of the upper confidence limit U_N. The histograms from Fig. 3 clearly reiterate that $P(U_N > 1) > 0$.

Tables 1, 2 and 3 and Figs. 1, 2 and 3 show that the existing purely sequential confidence interval methods described in (1.6) do not perform satisfactorily.

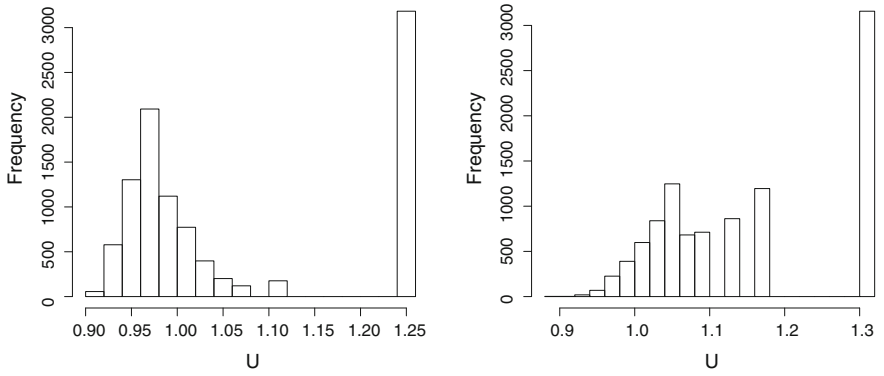


Fig. 3 Plots on the *left* and the *right* correspond respectively to the rows $d = 1.10$ and $d = 1.05$ in Table 3. The histograms plot values of upper confidence limit U_{N_d} using (1.6) part (c) under simulations with 10000 replications from Bernoulli($p = 0.9$) corresponding to $\gamma = 0.7$, $n_0 = 10$ and $\alpha = 0.05$

2.2 Drawbacks: Validation with Potato Beetle Data

Potato beetles were counted in a field near Ontario Beall [2]. Data were recorded from 16 strata each of size 144. Inside each strata, each of the 144 observations were sampling units for which a count of potato beetles was provided. Although the data were collected in strata, we combined all strata and treated such combined data as a single dataset of size 2304. Next we re-coded this count data as 1 for a sample unit if infestation was present (that is, the insect count was non-zero) and 0 otherwise.

This binary variable indicated the presence or absence of infestation. The corresponding parameter p quantifies a measure of the probability of infestation. To get an idea about p, we found $\widehat{p}_{n,\text{MLE}} = 0.918$. The full binary data so coded agreed extremely well with a Bernoulli distribution with $p = 0.918$. The size of the dataset, 2304, appears large and hence for the purpose of illustration, we pretend 0.918 to be the "true" value of p. The p-value for the chi square goodness of fit test is 1 (chi square statistic $= 1.503709 \times 10^{-11}$). A q-q plot shows a nearly perfect fit.

Tables 4, 5 and 6 summarize the performance of the purely sequential confidence interval procedures (N_c, G_{N_c}), (N_δ, H_{N_δ}) and (N_d, I_{N_d}) from (1.6). In the context of each methodology, with a few choices of c, δ and d respectively, we ran the

Table 4 Potato beetle data illustration with MLE $\widehat{p}_{2304,\text{MLE}} = 0.918$ treated as "true" p under a single run of the purely sequential rule (1.6) part (a) for each row: $\alpha = 0.05$, $\gamma = 0.7$, and (L_N, U_N) from (1.5) part (a)

c	$\widehat{n}_c^0$	n_0	N	$N/\widehat{n}_c^0$	$N - \widehat{n}_c^0$	$[L_N, U_N]$
0.10	29.067	10	39	1.342	9.933	(0.900, 1.100)
0.05	116.269	10	126	1.084	9.731	(0.904, 1.004)
0.02	726.684	10	600	0.826	−126.68	(0.926, 0.966)

Table 5 Potato beetle data illustration with MLE $\widehat{p}_{2304,\mathrm{MLE}} = 0.918$ treated as "true" p under a single run of the purely sequential rule (1.6) part (b) for each row: $\alpha = 0.05$, $\gamma = 0.7$, and (L_N, U_N) from (1.5) part (b)

δ	$\widehat{n}_\delta^0$	n_0	N	$N/\widehat{n}_d^0$	$N - \widehat{n}_\delta^0$	$[L_N, U_N]$
0.10	29.067	10	29	0.998	−0.067	(0.932, 1.140)
0.05	116.269	10	108	0.929	−8.269	(0.926, 1.024)
0.02	726.684	10	894	1.230	167.32	(0.905, 0.942)

Table 6 Potato beetle data illustration with MLE $\widehat{p}_{2304,\mathrm{MLE}} = 0.918$ treated as "true" p under a single run of the purely sequential rule (1.6) part (c) for each row: $\alpha = 0.05$, $\gamma = 0.7$, and (L_N, U_N) from (1.5) part (c)

d	$\widehat{n}_d^0$	n_0	N	$N/\widehat{n}_d^0$	$N - \widehat{n}_d^0$	$[L_N, U_N]$
1.10	38.009	10	31	0.816	−7.009	(0.932, 1.128)
1.05	145.043	10	102	0.703	−43.043	(0.934, 1.029)
1.02	880.471	10	726	0.824	−154.47	(0.923, 0.960)

sequential procedures (1.6) parts (a), (b), and (c) to come up with the corresponding confidence intervals for p. The second, third and fourth columns in these tables provide the estimated optimal sample size required (1.5), pilot sample size, and the purely sequential sample size upon termination. The next two columns measure asymptotic first and second-order efficiencies which are mostly unsatisfactory.

However, in the case of Tables 4, 5 and 6, we note that each row corresponds to a single run. The last row in Table 4 ($c = 0.02$), first two rows in Table 5 ($\delta = 0.05, 0.02$), and all rows in Table 6 ($d = 1.10, 1.05, 1.02$) fail to include the most plausible value of p, namely 0.918. Also by looking at the final confidence intervals in each case, it is apparent that the upper confidence limit has exceeded 1 a number of times.

2.3 A Naive and Not-So-Promising Resolution

In the case of each interval constructed from (1.6), one may feel tempted to propose a fine-tuning of the respective confidence interval as follows:

$$\left(L_N^*, U_N^*\right), \tag{2.1}$$

where $L_N^* = \max(0, L_N)$ and $U_N^* = \min(1, U_N)$ with $N (= N_c$ or N_δ or $N_d)$. However, such a modified confidence interval $\left(L_N^*, U_N^*\right)$ may not enjoy the desirable asymptotic consistency property.

Indeed, one should redefine the optimal fixed-sample-size n^* as follows:

$$\min n(\geq 1) \text{ such that } P_p\left\{p \in \left(L_n^*, U_n^*\right)\right\} \geq 1-\alpha \text{ holds approximately,} \quad (2.2)$$

assuming n large. Then, one may mimic such a revised expression of n^* from (2.2) in order to formulate an appropriate stopping time N in the context of a specific notion of desired accuracy (for example, fixed-width or fixed proportional closeness or fixed-accuracy). That way, one may genuinely expect to claim:

$$P_p\left\{p \in \left(L_N^*, U_N^*\right)\right\} \to 1-\alpha \text{ asymptotically,} \quad (2.3)$$

which will be the asymptotic consistency property. Corresponding analytical steps and the formulation of an ensuing purely sequential methodology may become very invasive.

2.4 A Synopsis

In summary, none of the existing purely sequential confidence interval estimation procedures from (1.6) performed satisfactorily for the Bernoulli parameter p. A possible naive approach pointed out briefly in Sect. 2.3 does not appear very promising. Ideally, we would like to come up with a new purely sequential confidence interval methodology (Q, J_Q), defined via (3.5) and (3.7), or a new two-stage confidence interval methodology (R, J_R), defined via (4.2)–(4.4), for estimating p directly in such a way that J_Q, J_R are surely sub-intervals of $(0, 1)$. We additionally demand that any such newly proposed confidence interval estimation methodology (Q, J_Q) or (R, J_R) should satisfy both asymptotic first-order efficiency and asymptotic consistency properties.

3 First New Fix: Purely Sequential Bounded-Length Confidence Interval Methodology

In this section, we propose a new way of estimating the Bernoulli parameter p such that we may ensure that the confidence bounds satisfy the requirement $0 < L_N < U_N < 1$ w.p. 1 while preserving the first-order asymptotic efficiency and asymptotic consistency properties.

We begin with the odds-ratio $\theta \equiv \theta(p) = p(1-p)^{-1}$ which is a one-to-one function of p. Clearly, the parameter θ is unknown with its parameter space $R^+ \equiv (0, \infty)$. Now, we revisit the fixed-accuracy confidence interval estimation problem for θ along the lines of Banerjee and Mukhopadhyay [1].

Based on $X_1, \ldots, X_n$, we recall the MLE, $\widehat{p}_{n,\text{MLE}} = \overline{X}_n$, for p. Using the invariance property of MLE Zehna [18], we let:

$$W_n \equiv \widehat{\theta}_{n,\text{MLE}} = \frac{\overline{X}_n}{1 - \overline{X}_n}. \tag{3.1}$$

Since $P_p(\overline{X}_n = 0) > 0$ and $P_p(\overline{X}_n = 1) > 0$ whatever $0 < p < 1$ may be, we define:

$$T_n = \frac{\overline{X}_n + n^{-\gamma}}{1 - \overline{X}_n + n^{-\gamma}}, \tag{3.2}$$

where $\gamma > \frac{1}{2}$ is a constant of our choice. Much in the spirit of (1.4), this T_n is a consistent estimator of θ.

Next, with a preassigned level of accuracy $d(> 1)$, Banerjee and Mukhopadhyay [1] constructed the following fixed-accuracy confidence interval for θ:

$$K_n = \{\theta : \theta \in [d^{-1}T_n, dT_n]\}. \tag{3.3}$$

Recall (1.3) which was considered in the same spirit. However, there is a significant difference. In (1.3), the parameter space was $(0, 1)$, but the parameter space in (3.3) is $(0, \infty)$.

Using CLT, we obviously have

$$n^{1/2}(T_n - \theta) \xrightarrow{\mathcal{L}} N(0, \sigma^2(\theta)) \text{ as } n \to \infty,$$

where the variance in the asymptotic distribution is given by $\sigma^2(\theta) \equiv \theta(\theta + 1)^2$. Thus, for K_n to include θ with an approximate preassigned probability $1 - \alpha, 0 < \alpha < 1$, the required optimal fixed-sample-size will reduce to:

$$\text{the smallest } n \geq n_d^* \equiv \left(\frac{z_{\alpha/2}}{\ln d}\right)^2 \theta^{-1}(\theta + 1)^2. \tag{3.4}$$

But, n_d^* is a function of θ and hence remains unknown.

Now, observe that $n_d^* \geq 4\left(\frac{z_{\alpha/2}}{\ln d}\right)^2$. Thus, we may define the pilot sample size as $\approx 4\left(\frac{z_{\alpha/2}}{\ln d}\right)^2$. So, we let $X_1, \ldots, X_{n_0}$ be our pilot data which are followed by a one at-a-time drawing of additional observations according to the stopping time:

$$Q \equiv Q_d = \inf\left\{n \geq n_0 : nW_n(W_n + 1)^{-2} \geq \left(\frac{z_{\alpha/2}}{\ln d}\right)^2\right\}, \tag{3.5}$$

$$\text{with } n_0 \equiv n_{0d} = \left\lfloor 4\left(\frac{z_{\alpha/2}}{\ln d}\right)^2 \right\rfloor,$$

where W_n comes from (3.1) and $\lfloor u \rfloor$ = the largest integer $< u$ for $u > 0$. Banerjee and Mukhopadhyay [1] arrived at the same stopping rule with n_0 fixed, that is, with n_0 not involving d.

One may show that $P_p\{Q_d < \infty\} = 1$ and $E_p[Q_d] < \infty$ by appealing to Chow and Robbins [5]. Else, Theorem 1 from Banerjee and Mukhopadhyay [1] could be applied.

Notice that as long as W_n is observed as zero, that is as long as $\overline{X}_n = 0$, (3.5) will not stop sampling. On the other hand, we interpret $W_n = \infty$ whenever $\overline{X}_n = 1$, but then the corresponding value of $W_n(W_n + 1)^{-2}$ used in the left-hand side of the inequality in (3.5) is interpreted as $\lim_{\psi \to \infty} \psi(1 + \psi)^{-2}$. However, this limit reduces to $\lim_{\psi \to \infty} \frac{1}{2}(1 + \psi)^{-1} = 0$, by L'Hôspital's rule. Thus, as long as $\overline{X}_n = 1$ is observed, (3.5) will not stop sampling. This argument may sound more convincing if we replaced the expression $W_n(W_n + 1)^{-2}$ used in the left-hand side of the inequality in (3.5) by an equivalent expression $\overline{X}_n(1 - \overline{X}_n)$.

At termination, we note that $0 < W_{Q_d} < \infty$ w.p.1. After implementing the purely sequential procedure (3.5), the final dataset at hand will be $\{Q_d, X_1, \ldots, X_{Q_d}\}$. Using this final data, Banerjee and Mukhopadhyay [1] proposed the confidence interval

$$K_{Q_d} \equiv [d^{-1}T_{Q_d}, dT_{Q_d}]$$

to estimate θ in the light of (3.3).

But now, let us define:

$$L_{Q_d} = (d + T_{Q_d})^{-1}T_{Q_d} \text{ and } U_{Q_d} = (1 + dT_{Q_d})^{-1}dT_{Q_d}. \tag{3.6}$$

Then, the associated coverage probability may be expressed as follows:

$$P_p\left\{\theta(p) \in K_{Q_d}\right\} = P_p\left\{d^{-1}T_{Q_d} \le \theta \le dT_{Q_d}\right\} = P_p\left\{L_{N_d} \le p \le U_{N_d}\right\},$$

which leads us to propose the following bounded-length purely sequential confidence interval for p:

$$J_{Q_d} \equiv [L_{Q_d}, U_{Q_d}]. \tag{3.7}$$

3.1 Properties of the Purely Sequential Confidence Interval (3.7)

Clearly, both lower and upper confidence limits L_{Q_d}, U_{Q_d} lie between 0 and 1 w.p.1. That is, the earlier criticisms labeled against G_N, H_N, and I_N from (1.3) no longer hold in the case of our newly proposed J_{Q_d}. Let us summarize a number of desirable theoretical properties that are associated with the methodology (Q_d, J_{Q_d}).

Theorem 3.1 *For the purely sequential estimation rule* $\left(Q_d, J_{Q_d}\right)$ *under (3.5)–(3.7), for each fixed* $0 < p < 1$ *and* $0 < \alpha < 1$, *we have as* $d \downarrow 1$*:*

(i) $Q_d/n_d^* \to 1$ *w.p.1;*
(ii) $E_p\left[Q_d/n_d^*\right] \to 1$ *[Asymptotic first-order efficiency]; and*
(iii) $P_p\left\{p \in J_{Q_d} : [L_{Q_d}, U_{Q_d}]\right\} \to 1 - \alpha$ *[Asymptotic consistency],*

where n_d^* *comes from (3.4) and* L_{Q_d}, U_{Q_d} *come from (3.6).*

A proof of this result follows directly from the proof of Theorem 1 in Banerjee and Mukhopadhyay [1]. We omit further details for brevity.

3.2 Bounded-Length for the Purely Sequential Confidence Interval (3.7)

The length of the proposed confidence interval for p from (3.7) is given by:

$$\text{Length}_Q \equiv \text{Length}_{Q_d} = U_{Q_d} - L_{Q_d} = \frac{(d^2-1)T_{Q_d}}{(1+dT_{Q_d})(d+T_{Q_d})}. \tag{3.8}$$

Theorem 3.2 *The length of the confidence interval* J_{Q_d} *from (3.7) satisfies the following inequality:*

$$Length_{Q_d} \leq \frac{d-1}{d+1} \text{ w.p.1}, \tag{3.9}$$

where the expression of $Length_{Q_d}$ *comes from (3.8).*

Proof We define

$$g(x) = \frac{(d^2-1)x}{(1+dx)(d+x)}, 0 < x < \infty$$

which implies:

$$h(x) \equiv \ln g(x) = \ln(d^2-1) + \ln x - \ln(1+dx) - \ln(d+x)$$

$$\Rightarrow h'(x) = \frac{d(1-x^2)}{x(1+dx)(d+x)}.$$

Thus, we have $h(x) >, =, < 0$ if and only if $(0 <) x <, =, > 1$ so that $h(x)$, and hence equivalently $g(x)$, is increasing (decreasing) from $0 < x < 1$ $(x > 1)$. Then, clearly, $g(x)$ has its maximum at $x = 1$ with $g(1) = \frac{d-1}{d+1}$. This proves the desired result. ■

3.3 Some Discussions

In (1.5) part (a), note that $c(>0)$ should ideally be chosen "small". Suppose that in (1.5) part (a), we pick $0 < c < \frac{1}{2}$ and accordingly fix:

$$d = \frac{1+2c}{1-2c} \text{in (3.3) so that} \frac{d-1}{d+1} = 2c. \tag{3.10}$$

Thus, our constructed purely sequential methodology $\left(Q_d, J_{Q_d}\right)$ *under* (3.5)–(3.7) will produce a final confidence interval $J_{Q_d} \equiv [L_{Q_d}, U_{Q_d}]$ for p with its length *bounded from above by the number* $2c$. The spirit of (1.5) part (a) was different because, there one intended to obtain a fixed-width $(= 2c)$ $1-\alpha$ confidence interval for p.

It is important to recall that while the confidence interval G_{N_c} from (1.5) part (a) had a fixed length $2c$, it did not perform very well. But, per (3.10), with a choice of $d = \frac{1+2c}{1-2c}$ when $0 < c < \frac{1}{2}$, J_{Q_d} produces a confidence interval for p whose width is shorter than $2c$. Hence, it is easy to grasp why n_d^* from (3.4) would exceed the corresponding n_c^0 from (1.5) part (a). In this scenario, that is with $d = \frac{1+2c}{1-2c}$ when $0 < c < \frac{1}{2}$, the discrepancy between n_d^*, n_c^0 is the largest when p is near 0 or 1, whereas n_d^*, n_c^0 are close to each other when p is near $\frac{1}{2}$. That said, we need to reiterate that the methodologies under discussion do not utilize any prior knowledge about p other than the fact that $p \in (0, 1)$.

4 Second New Fix: A Two-Stage Bounded-Length Confidence Interval Methodology

Recall that $\theta^{-1}(\theta+1)^2$ and $p^{-1}(1-p)^{-1}$ are identical and thus the expression of n_d^* from (3.4) is alternatively expressed as

$$n_d^* = \left(\frac{z_{\alpha/2}}{\ln d}\right)^2 \left\{p^{-1} + (1-p)^{-1}\right\}. \tag{4.1}$$

We again note that $n_d^* \geq 4\left(\frac{z_{\alpha/2}}{\ln d}\right)^2$ and thus, we define the pilot sample size as:

$$n_0 \equiv n_{0d} = \left\lfloor 4\left(\frac{z_{\alpha/2}}{\ln d}\right)^2 \right\rfloor, \tag{4.2}$$

in the spirit of (3.5). Now, having recorded the pilot data $X_1, \ldots, X_{n_0}$, we determine the final sample size by the stopping rule:

$$R \equiv R_d = \max\left\{n_0, \left\lfloor \left(\frac{z_{\alpha/2}}{\ln d}\right)^2 \left\{(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} + (1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1}\right\} \right\rfloor + 1\right\}, \tag{4.3}$$

where $\gamma(>0)$ plays a similar role as in (3.2).

Once R is determined, we sample the difference in the second stage by gathering $R - n_0$ additional observations $X_{n_0+1}, \ldots, X_R$ in a single batch if $R > n_0$. On the other hand, if $R \leq n_0$, we do not gather additional observations in the second stage. That is, after implementing the two-stage methodology (4.2)–(4.3), the final dataset $\{R_d, X_1, \ldots, X_{R_d}\}$ would become available. Thus, along the lines of (3.6)–(3.7), we may construct the following bounded-length confidence interval for p:

$$J_{R_d} \equiv [L_{R_d}, U_{R_d}], \tag{4.4}$$

with $L_{R_d} = (d + T_{R_d})^{-1} T_{R_d}$ and $U_{R_d} = (1 + dT_{R_d})^{-1} dT_{R_d}$, with T defined earlier in (3.2). The earlier bound (3.9) will continue to hold for the length of $J_{R_d} \equiv [L_{R_d}, U_{R_d}]$ obtained by implementing the proposed two-stage methodology (4.2)–(4.4).

One may feel some resemblance with the two-stage fixed-width confidence interval procedure of Zacks and Mukhopadhyay [17]. Upon close inspection, one will find immediately that Zacks and Mukhopadhyay [17] constructed a two-stage fixed-width confidence interval procedure for estimating ln (odds-ratio) which is an unknown parameter that belongs to the whole real line. Thus, a fixed-width confidence interval will look reasonable. We continue to investigate a two-stage bounded-length confidence interval procedure for estimating p.

Remark 4.1 Even though $\gamma > 0$ will suffice, for practical purposes, as in Sect. 3 earlier, we may choose γ to be larger than $\frac{1}{2}$.

Lemma 4.1 *For the two-stage estimation rule* $\left(R_d, J_{R_d}\right)$ *under (4.2)–(4.4), for each fixed* $0 < p < 1$ *and* $0 < \alpha < 1$, *we have:*

$$\lim_{d \downarrow 1} E_p \left[(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} \right] = p^{-1} \text{and} \lim_{d \downarrow 1} E_p \left[(1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1} \right] = (1 - p)^{-1},$$

where n_0 *comes from (4.2) and* $\gamma > 0$.

This lemma can be proved using techniques similar to those found in the proof of Lemma 2 in Mukhopadhyay and Diaz [11]. Further details are omitted. Next, we summarize another lemma and outline its proof.

Lemma 4.2 *For the two-stage estimation rule* $\left(R_d, J_{R_d}\right)$ *under (4.2)–(4.4), for each fixed* $0 < p < 1$, *and* $0 < \alpha < 1$, *we have:*

$$(i) \lim_{d \downarrow 1} P_p \left(R = n_0\right) = 0 \text{ when } p \neq \frac{1}{2},$$

$$(ii) \lim_{d \downarrow 1} P_p \left(R = n_0\right) = 1 \text{ when } p = \frac{1}{2},$$

where n_0 *comes from (4.2) and* $\gamma > 0$.

Proof First, we consider $0 < p < 1$, $p \neq \frac{1}{2}$. For sufficiently small d in a right neighborhood of 1, with $\varepsilon_p = p^{-1}(1 - p)^{-1} - 4 (> 0)$, we may write:

$$
\begin{aligned}
&P_p(R = n_0)\\
&\quad \le P_p\left\{(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} + (1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1} \le 4\right\}\\
&\quad = P_p\left\{\left[(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} - p^{-1}\right] + \left[(1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1} - (1-p)^{-1}\right] \le -\varepsilon_p\right\},\\
&\quad \le P_p\left\{\left|\left[(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} - p^{-1}\right] + \left[(1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1} - (1-p)^{-1}\right]\right| \ge \varepsilon_p\right\}.
\end{aligned}
$$

But, $(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} \xrightarrow{P} p^{-1}$ and $(1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1} \xrightarrow{P} (1-p)^{-1}$ as $d \downarrow 1$, and hence the result follows.

The case when $p = \frac{1}{2}$ is left out for brevity. ■

Theorem 4.1 *For the two-stage estimation rule* $\left(R_d, J_{R_d}\right)$ *under (4.2)–(4.4), for each fixed* $0 < p < 1$, *and* $0 < \alpha < 1$, *we have as* $d \downarrow 1$*:*

(i) $R_d/n_d^* \to 1$ *w.p.1;*
(ii) $E_p\left[R_d/n_d^*\right] \to 1$ *[Asymptotic first-order efficiency]; and*
(iii) $P_p\left\{p \in J_{R_d} : [L_{R_d}, U_{R_d}]\right\} \to 1 - \alpha$ *[Asymptotic consistency];*

where n_d^* *comes from (4.4) with* $\gamma\,(> 0)$ *arbitrary.*

Proof First, we again consider $0 < p < 1$, $p \ne \frac{1}{2}$.

Proof of part (i) We note the basic inequality:

$$
\begin{aligned}
\left(\frac{z_{\alpha/2}}{\ln d}\right)^2 \left\{(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} + (1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1}\right\} \le R_d \le\ & n_0 I(R = n_0)\\
+ \left(\frac{z_{\alpha/2}}{\ln d}\right)^2 \left\{(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} + (1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1}\right\} &+ 1\text{w.p.1.} \qquad (4.5)
\end{aligned}
$$

Next, dividing throughout (4.5) by n_d^*, we get:

$$
\begin{aligned}
p(1-p)\left[(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} + (1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1}\right] \le R_d/n_d^* \le\ & n_0 n_d^{*-1} I(R = n_0)\\
p(1-p)\left[(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} + (1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1}\right] &+ n_d^{*-1}\text{w.p.1.} \qquad (4.6)
\end{aligned}
$$

Recall that $(\overline{X}_{n_0} + n_0^{-\gamma})^{-1} \xrightarrow{P} p^{-1}$, $(1 - \overline{X}_{n_0} + n_0^{-\gamma})^{-1} \xrightarrow{P} (1-p)^{-1}$, $I(R = n_0) \xrightarrow{P} 0$ (in view of Lemma 4.2), $n_d^{*-1} \to 0$, and $n_0 n_d^{*-1} \to \frac{4}{p(1-p)}$ as $d \downarrow 1$. Thus, part (i) follows from (4.6).

Proof of part (ii) Lemma 4.1 shows that

$$\lim_{d\downarrow 1} E_p\left[(\overline{X}_{n_0}+n_0^{-\gamma})^{-1}+(1-\overline{X}_{n_0}+n_0^{-\gamma})^{-1}\right]=p^{-1}+(1-p)^{-1}=p^{-1}(1-p)^{-1}.$$

Now, we take expectations throughout (4.6). Then, part (ii) follows by applying Lemmas 4.1, 4.2 and taking limits as $d \downarrow 1$.

Proof of part (iii) The proof will move along the lines developed in the case of the sequential procedure (3.5)–(3.7). For more details one may refer to Banerjee and Mukhopadhyay [1], Mukhopadhyay and Banerjee [10].

The case when $p = \frac{1}{2}$ is left out for brevity. ■

4.1 Purely Sequential or Two-Stage: Which One to Implement?

An important and natural question may arise: How should one choose between the purely sequential methodology and the two-stage methodology? In all fairness, there is no one simple answer.

In some situations, observations may arrive naturally in a sequence. For example, in a clinical trial, patients may arrive one by one. In a production line, manufactured items may come off a conveyor belt one by one. In such situations, our newly proposed purely sequential confidence interval estimation methodology should be implemented.

In other situations, observations may arrive naturally in bulks or groups. Finished boxes of batteries may come off a conveyor belt in batches. In such situations, our newly proposed two-stage confidence interval estimation methodology should be implemented.

In yet another kind of a situation, if it so happens that either purely sequential or two-stage confidence interval estimation methodology can be implemented, then our recommendation will be to use the two-stage methodology because of its operational convenience and logistical simplicity.

5 Data Analysis

Now, we provide some interesting results from data analysis for the purely sequential methodology (3.5)–(3.7) and the two-stage methodology (4.2)–(4.4). In Sect. 5.1, we discuss illustrations and performances of both bounded-length confidence interval methodologies for a Bernoulli parameter p using extensive sets of computer simulations. Section 5.2 shows illustrations and implementations of the proposed methodologies (3.5)–(3.7) and (4.2)–(4.4) utilizing three kinds of real data sets.

All results so summarized are presented when we had fixed the following values: $\alpha = 0.05$ and $\gamma = 0.7$. However, the interesting features and performances that are highlighted here remain nearly same for many other choices of α and γ. We have deliberately omitted those for brevity.

5.1 Data Analysis from Simulations

Tables 7 and 8 summarize simulation results corresponding to the purely sequential procedure (3.5)–(3.7) when $p = 0.9$ and $p = 0.8$, respectively. The results in each row show averages from 10000 replications. Column 1 shows the fixed-accuracy level $d(> 1)$, and column 2 shows n_d^* from (3.4). We assumed the pilot sample size n_0 as in (3.5). Column 4 shows the average purely sequential sample size $\overline{q}$ and its estimated standard error $(s(\overline{q}))$ obtained from averaging all 10000 runs. Columns 5 and 6 respectively show the ratio and difference of $\overline{q}$ and n_d^*. The ratio gives us an idea about first-order efficiency measure in practice. This is supposed to be near 1. Column 6 briefly addresses second-order efficiency measure in practice.

For column 7, we defined w, an indicator variable taking the value 1(0) if a constructed confidence interval (L_q, U_q) from (3.7) upon termination when $Q = q$ included (did not include) the true value of p. We show the associated $\overline{w}$, the average from 10000 such observed values of w, and its estimated standard error $(s(\overline{w}))$. This $\overline{w}$ gives an idea about the achieved coverage probability which we expect to be close to the preset target, 0.95.

Table 7 Simulation results from 10000 replications for the purely sequential methodology (3.5)–(3.7) in a Bernoulli($p = 0.9$) population when $\alpha = 0.05$ and $\gamma = 0.7$ in (3.2)

d	n_d^*	n_0	$\overline{q}(s(\overline{q}))$	$\overline{q}/n_d^*$	$\overline{q} - n_d^*$	$\overline{w}(s(\overline{w}))$
1.12	3323.456	1196	3328.233 (1.544)	1.001	4.777	0.951 (0.003)
1.11	3919.236	1410	3922.729 (1.654)	1.001	3.493	0.949 (0.003)
1.10	4698.844	1691	4702.695 (1.832)	1.001	3.850	0.950 (0.003)
1.09	5747.512	2069	5753.848 (2.013)	1.001	6.336	0.949 (0.003)

Table 8 Simulation results from 10000 replications for the purely sequential methodology (3.5)–(3.7) in a Bernoulli($p = 0.8$) population when $\alpha = 0.05$ and $\gamma = 0.7$ in (3.2)

d	n_d^*	n_0	$\overline{q}(s(\overline{q}))$	$\overline{q}/n_d^*$	$\overline{q} - n_d^*$	$\overline{w}(s(\overline{w}))$
1.12	1869.444	1196	1871.911 (0.646)	1.001	2.467	0.948 (0.002)
1.11	2204.570	1410	2206.825 (0.710)	1.001	2.255	0.948 (0.002)
1.10	2643.109	1691	2647.002 (0.772)	1.001	3.902	0.949 (0.002)
1.09	3232.975	2069	3235.781 (0.854)	1.001	2.806	0.949 (0.002)

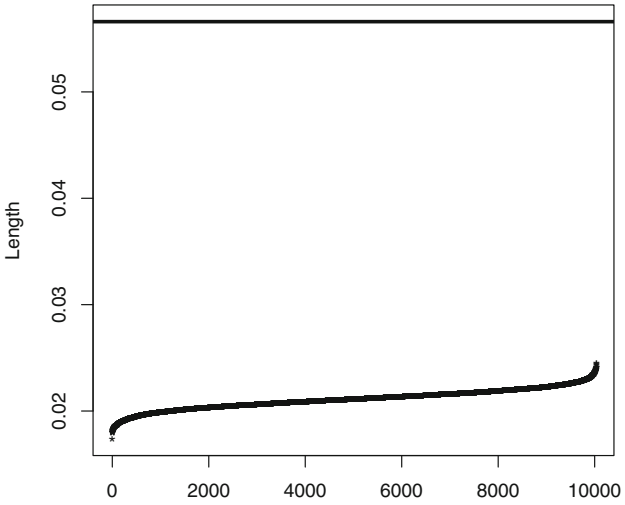


Fig. 4 The plot shows 10000 observed and sorted values of the Length_Q function (3.8) associated with the first row ($d = 1.12$) of Table 7 obtained via simulations using a Bernoulli($p = 0.9$) population under the purely sequential methodology (3.5)–(3.7). The *top horizontal line* corresponds to $(d+1)^{-1}(d-1) \approx 0.05660377$ when $\alpha = 0.05$; $\gamma = 0.7$

Going back to Theorem 3.1 parts (ii) and (iii), as d moves closer to 1, we expect the ratio $\overline{q}/n_d^*$ to move closer to 1 and $\overline{w}$ to move close to the target 0.95. These features are largely validated from simulations.

In the context of the first row of Table 7 (that is, when $d = 1.12$), we have shown a plot of 10000 values of the function Length_q from (3.8) to see an empirical validation (Fig. 4) of Theorem 3.2. The upper bound for Length_q, namely

$$(d+1)^{-1}(d-1) = 0.056603773,$$

is included as the top horizontal line in Fig. 4. The Length_q function clearly remains under the horizontal line with

$$\max(\text{Length}_q) = 0.0244884 < (d+1)^{-1}(d-1).$$

Similar features were empirically validated in all other cases, but we avoid providing such details or commenting on them in every other situation.

Tables 9 and 10 summarize simulation results corresponding to the two-stage procedure (4.2)–(4.4) when $p = 0.9$ and $p = 0.8$ respectively. The results in each row show averages from 10000 replications. The pilot sample size n_0 was computed according to (4.2). All other entities have the same interpretations as those in Tables 7 and 8.

The two-stage methodology is operationally more convenient and less time consuming. By comparing Tables 7 and 8 with Tables 9 and 10, we notice that in all

Table 9 Simulation results from 10000 replications for the two-stage methodology (4.2)–(4.4) in a Bernoulli($p = 0.9$) population when $\alpha = 0.05$ and $\gamma = 0.7$ in (4.3)

d	n_d^*	n_0	$\overline{r}(s(\overline{r}))$	$\overline{r}/n_d^*$	$\overline{r} - n_d^*$	$\overline{w}(s(\overline{w}))$
1.12	3323.456	1196	3145.672 (2.256)	0.946	–177.784	0.949 (0.003)
1.11	3919.236	1410	3737.626 (2.520)	0.951	–191.610	0.947 (0.003)
1.10	4698.844	1691	4495.730 (2.767)	0.957	–203.114	0.948 (0.003)
1.09	5747.512	2069	5522.576 (3.103)	0.961	–224.936	0.951 (0.003)

Table 10 Simulation results from 10000 replications for the two-stage methodology (4.2)–(4.4) in a Bernoulli($p = 0.8$) population when $\alpha = 0.05$ and $\gamma = 0.7$ in (4.3)

d	n_d^*	n_0	$\overline{r}(s(\overline{r}))$	$\overline{r}/n_d^*$	$\overline{r} - n_d^*$	$\overline{w}(s(\overline{w}))$
1.12	1869.444	1196	1820.784 (0.761)	0.974	–48.660	0.946 (0.002)
1.11	2204.570	1410	2154.071 (0.761)	0.978	–50.499	0.946 (0.002)
1.10	2643.109	1691	2588.066 (0.918)	0.979	–55.034	0.949 (0.002)
1.09	3232.975	2069	3174.276 (1.016)	0.982	–58.700	0.949 (0.002)

cases under consideration, the average two-stage sample size ($\overline{r}$) is smaller than the average purely sequential sample size ($\overline{q}$). Again, we note that the values of $\overline{r}/n_d^*$ and $\overline{w}$ confirm the asymptotic results of first-order efficiency and consistency properties proved in Theorem 4.1 parts (ii) and (iii).

5.2 Three Illustrations with Real Data

In this section, we implement the proposed purely sequential and two-stage methodologies with the help of real data. In the following sections, we handle three real data sets of size small to moderate to large covering interesting areas from health studies ((i) chance of relapse in bone marrow transplant patients, (ii) presence or absence of diabetes for Pima Indians) and from entomology ((iii) presence or absence of potato beetle infestation).

5.2.1 Illustration 1: Chance of Relapse in Bone Marrow Transplant Patients

A standard treatment for acute leukemia is bone marrow transplant (BMT). BMT is considered to be a failure if a patient relapses or dies during remission. Data were collected on 137 patients treated in 4 hospitals: The Ohio State University Hospital in Columbus (76 patients); Hahnemann University in Philadelphia (21 patients); St. Vincent's Hospital in Sydney, Australia (23 patients); and at Alfred Hospital in Melbourne (17 patients). This *small-size data* under consideration may be found in Klein and Moeschberger [8], Copelan et al. [6].

Table 11 Illustration 1 using BMT data with MLE $\widehat{p}_{137,\mathrm{MLE}} = 0.306$ treated as "true" p under a single run of the purely sequential methodology (3.5)–(3.7) when $\alpha = 0.05$, $\gamma = 0.7$, and with (L_Q, U_Q) from (3.7)

d	$\widehat{n}_d^*$	n_0	Q	$Q/\widehat{n}_d^*$	$Q - \widehat{n}_d^*$	$[L_Q, U_Q]$	$\frac{d-1}{d+1}$	Length_Q
1.65	72.060	61	66	0.916	−6.060	(0.280, 0.514)	0.245	0.234
1.60	81.805	69	76	0.929	−5.805	(0.267, 0.482)	0.231	0.215
1.55	94.086	80	87	0.924	−7.086	(0.282, 0.486)	0.216	0.204
1.50	109.919	93	103	0.937	−6.919	(0.273, 0.458)	0.200	0.185

Table 12 Illustration 1 using BMT data with MLE $\widehat{p}_{137,\mathrm{MLE}} = 0.306$ treated as "true" p under a single run of the two-stage methodology (4.2)–(4.4) when $\alpha = 0.05$, $\gamma = 0.7$, and with (L_R, U_R) from (4.4)

d	$\widehat{n}_d^*$	n_0	R	$R/\widehat{n}_d^*$	$R - \widehat{n}_d^*$	$[L_R, U_R]$	$\frac{d-1}{d+1}$	Length_R
1.65	72.060	61	68	0.944	−4.060	(0.196, 0.399)	0.245	0.203
1.60	81.805	69	76	0.929	−5.805	(0.208, 0.402)	0.231	0.194
1.55	94.086	80	89	0.946	−5.086	(0.215, 0.397)	0.216	0.182
1.50	109.919	93	104	0.946	−5.919	(0.225, 0.395)	0.200	0.170

Transplants were conducted between March 1, 1984 and June 30, 1989 with a maximum follow-up time of 7 years. Among all patients, 42 relapsed. Data included an indicator variable that took the value 1 if a patient relapsed or 0 otherwise. The parameter p indicates the chance or probability of relapse.

We may have an idea of p from the MLE, $\widehat{p}_{137,\mathrm{MLE}} = 0.306$ obtained from the full dataset. For the purpose of illustration, we may regard this as the "true" p even though this make-belief "true" p has no bearing on our methodologies. We fitted a Bernoulli distribution with $p = 0.306$ to this dataset. The q-q plot indicated a perfect fit and the Chi-square goodness of fit test (Chi-square statistic $= 1.195231 \times 10^{-12}$) gave a p-value of 1.

Table 11 includes the analysis of this data set for the purely sequential procedure (3.5)–(3.7). We consider 4 values of the accuracy d (> 1) as accommodated by the data set. Here $\widehat{n}_d^*$, n_0 and Q denote the estimated optimal sample size required (3.4), pretending that "true" p is 0.306, the pilot sample size and the sequential sample size required as in (3.5) respectively. The ratio of $Q/\widehat{n}_d^*$ are close to 1 as expected from Theorem 3.1 (ii). For all choices of d, the confidence interval (L_Q, U_Q) computed as in (3.6), include the most plausible value of p, 0.306. Also, in the light of Theorem 3.2, observe that in each case, the length of the obtained confidence interval is bounded by the quantity $(d+1)^{-1}(d-1)$.

In Table 12 we find analysis of the two-stage procedure (4.2)–(4.4). Here, we consider the same values of the accuracy d as in Table 11. The pilot sample size n_0 is computed as in (4.2). All other symbols have their usual meaning. The ratios of the two-stage and optimal fixed sample sizes $R/\widehat{n}_d^*$ are close to 1 and in each case,

the final two-stage confidence interval includes the most plausible value of p, 0.306. Here too, we see that the length of each interval is bounded by $(d+1)^{-1}(d-1)$.

5.2.2 Illustration 2: Presence or Absence of Diabetes For Pima Indians

This *medium size* dataset came from the National Institute of Diabetes and Digestive and Kidney Diseases. This data primarily comprises of 768 female patients, all of whom were at least 21 years old, and came from Pima Indian heritage. The data were collected from a population living in Phoenix, Arizona, USA. Smith et al. [14] used the data with a unique algorithm to forecast the onset of diabetes mellitus.

In this illustration, a binary variable would take the value 1 if a patient indicated signs of diabetes according to the World Health Organization's criteria (that is, if the 2 h post-load plasma glucose was at least 200 mg/dl in any survey examination or during routine medical care) and the value 0 otherwise. The dataset is publicly available from the following website: http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/.

We may have an idea of p, the proportion of patients who show signs of diabetes, from the MLE, $\widehat{p}_{768,\mathrm{MLE}} = 0.349$ obtained from the full dataset. For the purpose of illustration, we may regard this as the "true" p even though this make-belief "true" p has no bearing on our methodologies. We fitted a Bernoulli distribution with $p = 0.349$ to this dataset. The q-q plot indicated a perfect fit and the Chi-square goodness of fit test (Chi-square statistic $= 3.841358 \times 10^{-12}$) gave a p-value 1. We now proceed with the illustrations.

Table 13 includes the analysis under the purely sequential methodology (3.5)–(3.7). We consider 4 choices for the accuracy $d(>1)$ and $\widehat{n}_d^*$, n_0 and Q denote the estimated optimal fixed-sample-size required (3.4), pretending that "true" p is 0.349, the pilot sample size and the purely sequential sample size as in (3.5), respectively.

Table 14 highlights analogous performances of the two-stage methodology procedure (4.2)–(4.4) in the light of diabetes data. We used the same set of d values as in Table 13. The pilot sample size n_0 was determined using (4.2).

Table 13 Illustration 2 using diabetes data with MLE $\widehat{p}_{768,\mathrm{MLE}} = 0.349$ treated as "true" p under a single run of the purely sequential methodology (3.5)–(3.7) when $\alpha = 0.05$, $\gamma = 0.7$, and with (L_Q, U_Q) from (3.7)

d	$\widehat{n}_d^*$	n_0	Q	$Q/\widehat{n}_d^*$	$Q - \widehat{n}_d^*$	$[L_Q, U_Q]$	$\frac{d-1}{d+1}$	Length_Q
1.35	187.752	170	182	0.969	−5.752	(0.317, 0.458)	0.149	0.141
1.30	245.652	223	237	0.965	−8.652	(0.325, 0.448)	0.130	0.124
1.25	339.595	308	329	0.969	−10.595	(0.330, 0.435)	0.111	0.105
1.20	508.691	462	501	0.985	−7.691	(0.324, 0.408)	0.091	0.084

Table 14 Illustration 2 using diabetes data with MLE $\widehat{p}_{768,\mathrm{MLE}} = 0.349$ treated as "true" p under a single run of the two-stage methodology (4.2)–(4.4) when $\alpha = 0.05$, $\gamma = 0.7$, and with (L_R, U_R) from (4.4)

d	$\widehat{n}_d^*$	n_0	R	$R/\widehat{n}_d^*$	$R - \widehat{n}_d^*$	$[L_R, U_R]$	$\frac{d-1}{d+1}$	Length_R
1.35	187.752	170	202	1.076	14.248	(0.246, 0.374)	0.149	0.127
1.30	245.652	223	252	1.026	−6.348	(0.260, 0.372)	0.130	0.112
1.25	339.595	308	334	0.984	−5.595	(0.295, 0.396)	0.111	0.100
1.20	508.691	462	499	0.981	−9.691	(0.308, 0.391)	0.091	0.083

Table 15 Illustration 3 using potato beetle data with MLE $\widehat{p}_{2304,\mathrm{MLE}} = 0.918$ treated as "true" p under a single run of the purely sequential methodology (3.5)–(3.7) when $\alpha = 0.05$, $\gamma = 0.7$, and with (L_Q, U_Q) from (3.7)

d	$\widehat{n}_d^*$	n_0	Q	$Q/\widehat{n}_d^*$	$Q - \widehat{n}_d^*$	$[L_Q, U_Q]$	$\frac{d-1}{d+1}$	Length_Q
1.29	782.993	236	750	0.958	−32.993	(0.881, 0.925)	0.127	0.044
1.26	950.549	287	926	0.974	−24.549	(0.887, 0.925)	0.115	0.039
1.23	1184.727	358	1266	1.069	81.273	(0.900, 0.931)	0.103	0.032
1.20	1527.364	462	1554	1.017	26.636	(0.899, 0.927)	0.091	0.029
1.17	2059.674	623	2061	1.001	1.326	(0.900, 0.925)	0.078	0.025

5.2.3 Illustration 3: Presence or Absence of Potato Beetle Infestation

This classic dataset came from Beall [2] which we have described and used in Sect. 2.2. In what follows, we again use this large dataset to validate performances of the proposed purely sequential and two-stage methodologies.

Table 15 includes the analysis under the purely sequential methodology (3.5)–(3.7). We consider 4 choices for the accuracy $d(> 1)$ and $\widehat{n}_d^*$, n_0 and Q denote the estimated optimal fixed-sample-size required (3.4), pretending that "true" p is $\widehat{p}_{2304,\mathrm{MLE}} = 0.918$, the pilot sample size and the purely sequential sample size required as in (3.5), respectively.

Table 16 highlights analogous performances of the two-stage methodology procedure (4.2)–(4.4) in the light of potato beetle infestation data. We used the same set of d values as in Table 15. The pilot sample size n_0 was determined using (4.2).

5.2.4 Brief Comments on Illustrations 1–3

We reiterate that the size of data in illustrations 1–3 may respectively be considered small, medium, and large. For the implementation of either methodology, we treated each dataset as our universe, and drew observations from it, as dictated by the designed stopping times. We have used *simple random sampling without replacement* (SRSWOR) in order to gather observations from a universe under consideration. In a fixed universe, the SRSWOR created Bernoulli observations with the same p. The

Table 16 Illustration 3 using potato beetle data with MLE $\widehat{p}_{2304,\mathrm{MLE}} = 0.918$ as "true" p under a single run of the two-stage methodology (4.2)–(4.4) when $\alpha = 0.05$, $\gamma = 0.7$, and with (L_R, U_R) from (4.4)

d	$\widehat{n}_d^*$	n_0	R	$R/\widehat{n}_d^*$	$R - \widehat{n}_d^*$	$[L_R, U_R]$	$\frac{d-1}{d+1}$	Length_R
1.29	782.993	236	724	0.925	−58.993	(0.894, 0.933)	0.127	0.039
1.26	950.549	287	920	0.968	−30.549	(0.892, 0.929)	0.115	0.037
1.23	1184.727	358	1171	0.988	−13.727	(0.899, 0.931)	0.103	0.032
1.20	1527.364	462	1482	0.970	−45.364	(0.900, 0.928)	0.091	0.028
1.17	2059.674	623	2038	0.989	−21.674	(0.903, 0.927)	0.078	0.024

observations became dependent, however such dependence became weaker as the size of our universe moved from small to medium to large.

We emphasize that our reported $\widehat{p}_{n,\mathrm{MLE}}$ in a universe was exclusively used to obtain $\widehat{n}_d^*$ which provides a reasonable landmark with which we may want to compare Q or R found in single runs as provided in Tables 11, 12, 13, 14, 15 and 16. We feel encouraged by noting that they compare remarkably well as we walk through Tables 11, 12, 13, 14, 15 and 16 and make a special note of the fact that both ratios $Q/\widehat{n}_d^*$ and $R/\widehat{n}_d^*$ stay close to 1 across the board.

For all choices of d under consideration, both the purely sequential confidence interval (L_Q, U_Q) constructed from (3.7) and the two-stage confidence (L_R, U_R) constructed from (4.4) include the plausible value of p, namely the respective $\widehat{p}_{n,\mathrm{MLE}}$. In each situation, we also highlight that the observed length of (L_Q, U_Q) or (L_R, U_R) fell below $(d+1)^{-1}(d-1)$, the maximum width (Theorem 3.2). It is truly encouraging to see that our proposed methodologies deliver expected outcomes while withstanding mild dependence among recorded observations under SRSWOR.

Acknowledgments We remain grateful to the editors of this special volume for giving us an opportunity to contribute our work in celebration of Professor H.N. Nagaraja's 60th birthday. An anonymous referee gave a number of pointers which helped in revising the manuscript. We thank the editors and the anonymous referee.

References

1. Banerjee, S., and Mukhopadhyay, N. (2015). A general sequential fixed-accuracy confidence interval estimation methodology for a positive parameter: Illustrations using health and safety data. *Annals of Institute of Statistical Mathematics*, in press, pp. 1–30. doi:10.1007/s10463-015-0504-2
2. Beall, G. 1942. The transformation on data from entomological field of experiments so that the analysis of variance becomes applicable. *Biometrika* 32: 243–262.
3. Cho, H. 2007. Sequential risk-efficient estimation for the ratio of two binomial proportions. *Journal of Statistical Planning and Inference* 127: 2336–2346.
4. Cho, H. 2013. Approximate confidence limits for the ratio of two binomial variates with unequal sample sizes. *Communications for Statistical Applications and Methods* 20: 347–356.

5. Chow, Y.S., and H. Robbins. 1965. On the asymptotic theory of fixed width sequential confidence intervals for the mean. *Annals of Mathematical Statistics* 36: 457–462.
6. Copelan, E.A., J.C. Biggs, J.M. Thompson, P. Crilley, J. Szer, J.P. Klein, N. Kapoor, B.R. Avalos, I. Cunningham, K. Atkinson, K. Downs, G.S. Harmon, M.B. Daly, I. Brodsky, S.I. Bulova, and P.J. Tutschka. 1991. Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with Bu/Cy. *Blood* 78: 838–843.
7. Ehrenfeld, S., and S.B. Littauer. 1964. *Introduction to statistical methods*. New York: McGraw Hill.
8. Klein, J.P., and M.L. Moeschberger. 2003. *Survival analysis: techniques for censored and truncated data*. New York: Springer.
9. Khan, R.A. 1969. A general method of determining fixed-width confidence intervals. *Annals of Mathematical Statistics* 40: 704–709.
10. Mukhopadhyay, N., and S. Banerjee. 2014. Purely sequential and two-stage fixed-accuracy confidence interval estimation methods for count data from negative binomial distributions in statistical ecology: one-sample and two-sample problems. *Sequential Analysis* 33: 251–285.
11. Mukhopadhyay, N., and J. Diaz. 1985. Two stage sampling for estimating the mean of a negative binomial distribution. *Sequential Analysis* 4: 1–18.
12. Nadas, A. 1969. An extension of a theorem of chow and robbins on sequential confidence intervals for the mean. *Annals of Mathematical Statistics* 40: 667–671.
13. Robbins, H., and D. Siegmund. 1974. Sequential estimation of p in Bernoulli trials, In: Pitman Volume, E.J.G., E.J. Williams, (Eds.), *Studies in Probability and Statistics*, pp. 103–107. Jerusalem: Jerusalem Academic Press.
14. Smith, J.W., J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the Symposium on Computer Applications and Medical Care, pp. 261–265.
15. Willson, L.J., and J.L. Folks. 1983. Sequential estimation of the mean of the negative binomial distribution. *Communications in Statistics - Sequential Analysis* 2: 55–70.
16. Zacks, S. 1966. Sequential estimation of the mean of a log-normal distribution having a prescribed proportional closeness. *Annals of Mathematical Statistics* 37: 1439–1888.
17. Zacks, S., and N. Mukhopadhyay. 2007. Distributions of sequential and two-stage stopping times for fixed-width confidence intervals in Bernoulli trials: Application in Reliability. *Sequential Analysis* 26: 425–441.
18. Zehna, P.W. 1966. Invariance of maximum likelihood estimators. *Annals of Mathematical Statistics* 37: 744.

A Population Based Confidence Set Inference Method for SNPs that Regulate Quantitative Phenotypes

Charalampos Papachristou

Abstract The increased use of genome-wide association studies based on genetic maps consisting of hundreds of thousands of SNPs has prompted the need for methods that can be used in preliminary analyses to limit the number of SNPs investigated in follow-up studies. I introduce a Confidence Set Inference method for independent individuals that can be used as a first step in association studies to derive a set of SNPs that contribute at least a specific percentage to the total variance of a quantitative trait. The main advantage of the method is that it allows control over the confidence level with which one can identify genes with specific effects on the genetic variance of the trait of interest. Developed in the framework of linear models, the method can efficiently incorporate information on pertinent covariates. I investigate the properties of the method through an extensive simulation study under various simple inheritance models and compare its performance to that of a standard association approach as it is implemented in the software package Merlin.

Keywords GWAS · Association studies · Confidence sets · Fine mapping · CSI

1 Introduction

Genome-wide association studies (GWAS) have become a standard tool in the quest for loci that play an important role in the regulation of quantitative phenotypes or the development of qualitative traits of humans [9, 11]. They usually involve the scan of a large number, of the order of at least half a million, of single nucleotide polymorphisms (SNPs) densely covering the entire human genome. As such, the

C. Papachristou (✉)
Department of Mathematics, Rowan University,
201 Mullica Hill Road, Glassboro, NJ 08028, USA
e-mail: papachristou@rowan.edu

C. Papachristou
Department of Mathematics, Physics, and Statistics, University of the Sciences,
600 South 43rd Street, Philadelphia 19104, USA

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_14

performance of GWASs is plagued by impediments such as multiplicity adjustment for the number of SNPs examined and computational issues due to the large volume of data need to be handled [7, 9]. To ameliorate the effect of these two hurdles, researchers have been employing an efficient two-stage analysis scheme comprised of an initial screening step followed by a testing step [3, 7, 8, 11–13]. In the screening step, stage I, all SNPs on the genetic map are examined to identify those which appear to be associated with the trait of interest. This step can be based on all available individuals in the sample [4, 8] or on an appropriate subgroup of the sample such as all unrelated individuals [7]. In addition, the methods employed are usually single marker based so as to keep the computational requirements to a minimum [9, 14]. In the second step, a subset of SNPs prioritized by the screening step is analyzed using more elaborate and powerful methods, such as candidate genes [15], haplotype methods [2], or Least Absolute Shrinkage and Selection Operator (Lasso) penalized regression [4, 10]. These analyses can also include additional individuals that can be potentially related to those used in the screening step [7]. In general, the statistical methods used in stage two tend to be more computationally involved as they handle multiple markers at once and may also take into account (potential) relatedness among study participants [7, 10].

The two-stage strategy has the advantage that it reduces both the computation complexity of the analysis and the need for multiplicity adjustment for the number of markers tested. However, the inherent risk is that the overall power of the study to identify trait regulating genes largely depends on the ability of the method applied in the preliminary screening stage to correctly identify and advance such loci to stage II. Clearly, the choice of the method implemented in stage I is very important. Hence, methods that can guarantee a high level of power while keeping the computational intensity to minimum are highly desirable.

Recently, Papachristou and Lin [9] described a Confidence Set Inference (CSI) method that can be used in GWAS studies to identify, with a given predetermined confidence, loci contributing a certain level of heritability to a quantitative phenotype. The method uses data from related individuals to first split the overall variance of a quantitative trait to its genetic and environmental components. Then, using the estimate of the additive genetic component, it scans the entire genome in search of SNPs that contribute at least a certain percentage to the overall additive genetic variance of the trait. As the method is based on the additive genetic variance component of the trait, it cannot be used with independent individuals as estimation of genetic variance components is not possible with such data.

Here, I propose a modified version of the CSI method that can be used in GWAS studies based on independent individuals to identify, with a given predetermined confidence, loci that contribute at least a certain percentage to the overall variance of a quantitative phenotype. It is a regression based method. As such, it allows for the incorporation of fixed effects, such as covariates, that can potentially have an effect on the phenotypic value. The method postulates additivity of the effects across all qualitative loci (QTLs). Furthermore, it assumes that the particular locus of interest contributes only additive effects to the overall variance of the trait. I perform a simulation study to explore the properties of the method under a variety of

conditions, some of which violate the assumptions of the method, and I demonstrate that the method maintains a true positive rate close to the nominal one, even under mild violation of its assumptions. A comparison between the CSI method and the standard association method (fastAssoc) implemented in Merlin [1] reveals that the former can be an attractive alternative to the latter.

2 Methods

2.1 The Hypothesis

For the development of the method, we will assume that we perform a GWAS based on M binary markers (SNPs) spanning the entire genome. For each SNP m on the map we test at level α to see if its contribution to the variance of a quantitative trait is at least $h_0 \times 100\,\%$ of the total variance. That is, for each marker we perform the following hypothesis test:

$$H_0 : \sigma_m^2 \geq h_0 \sigma_T^2 \quad \text{versus} \quad \mathrm{H_a} : \sigma_{\mathrm{m}}^2 < \mathrm{h_0}\sigma_{\mathrm{T}}^2, \tag{1}$$

where σ_m^2 and σ_T^2 are the mth locus specific genetic variance at the trait and the total variance of the trait, respectively, and $h_0 (0 < h_0 < 1)$ is a predetermined constant corresponding to the desired heritability. We can use this hypothesis test to obtain a $1 - \alpha$ confidence set of loci that contribute to the trait at least $h \times 100\,\%$ of the total variance. Such a set consists of all SNPs for which the above null hypothesis is not rejected at level α. As we will demonstrate below, the above hypothesis test can be easily performed by constructing a test developed in the context of the linear regression models.

2.2 Modeling of the Phenotype

Suppose that we have a sample of n unrelated individuals and let $y_i, i = 1, \ldots, n$, be the value of a quantitative phenotype of interest of the ith person in the study. We assume that for person i, the value of a quantitative phenotype can be (potentially) expressed in terms of several known fixed covariates X_i (e.g., age, gender, etc.) and a random effect ω_i, as follows

$$y_i = X_i \beta + \omega_i, \tag{2}$$

where β is a vector of unknown coefficients. We further assume that the random effects ω_i's are independent and identically distributed (i.i.d.) coming from a normal distribution with mean zero and variance σ_T^2 and they do not interact in any fashion with the fixed effects. Note that the random effects ω_i's consist of both genetic and

non-genetic (environmental) effects. In fact, if we assume that the individuals are unrelated, it can be shown that $\sigma_T^2 = \sigma_g^2 + \sigma_e^2$, where σ_g^2 is the variance due to the genetic component of the trait and σ_e^2 is the variance due to environmental factors. Finally, in the absence of gene-gene interactions, it can be shown that the overall genetic variance of the trait (σ_g^2) is just the sum of the locus specific variances of all trait contributing genes, that is, $\sigma_g^2 = \sum_{\tau=1}^{\tau=G} \sigma_\tau^2$, where σ_τ^2 is the genetic variance attributed to locus τ, $\tau = 1, \ldots, G$.

Effects of major genes can be incorporated in (2) by adding in the model appropriate fixed factors. In particular, the effects of binary loci can be expressed in terms of the copies of their minor allele carried by the individual. Consider a trait contributing loci τ and let $z_{\tau i}$ be the number of copies of the minor frequency trait allele carried by the ith individual in the sample. Then, the phenotypic value of the person can be expressed as

$$y_i = X_i\beta + \gamma_t z_{\tau i} + \omega_{-\tau i}, \tag{3}$$

where y_i, X_i, β are defined as before, γ_τ denotes the effect of the trait locus τ, and $\omega_{-\tau i}$ is the residual random effect due to non-genetic and genetic factors other than locus τ [9].

2.3 Construction of the Confidence Sets

If we assume that the effect of the alleles at the trait locus τ interact additively, we can show [9] that the coefficient in the regression model is equal to: $\gamma_\tau = \sigma_\tau/\sqrt{2p_\tau(1-p_\tau)}$, where σ_τ^2 is the genetic variance due to locus τ and $p_\tau (< 0.5)$ is the minor allele frequency (MAF) of the trait locus. Using this fact, the hypothesis test in (1) is equivalent to testing

$$H_0 : \gamma_m \geq \sqrt{h_0\sigma_T^2/2p_m(1-p_m)} \quad \text{versus} \quad \mathrm{H_a} : \gamma_\mathrm{m} < \sqrt{\mathrm{h_0}\sigma_\mathrm{T}^2/2\mathrm{p_m}(1-\mathrm{p_m})}, \tag{4}$$

where p_m is the MAF of the marker m.

The model in (3) can be used to obtain estimates of the coefficient γ_m and its standard error, denoted by $\hat{\gamma}_m$ and $s(\hat{\gamma}_m)$, respectively, either by maximizing the appropriate likelihood function or by fitting the corresponding regression model. Standard asymptotic theory can be evoked in the presence of a large sample size to show that $\hat{\gamma}_m$ asymptotically follows a normal distribution. Thus the following statistic

$$T_m = \tfrac{1}{s(\hat{\gamma}_m)}[\hat{\gamma}_m - \sqrt{h_0\sigma_T^2/2p_m(1-p_m)}] \tag{5}$$

can be used to test the hypotheses in (4). Obviously, the collection of SNPs for which $|T_m| \geq z_\alpha$, where z_α is the upper α percentile of the standard normal distribution, provides an $(1-\alpha) \times 100\,\%$ confidence set of loci contributing at least $h \times 100\,\%$ to the total variance of the quantitative phenotype. Note, the estimate of the total

variance of the trait, σ_T^2, can be obtained by fitting the model without any major gene effects given in Eq. (2).

3 A Simulation Study

In this section I study the properties of the proposed method through a simulation study. In particular, I gauge the true coverage of the resulting confidence sets and how it relates to the nominal confidence level and whether model misspecification bears an effect on it. I also investigate the false positive rate of the method and how it is affected by the available sample size. Finally, I compare the performance of the CSI method to that of a standard association method as it is implemented in the software package Merlin [1].

For all simulations, I used the software package SIMPED [6] to simulate genotypes for the sample members. SIMPED allows for linkage disequilibrium (LD) among groups (blocks) of consecutive markers. I assumed a map of three chromosomes each carrying 10,000 SNPs. The average distance between consecutive SNPs was 0.1 cm while the MAF ranged between 0.10 and 0.48. Finally, each marker was part of a window of 20 SNPs in high LD with each other (D' between 0.6 and 1).

The simulated phenotypes were modeled after the cholesterol levels of the individuals in the Framingham Heart Study [5] and were set to have a total variance of 1,309.4, 33 % of which attributed to genetics, with 32 and 1 % corresponding to additive and dominance genetic effects, respectively. I explored several two-locus models with the loci assumed to segregate independently. Both QTLs were assumed to be diallelic, in Hardy-Weinberg and linkage equilibrium with each other, and interacted in an additive manner. I explored models where each locus contributed roughly 0.5 or 1 % of the total variance of the trait. Furthermore, I considered models in which the loci either only contributed to the additive genetic variance of the trait or both additive and dominance genetic variance.

The results under all simulation models qualitatively exhibited similar trends, so I present the results from the model under which each QTL contributed 1 % of the total variance of the trait. The first locus (Q1) contributed only additive genetic variance, while the second (Q2) contributed 0.9 % additive and 0.1 % dominance. The two QTLs were placed in the middle (5,010th SNP) of chromosomes 1 and 2, respectively. Q1 had MAF of 0.10 while Q2 had MAF of 0.20. Finally, chromosome 3 housed no trait contributing loci and it was used to gauge the false positive rate (FPR) of the methods.

For each simulating scenario, I generated genotypes and phenotypes for a total of 500 replicates. I considered three different values for the sample size: 500, 1,000, and 2,000 unrelated individuals. Every replicate was analyzed twice: once using the CSI approach and once using the "fastAssoc" option on Merlin [1]. For the CSI approach, 95 % confidence sets (CSs) were obtained for loci contributing at least 1 % of the total variance of the trait, that is, I set $h_0 = 0.01$, which corresponds to the actual contribution of each of the loci under the simulating model. The overall variance

Table 1 Simulation results using as threshold of the genetic contribution $h_0 = 0.01$, which reflects the actual contribution of each of the trait loci to the total variance under the simulation models

N_P^b	Chromosome 1[a]			Chromosome 2[a]		
	TPR[c]	$\bar{N}^d$	s_N^d	TPR[c]	$\bar{N}^d$	s_N^d
500	0.954	5,562.5	65.8	0.938	5,565.2	62.8
1,000	0.960	1,298.9	47.9	0.930	1,300.7	51.2
2,000	0.942	49.5	8.4	0.930	50.6	9.2

[a]QTL Q1 located on chromosome 1 contributes 1 % to the total variance, all due to additive effects, while QTL Q2 located on chromosome 2 contributes 1 % to the total variance, 0.9 % of which are due to additive genetic effects and 0.1 % are due to dominance genetic effects
[b]Number of unrelated individuals in the study
[c]Percentage of replicates, out of 500, that the 95 % confidence set included the trait locus
[d]Observed mean ($\bar{N}$) and standard deviation (s_N) of the number of SNPs included in the 95 % CS

of the trait needed for the construction of the confidence sets for each replicate was obtained from the data themselves as explained earlier, i.e., by fitting the model in (2). Finally, the MAF for each marker was also estimated from the sample data.

The analysis results are displayed in Table 1 and Fig. 1. Table 1 summarizes the observed coverage probabilities of the 95 % confidence sets for each chromosome. Columns $\bar{N}_S$ and s_{N_S} report the average and the standard deviation, respectively, of the number of SNPs included in the resulting CS from all 500 replicates. The column labeled "TPR" represents the locus specific discovery rate, which is defined as the proportion of the replicates, out of the 500, that the resulting CS included the particular simulated QTL.

From the table we can see that the observed coverage probability of the 95 % confidence sets on chromosome 1, which harbored locus Q1 that contributed only additive variance to the trait, matched very well the nominal one. Hence, it appears

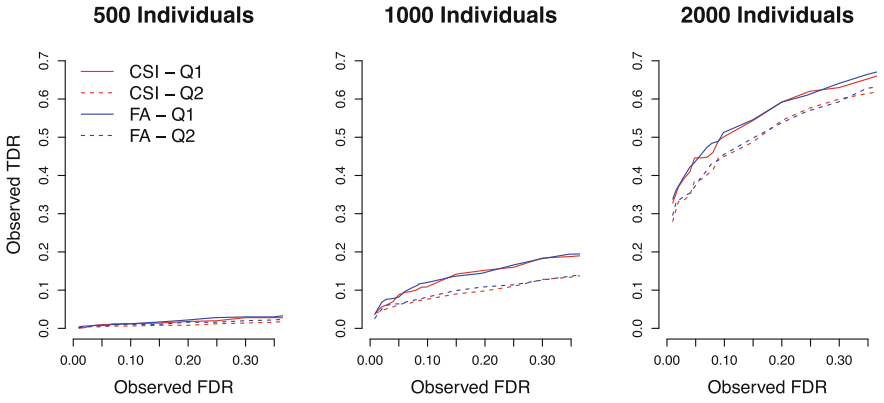


Fig. 1 Observed locus specific true positive rates (TPRs) of QTLs Q1 and Q2 for the CSI and the FA methods as functions of the corresponding observed false positive rates (FPRs)

that estimating the total variance as well as the MAF of the SNPs from the data themselves bears no significant effect on the coverage of the resulting sets. The observed coverage for locus Q2, which contributed both additive and dominance effects, tended to be slightly less than the nominal one, 95 %, as it ranged between 93 and 93.8 %. Thus, it seems that, even if the additivity between the marker alleles is mildly violated, the coverage of the resulting CSs is not severely affected.

On average, regardless of the sample size, the resulting CSs included a significant number of non-trait contributing loci. For example, even with 1,000 people in the sample, the CSs included about 1,300 SNPs per chromosome out of the total 10,000 residing on each of them. This was expected as the threshold for inclusion of a SNP in the CS was set very low, 1 % of the total variance. As such, the method was not powered enough by the given sample size to control false discoveries. When the sample size was increased to 2,000 individuals, though, the number of false discoveries was significantly reduced, as the CSs included about 50 SNPs on each chromosome, on average, underlining the ability of the method to localize trait regulating loci.

Figure 1 displays the observed locus specific TPR for both the CSI and the analysis using Merlin, denoted as FA, as functions of the observed FPR for various thresholds of significance. The FPR for the CSI approach was defined as the proportion of replicates for which the resulting CS included *at least* one SNP on chromosome 3 that harbored no loci regulating the phenotype. For the analyses using the FA, the locus specific TPR was computed as the proportion of replicates, out of the 500, for which the p-value of the specific QTL was less than the predetermined threshold α. Similarly, the FPR for the FA approach was defined as the proportion of replicates for which the analysis resulted in *at least* one SNP on chromosome 3 with a p-value less than the chosen threshold.

Clearly, both methods behave very similarly having almost identical TPRs across all sample sizes, when they have the same FPR. However, CSI has the advantage that it yields CSs of markers with known statistical properties. As such, when the researcher applies the CSI method using a certain threshold h_0 for the contribution of a locus to the heritability of the trait, he/she can be sure that, if there is one, it will be identified by the method with a predetermined confidence.

4 Discussion

Two-stage analyses in genetic studies have become increasingly common. First, a single marker GWAS based on a dense genetic map is performed to identify a small subgroup of SNPs whose effect is worth further investigation. A follow up analysis is then performed on only the few selected SNPs using more elaborate, and usually more powerful, analyses. By design, the power of the two-stage strategies rely heavily on the ability of the first stage analysis to forward to the second stage loci that truly play a role in the regulation of the trait. I have described a Confidence Set Inference association approach that can be used at stage I to identify such SNPs. The method gives the researcher the flexibility to target QTLs with specific contribution to the

variance of the trait with a fixed predetermined coverage probability also selected a priori by the researcher. Thus, the CSI method is an ideal tool for use at the screening stage of two-step analysis designs, as it can significantly reduce the number of SNPs analyzed at the second stage while at the same time it can guarantee, with high confidence, that any QTL with certain contribution to the trait will be identified for further investigation.

I performed a simulation study to investigate the performance of the method under a variety of relevant factors that can have an effect on its ability to identify true QTLs. Specifically, I studied the effects that sample size, estimation of the MAF of the markers from the data, and estimation of the overall variance of the trait have on the TPR and FPR of the method. The results demonstrated that estimating the necessary quantities for the implementation of the method from the data themselves seems to have a negligible effect on the behavior of the method, as long as the threshold h_0 used in the construction of the CS accurately reflects the contribution of the QTL to the trait. When the value of h_0 is higher than the actual contribution, the actual coverage of the resulting confidence set is expected to be less than its nominal one. How far off the true coverage would be from the nominal would depend on a host of factors such as sample size and the difference between the actual and the selected value of the threshold h_0; the larger the difference between the two values, the more severe the effect on the coverage of the resulting interval. The sample size can ameliorate or exacerbate this effect. Generally speaking, smaller sample sizes will allow the method to maintain coverage close to the nominal, while larger samples will significantly reduce it.

Another important factor that can affect the performance of the CSI method is the mode of interaction of the alleles at the trait locus. CSI assumes that the trait locus only contributes to the additive component of the trait variance. The simulation results indicate that, even if this assumption is (mildly) violated, the confidence sets still maintain a true coverage probability very close to the nominal one. Thus, deviations from the additivity assumption seem to bear a minimal effect on the performance of the method.

In comparing the CSI approach to the standard association method implemented on Merlin [1], we saw that both methods seem to perform similarly, having almost identical powers when their FPRs were set to be the same. However, CSI emerges as a more attractive alternative to the standard method as it has the advantage that it produces CSs of SNPs with known statistical properties. Furthermore, it allows the researcher to control the probability of capturing a specific locus on the GWAS step of the analysis, thereby ensuring its advancement to the follow up analysis, and ultimately increasing its chances of being discovered.

The FPR of the CSI method depends on the choice of the threshold h_0 for the contribution of the putative QTL on the total variance of the quantitative phenotype. Large values of h_0 will lead to a lower FPR, while smaller values to a higher FPR. Thus, in any given situation, the optimal choice of the threshold will depend on the available sample size. To circumvent the need to specify the value of the threshold h_0, one can follow the practical approach described by Papachristou and Lin [9]. In short, for each SNP on the map one can use the test statistic T_m in (5) to compute an

upper confidence bound for its contribution to the phenotype with a given confidence, say 95 %. Then, the SNPs can be ranked based on these confidence bounds and the top ones can be selected for additional analysis. The research, then, based on the available resources, can determine the exact number of SNPs to be advanced to the next stage for further exploration.

Finally, for the simulations, I considered only SNPs with relative common minor allele. This was done to avoid potential problems with unstable estimates during the maximization process, especially because the sample sizes were fairly small. With larger sample sizes, one would be able to utilize markers with small MAFs. Nevertheless, the CSI principles can be readily applied to methods specifically designed for rare variants. The only requirement is the ability to derive the distribution of the test statistic under the non-traditional null hypothesis tested by the CSI approach, which may not be a trivial task.

5 Software

The software package CSI-QTL for implementing the described method is freely available at http://code.google.com/p/papachristou-free-genetics-software/downloads/list.

References

1. Abecasis, G.R., S.S. Cherny, W.O. Cookson, and L.R. Cardon. 2002. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 30: 97–101.
2. Biswas, S., S. Xia, and S. Lin. 2014. Detecting rare haplotype-environment interaction with logistic Bayesian LASSO. *Genetic Epidemiology* 38(1): 31–41.
3. Bull, S.B., S. John, and L. Briollais. 2005. Fine mapping by linkage and association in nuclear family and case-control designs. *Genetic Epidemiology* 29(Suppl 1): 48–58.
4. Ding, X., S. Su, K. Nandakumar, X. Wang, and D.W. Fardo. 2014. A 2-step penalized regression method for family-based next-generation sequencing association studies. *BMC Proceedings* 8(Suppl 1): S25.
5. Kraja, A.T., R. Culverhouse, E.W. Daw, J. Wu, A. Van Brunt, M.A. Province, et al. 2009. The genetic analysis workshop 16 problem 3: Simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham heart study. *BMC Proceedings* 3(Suppl 7): S4.
6. Leal, S.M., K. Yan, and B. Muller-Myhsok. 2005. SimPed: A simulation program to generate haplotype and genotype data for pedigree structures. *Human Heredity* 60: 119–122.
7. Murphy, A., S.T. Weiss, and C. Lange. 2010. Two-stage testing strategies for genome-wide association studies in family-based designs. *Methods in Molecular Biology* 620: 485–496
8. Papachristou, C., and S. Lin. 2006. A comparison of methods for intermediate fine mapping. *Genetic Epidemiology* 30: 677–689.
9. Papachristou, C., and S. Lin. 2012. A confidence set inference method for identifying SNPs that regulate quantitative phenotypes. *Human Heredity* 73(3): 174–183.

10. Papachristou, C., C. Ober, and M. Abney. (in press). A Lasso penalized regression approach for genome-wide association analyses using related individuals: Application to the Genetic Analysis Workshop 19 simulated data. *BMC Proceedings*
11. Skol, A.D., L.J. Scott, G.R. Abecasis, and M. Boehnke. 2007. Optimal designs for two-stage genome-wide association studies. *Genetic Epidemiology* 31(7): 776–788.
12. Wason, J.M., and F. Dudbridge. 2012. A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. *American Journal of Human Genetics* 90(5): 760–773.
13. Yang, H.H., N. Hu, P.R. Taylor, and M.P. Lee. 2008. Whole genome-wide association study using affymetrix SNP chip: A two-stage sequential selection method to identify genes that increase the risk of developing complex diseases. *Methods in Molecular Medicine* 141: 23–35.
14. Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44(7): 821–824.
15. Zhu, M., and S. Zhao. 2007. Candidate gene identification approach: Progress and challenges. *International Journal of Biological Sciences* 3(7): 420–427.

Statistical Inference on Three-Dimensional Structure of Genome by Truncated Poisson Architecture Model

Jincheol Park and Shili Lin

Abstract In recent years, next generation sequencing technology, coupled with an assay that is capable of detecting genome-wide chromatin interactions, has produced a massive amount of data and led to a greater understanding of long-range, or spatial, gene regulation mechanisms. Hence, the traditional one-dimensional linear view of a genome, which is especially prevalent in statistical and mathematical modeling, is inadequate in many genomic studies. Instead, it is essential, in studying genomic functions, to estimate the three-dimensional (3D) structure of a genome. The availability of genome-wide interaction data necessitates the development of analytical methods to recover the underlying 3D spatial chromatin structure, but challenges abound. One particular issue is the excess of zeros, especially with higher resolution, or inter-chromosomal, data. This leads to questions concerning the appropriateness of using the Poisson distribution to model such data. In this article, we introduce a truncated Poisson Architecture Model (tPAM) to directly model sequencing counts with many zeros. We carried out an extensive simulation study to evaluate tPAM and to compare its performance with an existing method that uses the Poisson distribution to model the counts. We applied tPAM to reconstruct the underlying 3D structures of two data sets, one of human and one of mouse, to demonstrate its utility. The analysis of the human data set considered chromosomes 14 and 22 jointly, thereby illustrating tPAM's capability of analyzing inter-chromosomal data. On the other hand, the mouse analysis was focused on a region on chromosome 2 to evaluate tPAM's performance for recovering structure with loci in different topologically associated domains.

Keywords Spatial interactions · Hi-C · Excess of zeros · Chromatin looping · Data resolution

J. Park
Department of Statistics, Keimyung University, 1095 Dalgubeol-daero,
Daegu 704-701, South Korea
e-mail: park.jincheol@gw.kmu.ac.kr

S. Lin (✉)
Department of Statistics, Ohio State University, 1958 Neil Avenue,
Columbus 43210, USA
e-mail: shili@stat.osu.edu

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3_15

1 Introduction

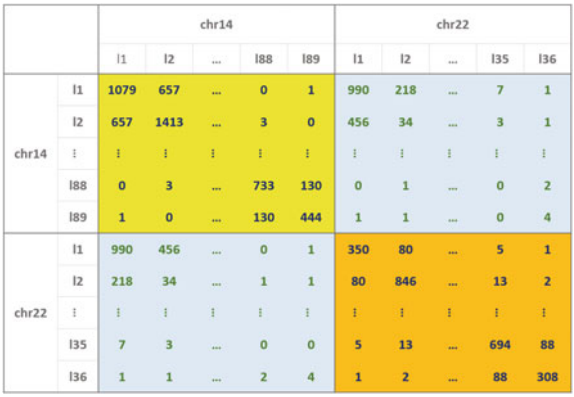
The spatial (three-dimensional, or 3D) organization of a genome is closely linked to its biological function, and thus, full understanding of the genomic structure is essential. In recent years, the ability to identify long-range chromatin interactions genome-wide, known as looping, aided by next generation sequencing technology, has been truly revolutionary in genomic and epigenetic research. The most well-known assay for detecting chromatin interaction, Hi-C [14], produces a library of products that are pairs of fragments in close proximity to each other in the cell nucleus but may be far apart in terms of their chromosomal locations (and may even be on different chromosomes). The library is then analyzed through massively parallel DNA sequencing, producing a catalog of interacting fragments that can be organized into a two-dimensional matrix (known as a contact matrix) of contact counts. Figure 1 provides an example of a contact matrix for chromosomes 14 and 22 based on data from [14], showing only some of the contact counts for illustration purposes. In addition to Hi-C, other assays for detecting genome-wide long-range interactions have also been developed, such as ChIA-PET [6] and TCC [12].

Despite spectacular advances in molecular technologies that allow for unprecedented identifications of genome-wide chromatin interactions, our understanding of 3D organization of genomes is still coarse and incomplete, especially for complex organisms such as humans and mice. This is partly due to the massive amount of data that prove to be extremely difficult to analyze. In addition to its size, the features of the data also pose challenges, rendering conventional statistical methods ineffective. To tackle these issues, analytical approaches have been proposed to understand the spatial organization of the genome based on Hi-C long-range looping data. The approaches can be classified into optimization-based and modeling-based.

For optimization-based approaches, the idea is to first translate each pairwise contact count into a distance using a biophysical property. One then obtains a consensus 3D structure by minimizing some objective function, such as the total "differences" between the translated distances and those inferred from the hypothesized 3D architecture [1, 4, 5, 13, 17, 21]. Many of the optimization methods are based on metric or non-metric multi-dimensional scaling [2, 4, 17]. For this type of approach, normalization of the data is key [11].

Modeling-based approaches, on the other hand, are all based on probability models that describe the relationship between the contact counts with the 3D physical distance. The contact counts are modeled either by a normal distribution to account for variability in the estimation [16] or by a Poisson distribution [10, 18] with its intensity parameter assumed to be related to the physical distance by an inverse relationship. Statistical inferences on the 3D structure (together with other model parameters) are made either by maximum likelihood [18] or through casting the problem into a Bayesian framework [10, 16].

As discussed earlier, a Hi-C experiment produces contact counts that are organized as a 2D matrix for a given resolution. For example, the data matrix shown in Fig. 1 is based on a 1 Mb (megabases) resolution. If there is sufficient sequencing depth,



		chr14					chr22				
		l1	l2	...	l88	l89	l1	l2	...	l35	l36
chr14	l1	1079	657	...	0	1	990	218	...	7	1
	l2	657	1413	...	3	0	456	34	...	3	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	l88	0	3	...	733	130	0	1	...	0	2
	l89	1	0	...	130	444	1	1	...	0	4
chr22	l1	990	456	...	0	1	350	80	...	5	1
	l2	218	34	...	1	1	80	846	...	13	2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	l35	7	3	...	0	0	5	13	...	694	88
	l36	1	1	...	2	4	1	2	...	88	308

Fig. 1 Contact matrix of Hi-C data. The two *diagonal blocks* correspond to intra-chromosomal contacts among loci in chromosome 14 and 22, respectively, while the two *off-diagonal blocks* depict inter-chromosomal contacts between loci in chromosomes 14 and 22. Note that the matrix is symmetric

a higher resolution matrix can lead to a finer and more useful 3D structure, but there tends to be more zero entries in the contact matrix, rendering the Poisson distribution inadequate for modeling the data. To remedy the problem, in this paper, we propose a truncated Poisson Architecture Model (tPAM) by using a truncated Poisson distribution without the zero counts. We carried out an extensive simulation study to evaluate tPAM and to compare its performance with an existing method [10] that uses the Poisson distribution to model the counts. We applied tPAM to reconstruct the underlying 3D structures of two data sets, one of human and one of mouse, to demonstrate its utility. The analysis of the human data set considered chromosomes 14 and 22 jointly, thereby illustrating its capability of analyzing inter-chromosomal data. On the other hand, the mouse analysis was focused on a region on chromosome 2 to evaluate tPAM's performance for recovering a structure with loci in different topologically associated domains (TADs).

2 Methods

2.1 The tPAM Model

Consider a set of n fragments (also referred to as loci), each being represented by a point in the 3D space. Collectively, they are denoted by $\Omega \equiv \{\mathbf{p}_i = (p_i^x, p_i^y, p_i^z);\ i = 1, \ldots, n\}$. Let d_{ij} denote the Euclidean distance between loci i and j, that is,

$$d_{ij} = \sqrt{(p_i^x - p_j^x)^2 + (p_i^y - p_j^y)^2 + (p_i^z - p_j^z)^2}. \tag{1}$$

The contact counts of these n loci are organized into a 2D matrix, with y_{ij} denoting the contact count (the (i, j) entry of the matrix), which represents the interaction intensity between loci i and j. Based on these data ($\mathbf{y} = \{y_{ij}, 1 \le i < j \le n\}$; note

that the matrix is symmetric), the goal is to make inference about the coordinates, Ω, of the 3D structure.

We assume that the contact counts follow a truncated Poisson distribution, with its intensity parameter linked to the 3D distance and other covariates through a log-linear model. More specifically, the Poisson model was built under the assumption that two loci in close proximity in 3D space are likely to interact more, which leads to the following model for the Poisson intensity parameter λ_{ij}:

$$\log \lambda_{ij} = \alpha_0 + \alpha_1 \log d_{ij} + \mathbf{x}_{ij}^T \beta, \tag{2}$$

where $\mathbf{x}_{ij}^T = (x_{ij}^1, \ldots, x_{ij}^K)$ and $\beta = (\beta_1, \ldots, \beta_K)^T$ denote the vector of K covariates and its associated vector of coefficients, respectively. Typical covariates include GC content, fragment length, mappability score, and potentially also restriction enzyme to take care of systematic bias and to normalize data [10, 20]. Under the assumption that the physical 3D distance between two loci is inversely related to the contact counts [14], the restriction of $\alpha_1 < 0$ is imposed in the model.

Letting θ denote the collection of all model parameters, we have the following log-likelihood function:

$$\log p(\mathbf{y}|\theta, \Omega) \propto \sum_{(i,j)\in\mathscr{I}} \sum \left\{ y_{ij} \log \lambda_{ij} - \log(e^{\lambda_{ij}} - 1) \right\}, \tag{3}$$

where $\mathscr{I}$ denotes the index set of non-zero contact counts, that is, $\mathscr{I} = \{(i, j); y_{ij} \neq 0, 1 \leq i < j \leq n\}$. This model, which excludes the zero contact counts, is referred to as the truncated Poisson Architecture Model (tPAM).

We remark that model (2) suffers from non-identifiability because the estimated structure, $\hat{\Omega}$, is not invariant to scale, rotation, reflection, and translation. To resolve this issue, without loss of generality, we can fix α_0 to be an arbitrarily predefined quantity. Note that α_0 controls the scale of the 3D structure, thus fixing α_0 will effectively lead to the structure being estimated only up to a scale. However, this is not an issue since the relative distance does not affect the predicted structure and its correlation with genomic functions [21]. Following [10], we further place the following restrictions on Ω to make it estimable, as four conditions on the structure are sufficient to uniquely determine the 3D structure: $\mathbf{p}_1 = (0, 0, 0)$, $\mathbf{p}_2 = (p_2^x, 0, p_2^z)$ with $p_2^z > 0$, $\mathbf{p}_3 = (p_3^x, p_3^y, p_3^z)$ with $p_3^y > 0$, and $\mathbf{p}_n = (p_n^x, 0, 0)$ with $p_n^x > 0$.

2.2 MCMC Procedure for Parameter Estimation

To make inferences about the 3D coordinates, we devise a Markov chain Monte Carlo (MCMC) sampling procedure as follows. We write the posterior distribution of Ω (main parameters of interest), together with nuisance parameters θ, as

$$p(\Omega, \theta|\mathbf{y}) \propto p(\mathbf{y}|\Omega, \theta) p(\Omega) p(\theta). \tag{4}$$

The first component of Eq. (4) corresponds to the likelihood as given in (3), that is,

$$p(\mathbf{y}|\Omega, \theta) = \prod_{(i,j)\in\mathscr{I}} \prod \mathscr{L}_P\{\lambda_{ij}(\Omega, \theta)\}, \tag{5}$$

where $\mathscr{L}_P(.)$ denotes the zero-truncated Poisson distribution and

$$\lambda_{ij}(\Omega, \theta) = \exp\left(\alpha_0 + \alpha_1 \log d_{ij} + \mathbf{x}_{ij}^T \beta\right). \tag{6}$$

The remaining parts of (4) describe the distributions for $\mathbf{p}$ and θ, which are assigned non-informative priors: $p(\Omega) \propto 1$, $p(\alpha_1) \propto I(\alpha_1 < 0)$, and $p(\beta) \propto 1$.

To accommodate the estimable conditions imposed on Ω, we consider an isometric transformation, with details provided in Appendix A. To sample from the posterior distributions of θ, we use Metropolis-Hastings algorithms, and in particular the Gibbs sampler whenever the conditional distribution of a parameter is of a commonly known one. In sampling the posterior of Ω, we employ Hamiltonian MCMC to more effectively handle the high correlations among the samples [7]. In the following, we briefly describe the updating schemes. Let ϑ denote the current estimates of (Ω, θ) at iteration t, and ϑ_{-a} denote ϑ without the element a.

- Updating of α_1.
  We base on the current α_1^t to sample a candidate α_1^* from proposal distribution $J_\alpha(\alpha_1^*|\alpha_1^t)$, a normal distribution with mean α_1^t and predefined proposal $\sigma_{\alpha_1}^2$, and calculate the ratio of the densities

$$r = \frac{p(\alpha_1^*|\mathbf{y}, \vartheta_{-\alpha_1})}{p(\alpha_1^t|\mathbf{y}, \vartheta_{-\alpha_1})}, \tag{7}$$

  where $p(\alpha_1^*|\mathbf{y}, \vartheta_{-\alpha_1}) \propto p(\mathbf{y}|\vartheta_{-\alpha_1}, \alpha_1^*)$. Accept α_1^* as α_1^{t+1} with probability equal to $\min(r, 1)$; otherwise $\alpha_1^{t+1} = \alpha_1^t$.
- Updating of $\beta_k, k = 1, \ldots, K$.
  We base on the current β_k^t to sample a candidate β_k^* from proposal distribution $J_\beta(\beta_k^*|\beta_k^t)$, a normal distribution with mean β_k^t and predefined proposal σ_β^2, and calculate the ratio of the densities

$$r = \frac{p(\beta_k^*|\mathbf{y}, \vartheta_{-\beta_k})}{p(\beta_k^t|\mathbf{y}, \vartheta_{-\beta_k})}, \tag{8}$$

  where $p(\beta_k^*|\mathbf{y}, \vartheta_{-\beta_k}) \propto p(\mathbf{y}|\vartheta_{-\beta_k}, \beta_k^*)$. Accept β_k^* as β_k^{t+1} with probability equal to $\min(r, 1)$; otherwise $\beta_k^{t+1} = \beta_k^t$.
- Updating of Ω.
  Based on an analogy with physical systems, Hamiltonian Monte Carlo introduces an additional parameter vector $\mathbf{v}_i = (v_i^x, v_i^y, v_i^z)^T$ corresponding to parameter $\mathbf{p}_i$ and updates both of them together in a new Metropolis-Hastings algorithm. Specifically, we use Hamiltonian functions defined by $H(\mathbf{p}_i, \mathbf{v}_i) = U(\mathbf{p}_i) +$

$K(\mathbf{v}_i)$, where $U(\mathbf{p}_i)$, a potential energy, is assigned $-\log\{p(\mathbf{p}_i|\mathbf{y}, \vartheta_{-\mathbf{p}_i})\}$, while $K(\mathbf{v}_i)$, a kinetic energy, is defined as $\mathbf{v_i}^T\mathbf{v_i}/2$. Then we consider the following joint density of $(\mathbf{p}_i, \mathbf{v}_i|\mathbf{y}, \vartheta_{-\mathbf{p}_i})$ using the Hamiltonian function $H(\mathbf{p}_i, \mathbf{v}_i)$:

$$p(\mathbf{p}_i, \mathbf{v}_i|\mathbf{y}, \vartheta_{-\mathbf{p}_i}) \propto \exp\{-H(\mathbf{p}_i, \mathbf{v}_i)\} = \exp\{-U(\mathbf{p}_i)\}\exp\{-K(\mathbf{v}_i)\}. \quad (9)$$

Hamiltonian MCMC then proceeds in three stages. First, we sample random auxiliary variables v_i^x, v_i^y, and v_i^z from $N(0, 1)$. Then we simultaneously update $(\mathbf{p}_i, \mathbf{v}_i)$ to obtain a proposal vector $(\mathbf{p}_i^*, \mathbf{v}_i^*)$ using a leapfrog method (see Appendix B). In the last stage, we accept the proposed vector $(\mathbf{p}_i^*, \mathbf{v}_i^*)$ using the Metropolis-Hastings method where the ratio is given by

$$r = \exp\{-H(\mathbf{p}_i^*, \mathbf{v}_i^*) + H(\mathbf{p}_i, \mathbf{v}_i)\}. \quad (10)$$

Accept $\mathbf{p}_i^*$ as $\mathbf{p}_i^{t+1}$ with probability $\min(r, 1)$; otherwise $\mathbf{p}_i^{t+1} = \mathbf{p}_i^t$.

3 Application to Two Hi-C Datasets

We demonstrate the utility of tPAM by applying it to two Hi-C datasets. The application to the first dataset illustrates tPAM's ability of analyzing inter-chromosomal data with many zero contact counts. Its performance is also evaluated by comparing the structure inferred to distances obtained from limited experimental validation data. The second application aims to explore how tPAM performs with modularized structures, the TADs, also known as topological domains [3].

3.1 Human Lymphoblastoid Cell Line Hi-C Data

We applied tPAM to the Hi-C data produced by [14]. In fact, there are two Hi-C experiments performed on the same karyotypical normal human lymphoblastoid cell line, which are combined into a single data set in our analysis given their high reproducibility [14]. We focused on chromosome 14 and 22, as experimental validation data based on Fluorescence In Situ Hybridization (FISH) are available for several loci on these two chromosomes and are publicly available [14]. Specifically, [14] discussed interesting features of spatial interactions, based on the FISH measures, among 4 loci on chromosome 14 (L_1, L_2, L_3, and L_4, located in that linear order) and 4 loci on chromosome 22 (L_5, L_6, L_7, and L_8, in that linear order) using the FISH experiment. In particular, the spatial 3D distance between L_2 and L_4 was observed by FISH experiments to be smaller than that between L_2 and L_3, despite the fact that L_2 is farther apart from L_4 than from L_3 in terms of their linear 1D distances. A similar observation was made for (L_6, L_7, L_8), in that the spatial 3D distance between L_6

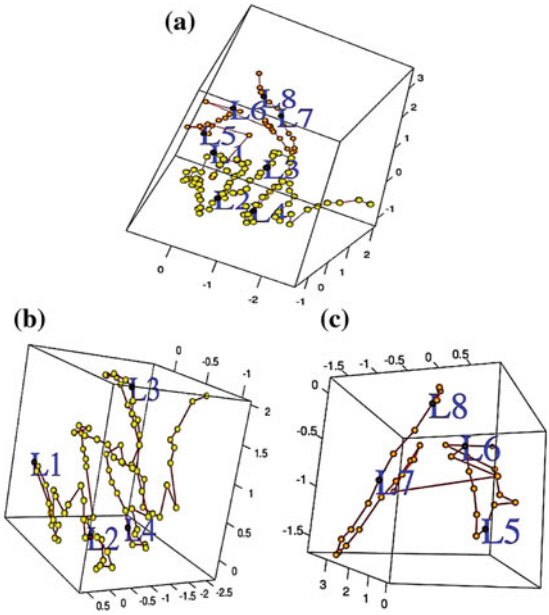


Fig. 2 Reconstructed 3D structure of chromosomes 14 and 22. **a** Joint 3D structure of chromosomes 14 and 22, with each loci marked by a ball, among them positions of L_1 through L_8 are labeled and marked by *black balls*; **b** 3D structure of chromosome 14, with a different orientation than that of the joint structure for better visualization; **c** 3D structure of chromosome 22, with a different orientation than that of the joint structure for better visualization. These figures were drawn using the R package 'rgl'

and L_8 is significantly smaller than that between L_6 and L_7. The resolution used is 1 Mb, which leads to 89 loci in chromosome 14 and 36 loci in chromosome 22.

We ran the MCMC procedure for 1.1×10^6 iterations, with the first 10^5 iterations for burn-in and the remaining 10^6 iterations for obtaining 10,000 posterior samples after thinning. The convergence of the posterior samples was confirmed by several diagnostic statistics, including those developed by [8, 9, 15]. The 3D structure identified by tPAM is given in Fig. 2a. For a better visualization of the structure in each of the chromosomes, we also provide Fig. 2b, c with different orientations. We can see from these figures that, indeed, L_2 and L_4 are much closer in terms of their spatial distance compared to L_2 and L_3, and L_6 and L_8 are closer compared to L_6 and L_7. These observations are consistent with the results of [14] that the pairs of (L_2, L_4) and (L_6, L_8) are brought to close proximity through chromatin looping.

To further evaluate the performance of tPAM, we compare its estimates of pairwise distances to those of FISH, the gold standard measurements. To make it possible to compare due to scale differences (recall we set α_0 arbitrarily), we first calculated a unitless distance $\tilde{d}(L_i, L_j)$ by dividing each distance $d(L_i, L_j)$ by the median distance between L_3 and L_4 (the largest distance among all pairs). Note that the median is taken over 100 measurements for FISH and 10,000 estimates for tPAM. The results, given in Fig. 3, show that the tPAM estimates agree well with the FISH measurements. In fact, the FISH measurements (100 measures for each pair) are much more variable compared to the tPAM estimates, as evident from the larger

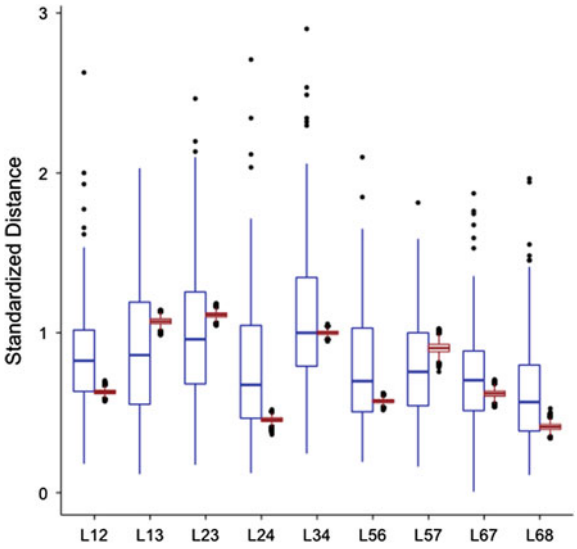


Fig. 3 Assessment of performance of tPAM in comparison with FISH measurements. For each pair of loci for which FISH measurements are available, boxplots are used to summarize the results for the 100 FISH measurements (*left box*) and 10,000 tPAM estimates (*right box*)

boxes, longer whiskers, and existence of outliers in the boxplots. The results also confirm that the distance between L_2 and L_4 is indeed smaller than that between L_2 and L_3 or L_3 and L_4, and L_6 is located closer to L_8 than to L_7.

3.2 Mouse Embryonic Stem Cell Hi-C Data

We applied tPAM to a mouse embroyonic stem cell line [3] generated at 40 Kb resolution (i.e. interaction frequencies are available for regions of 40 Kb in length). We used the bias-corrected Hi-C count data directly, as libraries of factors that are known to cause systematic biases are not available to us. In particular, we focused on the segment of chromosome 2 from base pair (bp) 73720001 to bp 75440000, as this segment is believed to contain two TADs [3]. Loci within the same domain interact with each other much more than across domains, and thus the two domains should be well separated in 3D space. The data based on a 40 Kb resolution lead to a contact matrix of dimensions 43 by 43. Application of tPAM yielded the estimated 3D structure depicted in Fig. 4. We can see, from the figure, that the 19 loci within the segment from bp 73720001 to bp 74480000 are located close to one another in 3D space (red balls), whereas the remaining 24 loci within the segment from bp 74480001 to bp 75440000 make up the other cluster (green balls) in 3D space. As it turns out, these two clusters of loci do correspond to the two TADs discussed in [3]. In MCMC sampling, 3×10^5 and 7×10^5 iterations were executed respectively for burn-in and statistical inference. Thinning resulted in 10,000 posterior samples for structure estimation. Convergence of the sample was confirmed by the diagnostic measures described in Sect. 2.

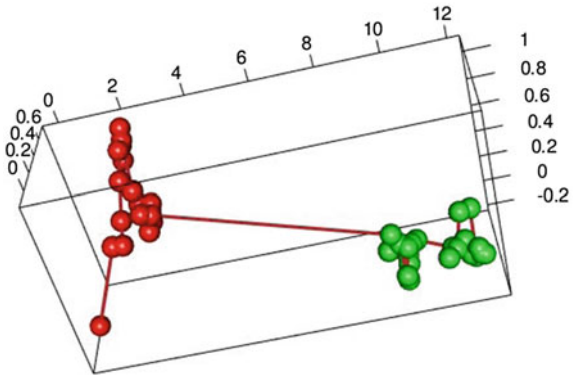


Fig. 4 Reconstructed 3D structure of mouse data. Loci within the two topological domains are denoted by two different colors

4 Simulation Study

As we can see from the analysis results of the human Hi-C data, the inferred 3D structure from tPAM leads to consistent results with FISH experimental data. Nevertheless, the aptness of the 3D structure as a whole was not adequately assessed due to the limited number of loci involved in the FISH experiment. Similarly, although the analysis of the Hi-C mouse data yielded results that support the concept of compartmentalization of a chromosome [3, 14], the within compartment (domain) organization was not assessable. Therefore, to more fully evaluate the performance of tPAM, we conducted a simulation study in this section using two underlying 3D structures, which will serve as the "gold standard". We further compared the performance of tPAM with BACH, a Bayesian inference method proposed by [10] based on the Poisson model. The simulation settings and results are presented in two subsections below, but we first describe several assessment criteria for comparing the performances between tPAM and BACH.

4.1 Performance Assessment

We consider three criteria to assess the performance of the methods. The first is the overall goodness of fit of a model by comparing the observed with their predicted values from the model. More specifically, our measure is the Pearson χ^2 goodness of fit statistic, which is given by

$$\chi^2 = \sum_{(i,j)\in\mathscr{I}} \sum \frac{(y_{ij} - \hat{\lambda}_{ij})^2}{\hat{\lambda}_{ij}} / n(\mathscr{I}), \tag{11}$$

where $\mathscr{I}$ is the index set denoting all non-zero contact counts as defined in Sect. 2 and $n(\mathscr{I})$ denotes a size of the set $\mathscr{I}$.

Given that, in our simulation, the underlying structure is known, we can also devise two other criteria that make use of the true underlying distance between a pair of loci. Recall that the structure estimated is accurate up to a scaling factor, γ, which is estimated by the least squares model as follows:

$$\hat{\gamma} = \arg\min_{\gamma} \left\{ \sum_{1 \le i < j \le n} \sum (d_{ij} - \gamma \hat{d}_{ij})^2 \right\}. \tag{12}$$

Note that, as mentioned above, the fact that tPAM or BACH can only estimate the structure up to a scale is not an issue, because the relative distance does not affect the predicted structure nor its correlation with genomic functions [21]. After scaling the estimated structure $\hat{\Omega}$ by the factor estimate $\hat{\gamma}$, we can compare the true structure with the estimated structure after appropriate isometric transformation. This leads to the proposal of the following two measures:

$$\mathscr{D}_{mean} = \frac{1}{n} \sum_{i=1}^{n} \frac{||\mathbf{p}_i - \hat{\gamma}\hat{\mathbf{p}}_i||}{\bar{d}_{\mathbf{p}}} \times 100 \tag{13}$$

$$\mathscr{D}_{max} = \max_{1 \le i \le n} \frac{||\mathbf{p}_i - \hat{\gamma}\hat{\mathbf{p}}_i||}{\bar{d}_{\mathbf{p}}} \times 100, \tag{14}$$

where $\bar{d}_{\mathbf{p}}$ is the average pairwise distance derived from the true underlying structure Ω. Thus, these two measures compute respectively the average- and the maximum-coordinate departure of loci (based on the estimated architecture) from the corresponding true ones (based on the true architecture). As we will see below, the true structures are being specified completely either based on the helix model or the estimated mouse model for the purpose of the simulation study.

4.2 Helix Structure

We consider a helix model with 50 loci. We chose this model for our first simulation as a helix structure has been used as a means of modeling chromatin in the statistical literature [19]. We denote the helix structure by $\Omega^h = \{\mathbf{p}_i, i = 1, \ldots, 50\}$. The 3D location of each locus, $\mathbf{p}_i = (p_i^x, p_i^y, p_i^z)$, is constructed as

$$p_i^x = \cos(\theta_i),\ p_i^y = \sin(\theta_i),\ p_i^z = L\theta_i/(2\pi), \tag{15}$$

where $L = 0.2$ and $\theta_i = \pi i/4$. To mimic real data, we also include three covariates, $\{x_{l,i}, x_{g,i}, x_{m,i}, i = 1, \ldots, 50\}$, to capture systematic bias, leading to the following simulation model:

$$\log \lambda_{ij} = \alpha_0 + \alpha_1 \log d_{ij} + \beta_l \log(x_{l,i} x_{l,j}) + \beta_g \log(x_{g,i} x_{g,j}) + \log(x_{m,i} x_{m,j}). \tag{16}$$

We set $\alpha_0 = 3.5$, and $\alpha_1 = -1.5, \beta_l = \beta_g = 0.3$ and simulated $x_{l,i} \sim \text{Unif}(0.2, 0.3)$, $x_{g,i} \sim \text{Unif}(0.4, 0.5)$ and $x_{m,i} \sim \text{Unif}(0.9, 1)$, where Unif(.) denotes a uniform distribution. To simulate the excess of zero situation in real data, we considered the following zero-inflated Poisson model:

$$\begin{aligned} P(Y_{ij} = 0) &= \pi + (1-\pi)e^{-\lambda_{ij}}, \\ P(Y_{ij} = y_{ij}) &= (1-\pi)\frac{\lambda_{ij}^{y_{ij}} e^{-\lambda_{ij}}}{y_{ij}!}, \ y_{ij} = 1, 2, \ldots. \end{aligned} \tag{17}$$

In other words, the above represents a mixture of a point mass at 0 and a Poisson distribution with intensity parameter λ_{ij}, with the mixing proportion being π. In our simulation, we considered four mixing proportions: $\pi = 0.0, 0.1, 0.2$, and 0.3. Note that the setting with $\pi = 0.0$ corresponds to the BACH model of [10] and as such, BACH is expected to perform well.

The results are presented in Table 1. In MCMC sampling, $10^5 \sim 10^6$ iterations were run for burn-in and an additional $10^6 \sim 2 \times 10^6$ iterations were executed for posterior sampling to obtain 10^4 realizations for inference after thinning. The convergences of the posteriors were confirmed by the diagnostics described in Sect. 2. As we can see from the table, across all three criteria, tPAM performs significantly better than BACH for the settings when $\pi \neq 0$. More specifically, tPAM yielded significantly smaller average and maximum relative departure from the true Ω^h (all p-values $<10^{-3}$ based on paired-t tests). This is to be expected as BACH, based on Poisson, cannot adequately accommodate the excess of zeros. We are also reassured to see that, even when $\pi = 0$, the underlying setting of BACH, tPAM still performs as well as BACH or may even be viewed as slightly better based on all three criteria. We can further observe that the results of tPAM are fairly consistent for different zero inflation proportions (i.e. similar values under the same criterion), demonstrating the robustness of tPAM to excess of zeros in the observed data, and hence data with different resolutions. In contrast, BACH's performance gets worse (with larger criterion value) as the inflation proportion becomes larger.

Table 1 Performance evaluation of tPAM and BACH with the Ω^h 3D structure

π	Model	$\mathscr{D}_{mean}$ (%)	$\mathscr{D}_{max}$ (%)	χ^2
0.0	BACH	26.37 (17.70)	63.99 (39.26)	1.04 (0.13)
	tPAM	23.70 (11.15)	60.11 (29.99)	0.98 (0.11)
0.1	BACH	39.14 (17.79)	96.66 (35.83)	2.03 (0.24)
	tPAM	23.65 (12.94)	57.12 (32.38)	0.98 (0.13)
0.2	BACH	61.07 (25.41)	140.84 (51.43)	3.94 (0.44)
	tPAM	25.79 (11.74)	59.96 (28.19)	0.95 (0.19)
0.3	BACH	62.65 (20.06)	142.05 (40.83)	7.16 (0.70)
	tPAM	26.49 (16.67)	65.56 (46.17)	0.88 (0.07)

Table 2 Performance evaluation of tPAM and BACH with the Ω^m 3D structure

π	Model	$\mathscr{D}_{mean}$ (%)	$\mathscr{D}_{max}$ (%)	χ^2
0.0	BACH	49.85 (5.14)	93.60 (7.43)	1.23 (0.04)
	tPAM	39.57 (7.55)	74.16 (15.08)	1.80 (0.76)
0.1	BACH	65.26 (11.20)	109.40 (15.63)	1.65 (0.17)
	tPAM	42.51 (9.45)	77.26 (14.93)	1.42 (0.56)
0.2	BACH	77.65 (13.52)	124.67 (19.56)	3.43 (0.36)
	tPAM	43.00 (8.63)	79.56 (15.25)	1.62 (0.71)
0.3	BACH	84.41 (15.36)	139.94 (23.14)	6.90 (0.88)
	tPAM	46.67 (20.76)	89.78 (45.03)	1.36 (0.52)

4.3 Mouse Model

Using the mouse structure $\hat{\Omega^m}$ and the $\hat{\alpha_1}$ value estimated by tPAM in Sect. 3.2, we let $\log \lambda_{ij} = 3 + \hat{\alpha_1} \log d_{ij}$, where d_{ij} is the pairwise distance inferred from the estimated structure $\hat{\Omega^m}$. We simulated datasets of $\{Y_{ij}\}$ from the zero-inflated Poisson model (17) with $\pi = 0.0, 0.1, 0.2$, and 0.3. In MCMC sampling, $7 \times 10^5 \sim 10^6$ iterations were run for burn-in, and afterward $5 \times 10^5 \sim 10^6$ iterations were run to obtain 10^4 realizations for inference after thinning. As with the helix simulation, the convergences of the posteriors were confirmed by the diagnostics described in Sect. 2. The results are given in Table 2, from which, one can see that tPAM clearly outperforms BACH for $\pi \neq 0$ (all p-values $\leq 10^{-4}$ based on paired-t tests), consistent with the results for the helix model. Similarly, when $\pi = 0.0$, the underlying model for BACH, tPAM is seen to perform just as well. The robustness of tPAM to the proportion of zero-inflation component, and the lack of such for BACH, is once again observed.

5 Conclusion and Discussion

The spatial organization of a genome has gained a great deal of continuing attention in recent years, as the structure is intimately linked to the biological functions of the genome, especially on long-range gene regulation. To turn experimental data into accurate estimates of spatial chromatin structures, a number of analytical methods have been proposed, including those that make use of the Poisson distribution to model the contact counts. Recognizing the sparsity of the contact matrix for inter-chromosomal interactions and with higher resolutions, in this paper, we propose a truncated Poisson model as a solution to accommodate this feature of data so that it is robust to resolution specification. Applications of tPAM to two existing data sets, one human and one mouse, illustrate its utility, as the results are consistent with those obtained from the limited FISH validation data. For the mouse data, with a 40

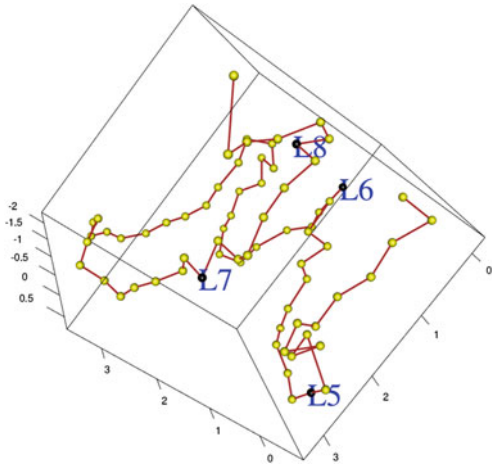


Fig. 5 Reconstructed 3D structure of chromosome 22 with 500 Kb resolution

Kb resolution, we see two clear TADs, reflecting chromatin long-range interaction in a "domain scale". Within each domain, with such an intermediate resolution, we can see looping within each domain, perhaps representing spatial interaction within a gene structure. For the human data, the analysis was performed at a 1 Mb resolution following the original analysis [14], which appears to capture the broad looping feature of chromatin organization, but fine scale looping within gene structures are largely unobserved. Inspired by the mouse data results with intermediate resolution, we carried out an additional analysis for constructing the 3D structure of chromosome 22 at a 500 Kb resolution. We observe that the result (Fig. 5) preserves the "domain level" looping, with locus L_6 still closer to L_8 than to L_7. Furthermore, the finer structure now also depicts more "local level" looping. Nevertheless, a more comprehensive study with even higher resolution is needed to study spatial interactions within gene structures, especially between promoters and enhancers.

Our simulation study, with two underlying structures, further substantiates the appropriateness of tPAM for analyzing Hi-C data, and more clearly showcases its ability to handle the sparsity of the contact matrix. The different mixing proportions in the zero-inflated model can be viewed as representing different resolutions, thus clearly demonstrating the robustness of tPAM to varying resolution level. This is in contrast to an existing method based on the Poisson model, in which one can see that the results are quite sensitive to the level of resolution: as the resolution gets finer and finer, the deviation from the "true" gets larger and larger for each of the evaluation criteria, compared to the stable feature of the tPAM values.

Computational feasibility is a major concern for genomic data, but the concern is even greater for chromatin interaction data as the size of the data is $O(n^2)$ when there are n genomic loci, an order of magnitude increase compared to analysis of linear chromosomal data. In this regard, tPAM has the added advantage as its computational cost is greatly reduced by excluding the zero counts. As such, higher resolution data, which lead to a much larger contact matrix (i.e. larger n), does not necessarily result

in more computational cost due to the sparsity nature of the matrix. In contrast, for methods based on the Poisson distribution, the computational cost increases with higher resolution data.

Acknowledgments This work was supported in part by the National Science Foundation grants DMS-1042946 and DMS-1220772, the National Institute of Health Grant IROIGM114142-01, and by the Bisa Research Grant of Keimyung University in 2014. This material was also based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendices

A. Isometric Transformation

To make Ω uniquely estimable, instead of incorporating the restrictions on Ω into prior, we employed a group of isometric (distance preserving) mappings. Suppose we sample Ω^t at iteration t. For simplicity, we let Ω denote the transformed one throughout the rest of this appendix.

Step 1. $\mathbf{p}_1 \to (0, 0, 0)$.
To place $\mathbf{p}_1^t$ at the origin $(0, 0, 0)$, we apply a translation operation $\mathscr{R}_\tau$ such that

$$\mathscr{R}_\tau : \mathbf{p}_i^t \to \mathbf{p}_i^t - \mathbf{p}_1^t. \tag{18}$$

Let $\Omega = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ be the translated architecture.

Step 2. $\mathbf{p}_n \to (p_n^x, 0, 0)$ with $p_n^x > 0$.

a. $\mathbf{p}_n \to (p_n^x, 0, p_n^z)$.
To place $\mathbf{p}_n$ on the xz-plane, we apply a rotation operation $\mathscr{R}_z^z$ with associated matrix R_z^z, clockwise-rotation matrix on $\mathbf{p}_n$ about the z-axis, sending it to the xz-plane:

$$R_z^z = \begin{bmatrix} \cos\phi_1 & \sin\phi_1 & 0 \\ -\sin\phi_1 & \cos\phi_1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where,

$$\cos\phi_1 = p_n^x / \sqrt{(p_n^x)^2 + (p_n^y)^2},$$
$$\sin\phi_1 = p_n^y / \sqrt{(p_n^x)^2 + (p_n^y)^2}.$$

Let $\Omega = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ be the rotated architecture.

b. $\mathbf{p}_n \to (p_n^x, 0, 0)$.
To place $\mathbf{p}_n$ on the x-axis, we apply a rotation operation $\mathscr{R}_{\mathring{y}}$ with associated matrix $R_{\mathring{y}}$, a clockwise-rotation matrix around the y-axis:

$$R_{\mathring{y}} = \begin{bmatrix} \cos\phi_2 & 0 & \sin\phi_2 \\ 0 & 1 & 0 \\ -\sin\phi_2 & 0 & \cos\phi_2 \end{bmatrix},$$

where

$$\cos\phi_2 = p_n^x / \sqrt{(p_n^x)^2 + (p_n^z)^2},$$
$$\sin\phi_2 = p_n^z / \sqrt{(p_n^x)^2 + (p_n^z)^2}.$$

Let $\Omega = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ be the rotated architecture.

Step 3. $\mathbf{p}_2 \to (p_2^x, 0, p_2^z)$ with $p_2^z > 0$.
To place $\mathbf{p}_2$ on the xz-plane, we apply a counter-clockwise rotation about the x-axis $\mathscr{R}_{\mathring{x}}$ with associated matrix $R_{\mathring{x}}$:

$$R_{\mathring{x}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_3 & -\sin\phi_3 \\ 0 & \sin\phi_3 & \cos\phi_3 \end{bmatrix},$$

where

$$\cos\phi_3 = p_2^z / \sqrt{(p_2^y)^2 + (p_2^z)^2},$$
$$\sin\phi_3 = p_2^y / \sqrt{(p_2^y)^2 + (p_2^z)^2}.$$

Let $\Omega = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ be the rotated architecture.

Step 4. $\mathbf{p}_3 \to (p_3^x, p_3^y, p_3^z)$ such that $p_3^y > 0$.
To satisfy $p_3^y > 0$, if $p_3^y < 0$, reflect $\mathbf{p}$ as

$$\mathscr{R}_{rfl} : p_i^y \to -p_i^y. \tag{19}$$

Let transformation $\mathscr{I}$ be the composite of the five isometric transformations, $\mathscr{R}_\tau$, $\mathscr{R}_{\mathring{z}}$, $\mathscr{R}_{\mathring{y}}$, $\mathscr{R}_{\mathring{x}}$, and $\mathscr{R}_{rfl}$ in the following way: $\mathscr{I} \equiv R_{rfl}\mathscr{R}_{\mathring{x}}\mathscr{R}_{\mathring{y}}\mathscr{R}_{\mathring{z}}\mathscr{R}_\tau$. Then $\mathscr{I}$ is an isometric (distance-preserving) transformation and the transformed coordinates satisfy the following estimability conditions on $\mathbf{p}$: $\mathbf{p}_1 = (0, 0, 0)$, $\mathbf{p}_2 = (p_2^x, 0, p_2^z)$ with $p_2^z > 0$, $\mathbf{p}_3 = (p_3^x, p_3^y, p_3^z)$ with $p_3^y > 0$, and $\mathbf{p}_n = (p_n^x, 0, 0)$ with $p_n^x > 0$.

B. Leapfrog Method for Hamiltonian MCMC

In the second stage of Hamiltonian MCMC, we simultaneously update $(\mathbf{p}_i, \mathbf{v}_i)$ to obtain a proposal vector $(\mathbf{p}_i^*, \mathbf{v}_i^*)$ using a leapfrog method which involves a leap scale ε and a repetition number L:

(1) For each of x, y, z, update v_i^x, v_i^y, v_i^z as

$$v_i^{(.)} \leftarrow v_i^{(.)} + \frac{1}{2}\varepsilon \frac{d \log p(p_i^{(.)} | \mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^{(.)}}. \tag{20}$$

(2) Repeat the following updates $L - 1$ times:

$$v_i^{(.)} \leftarrow v_i^{(.)} + \frac{1}{2}\varepsilon \frac{d \log p(p_i^{(.)} | \mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^{(.)}}, \qquad p_i^{(.)} \leftarrow p_i^{(.)} + \varepsilon v_i^{(.)}. \tag{21}$$

(3) Update v_i^x, v_i^y, v_i^z as

$$v_i^{(.)} \leftarrow v_i^{(.)} + \frac{1}{2}\varepsilon \frac{d \log p(p_i^{(.)} | \mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^{(.)}}. \tag{22}$$

(4) The updated $\mathbf{p}_i$ and $\mathbf{v}_i$ constitute a proposal vector $(\mathbf{p}_i^*, \mathbf{v}_i^*)$.
In the leapfrog method, the essential quantities to evaluate are

$$\frac{d \log p(p_i^x | \mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^x} = \sum_{j \neq i} \left(y_{ij} - \lambda_{ij} \frac{e^{\lambda_{ij}}}{e^{\lambda_{ij}} - 1} \right) \alpha_1 \frac{p_i^x - p_j^x}{\delta_{ij}^2}, \tag{23}$$

$$\frac{d \log p(p_i^y | \mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^y} = \sum_{j \neq i} \left(y_{ij} - \lambda_{ij} \frac{e^{\lambda_{ij}}}{e^{\lambda_{ij}} - 1} \right) \alpha_1 \frac{p_i^y - p_j^y}{\delta_{ij}^2}, \tag{24}$$

$$\frac{d \log p(p_i^z | \mathbf{y}, \vartheta_{-\mathbf{p}_i})}{dp_i^z} = \sum_{j \neq i} \left(y_{ij} - \lambda_{ij} \frac{e^{\lambda_{ij}}}{e^{\lambda_{ij}} - 1} \right) \alpha_1 \frac{p_i^z - p_j^z}{\delta_{ij}^2}. \tag{25}$$

References

1. Baù, D., A. Sanyal, B.R. Lajoie, E. Capriotti, M. Byron, et al. 2011. The three-dimensional folding of the a-globin gene domain reveals formation of chromatin globules. *Nature Structural and Molecular Biology* 18: 107–114.
2. Ben-Elazar, S., et al. 2013. Spatial localization of co-regulated genes exceeds genomic gene clustering in the saccharomyces cerevisiae genome. *Nucleic Acids Research* 41: 2191–2201.
3. Dixon, J.R., S. Selvaraj, F. Yue, et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485: 376–380.

4. Duan, Z., M. Andronescu, K. Schutz, S. McIlwain, et al. 2010. A three-dimensional model of the yeast genome. *Nature* 465: 363–367.
5. Fraser, J., M. Rousseau, S. Shenker, M.A. Ferraiuolo, et al. 2009. Chromatin conformation signatures of cellular differentiation. *Genome biology* 10: R37+.
6. Fullwood, M.J., M.H. Liu, Y.F. Pan, J. Liu, et al. 2011. TAn oestrogen-receptor-[agr]-bound human chromatin interactome. *Nature* 462: 58–64.
7. Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, et al. 2013. *Bayesian Data Analysis, Third Edition (Chapman and Hall/CRC Texts in Statistical Science).* Chapman and Hall/CRC
8. Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian Statistics (Vol. 4, pp. 169–193). Oxford: Oxford University Press.
9. Heidelberger, P., and P.D. Welch. 1983. Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research* 31: 1109–1145.
10. Hu, M., K. Deng, Z. Qin, et al. (2013). Bayesian inference of spatial organizations of chromosomes. *PLOS Computational Biology* 9: e1002893+.
11. Imakaev, M., G. Fudenberg, R. McCord, et al. 2012. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature Methods* 9: 999–1003.
12. Kalhor, R., H. Tjong, N. Jayathilaka, et al. 2012. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology* 30: 90–98.
13. Lesne, A., J. Riposo, P. Roger, et al. (2014). 3D genome reconstruction from chromosomal contacts. *Nature Biotechnology*, advance online publication.
14. Lieberman-Aiden, E., N.L. van Berkum, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326: 289–293.
15. Raftery, A.E., and S.M. Lewis. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms, *In Practical Markov Chain Monte Carlo*, (pp. 115–130).
16. Rousseau, M., J. Fraser, M. Ferraiuolo, J. Dostie, and M. Blanchette. (2011). Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling, *BMC Bioinformatics* 12: 414+.
17. Tanizawa, H., O. Iwasaki, A. Tanaka, et al. 2010. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research* 38: 8164–8177.
18. Varoquaux, N., F. Ay, W.S. Noble, and J. Vert. 2014. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* 30: 26–33.
19. Xiao, G., X. Wang, and A.B. Khodursky. 2011. Modeling three-dimensional chromosome structures using gene expression data. *Journal of the American Statistical Association* 106: 61–72.
20. Yaffe, E., and A. Tanay. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics* 43: 1059–1065.
21. Zhang, Z., Li, G., K. Toh, and W. Sung. 2013. Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-c data. *Proceedings of the 17th International Conference on Research in Computational Molecular Biology* 16: 317–332.

Index

P.K. Choudhary et al. (eds.), *Ordered Data Analysis, Modeling and Health Research Methods*, Springer Proceedings in Mathematics & Statistics 149,
DOI 10.1007/978-3-319-25433-3