Festschrift

In Honor of

George and Frances Ball Distinguished Professor of Statistics

Mir Masoom Ali

On

The Occasion of his Retirement

B A L L S T A T E U N I V E R S I T Y.

Muncie, Indiana, USA
May 18-19, 2007
George and Frances Ball Distinguished Professor Mir Masoom Ali
Photograph with Sir R. A. Fisher, November 18, 1954, Dhaka University Statistics Department

Professor Mir Masoom Ali seated on chair first from the right, Sir R. A. Fisher seated on chair sixth from right and Professor Qazi Motahar Husain seated on chair seventh from the right
May 18, 2007

Dr. Mir Masoom Ali  
Department of Mathematical Sciences  
Ball State Campus

Dear Mir:

On the occasion of your retirement, and on behalf of the Ball State community, I congratulate you on an outstanding career that has brought recognition to our University, and has profoundly influenced many people throughout your decades of service.

As an educator, you founded the Master’s program in statistics at Ball State University, where you have taught and mentored countless students. You have prepared and guided many of the program’s graduates toward doctoral programs at distinguished universities, and to influential careers in academia and industry. You have shared your expertise beyond our campus through visiting positions at universities around the world. Your tireless efforts at recruiting students have been critical to the success of the statistics program, and have contributed to the internationalization of our campus community.

Your scholarly contributions are immense, as demonstrated by your voluminous record of published research in statistics, and the long list of scholars with whom you have successfully collaborated. The esteem of your colleagues is clear from your recognitions within the University, as recipient of the 1992-93 Outstanding Faculty Award and the 1985 Outstanding Researcher Award, as well as by your peers in statistics, who have honored you through your election as Fellow to several national and international learned societies and through special journal issues dedicated to your honor. The State of Indiana also has recognized your academic achievements and contributions to the area of higher education by naming you a Sagamore of the Wabash, the highest award given by the State.

You have generously contributed to the scholarly enterprise through editorship of journals, and your work as cofounder of the Midwest Biopharmaceutical Statistics Workshop, the longest running conference in America on health and pharmaceutical statistics, which has been held here on campus annually for the past thirty years, has led to fruitful interaction among literally thousands of statisticians who have come to our campus from both academia and industry. This conference has increased the visibility of our University to the professional community of statisticians.
I thank you for your commitment to the advancement of Ball State University and to the profession of statistics. In your retirement, I wish you continued scholarly excitement, and trust that as you take some well-deserved rest, you will look back on your career here with pride and satisfaction.

Sincerely,

[Signature]

To Ann M. Gora
President

mdf
May 2007

Dr. Mir Masoom Ali
George and Frances Ball Distinguished Professor
of Statistics and Professor of Mathematical Sciences
Ball State University

Dear Dr. Ali:

On behalf of Ball State University, I am pleased to extend congratulations upon your retirement! It is my understanding that as many as 80 prominent statisticians from around the world join me in marking the occasion of your retirement and your commitment to the profession. This is truly a well-deserved honor.

As you look back over your career, I hope that you will take satisfaction in your many achievements. At Ball State University, you have given 38 years of dedicated service to our students, faculty, and staff. You founded our graduate program in statistics and have served as its director since 1971. Your Ball State colleagues have recognized your excellence in teaching, research, and professional service several times over by awarding you both the Outstanding Researcher Award and the Outstanding Faculty Award.

As a newcomer to Ball State University this year, I became rapidly acquainted with your reputation on our campus, but I know also that your reputation extends far beyond its boundaries. In 1990, the Bangladesh Statistical Association awarded you the first Qazi Motahar Husain Gold Medal, and you are an elected Fellow of the American Statistical Association, the Royal Statistical Society, the Institute of Statisticians, and the Bangladesh Academy of Sciences. Few of our faculty have ever amassed such honors.

However, although you are an accomplished scholar, you also serve others. Particularly I note your role as co-founder, program co-chair, and local arrangements chair of the Midwest Biopharmaceutical Statistics Workshop, which has awarded you the Meritorious Service Award on three separate occasions. You have served on several committees of the American Statistical Association, and you also have been editor of several international statistical journals.
In your retirement, I am certain that you will continue your research and service endeavors. However, I hope that it will allow you to enjoy other long-deferred activities as well. My gratitude, appreciation, and best wishes go with you.

Sincerely,

Terry King
Provost and Vice President for Academic Affairs
April 16, 2007

Dr. Mir Masoom Ali
Department of Mathematical Sciences
Ball State University
Muncie, IN 47306

Dear Mir:

I am very pleased to write this letter of congratulations to you on the eve of your retirement. The record of your achievements during the 38 year career you have had at Ball State is extraordinary. As the George and Frances Ball Distinguished Professor of Statistics you have brought international attention to your department and to Ball State University. Being presented Ball State’s Outstanding Faculty Award and Outstanding Researcher Award recognizes that your colleagues at this university hold you in high esteem. Awards from the Bangladesh Statistical Association and the Midwest Biopharmaceutical Statistics Workshop, and being elected a Fellow of the American Statistical Association, the Royal Statistical Society, the Institute of Statisticians, and the Bangladesh Academy of Sciences confirms this high esteem goes far beyond the boundaries of Ball State University.

Even with all the international recognitions you have received, throughout your career you have served the Department of Mathematical Sciences with distinction. I appreciate your efforts in teaching Ball State students, publishing dozens and dozens of scholarly works, being an exemplary mentor to young colleagues, serving as editor and on editorial boards of statistical journals, and involving yourself in university, college, and departmental activities. Thank you for all of this!

I wish you the very best in retirement and hope you will always find ways to keep in touch with your friends and colleagues at the university. Your positive influence on the Department of Mathematical Sciences has helped make it an invaluable asset to the college and university. We will miss you not being a full-time faculty member, but we will never forget all you have done for us.

Sincerely,

Michael A. Maggiotto
Dean
May 19, 2007

Dr. Mir Masoom Ali  
George and Frances Ball Distinguished Professor of Statistics  
Department of Mathematical Sciences  
Ball State University

Dear Dr. Ali:

I write to offer the sincere gratitude and appreciation of the Department of Mathematical Sciences for your many years of dedication to the Department, its students, and the statistics profession. Your devotion to your students and their success continues to reflect in the achievements of our alumni. Your prolific, respected, and far-reaching record of scholarship and publication continues to bring honor and recognition to the Department. Your valued commitment to the profession is affirmed by the many honors conferred by your colleagues.

The popularity of this Festschrift and its accompanying conference is further compelling evidence of the esteem and regard so many people hold for you. I join with them, celebrating with admiration your distinguished career at Ball State University. As a new chapter of life begins, may you continue to find fulfillment and joy in your pursuits.

Best wishes to you and your family,

\[Signature\]

Dr. John Emert  
Professor and Chair  
Department of Mathematical Sciences  
Ball State University
April 15, 2007

Last Fall, I happily took on the task of organizing a conference to honor Mir Masoom Ali on the occasion of his retirement after 38 years of service to Ball State University. Unfortunately, I was not able to complete the task due to a serious illness, from which I am recovering well. I am very grateful to the Department of Mathematical Sciences for carrying through with the conference and *Festschrift*. I am especially indebted to Ralph Bremigan for taking over the task of facilitating production of the *Festschrift* and to the Department Chair, John Emert, for taking over the conference organization for me.

I would like to extend my thanks to Mir Masoom Ali, George and Frances Ball Distinguished Professor of Statistics. He has been a generous colleague, not only to me but to all of our faculty. He and his wife and children have been as close as family to me. His service to the University has been deep and broad. His contributions to the Department over the past 38 years have been invaluable in helping make the Department that we are today. His contributions to the profession, both through service and research accomplished, are truly outstanding. Mir, you are quite deserving of this honor. I thank you for being my friend and colleague and the Department for carrying through in my time of need.

Dale Umbach  
Professor
Preface

INTRODUCTION

This Festschrift, consisting of forty papers contributed by eighty-one authors and co-authors, honors Mir Masoom Ali, George and Frances Ball Distinguished Professor of Statistics and Professor of Mathematical Sciences, on the occasion of his retirement from Ball State University. Many of these papers were presented at the Conference marking Dr. Ali’s retirement, May 18-19, 2007, on the Ball State campus. This outpouring from colleagues is a fitting tribute to Dr. Ali’s record of research, his dedication to his students, and his contributions to the University and the profession.

MIR MASOOM ALI

Mir Masoom Ali joined the Department of Mathematical Sciences at Ball State University in Muncie, Indiana in 1969 after completing his doctoral work at the University of Toronto. He had obtained his B.Sc. (Honours) degree in 1956 and M.Sc. degree in 1957, both in Statistics, from the University of Dhaka. He came to the University of Toronto, Canada on leave of absence from the Government of Pakistan in 1966 for graduate studies and obtained a second Master’s degree in 1967 and Ph.D. degree in 1969, both in Mathematical Statistics. Prior to coming to the United States in 1969, he worked for a brief period in 1957 for the Socio-Economic Survey Board of the University of Dhaka and then served in various statistical positions with the Government of Pakistan.

Professor Ali has been the director of the graduate program in statistics at Ball State University since 1971, the year in which he founded the program. His students have gone on to doctoral programs at many universities, including Bowling Green, Brown, Colorado State, Indiana-Bloomington, North Carolina-Chapel Hill, North Carolina State, Ohio State, Oregon State, Pittsburgh, Purdue, and Southern Methodist. In recognition of his excellence in teaching, research and professional service, Dr. Ali received the 1992-93 Outstanding Faculty Award and the 1985 Outstanding Researcher Award from Ball State University. In 1990, the Bangladesh Statistical Association awarded him the first Qazi Motahar Husain Gold Medal for his outstanding contributions in the field of statistics. He was also awarded the Meritorious Service Awards in 1987, 1997, and 2002 by the Midwest Biopharmaceutical Statistics Workshop, which is co-sponsored by the American Statistical Association, for his role as a co-founder, program co-chair, and local arrangements chair.

The thrust of Professor Ali’s research has been in the areas of finite sampling, statistical inference, and order statistics. He has published extensively in leading statistical journals. Dr. Ali is an elected Fellow of the American Statistical Association, the Royal Statistical Society, the Institute of Statisticians, and the Bangladesh Academy of Sciences. He is also an elected member of the International Statistical Institute. He is the founding president of the North America Bangladesh Statistical Association.

Professor Ali has served or is serving as an editor/associate editor of several international statistical journals published in Bangladesh, India, Pakistan, and South Korea. He has served on several committees of the American Statistical Association including two terms as President of the Central Indiana Chapter. He also held visiting appointments at a number of universities and
statistical institutes in Canada, United States, Bangladesh, India, Korea, and Japan. The Journal of Statistical Studies and the Pakistan Journal of Statistics published Special Volumes in Professor Ali’s honor in 2002 and 2004, respectively. Currently, the International Journal of Statistical Sciences is in the process of publishing a Special Volume in Dr. Ali’s honor on the occasion of his 70th birthday. On October 24, 2002, Indiana Governor Frank O’Bannon named Dr. Ali a Sagamore of the Wabash, the State of Indiana’s highest award, in recognition of Dr. Ali’s tremendous contributions to Ball State University, to the statistics profession and, especially to higher education in the State of Indiana for over three decades. In 2005, Professor Ali was awarded the “Our Pride Award” by the Bangladeshi-American Foundation, Inc. in Washington, DC, for his distinguished achievements in the field of statistics. Later that same year he was awarded a Gold Medal by the Islamic Society of Statistical Sciences (ISOSS) for his outstanding contributions in the field of statistics and ISOSS affairs.

With his retirement in June 2007, Dr. Ali will have completed fifty years of service as a statistician, including thirty-eight years at Ball State University.

FROM THE ORGANIZERS

This Conference was organized by Dr. Dale E. Umbach, Professor of Statistics and former Chairman of the Department of Mathematical Sciences at Ball State University, as a personal tribute to Mir Ali, as friend, colleague, and collaborator. In the later stages of preparation, Dale was assisted by Drs. Ralph J. Bremigan and John W. Emert of the Department of Mathematical Sciences.

We thank Mr. M. Mahbubul Majumder, a graduate student in statistics and research assistant in our Department, for his many hours of hard work in formatting the soft copies of the papers and putting them together in printable form for the Festschrift. Without his expertise and generous gift of time, it would have been impossible to meet the printing deadline.

Our efforts relied on the constant assistance of our Department’s outstanding Administrative Coordinator, Mrs. Susan Bourne. We deeply appreciate her cheerful and knowledgeable support in working through the logistics of hosting the Conference.

We gratefully acknowledge financial support from the College of Sciences and Humanities of Ball State University, its Department of Mathematical Sciences, and the generous private donors to the Ball State University Foundation.

Our profound gratitude and appreciation go to the many colleagues of Mir who spoke at the Conference or who contributed a manuscript to the Festschrift. Obviously, their work represents the scholarly substance of the Conference and Festschrift. It is our hope that those in attendance at the Conference received some measure of return on their generosity, through new and renewed friendships, stimulating discussion, and enjoyment of our Department’s hospitality.

To have worked through the years with such an inspiring and kind colleague as Mir Masoom Ali has been an honor, and to honor him through this Conference and Festschrift has been a joy.

D.E.U. R.J.B. J.W.E.
CONTENTS

Let’s Get Together and Have a Conference ................................. CHARLES B. SAMPSON 1-2
Flexible Univariate and Multivariate Models Based on Hidden
Truncation ................................................................. BARRY C. ARNOLD 3-13
Using Control Information to Design Type I Censored Treatment Versus Control Clinical
Trials ................................................................. P. L. GRAHAM, S. N. MACEACHERN AND D. A. WOLFE 14-22
A Comparison of Graphical Methods for Assessing the Proportional Hazards
Assumption in the Cox Model ................................. INGER PERSSON, HARRY KHAMIS 23-43
A Multistage Model for Analyzing Repeated Observations on Depression In
Elderly .......... M. ATAHARUL ISLAM, RAFIQUIL I. CHOWDHURY, SHAHARIAR HUDA 44-54
The Stochastic Analysis of Minimum $x^2$ for Various Clumping Models ............... M.I. A. AGEEL 55-60
Two Nonlinear Models for Time Series ............... DAVID A. DICKEY AND SANGIL HWANG 61-70
On A- and D-Rotatability of Two-Dimensional
Third-Order Designs ................ S. HUDA, L. BENKHEROUF AND F. ALQALLAF 71-77
Weibull-Based Approaches to Survival Analysis: An Application to a Breast
Cancer Data Set .... ABDUS S. WAHED, THE MINH LUONG AND JONG-HYEON JEONG 78-95
Optimum Designs for Estimation in Accelerated Life
Testing Problems .................. MANISHA PAL AND NRIPEKS KUMAR MANDAL 96-102
Some Variations of Blackwell Martingate Inequality ................. RASUL A. KHAN 103-105
A Note on the Modified Box-Cox
Transformation ....................................... MEZBAHUR RAHMAN AND LARRY M. PEARSON 106-115
Inference About a Common Mean in One Way
Random Effects Model ........ GUIDO KNAPP, PRANAB K. MITRA, BIMAL K. SINHA 116-140
LM-Stationary Processes to Analyse Time Series Data With Linearly Compressing Periodic
Behavior .... MD. JOBAYER HOSSAIN, WAYNE A. WOODWARD AND HENRY L. GRAY 141-151
A Permutation Test for the Stimuli Effect in the Spatio-Spectral
Profile of Brain Signals ............................... ZHEWEN FAN AND HERNANDO OMBAO 152-162
Inference on Limiting Availability of a
One-Unit Repairable System ................................ FANG LI AND JYOTIRMROY SARKAR 163-174
On The UMVUE of Reliability in Some Discrete Distributions .............. JUNGSOO WOO 175-178
A Stochastic Representation of Matrix Variate Skew Normal Models ............... JOHN T. CHEN 179-184
Prior Statistics Coursework and Student Expectations of a
Graduate Statistics Class ................................ HOLMES FINCH, MOLLY M. JAMESON 185-189
Price Dispersion and Border Effect: A Survey Paper ...... JULIUS HORVATH, BALINT HERCEZ 190-196
Identities Based on Probability Mass Functions ... M. FORHAD HOSSAIN, ANWAR H. JOARDER 197-200
Mahalanobis Moments of Bivariate Distributions .......................... ANWAR H. JOARDER 201-206
Minimax Estimation of the Parameter of the Rayleigh Distribution
M. KAMRUJ JAMAN BHUIYAN, MANINDRA KUMAR ROY, M. FAROUQ IMAM 207-212
<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>An Alternative Method for Selecting the Number of Components for Smooth Lack of Fit Tests</td>
<td>P.P.B. EGGERMONT AND V.N. LARICCIA</td>
<td>225-234</td>
</tr>
<tr>
<td>On the Asymptotic Variance of the Estimator of Attributable Risk and Testing Association in a 2X2 Contingency Table</td>
<td>TANWEER J. SHAPLA</td>
<td>235-241</td>
</tr>
<tr>
<td>Thoughts on Actuarial Science, Demography, Probability, Statistics and Stochastic Processes</td>
<td>JOHN A. BEEKMAN</td>
<td>242-255</td>
</tr>
<tr>
<td>Estimating a Population Median with a Small Sample</td>
<td>BORIS SHULKIN AND SHLOMO SAWIOWSKY</td>
<td>256-267</td>
</tr>
<tr>
<td>Sum of Hypergeometric Series Functions Using Probability Distributions</td>
<td>MUNIR AHMAD AND AYESHA ROOHI</td>
<td>268-273</td>
</tr>
<tr>
<td>Fuzzy Set Representation of a Prior Distribution</td>
<td>GLEN MEEDEN</td>
<td>274-275</td>
</tr>
<tr>
<td>The Impact of Different Math Prerequisites on the Performance of Business Statistics Students</td>
<td>JEFFREY J. GREEN, COURTENAY C. STONE, ABERA ZEGEYE AND THOMAS A. CHARLES</td>
<td>276-285</td>
</tr>
<tr>
<td>Training Environmental Statisticians – Tomorrow’s Problem Solvers</td>
<td>WILLIAM F. HUNT, JR., KIMBERLY WEEMS, WILLIAM SWALLOW, ANDREW MOORE, NAGAMBAL SHAH AND MONICA STEPHENS</td>
<td>286-295</td>
</tr>
<tr>
<td>Joint Distribution of Some Random Variables in Crossing of Two-Stage Erlang Processes</td>
<td>MIR GHULAM HYDER TALPUR AND IFFAT ZAMEER</td>
<td>296-304</td>
</tr>
<tr>
<td>Sampling Plans Excluding Certain Neighboring Units</td>
<td>KYOUNGAH SEE AND A. JOHN BAILER</td>
<td>305-312</td>
</tr>
<tr>
<td>Difference Based Variance Estimators for Partially Linear Models</td>
<td>KARON KLIPPLE AND R. L. EUBANK</td>
<td>313-323</td>
</tr>
<tr>
<td>Bootstrap Bias Reduction and Estimator Decomposition</td>
<td>KRISTOFER JENNINGS</td>
<td>324-327</td>
</tr>
<tr>
<td>Bayesian Prediction for the Linear Model With Equi-correlated Responses: An Application of the Generalized Multivariate Modified Bessel Distribution</td>
<td>LEHANA THABANE AND B. M. GOLAM KIBRIA</td>
<td>328-335</td>
</tr>
<tr>
<td>Middle Censoring for Circular Data</td>
<td>S. RAO JAMMALAMADAKA AND VASUDEVAN MANGALAM</td>
<td>336-339</td>
</tr>
<tr>
<td>Highly Efficient Nonlinear Regression Based on the Wilcoxon Norm</td>
<td>ASHEBER ABEBE AND JOSEPH W. MCKEAN</td>
<td>340-357</td>
</tr>
<tr>
<td>On Tail Probabilities of the t-Distribution for Samples from the Uniform Distribution</td>
<td>R. AYESHA ALI AND M. M. ALI</td>
<td>358-363</td>
</tr>
</tbody>
</table>
During the spring of 1976, I was sitting in my office at Eli Lilly and Company, minding my own business, trying to get some work done, when a coworker came by and asked me to do a favor for him. He had committed Lilly to participate in some sort of statistics conference at Ball State. He told me he could not do it and asked if I would help out. Well, I said, I am not exactly in the business of proving theorems and that is probably what was desired at such a conference. But I agreed to call Dr. Mir Ali, whom I had never met, to negotiate a topic which I could handle. I subsequently agreed to participate. Mir can sell, when he wants to, and he wanted to sell for his conference. It was a fine conference with a balance among mathematical statistics topics, applied statistics, career possibilities in statistics and so forth. The conference was appropriately sized for interaction between all parties should they so choose and indeed the interaction was very good. I was impressed. Mir Ali had done a wonderful job with his “statistics days” conference.

But I asked Mir if he wanted to do a “statistics conference” again with a different theme. Mir said we should talk about it. I told Mir that I had been watching the Princeton Conference, held sometimes at Princeton, and would like to have a similar, but smaller, conference in the Midwest emphasizing health statistics and the Pharmaceutical Industry. I said we had lots of Pharmaceutical Companies in the midwest (which was true in 1976) and I thought there was a desire to have a forum to discuss problems relevant to the Pharmaceutical Industry and the FDA.

We decided to seek the approval, participation, and financial backing of the leaders of the Midwest Pharmaceutical Companies’ statistics groups. Those who attended the first meeting (7/14/1977) to discuss the feasibility of what is now known as the Midwest Biopharmaceutical Statistics Workshop (MBSW) were Wen-Dar Chang (Dow), Ken Falter (Searle), Saul Gitomer (Marion), Bernie McDonagh (Riker), Tony Orlando (Mead Johnson), Lyman Ott (Merrill-National), Ron Platt (Miles), Alan Sampson (Abbott), Charles Sampson (Lilly), Roy Sanford (Baxter), John Schultz (Upjohn), Ron Schwartz (Arnar-Stone), Mir Ali (Ball State University) and Tom Spradlin (Lilly). Please note that Dow, Searle, Marion, Miles, Mead-Johnson, Merrill-National, Upjohn, Arnar-Stone and 3-M Pharmaceuticals do not exist anymore as they were merged or bought out by other companies. Mergers were a potential threat to the continued existence of the MBSW. The important home bases of a number of companies were to disappear -- from Chicago, Kalamazoo, Ann Arbor, Cincinnati, Evansville and Indianapolis.

Those in attendance at the organization meeting of the first MBSW were unanimous in not wanting to replicate the format and style of the already popular national statistical meetings of the ASA. In addition to an emphasis on practical problems, it seemed that the format of the meeting could be varied so as to enhance interest and to promote participation by the younger members of the statistical profession. A planning meeting was then held in Chicago on August 15, 1977 and the program content for the first MBSW was determined. One additional meeting was held in October of 1977 and it appeared very likely that the first conference would be held in May of 1978. The statistical heads of the Midwest Pharmaceutical Companies agreed to hold the conference once, for a test, and each provided $500 for the kitty in case we bombed out completely and lost money. It was agreed that the MBSW was to include roundtable discussions, pedagogical lectures, workshops, poster sessions, analysis of data sets, panel discussions, methodology sessions, and discussion groups.

The first meeting of the MBSW was held May 23-24, 1978 and was a great success.
The location of the Workshop was a heated discussion topic during its formative years. The urge for an urban setting was strong among some of the early charter members, but the hospitality of Ball State University and the Muncie community has served the workshop well for the last 30 years. Keeping the workshop in Muncie was enhanced by low fees and lots of “bang for the buck”. For example, the first registration fee was $40 and included two luncheons as well as a banquet at the Morris Bryant Restaurant which featured all the lobster one could eat. The motel rooms ranged from $13 to $25. Attendees came from as far away as Palo Alto and Europe. We joked that the operating definition of the “Midwest” was everywhere in the world but New Jersey, Manhattan, and perhaps Philadelphia. However, as one can observe with this year’s workshop, most of our organizing committee is from the East. Mergers, mergers, and more mergers. Those of us associated with MBSW over the years have witnessed the effect of numerous mergers in our registration documents. The 30th MBSW will be held this year (2007). Mir Ali should be proud of this contribution to the statistical community. Without those statistics days held in 1976 there would have been no MBSW, I am sure.

I wish to offer some observations and comments regarding Professor Ali’s professional life. Professor Ali is an outstanding teacher. You ask how would I know this? I have interviewed (for employment) over 150 students from Ball State University over the years. These students were computer scientists, mathematicians, statisticians, and information scientists. As part of my interview, and as a warmup exercise, I would ask these applicants to name the 2-3 best teachers they have had so far in their post secondary education. Professor Ali was almost always mentioned, conditional upon the fact that the students had had him. I found this tremendously impressive. This helps validate yet another positive dimension in Professor Ali’s career, that of being a very fine teacher.

Now I would like to discuss Mir Ali, as my friend. These are my observations only, after 30 years of friendship and working on the MBSW together and after having similar health (heart) experiences and so on. Over the years my wife and I have been to parties in his house, met his kids, and had him to our house. We enjoyed their wedding anniversary celebration in Indianapolis arranged by his children.

Mir Ali is a gentleman, he is tenacious, and I believe he can be a task master. But he is also kind and patient, at least with me. He is a wonderful friend and we have shared many thoughts about our personal lives, something I do not do easily. In early Y-2002, I went in for what I thought was going to be a routine heart exam and 3 days later I was coming out of a anesthetic fog after open heart surgery and one of the first, maybe the first, persons I recognized was Mir Ali and his wife, Leena. Mir and Leena were standing there looking at me and I was beginning to wonder where I might be. Maybe I was in some sort of after life and Mir and Leena were already there. I asked my wife if I was in heaven or in Muncie. She said I was in St. Vincent’s Hospital in Indianapolis. I really appreciated the attention he gave me, and the counsel he offered, having had his experiences as a heart victim.

Mir Ali was, and still is, a “buttoned down” gentleman. I worked with him on the conference for 25 years before I saw him without a tie. I once asked him if he slept in a suit and he looked at me with that incredulous but controlled concerned look not knowing if he should reply to such a statement.

Here is one experience I will never forget. Some years ago, Mir and I were walking through some local arrangement issues at Ball State talking intently. We were heading out of the Student Union on the west side and going down the stairs when Mir stopped cold to stare at a handsome young boy coming up the stairs into the union. Apparently this was not the appropriate time for the son to be coming to the union and he was perhaps playing hooky. The boy looked up and froze. No words were exchanged but there was a great deal of intense eyeball to eyeball communication. The boy turned around and headed back to school He must have gotten back to school since Ishii is now a successful cardiologist. Everything worked out fine, it seems. What I learned from that experience was to try my hardest never to displease Mir.

I am very pleased to have known Mir Masoom Ali for all these years. My wife and I value Mir’s and Leena’s friendship greatly and we wish Mir’s retirement to be exciting and fulfilling.
Flexible Univariate and Multivariate Models
Based on Hidden Truncation

Barry C. Arnold

Department of Statistics, University of California, Riverside, USA.

Abstract

A broad spectrum of flexible univariate and multivariate models can be constructed by using a hidden truncation paradigm. Such models can be viewed as being characterized by a basic marginal density, a family of conditional densities and a specified hidden truncation point, or points. The resulting class of distributions includes the basic marginal density as a special case (or as a limiting case), but also includes an array of models that may unexpectedly include many well known densities. Most of the well known skew normal models (developed from the seed distribution popularized by Azzalini (1985)) can be viewed as being products of such a hidden truncation construction. However, the many hidden truncation models with non-normal component densities undoubtedly deserve further attention.

Key Words: Skew-normal distribution, conditional specification, weighted distribution, multivariate normal, normal conditionals, exponential conditionals, Pareto distribution.

1 Introduction

The skew-normal model, popularized and studied by Azzalini (1985) and his coworkers, is a one parameter family of densities of the form

\[ f(x; \lambda) = 2\phi(x)\Phi(\lambda x), \quad -\infty < x < \infty, \]  

where \( \phi \) (respectively \( \Phi \)) denotes the standard normal density (respectively distribution) function and \( \lambda \in \mathbb{R} \) is a parameter which controls skewness. The addition of a location parameter (\( \mu \)) and a scale parameter (\( \sigma \)) yields the following flexible family, useful for fitting asymmetric data sets.

\[ f(x; \mu, \sigma, \lambda) = \frac{2}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right)\Phi\left(\lambda\left(\frac{x - \mu}{\sigma}\right)\right) \]  

A good survey of results related to this skew-normal model may be found in Genton(2004).

A scenario which leads to data satisfying the model (2) is one involving hidden truncation of the following form (see Arnold and Beaver (2000) for more details on hidden truncation models which are also known as frontier models in the Economics literature, for example, in Kumbhakar and Knox Lovell (2000)) . In this setting, we begin with a two dimensional random variable \((X, Y)\) which has a classical bivariate normal distribution with mean vector \((\mu_X, \mu_Y)\) and variance-covariance matrix \(\Sigma\). Only the \(X\) coordinate random variable is observed and it is only observed if the concomitant \(Y\) exceeds \(\mu_Y\) (i.e. if \(Y\) is above average). Data collected in this fashion are readily confirmed to have density (2).

A slight variant of this model involves observing \(X\)'s only if their concomitant \(Y\) variable exceeds a given level \(y_0\) (not necessarily equal to \(\mu_Y\)). This extension was alluded to by Azzalini (1985), Henze (1986) and Arnold, Beaver et al. (1993). It is discussed in more detail in Arnold and Beaver (2000). The resulting extensions of (1) and (2) now involves two skewing parameters \( \lambda_0 \) and \( \lambda_1 \) and are of the forms:

\[ f(x; \lambda_0, \lambda_1) = \frac{\phi(x)\Phi(\lambda_0 + \lambda_1 x)}{\Phi\left(\frac{\lambda_0}{\sqrt{1-\lambda_1^2}}\right)} \]  

(3)
and
\[ f(x; \mu, \sigma, \lambda_0, \lambda_1) = \frac{\phi(x - \mu)}{\Phi(\frac{x - \mu}{\sigma})} \frac{\lambda_0 + \lambda_1}{\sqrt{1 - \lambda_1^2}} \] (4)

Many extensions of these models have been proposed. For example, we could select an arbitrary density \( \psi_1(x) \) to play the role of \( \phi(x) \) in (3) and an equally arbitrary distribution function \( \Psi_2(x) \) to play the role of \( \Phi(x) \). The resulting family of models, before introducing location and scale parameters, is of the form:
\[ f(x; \lambda_0, \lambda_1) \propto \psi_1(x)\Psi_2(\lambda_0 + \lambda_1 x). \] (5)

Computation of the required normalizing constant in (5) may be troublesome and indeed it will frequently be necessary to determine the constant by numerical integration.

There are multivariate extensions of the model (3) which may be viewed as having begun with a \( k + m \) dimensional random vector \((\bar{X}, \bar{Y})\) (here \( \bar{X} \) is of dimension \( k \) and \( \bar{Y} \) of dimension \( m \) ) and only observing \( X \) if \( Y > y_0 \), where \( y_0 \) is a pre specified vector in \( \mathbb{R}^m \). Most skew models of this genre begin by assuming a classical \( k + m \) dimensional normal distribution for \((\bar{X}, \bar{Y})\). An extensive survey of such models may be found in Azzalini (2006). In the present paper we will focus on general hidden truncation models (which of course include skew-normal models) beginning with a completely general distribution for \((X, Y)\) (or for \((\bar{X}, \bar{Y})\) in higher dimensional settings). This general hidden truncation paradigm will be shown to yield a remarkably rich vein of models which may profitably be used to fit univariate and multivariate data sets. Naturally, it would be desirable to identify a stochastic mechanism involving hidden truncation which can plausibly be argued to have played a role in generating the data set that is fitted by such a model. However, absent such identification, the hidden truncation model, provided that it fits well, may still be useful for prediction purposes.

Returning to the simple hidden truncation model in which we observe \( X \) only if \( Y > y_0 \), it is evident that the density of the observed \( X \)'s will have a (conditional) distribution of the form:
\[ F_{X|Y>y_0}(x_0) = P(X \leq x_0|Y > y_0) = \frac{\int_{-\infty}^{x_0} \int_{y_0}^{\infty} f_{X,Y}(x,y)dydx}{\int_{y_0}^{\infty} f_{Y}(y)dy} \] (6)

Assuming the existence of conditional densities (as will be done throughout this paper) we can write the corresponding conditional density as
\[ f_{X|Y>y_0}(x) = \frac{\int_{y_0}^{\infty} f_{X|Y}(x|y)dy}{F_{Y}(y_0)} \] (7)

In this formulation, the marginal density of \( Y \) and the conditional density of \( Y \) given \( X \) will determine the resulting hidden truncation model. In a sense, the model is parameterized by \( y_0 \in \mathbb{R} \) and \( f_Y(y) \) (or, more generally, a parametric family of densities \( f_Y(y; \theta) \)) and by \( f_{X|Y}(x|y) \) (or, more generally, by a parametric family of densities \( f_{X|Y}(x|y; \xi) \)). Clearly this represents an enormously flexible family of models. For example, we could take \( f_Y(y) \) to be a normal density and take \( f_{X|Y}(x|y) \) to be normal with linear regression and constant conditional variance. Inexorably we are led to the skew-normal model (4). But we could get a richer family by allowing \( f_{X|Y}(x|y) \) to have a more general regression function and perhaps a non-constant conditional variance function. This approach merits further investigation. However it is not the approach that will be followed in the rest of the present paper.

A joint density for \((X, Y)\) can be written as the product \( f_Y(y)f_{X|Y}(x|y) \), but equally well it can be written as \( f_X(x)f_{Y|X}(y|x) \). Using this expression for the joint density, it is readily verified that
\[ f_{X|Y>y_0}(x) = f_X(x) \frac{P(Y > y_0|X = x)}{P(Y > y_0)} \] (8)

In this formulation, the skewed distribution obtained by hidden truncation is clearly shown to be a weighted version of the original density for \( X \). The weight function, \( P(Y > y_0|X = x) \), depends on \( y_0 \) and on the conditional density of \( Y \) given \( X \). The representation of the hidden truncation
density in the form (8) may be found in Arellano-Valle et al. (2002) (their equation (5.1)) in the case in which \( y_0 = 0 \) but, as they remark, it is likely that it had appeared elsewhere at some time previous to 2002.

In subsequent sections, we will investigate hidden truncation models of the form (8) (truncation from below) as well as other truncation paradigms. In all cases, the basic components of the models will be a given density for \( X \) (or \( \mathbf{X} \) in higher dimensions) and a given conditional density for \( Y \) given \( X \) (or for \( Y \) given \( \mathbf{X} \)).

2 Basic hidden truncation models

Begin with a two dimensional absolutely continuous random vector \((X, Y)\). We might focus on the conditional distribution of \( X \) given \( Y \in C \) where \( C \) is a Borel set in \( \mathbb{R} \). Indeed we could write

\[
    f_{X|Y \in C}(x) = f_X(x) \frac{P(Y \in C | X = x)}{P(Y \in C)} \tag{9}
\]

(see Arellano-Valle, Branco and Genton (2006), where such general models are introduced). However, we will concentrate on hidden truncation of one of three forms only:

1. Lower truncation, where \( C = (y_0, \infty) \).
2. Upper truncation, where \( C = (-\infty, y_0] \).
3. Two sided truncation where \( C = (a, b] \).

For upper truncation at \( y_0 \), in which observations are only available for \( X \)'s whose concomitant variable \( Y \) is less than \( y_0 \), equation (9) becomes

\[
    f_{y_0^-}(x) = f_X(x) \frac{P(Y > y_0 | X = x)}{P(Y > y_0)} \tag{10}
\]

Models of this type are thus characterized by

1. \( f_X(x) \), the density assumed for \( X \).
2. The conditional density of \( Y \) given \( X \), \( f_{X|Y}(x|y) \).
3. The specific truncation point, \( y_0 \).

Note that the distribution function corresponding to (10) is of the form

\[
    F_{y_0^-}(x) = P(X \leq x | Y \leq y). \tag{11}
\]

Consequently, a convenient way to generate models of this type is to begin with a joint distribution for \((X, Y)\) for which \( P(X \leq x | Y \leq y) \) is available in a simple form (discuss of such bivariate distributions may be found in Arnold, Castillo and Sarabia (1999) and Arnold (1995)).

Models involving lower truncation will be of the form

\[
    f_{y_0^+}(x) = f_X(x) \frac{P(Y > y_0 | X = x)}{P(Y > y_0)} \tag{12}
\]

with corresponding survival function

\[
    F_{y_0^+}(x) = P(X > x | Y > y). \tag{13}
\]

Technically, models of the form (12) could be viewed as equivalent to those given by (10). One merely needs to replace the concomitant variable \( Y \) by \(-Y\) (or for non-negative variables, by \( 1/Y \) to go from one to the other. In practice, such a transformation may not seem to be natural and the concepts of upper and lower truncation are best dealt with separately.

Two sided truncation models are of the form
\[
    f_{a,b}(x) = f_X(x) \frac{P(a < Y \leq b | X = x)}{P(a < Y \leq b)}
\]

(14)

Such models are determined by the choice of the basic marginal density \( f_X(x) \), the choice of conditional density \( f_{Y|X}(y|x) \) and the truncation points, \( a \) and \( b \).

It will be observed that the upper and lower truncation models can be obtained as limiting cases of two sided truncation models, so in a sense we need only to deal with two sided truncation models. Typically the one sided models are simpler in structure and they sometimes can be obtained directly more easily, without first considering a two sided model. Note that, in order for any of these models to assume attractive form, it is necessary that the conditional distribution of \( Y \) given \( X \) should have an analytic expression for its distribution function, or at least that the conditional distribution can be evaluated by reference to available tables.

When using the formulations (10), (12) and (14) to construct flexible families of densities it will, as remarked earlier, typically be the case that the density of \( X \) is assumed to be a member of some parametric family of densities \( f_X(x, \theta) \) and that the conditional density of \( Y \) given \( X \) is a member of another parametric family of densities \( f_{Y|X}(x|y; \eta) \). We will consider some examples in which the family of marginal densities for \( X \) and the family \( \mathcal{F} \) of conditional densities for \( Y \) given \( X \) are of the same form (e.g. they might both be normal), but we can flexibility by mixing and matching (e.g. one family might be Weibull and the other gamma).

Before embarking on an investigation of some of the many parametric families of models that can be generated by such hidden truncation constructions, it is appropriate to remark that, beginning with a given choice of density function for \( X \), say \( f_0(x) \), it is possible to generate, via hidden truncation, an extremely broad class of densities by judicious choice of the conditional density of \( Y \) given \( X \). Just about any density with the same support as \( f_0(x) \) and lighter tails than \( f_0(x) \) can be generated in this fashion. For example, suppose that we wish to generate the density \( f_1(x) \) by applying hidden truncation to \( f_0(x) \). In order to achieve this in a simple fashion, we need to assume that there exists \( c > 0 \) such that \( f_1(x)/f_0(x) \leq c, \forall x \). If such a \( c \) exists, we can choose a family of conditional densities of \( Y \) given \( X \) in such a fashion that

\[
    P(Y \leq 0 | X = x) = \frac{1}{c} \frac{f_1(x)}{f_0(x)}.
\]

(15)

With this choice of conditional distributions of \( Y \) given \( X \) and by setting \( y_0 = 0 \), we may verify that hidden truncation above at 0, applied to \( f_0(x) \) will yield, via equation (10), the desired density \( f_1(x) \).

3 Hidden truncation using normal component densities

We begin by considering hidden truncation applied to classical bivariate normal data. In this case two sided hidden truncation will be considered (from which, results for upper and lower truncation can be readily derived). Thus we begin with \( X \sim N(\mu, \sigma^2) \) and we will assume the linear regression and constant conditional variance that is associated with the classical bivariate normal distribution for \( (X, Y) \). Thus we assume that \( Y|X = x \sim N(\alpha + \beta x, \tau^2) \). Referring to (14), for hidden truncation points \( a \) and \( b \), we have

\[
    f_{a,b}(x) = f_X(x) \frac{P(a < Y \leq b | X = x)}{P(a < Y \leq b)}
\]

\[
    = \frac{1}{\sigma \Phi\left(\frac{b - \alpha - \beta x}{\tau}\right) - \Phi\left(\frac{a - \alpha - \beta x}{\tau}\right)}
\]

(16)

In this expression, \( \mu, \alpha, \beta, \sigma, \tau \in \mathbb{R} \) and \( a < b \). It is convenient to introduce new parameters \( \delta_1, \delta_2 \) and \( \lambda_1 \) where \( -\infty < \delta_1 < \delta_2 < \infty \) and \( \lambda_1 \in \mathbb{R} \), allowing us to rewrite the model (16) as
\[ f_{a,b}(x) = \frac{1}{\sigma} \phi\left( \frac{x - \mu}{\sigma} \right) \frac{[\Phi(\delta_2 + \lambda_1 \frac{x - \mu}{\sigma}) - \Phi(\delta_1 + \lambda_1 \frac{x - \mu}{\sigma})]}{[\Phi(\frac{\delta_2}{\sqrt{1 + \lambda_2^2}}) - \Phi(\frac{\delta_1}{\sqrt{1 + \lambda_1^2}})]} \] (17)

This is the model involving two sided hidden truncation that was discussed, for example, in Arnold, Beaver et al (1993). If we consider upper truncation (letting \( \delta_1 \to -\infty \) in (18) we obtain the Henze-Arnold-Beaver skew-normal model (4) (where \( \delta_2 \) is replaced by \( \lambda_0 \)).

Instead of using a conditional distribution for \( Y \) given \( X \) that is normal with a linear regression function and a constant conditional variance function, we could consider a normal distribution with more general regression and conditional variance functions. Thus, if we assume that \( Y | X = x \sim N(\mu(x), \tau^2(x)) \), our two sided hidden truncation model becomes

\[ f_{a,b}(x) \propto \frac{1}{\sigma} \phi\left( \frac{x - \mu}{\sigma} \right) \Phi\left( \frac{b - \mu(x)}{\tau(x)} \right) - \Phi\left( \frac{a - \mu(x)}{\tau(x)} \right) \] (18)

The model (18) includes (as limiting cases) densities of the form

\[ f(x; \Delta) \propto \phi(x) \Phi\left( \frac{\lambda_{00} + \lambda_{10}x}{\sqrt{1 + (\lambda_{01} + \lambda_{11}x^2)}} \right). \] (19)

Such densities have been studied earlier in the literature. They are identifiable as marginal densities of the following class of bivariate distributions with conditionals in the skew normal family (4)

\[ f(x, y; \Delta) \propto \phi(x) \phi(y) \Phi(\lambda_{00} + \lambda_{10}x + \lambda_{01}y + \lambda_{11}xy), \] (20)

(See, for example, Arnold, Castillo and Sarabia (2002)). Such models can also be obtained as mixtures of univariate skew-normal densities (see, for example, Arellano-Valle et al. (2004)).

In fact, model (18) is, in a sense, completely general. Any weighted version of the normal(\( \mu, \sigma^2 \)) density can be represented in the form (18). Suppose that we wish to have

\[ f_{a,b}(x) \propto w(x) \frac{1}{\sigma} \phi\left( \frac{x - \mu}{\sigma} \right) \] (21)

for some specified weight function \( w(x) \). We can choose \( a = -\infty, b = 0 \) and \( \tau(x) = 1 \). The choice of \( \mu(x) \) which will then enable us to identify (21) as a special case of (18) will be such that \( \Phi(-\mu(x)) = w(x) \), i.e. we should choose \( \mu(x) = -\Phi^{-1}(w(x)) \).

4 Hidden truncation applied to normal conditionals distributions

Following early work by Bhattacharyya (1943), Arnold, Castillo and Sarabia (1999) provided detailed discussion of the class of bivariate densities, \( f_{X,Y}(x, y) \), which have all of their conditional densities (of \( X \) given \( Y \) and of \( Y \) given \( X \)) of the normal form. Such bivariate densities are necessarily of the form

\[ f_{X,Y}(x, y) = \exp - \left\{ (1, x^2) \begin{pmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ m_{20} & m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} 1 \\ y \\ y^2 \end{pmatrix} \right\} \] (22)

where the \( m_{ij} \)'s satisfy certain constraints to ensure integrability. For our hidden truncation constructions, we need expressions for the corresponding marginal \( f_X(x) \) and for the conditional densities \( f_{Y|X}(y|x) \). It is not difficult to verify that, if \( (X, Y) \) has density (22), then

\[ f_X(x) = \frac{\exp - \left( \frac{1}{2}(2(m_{20}x^2 + m_{10}x + m_{00}) - \frac{(m_{21}x^2 + m_{11}x + m_{01})^2}{2(m_{22}x^2 + m_{12}x + m_{02})})}{\sqrt{2(m_{22}x^2 + m_{12}x + m_{02})}} \] (23)
while $Y|X = x \sim N(\mu(x), \sigma^2(x))$ in which

$$\mu(x) = -\frac{(m_{21}x^2 + m_{11}x + m_{01})}{2(m_{22}x^2 + m_{12}x + m_{02})} \tag{24}$$
and

$$\sigma^2(x) = \frac{1}{2(m_{22}x^2 + m_{12}x + m_{02})}. \tag{25}$$

The corresponding two sided hidden truncation model will be

$$f_{a,b}(x) \propto f_X(x)\Phi\left(\frac{b - \mu(x)}{\sigma(x)}\right) - \Phi\left(\frac{a - \mu(x)}{\sigma(x)}\right) \tag{26}$$

where $f_X(x), \mu(x)$ and $\sigma(x)$ are as defined in (23), (24) and (25) respectively.

The centered normal conditionals model is considerably simpler. For it, we set $m_{01} = m_{10} = m_{11} = m_{12} = m_{21} = 0$ in (22). This leaves us with a 3 parameter bivariate density for which

$$f_X(x) = \frac{e^{-m_{20}x^2}}{\sqrt{2(m_{22}x^2 + m_{02})}} \tag{27}$$
and

$$Y|X = x \sim N(0, \frac{1}{2(m_{22}x^2 + m_{02})}), \tag{28}$$
so that

$$f_{a,b}(x) \propto \frac{e^{-m_{20}x^2}}{\sqrt{2(m_{22}x^2 + m_{02})}}[\Phi(b\sqrt{2(m_{22}x^2 + m_{02})}) - \Phi(a\sqrt{2(m_{22}x^2 + m_{02})})]. \tag{29}$$

## 5 Hidden truncation with exponential component densities

Suppose now that $X$ has an exponential distribution, i.e.

$$P(X > x) = e^{-\alpha x}, \quad x > 0. \tag{30}$$

Now assume that, for each $x > 0$, the conditional density of $Y$ given $X = x$ is also an exponential density with a constant failure rate which depends linearly on $x$. Thus

$$P(Y > y|X = x) = e^{-(\beta + \gamma x) y}, \quad y > 0. \tag{31}$$

The resulting joint density is of the form

$$f(x, y) = (\alpha \beta + \alpha \beta x) e^{\alpha x} e^{-(\alpha x + \beta y + \gamma xy)}, \quad x > 0, y > 0. \tag{32}$$

The corresponding two sided hidden truncation model will then be

$$f_{a,b}(x) \propto \alpha e^{-\alpha x} [e^{-(\beta + \gamma x)x} - e^{-(\beta + \gamma x)b}], \quad x > 0, \tag{33}$$
a linear combination of two exponential densities. The lower hidden truncation model is obtained from (33) by setting $b = \infty$ and $a = y_0$, in this manner we find

$$f_{y_0 +}(x) = (\alpha + \gamma y_0) e^{-(\alpha + \gamma y_0)x}, \quad x > 0, \tag{34}$$
i.e. again an exponential density. Thus, in this situation, lower hidden truncation does not lead to an enrichment of the class of distributions for $X$. A similar phenomenon is observable if we begin with $(X, Y)$ having an exponential conditionals distribution (see Arnold and Strauss (1988)). The corresponding joint density is of the form
\[ f(x, y) \propto \exp(-\alpha x + \beta y + \gamma xy), \quad x > 0, y > 0. \]  

(35)

In this case the marginal density for \( X \) is

\[ f_X(x) \propto (\beta + \gamma x)^{-1} e^{-\alpha x}, \quad x > 0, \]

and the conditional survival function of \( Y \) given \( X = x \) is of the form

\[ P(Y > y|X = x) = e^{-(\beta + \gamma x)y}, \quad x > 0, \]

(37)

( the same as (31). The corresponding lower hidden truncation model will be

\[ f_{y_0+} \propto (\beta + \gamma x)^{-1} e^{-(\alpha + \gamma y_0)x}, \quad x > 0. \]

(38)

Observe that (38) is obtainable from (36) by a simple change of one of the parameters and the family of lower hidden truncation models coincides with the original family of densities for \( X \).

It becomes evident that the use of lower hidden truncation with a conditional distribution given by (31) will be ineffective in enriching the class of densities assumed for \( X \) whenever \( f_X(x) \) includes a factor of the form \( e^{-g(\theta)x} \), where \( g(\theta) > 0 \). Thus, for example, if we begin with an assumption that \( X \sim N(\mu, \sigma^2) \) where \( \mu < 0 \) and assume that \( Y|X = x \) has a distribution satisfying (31) (i.e. an exponential conditional distribution with a constant failure rate that is a linear function of \( x \)), then the resulting lower hidden truncation models will again be normal with negative means.

If we allow the failure rate for the conditional distribution of \( Y \) given \( X = x \) to depend on \( x \) in a non-linear fashion, we can, of course, get new densities by using the lower hidden truncation paradigm as the following examples show.

Let us begin with \( X \sim \exp(\alpha) \), i.e. \( P(X > x) = e^{-\alpha x} \). Now assume that \( P(Y > y|X = x) = e^{-\gamma(x)y} \) for some positive function \( \gamma(x) \) defined on \( \mathbb{R}^+ \). It follows that

\[ f_{y_0+} \propto e^{-(\alpha x + \gamma(x)y_0)}, \quad x > 0. \]

(39)

As a special case, consider \( \gamma(x) = \gamma x^2 \) for some \( \gamma > 0 \). In this case, we find that

\[ f_{y_0+} \propto e^{-(\alpha x + \gamma y_0 x^2)}, \quad x > 0. \]

(40)

It is perhaps surprising that a truncated normal density such as (40) can arise via hidden truncation applied to a model with an exponential marginal distribution for \( X \) and exponential conditionals for \( Y \) given \( X \).

However, it is true that a very broad class of densities can be obtained by hidden truncation, even when we restrict attention to pre-truncated models involving an exponential marginal and exponential conditionals. Suppose that we wish to obtain a specific target density \( g(x)I(x > 0) \) in this manner. To do this, we must select \( y_0, k, \alpha \) and \( \gamma(x) \) such that, for \( x > 0 \),

\[ g(x) = \exp(m - \alpha x - y_0 \gamma(x)). \]

(41)

So \( \gamma(x) \) must satisfy

\[ \gamma(x) = \frac{m - \alpha x - \log g(x)}{y_0}. \]

(42)

For certain choices of \( g(x) \), the corresponding function \( \gamma(x) \) given by (42) will not be always positive, so that some mild conditions must be imposed on the form of \( g(x) \) in order for it to be obtainable via hidden truncation using exponential model components. Nevertheless, an extremely broad class of densities \( g(x) \) on \( \mathbb{R}^+ \) can be so constructed.
6 Hidden truncation with Pareto component densities

A bivariate Pareto conditionals density is of the form (Arnold, Castillo and Sarabia (1999))

\[ f(x,y) = (\alpha + \beta x + \gamma y + \delta xy)^{-(\tau + 1)}, \] (43)

where \(\alpha, \beta, \gamma, \delta, \tau > 0\). The corresponding marginal and conditional densities are:

\[ f_X(x) \propto [\alpha + \beta x - \gamma - \delta x^2]^{\tau}, \quad x > 0, \] (44)

and for \(x > 0\),

\[ f_{Y|X}(y|x) \propto \left[ 1 + \gamma + \delta x + \frac{\gamma y + \delta xy}{\alpha + \beta x} \right]^{-(\tau + 1)}, \quad y > 0, \] (45)

( i.e. \(Y|X = x \sim \text{Pareto}(\frac{\gamma + \delta x}{\alpha + \beta x}, \tau)\) ). The two sided hidden truncation distribution derived from this joint density will then be given by

\[
\begin{align*}
    f_{a,b}(x) & \propto f_X(x)P(a < X \leq b | X = x) \\
    & \propto [(\alpha + \beta x)^{\gamma} - (\gamma + \delta x)]^{\tau} - [(\gamma + \delta x)^{\tau}] \\
    & \propto \frac{1}{\gamma + \delta x} \left\{ \frac{1}{(\alpha + \beta x + \gamma + \delta x)^{\tau}} - \frac{1}{(\alpha + \beta x + \gamma + \delta x)^{\tau}} \right\}
\end{align*}
\] (46)

for \(x > 0\). Recalling that \(\alpha, \beta, \gamma, \delta > 0\) and \(0 \leq a < b\), we may write this as

\[
\begin{align*}
    f_{a,b}(x) & \propto \frac{1}{\gamma + \delta x} \left[ \frac{1}{(\alpha_1 + \beta_1 x)^{\tau}} - \frac{1}{(\alpha_2 + \beta_2 x)^{\tau}} \right], \quad x > 0, \tag{47}
\end{align*}
\]

a linear combination of two densities of the same form as the original marginal density of \(X\) (as in (44).

To obtain \(f_{y_0-}(x)\) we just set \(a = 0\) and \(b = y_0\) in (46) and we obtain a density, also of the form (47), with \(\alpha_1 = \alpha\) and \(\beta_1 = \beta\).

The lower hidden truncation model is simpler. To get \(f_{y_0+}(x)\) we must set \(a = y_0\) and \(b = \infty\) in (46) to obtain

\[ f_{y_0+}(x) \propto \frac{1}{\gamma + \delta x} \left[ \frac{1}{(\alpha_1 + \beta_1 x)^{\tau}} - \frac{1}{(\alpha_2 + \beta_2 x)^{\tau}} \right], \quad x > 0, \] (48)

or equivalently

\[ f_{y_0+}(x) \propto \frac{1}{\gamma + \delta x} \left[ \frac{1}{(\alpha_1 + \beta_1 x)^{\tau}} - \frac{1}{(\alpha_2 + \beta_2 x)^{\tau}} \right], \quad x > 0. \] (49)

So in this case we again observe the phenomenon in which lower hidden truncation fails to augment the class of models already assumed for \(X\).

If we begin with a Pareto distribution for \(X\) and Pareto conditional distributions for \(Y\) given \(X\), as in (45), then lower hidden truncation will lead to an enriched family of densities. We will have

\[ f_X(x) \propto (\alpha + \beta x)^{-(\eta + \tau + 1)}, \quad x > 0. \] (50)

and

\[ P(Y > y_0|X = x) = \left[ 1 + \frac{\gamma + \delta x}{\alpha + \beta x y_0} \right]^{-\tau}, \] (51)

so that, after reparameterization, we have

\[ f_{y_0+}(x) \propto (\alpha + \beta x)^{-(\eta + 1)}(\alpha' + \beta' x)^{-\tau}, \] (52)

where \(\alpha \leq \alpha'\) and \(\beta \leq \beta'\).
7 Multivariate cases

In the development thus far, both X and Y have been scalar variables. Of course, analogous arguments can be advanced when the variables are of higher dimensions. Thus one may consider a $k + m$ dimensional random vector $(\mathbf{X}, Y)$ where $\mathbf{X}$ is of dimension $k$ and $Y$ is of dimension $m$. We will consider the distribution of $\mathbf{X}$ subject to hidden truncation on $Y$ of the form $Y \leq y_0$. We can write this conditional density (in a form analogous to (10)) as:

$$f_{Y|X}(y) = \frac{f_{X,Y}(y, x)P(Y \leq y_0 | X = x)}{P(Y \leq y_0)}$$

(53)

Hidden truncation models of this type will be determined by the marginal density $f_X(x)$, the conditional density of $Y$ given $X$, $f_{Y|X}(y|x)$, and the truncation point $y_0$.

In this section, we will restrict attention to an illustrative case in which the component densities $(f_X$ and $f_{Y|X})$ are multivariate normal although, of course, the ideas discussed can be extended readily to deal with other examples, perhaps with component densities of different (non-normal) types.

We consider such hidden truncation in a setting in which $(\mathbf{X}, Y)$ has a classical $k + m$ dimensional normal distribution. Thus we begin with

$$\left( \begin{array}{c} \mathbf{X} \\ Y \end{array} \right) \sim N^{(k+m)} \left( \left( \begin{array}{c} \mu \\ \Sigma_{11} \Sigma_{12} \\ \Sigma_{21} \Sigma_{22} \end{array} \right) \right).$$

(54)

In this case $\mathbf{X} \sim N^{(k)}(\mu, \Sigma_{11})$ and the conditional distribution of $Y$ given $X = x$ is of the form:

$$Y|X = x \sim N^{(m)}(\mu + \Sigma_{21}\Sigma_{11}^{-1}(x - \mu), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}).$$

(55)

We will introduce notation as follows:

$$\phi^{(k)}(x) = \prod_{i=0}^{k} \phi(x_i),$$

(56)

and

$$\Phi^{(m)}(y; \delta, \Lambda) = P(Y \leq y),$$

(57)

where $Y \sim N(\delta, \Lambda)$. With this notation, referring to (53), we will have

$$f_{Y|X}(y) = |\Sigma_{11}|^{-1/2}\phi^{(k)}(\Sigma_{11}^{-1/2}(x - \mu)) \frac{\Phi^{(m)}(y_0 - \mu - \Sigma_{21}\Sigma_{11}^{-1}(x - \mu); 0, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})}{\Phi^{(m)}(y_0 - \mu; 0, \Sigma_{22})}.$$  

(58)

At this point it is convenient to make a change of variables, defining $Z = \Sigma_{11}^{-1/2}(X - \mu)$ so that $Z \sim N^{(k)}(0, I)$. If $X$ has density (58), then $Z$ will have a density of the following form

$$f_{Y|X}(z) = \phi^{(k)}(z) \frac{\Phi^{(m)}(\mu_0 + \Lambda z; 0, \Delta)}{\Phi^{(m)}(\mu_0; 0, \Delta + \Lambda^T \Lambda)}$$

(59)

for suitably defined $\mu_0$, $\Delta$ and $\Lambda$ (which will depend on the choice of $y_0$). The model (59) is known in the literature under a variety of names with variations in the labeling of the parameters. For example Gonzalez-Farias et al (2004) call it the closed skew normal family, while Arellano-Valle and Genton (2005) refer to it as the fundamental skew-normal distribution. See Azzalini (2005) for further discussion of these and other aliases.

It is not difficult to deal with analogous lower and two sided hidden truncation models. It is convenient to use the notation $\Phi^{(m)}_{\Lambda}(y; \delta, \Lambda)$ to denote $P(Y > y)$ where $Y \sim N^{(m)}(\delta, \Lambda)$. The resulting models obtained by hidden truncation on Y applied to the density of $Z$ are:
\[
f_{Y^+, z} = \phi^{(k)}(z) \frac{\Phi^{(m)}(\Delta_0 + \Lambda z; 0, \Delta)}{\Phi^{(m)}(\Delta_0; 0, \Delta + \Lambda^T \Lambda)}
\]  

(60)

and

\[
f_{Y^-, z} = \phi^{(k)}(z) \frac{\Phi^{(m)}(\delta_0 + \Lambda z; 0, \Delta) - \Phi^{(m)}(\delta_0 + \Lambda z; 0, \Delta)}{\Phi^{(m)}(\delta_0; 0, \Delta + \Lambda^T \Lambda)} - \Phi^{(m)}(\delta_0; 0, \Delta + \Lambda^T \Lambda).
\]

(61)

There is a considerable literature devoted to the discussion of the distribution of \(X = \mu + \Sigma^{1/2} Z\) where \(Z\) has a hidden truncation density of the form (59). Of course, (60) can be viewed as a special case of (59) in which \(Y\) has been replaced by \(-Y\) and \(y\) by \(-y\). Densities of the form (61) have received less attention, even though such two sided hidden truncation can be expected to be encountered in many real world data configurations.

8 Envoi

Generally speaking, hidden truncation models will be difficult to deal with analytically unless the joint density of \((X, Y)\) ( or of \((X, Y)\) in higher dimensions ) is a member of some tractable family of multivariate distributions. Even in such cases, an awkward normalizing constant may be associated with the hidden truncation distribution. Techniques for dealing with inference problems, even for hidden truncation models as simple as the basic Azzalini model (1), still require refinement. Much work remains to be done before the more complicated hidden truncation models can be expected to enter into the applied statistician’s toolkit.

9 References

References


Using Control Information to Design Type I Censored Treatment versus Control Clinical Trials

P. L. Graham¹, S. N. MacEachern², and D. A. Wolfe²*

¹CSIRO Mathematical and Information Sciences, North Ryde, NSW 2113, Australia
²Department of Statistics, Ohio State University, Columbus, OH 43210 USA

*Corresponding author: Email: daw@stat.ohio-state.edu

Summary

Constraints on time and finances are a continuing problem for researchers conducting clinical trials in which a new treatment is being compared to a standard (control) treatment. Information is often available, however, on the performance of the control from previous studies. We propose a way to utilize this previous knowledge about the control to provide a mechanism for completing a study earlier than might otherwise be possible. This approach can lead to significant savings in time and expense, while still retaining good power for detecting a treatment effect.

Keywords: Designed censoring percentage; Early stopping; Termination time; Wilcoxon rank sum test.

1 Introduction and Notation

Controlled clinical trials are expensive and time consuming, and as such are generally only conducted if there is some evidence of potential benefit to patients. In addition, for treatments that are found to be effective (particularly if they prevent death and increase survival time) there is an ethical question of whether it is appropriate to wait until the end of the trials before offering the treatment to all patients who could benefit from it.

For many clinical trials, we are faced with evaluation of a new treatment relative to a standard current (control) treatment about which we have considerable knowledge from previous studies. When the clinical trials are to be conducted within a fixed time frame (i.e., subject to Type I censoring), this information from previous studies about the control can be used to provide for early completion of the study while still maintaining adequate power for hypothesis tests. This can result in significant savings in both time and expense and lead to earlier availability of an effective new treatment.

Consider a clinical trial designed to compare a control (i.e., the standard current treatment), C, with a new treatment, V. We have N subjects available for the study and n of them are randomly assigned to the control C, with the remaining N - n assigned to the new treatment V. The measurement of interest is the time to event (for example, relapse or death) for the subjects. However, the clinical trials will be subject to Type I Censoring in that at the end of a pre-specified period of time, T, the study will be terminated and any subjects remaining in the study at that time (treatment or control) will be assigned censored measurement values T. Thus, for these censored subjects all that we know is that their true time-to-event values are greater than the censoring time T.

Now, let C₁, ..., Cₙ denote the measurements for the n control subjects and V₁, ..., Vₙ denote the measurements for the N - n treatment subjects. Note that these measurements are true times-to-event values for those subjects who experienced an event prior to termination of the study at time T and they are equal to T for those subjects who had not yet experienced an event by time T.

In this paper, we propose an approach for determining an optimal termination time for studies with Type I censoring. For illustration, we demonstrate how to implement this approach with the Mann-Whitney-Wilcoxon rank sum test procedure in the two-sample setting. We note, however, that the same technique can be used with other two-sample test procedures (such as the two-
sample t-test) and could be extended easily to the multiple treatments versus control setting.

In Section 2 we discuss the model and hypotheses of interest in the two-sample setting. In Section 3 we describe the application of the Mann-Whitney-Wilcoxon rank sum test procedure for these Type I censored data. The necessary details of the null distribution and critical values are provided in Section 4. Section 5 is devoted to simulation results related to the optimal choice of termination time \( T \). An example is presented in Section 6 and we conclude with a short discussion in Section 7.

2 Model and Hypotheses

We consider a general continuous model for this treatment versus control setting, with our interest being in making inferences about the relative effectiveness of the treatment and control. We assume that the control observations \( C_1, \ldots, C_n \) are a random sample of size \( n \) from a continuous distribution with c.d.f. \( F(x|\theta_c) \) and the treatments observations \( V_1, \ldots, V_{N-n} \) are a random sample of size \( N-n \) from the distribution with c.d.f. \( F(x|\theta_v) \), where \( \theta_c \) and \( \theta_v \) are unknown, real-valued parameters. Thus the control and treatment populations are both members of a family of distributions, with the difference between the two distributions captured in the values of the parameters \( \theta_c \) and \( \theta_v \). We presume that the family of distributions is strictly stochastically ordered in the parameter \( \theta \). We note that location families, scale families of positively-valued random variables, and many other common families satisfy this assumption. We also assume that the observations are mutually independent, but we take a nonparametric modeling approach and make no additional assumptions (other than continuity) about the form of the common \( F(\cdot) \).

Let \( \mu_c = E[g(C)] \) and \( \mu_v = E[g(V)] \) be the expected values of some increasing function \( g(\cdot) \) for which both expectations exist and for which there is a 1-1 map between \( \theta \) and \( \mu \). There is little loss in thinking of \( \mu_c \) and \( \mu_v \) as the means for the \( C \) and \( V \) distributions, respectively, and we will refer to them in this way throughout the paper. (These means are, of course, functions of the underlying parameters \( \theta_c \) and \( \theta_v \); stochastic ordering of the distributions in \( \theta \) results in the same ordering of the distributions in \( \mu \).) We are interested in making inferences about the difference in the means \( \Delta = \mu_v - \mu_c \). In particular, in this paper we will concentrate on testing the null hypothesis \( H^d_0: \Delta = 0 \), which corresponds to \( \theta_c = \theta_v \) and \( C = V \), against the alternative \( H^d_1: \Delta > 0 \), corresponding to the treatment mean being larger than the control mean. We will also discuss how to test for a negative treatment effect, \( H^d_2: \Delta < 0 \), or a general two-sided alternative, \( H^d_3: \Delta \neq 0 \).

3 Tests based on the Mann-Whitney-Wilcoxon Statistic

Since we have not made any assumption about the form of \( F(\cdot) \), other than the distribution being continuous, we will utilize the Wilcoxon form of the Mann-Whitney-Wilcoxon statistic to construct distribution-free procedures for testing \( H^d_0: \Delta = 0 \). For this purpose, we jointly rank the \( n \) control and \( N - n \) treatment time-to-event observations from least to greatest. Thus, the rank 1 is assigned to the first subject (treatment or control) to experience an event, and so on until \( M \) is assigned to the last subject to actually experience an event prior to the censoring time \( T \). Finally, each of the \( N - M \) subjects who has not yet experienced an event by time \( T \) is assigned the average, \( A \), of the unassigned ranks given by

\[
A = \frac{\sum_{i=1}^{N} i - \sum_{j=1}^{M} j}{N - M} = \frac{N(N+1)}{2} - \frac{M(M+1)}{2}.
\]

(1)

Thus, for the \( N \) combined control and treatment observations \( (C_1, \ldots, C_n, V_1, \ldots, V_{N-n}) \) we obtain a joint rank vector \( R = (Q_1, \ldots, Q_n, R_1, \ldots, R_{N-n}) \), where \( Q_i \) is the joint rank of \( C_i \), for \( i = 1, \ldots, n \), and \( R_j \) is the joint rank of \( V_j \), for \( j = 1, \ldots, N - n \). For given \( M \) (the number of subjects to experience an event prior to time \( T \)), \( R \) will be some permutation of the vector \( (1, \ldots, M, A, \ldots, A) \). Note that the support for \( M \) is \( m \in \{0, 1, \ldots, N\} \).

For our test statistic, we consider the Wilcoxon form of the Mann-Whitney-Wilcoxon statistic, \( G \), given by the sum of the joint ranks for the sample (control or treatment) with the fewer number of observations. Thus, if \( n \leq N - n \), we have

\[
G = \sum_{i=1}^{n} Q_i = \text{(sum of the ranks for the control sample)},
\]

while if \( n > N - n \) we take

\[
G = \sum_{j=1}^{N-n} R_j = \text{(sum of the ranks for the treatment sample)}.
\]

We first describe the test procedures based on \( G \) for the setting where \( n > N - n \) so that
\[ G = \sum_{j=1}^{N-n} R_j. \] In this situation, the test statistic \( G \) is the sum of the treatment sample ranks and large (small) values of \( G \) are indicative of \( \Delta > (\prec) 0. \) Hence, the appropriate level \( \alpha \) tests based on \( G \) are as follows:

Reject \( H_0: \Delta = 0 \) in favor of \( H_1: \Delta > 0 \) if and only if \( G \geq g_{\alpha,n} \) \hspace{1cm} (2)

and

Reject \( H_0: \Delta = 0 \) in favor of \( H_1: \Delta < 0 \) if and only if \( G < g_{1-\alpha,n} \), \hspace{1cm} (3)

where \( g_{\alpha,n} \) is the upper \( \alpha^{th} \) percentile for the null (\( H_0 \)) distribution of \( G \). The corresponding two-sided test is given by

Reject \( H_0: \Delta = 0 \) in favor of \( H_1: \Delta \neq 0 \) if and only if \( G \geq g_{\alpha,n} \) or \( G < g_{1-\alpha,n} \). \hspace{1cm} (4)

where \( \alpha \) and \( \alpha \) are chosen to satisfy \( \alpha_1 + \alpha_2 = \alpha \). Unless there are compelling reasons not to, we recommend choosing both \( \alpha_1 \) and \( \alpha_2 \) as close as possible to \( \frac{\alpha}{2} \).

For the setting where \( n \leq N - n \), the test statistic \( G \) is the sum of the ranks for the control sample and small (large) values of \( G \) are indicative of \( \Delta > (\prec) 0 \). The corresponding level \( \alpha \) one-sided tests for \( n \leq N - n \) are then given by:

Reject \( H_0: \Delta = 0 \) in favor of \( H_1: \Delta > 0 \) if and only if \( G < g_{1-\alpha,n} \) \hspace{1cm} (5)

and

Reject \( H_0: \Delta = 0 \) in favor of \( H_1: \Delta < 0 \) if and only if \( G \geq g_{\alpha,n} \). \hspace{1cm} (6)

The level \( \alpha \) two-sided test has the same form as in equation (4), except that here \( G \) is the sum of the control sample ranks.

Now to conduct any of these test procedures we need to know the null (\( H_0 \)) distribution of the test statistic \( G \). This, of course, depends on the censoring time \( T \) and the common underlying distribution under \( H_0 \). In the next section we derive an expression for the null distribution of \( G \) as a function of the termination time \( T \) and show how we can use information from previous studies on the control to design our clinical trial to provide for early termination when the treatment is substantially more effective than the control.

### Null Distribution of the Test Statistic \( G \)

As defined in Section 3, let \( \mathbf{R} = (Q_1, \ldots, Q_n, R_1, \ldots, R_{N-n}) \) be the joint rank vector for the combined control and treatment observations \( (C_1, \ldots, C_n, V_1, \ldots, V_{N-n}) \) for given termination time \( T \). Then the marginal probability distribution for \( \mathbf{R} \) can be expressed as

\[
P(R = r) = \sum_{m=0}^{N} P(R = r | M = m)P(M = m) \hspace{1cm} (7)
\]

where, as before, \( M \) is the number of subjects (treatment or control) who experience an event prior to the termination time \( T \).

On the other hand, the marginal distribution of \( M \) is determined by the termination time \( T \) and the lifetime distributions for the treatment and control populations. Let

\[
p_c = P(\text{subject on control C has an event before time } T) \hspace{1cm} (8)
\]

\[
p_v = P(\text{subject on treatment V has an event before time } T). \hspace{1cm} (9)
\]

Then, it follows that

\[
P(M = m) = \sum_{u_1=0}^{n} \sum_{u_2=0}^{n} \binom{N-n}{u_1} \binom{m}{n-u_1-u_2} \binom{n}{u_1} \binom{m}{n-u_1-u_2}, \hspace{1cm} (10)
\]

Now, when the null hypothesis \( H_0: \Delta = 0 \) is true, the \( C \)'s and \( V \)'s are i.i.d. random variables so that

\[
P_c(R = r | M = m) = \frac{(N - m)!}{N!}, r \in H_m. \hspace{1cm} (11)
\]

where \( H_m \) is the set of all distinct permutations of the vector \( (1, \ldots, m, A, \ldots, A) \).

Also, when the null hypothesis \( H_0: \Delta = 0 \) is true, we have \( p_c = p_c \) and equation (10) simplifies to

\[
P_c(M = m) = \sum_{u_1=0}^{n} \sum_{u_2=0}^{n} \binom{N-n}{u_1} \binom{m}{n-u_1-u_2} \binom{n}{u_1} \binom{m}{n-u_1-u_2}, \hspace{1cm} (12)
\]

Thus (not surprisingly), the marginal distribution of \( M \) is binomial with parameters \( N \) and \( p_c \) under \( H_0: \Delta = 0 \), which depends on the termination time \( T \) only through \( p_c \). Combining the expressions in equations (7), (11), and (12), it follows that the null marginal distribution of the rank vector \( \mathbf{R} \) has p.m.f

\[
P_c(R = r) = \sum_{m=0}^{N} \frac{p_c^m (1 - p_c)^{N-m}}{m!}, \hspace{1cm} r \in \bigcup_{m=0}^{N} H_m, \hspace{1cm} (13)
\]
where, for \( m = 1, \ldots, N \), \( H_m \) is once again the set of all distinct permutations of the vector \((1, \ldots, m, A, \ldots, A)\); that is, the null distribution of \( R \) is uniform over \( \bigcup \limits_{m=0}^{N} H_m \) and it depends on the termination time \( T \) only through \( p_c \).

The null distribution of our test statistic \( G \) is then obtained immediately from the expression in (13) by noting that
\[
P(G = g) = \sum_{m=0}^{N} \frac{p_c^m(1 - p_c)^{N-m}}{m!} \times Q(g),
\]
where \( Q(g) \) is the number of rank vectors \( r \in N \bigcup \limits_{m=0}^{N} H_m \) for which \( G = g \), and \( S \) is the set of possible values for \( G \) over all combinations of \( m \in \{0, \ldots, N\} \) and \( r \in H_m \).

Null distribution tables for \( G \) were calculated using an R program (R Development Core Team, 2006) that is available from the authors on request. Some examples of these null distribution tables can be found in Appendix A and a more extensive set of tables for \( N \leq 20 \) is given in Graham (2001) and Graham et al. (2001). In the event that \( N \) exceeds 20 a Monte-Carlo simulation program, also available from the authors on request, can be used to calculate an approximate p-value for an observed value of \( G \).

5 Choice of Termination Time \( T \)—Simulation Studies

As noted previously, properties of the tests based on \( G \) depend on the termination time \( T \) only through the probabilities of event occurrence, \( p_c \) and \( p_p \), for the control and treatment populations, respectively. Clearly smaller choices of \( T \) lead to shorter and less expensive clinical trials. On the other hand, for typical distributions the powers of the test procedures based on \( G \) are increasing functions of \( T \). We shall see that the use of prior information about the control can be useful in selecting a termination time that can provide dramatic shortening of the study length while still maintaining effective power properties for the test.

The goal in our setting is to select as small a \( p_c \) value (with associated early termination time) as possible without sacrificing too much power for the test procedure. To investigate this feature of the proposed test we conducted a Monte Carlo simulation power study as follows. For a given replicate, two random samples, each of size ten, of survival times are generated, one from the control distribution and the second from the treatment distribution. We consider a variety of different control proportions \( p_c \) (and associated termination times \( T \)) and treatment proportions \( p_p \), corresponding to different alternative values \( \Delta \). For each replicate the censored rank sum statistic \( G \) is calculated as described in Section 3 and we record whether it leads to rejection of \( H_0: \Delta = 0 \) in favor of the appropriate alternative at the available significance level closest to 0.045. (The level is chosen to be 0.045 so that we can make direct comparisons with the usual uncensored Mann-Whitney-Wilcoxon rank sum test for these sample sizes.) This process is repeated for 100,000 replicates for each \((p_c, p_p)\) combination and the power of the test is estimated by the observed percentage of rejections.

These simulations were carried out for underlying normal, exponential, and Weibull distributions. The results are similar for all three distributions, although the observed improvement in power is a bit greater for the normal distribution when more subjects are censored in the control than in the treatment and the time savings from early termination is a bit lower for the Weibull distribution when \( p_c \) and \( p_p \) do not differ by much. Thus, for brevity, in this paper we present details of the simulation results only for the underlying exponential distribution with c.d.f.
\[
F(x \mid \lambda) = 1 - e^{-\lambda x}, \lambda > 0.
\]

For this exponential setting, we take \( \lambda = 1 \) for the control population and use it to set the termination times \( T \) corresponding to a variety of values for \( p_c \). (At first glance, the choice of \( \lambda = 1 \) for our simulations appears to be somewhat arbitrary. However, its role is simply to serve as a baseline against which to compare treatment alternative \( \lambda \) values and associated \( p_p \) percentages. Starting with a different baseline, say \( \lambda = 2 \), would certainly lead to different termination times \( T \), but the observed power against comparable alternative \( \lambda \) and \( p_p \) values would have been similar to what we found for \( \lambda = 1 \). It is solely the difference between the \( p_c \) and \( p_p \) percentages that controls the power.)

Figure 1 shows how the power of the test varies with different combinations of \( p_c \) and \( p_p \). As expected, the power is a monotonically increasing function of the absolute difference \([p_c - p_p]\).

Our proposed rank sum procedure for censored data will typically have lower power than the corresponding Mann-Whitney-Wilcoxon rank sum procedure based on full and complete measurements. To provide some insight as to how much power is lost by using this early termination approach, we compare the difference between the power of our \( G \) test and the uncensored rank sum test. These simulation comparisons are presented in Figure 2.
The biggest differences in power between the uncensored rank sum test and our $G$ test occur when $p_v \leq 0.3$. In this region, the power of the uncensored rank sum test is frequently more than 0.15 higher than that for the $G$ test, and this difference reaches nearly 0.5 when $p_v$ is as low as 0.1. Of course, this is not really surprising since a great deal of treatment information is being lost if we are censoring more than 70% of the treatment observations. Also as expected, when both $p_c$ and $p_v$ are large (so there is very little censoring in either sample), the two tests are very similar in power.

We also investigated the time (and resultant expense) saved by using the early termination approach described in this paper. For each replicate of a given simulation (combination of $p_c$ and $p_v$), we recorded the following information:

1. Length of time to complete the uncensored experiment in which every subject experiences the event of interest—this, of course, is just the largest observation (time to event) in that iteration.
2. Length of time to complete the censored experiment—this is most often just the early termination time for the study. However, there were some simulations where all of the observations (both treatment and control) were less than the termination value (that is, all the subjects observed events prior to the termination time). For such iterations, this measure is then the largest observation.

These two pieces of information were then averaged over the 100,000 replicates in the simulation to obtain the summary statistics "Average time of uncensored experiment" and "Average time of censored experiment". We also computed the average of all of the actual values for observations that would have been censored across the entire 100,000 replicates in the simulation. This is reported as the "Conservative average of uncensored times", since it represents a more conservative estimate of the "length" of an uncensored experiment. We have chosen to include this calculation as well because outliers among the maximum observations could heavily influence the "Average time of uncensored experiment".

For each $(p_c, p_v)$ combination in our simulation study, we estimate the typical time saved from using the early termination approach by the observed difference

$$\text{Time saved} = \text{"Average time of uncensored experiment"} - \text{"Average time of censored experiment"}.$$  \hspace{1cm} (15)

To take into account the possible misleading effect of outliers among the maximum observed values, we also compute a conservative estimate of the time saved to be

$$\text{Conservative time saved} = \text{"Conservative average of uncensored times"} - \text{"Average time of censored experiment"}.$$ \hspace{1cm} (16)

In Figures 3 and 4 we graphically display both of these "time saved" measures for $p_c = 0.8$ (Figure 3) and $p_c = 0.2$ (Figure 4) and $p_v$ values ranging from 0.1 to 0.9. Figure 3 shows that when $p_c$ is large (0.8) and $p_v$ is small the conservative measure for uncensored experiments ranges, on average, from 5 to 15 times longer than the censored experiment even when we use the conservative average of uncensored times. For larger $p_v$ values the conservative time saved is less pronounced although the typical conservative measure for uncensored experiments is still 50% longer than the typical censored experiment. The savings are substantially greater when we compare against the average uncensored maximum time measurements. When $p_c$ is small as in Figure 4 ($p_c = 0.2$), the differences between the lengths of time for
censored and uncensored (using either the conservative or maximum measure) experiments is less striking (than when \( p_c = 0.8 \)) for small \( p_v \) values. However, a substantial timesaving is maintained over all studied values of \( p_v \). Similar results were obtained for other values of \( p_c \) ranging from 0.9 to 0.1 and are presented in Figures 5-11 in Appendix B.

Figure 3. Time saved with use of \( G \) test when \( p_c = 0.8 \) for various values of \( p_v \).

![Figure 3](image)

Figure 4. Time saved with use of \( G \) test when \( p_c = 0.2 \) for various values of \( p_v \).

![Figure 4](image)

6 Example

The data used in this example are adapted from a subset of a University of Massachusetts Aids Research Unit (UMARU) IMPACT study by McCusker et al. (1997) and described in Hosmer and Lemeshow (1999). In that study patients were randomized to one of two drug treatment programs of differing lengths. For the purposes of our example we use the 360 patients from treatment site A who returned to drug use within 500 days; as such there is no censoring in the basic data set. The aim of the study was to determine which of the two treatment programs leads to a longer typical time until return to drugs. For our example we consider the short program to be the control. For our analysis the data have been slightly modified so that there are no ties in the time until relapse values.

This example is typical of two-sample survival studies in which a control and a treatment are compared; a common feature of survival analysis data is censoring. As noted previously, Type I censoring occurs when the censoring of an event is dictated by a predetermined censoring time \( T \). If an event has not occurred by time \( T \) the observation is censored and assigned an "observed" value of \( T \).

This study was the first of its kind so that previous data about the control are not available. Thus we randomly divided the 360 patients into a "prior data" set of 60 observations to facilitate the application of our procedure and a second test set consisting of the remaining 300 observations. The second set is used to examine the effect of various termination (censoring) times on the proposed hypothesis tests.

Using subjects on the short program as controls we begin by looking at the "prior" data and determining times that correspond to different censoring proportions for the control group. The first column of Table 1 presents eleven potential termination times, \( T \), for the study. The numbers of controls in the prior data set that would have been censored with these termination times are given in the second column of the table and the corresponding percentages, \( p_c \), of controls with observed times to event (not censored) in each case are given in the third column. Then using each of these termination times to analyze the second test data set, the p-values associated with tests of the null hypothesis of no difference in the two treatments against the alternative that the long treatment program has a greater average time until relapse are given in the final column of the table.

### Table 1: Setting Termination Time Using the Prior Data and its Effect on the P-value for the Test Data.

<table>
<thead>
<tr>
<th>Termination Time ( T ) (days)</th>
<th>Prior Data Number Censored</th>
<th>( p_c )</th>
<th>Test Data P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>499.00</td>
<td>0</td>
<td>0.99</td>
<td>.00287</td>
</tr>
<tr>
<td>471.01</td>
<td>1</td>
<td>0.95</td>
<td>.00280</td>
</tr>
<tr>
<td>306.14</td>
<td>3</td>
<td>0.90</td>
<td>.00263</td>
</tr>
<tr>
<td>292.03</td>
<td>4</td>
<td>0.85</td>
<td>.00270</td>
</tr>
<tr>
<td>136.11</td>
<td>6</td>
<td>0.80</td>
<td>.00580</td>
</tr>
<tr>
<td>130.98</td>
<td>7</td>
<td>0.75</td>
<td>.00576</td>
</tr>
<tr>
<td>126.98</td>
<td>8</td>
<td>0.70</td>
<td>.00753</td>
</tr>
<tr>
<td>111.08</td>
<td>10</td>
<td>0.65</td>
<td>.00723</td>
</tr>
<tr>
<td>95.02</td>
<td>11</td>
<td>0.60</td>
<td>.09998</td>
</tr>
<tr>
<td>90.05</td>
<td>11</td>
<td>0.55</td>
<td>.15168</td>
</tr>
<tr>
<td>85.98</td>
<td>14</td>
<td>0.50</td>
<td>.17829</td>
</tr>
</tbody>
</table>

The results in Table 1 show that it would have been possible to conduct the clinical trial on the second set of data with \( p_c \) as low as 0.65 and still obtain a significant p-value in support of the conclusion that the long treatment program leads to
a greater average period of time before return to drugs. The "previous pilot" study ran for 500 days, but using this previous data to determine earlier stopping times we could have terminated the "new" clinical trial after only 115 days, for example, and still have reached the conclusion that the longer treatment time is preferred. Since our power studies indicated that larger values of $p_c$ lead to tests with power comparable to that of the usual Mann-Whitney-Wilcoxon rank sum test, we might want to be more conservative and take $p_c = 0.85$, for example. Even with this more conservative $p_c$ value, however, the termination time of 293 days is still 30% shorter than the full 500 days and the conclusion is the same. This would have permitted the investigators an opportunity to recognize that the longer treatment is more effective at least 200 days sooner than in the reported study.

7 Discussion
In this paper we proposed a Mann-Whitney-Wilcoxon type test procedure that allows us to use previously available data on a control to design treatment versus control clinical trials with early stopping times while still maintaining reasonable power for detecting effective treatments. We found that the procedure had good power when the control and treatment groups are sufficiently far apart. Moreover, when $p_v < p_c$ or when $p_c$ and $p_v$ are both large, the power is very close to that of the uncensored Mann-Whitney-Wilcoxon procedure even with early stopping of the study.

The example showed that it is possible to decrease the length of the trial by as much as 60% and yet still retain sufficient power. Of course, the effectiveness of our procedure will vary from one setting to another. However, we feel that it provides another useful tool for reducing the length of many clinical trials when prior information is available on the baseline control. It is evident that the associated savings in trial time and expense can be considerable.

Acknowledgments
This work was supported in part by the National Science Foundation under Award Numbers DMS-0072526 and DMS-9802358. The authors thank Tim Keighley for his assistance with the R functions.

References


Appendices
A Null Distribution Tables for $G$
For the following tables, $N$ is the total sample size (control plus treatment) and $n$ is the smaller of the two sample sizes. The proportion of the control population that is designated to be fully ranked is denoted by $p_c$. Tabulated entries are the upper tail critical values for the null distribution of $G$ that are closest to 0.01 and 0.05. More comprehensive tables containing roughly the entire upper and lower fifteen percent of the null distribution of $G$ are available in Graham (2000) and Graham et al. (2001).

<table>
<thead>
<tr>
<th>$p_c$</th>
<th>$G$</th>
<th>$P(G \geq g)$</th>
<th>$p_c$</th>
<th>$G$</th>
<th>$P(G \geq g)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>20</td>
<td>0.2367</td>
<td>0.6</td>
<td>20.5</td>
<td>0.0402</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.0421</td>
<td></td>
<td></td>
<td>0.0349</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0.0118</td>
<td></td>
<td></td>
<td>0.0349</td>
</tr>
</tbody>
</table>
Table 3: Upper Tail Critical Values for the $G$ Statistic, $N = 14, n = 7$

<table>
<thead>
<tr>
<th>$p_c$</th>
<th>$G$</th>
<th>$P(G \geq g)$</th>
<th>$p_c$</th>
<th>$G$</th>
<th>$P(G \geq g)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>22</td>
<td>0.0809</td>
<td>0.7</td>
<td>23.5</td>
<td>0.0555</td>
</tr>
<tr>
<td></td>
<td>22.5</td>
<td>0.0174</td>
<td></td>
<td>24</td>
<td>0.0477</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>0.0116</td>
<td></td>
<td>25</td>
<td>0.0135</td>
</tr>
<tr>
<td>0.3</td>
<td>22</td>
<td>0.1083</td>
<td>0.8</td>
<td>25</td>
<td>0.0136</td>
</tr>
<tr>
<td></td>
<td>22.5</td>
<td>0.0416</td>
<td></td>
<td>24</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>0.0227</td>
<td></td>
<td>25</td>
<td>0.0036</td>
</tr>
<tr>
<td>0.4</td>
<td>22</td>
<td>0.0533</td>
<td>0.9</td>
<td>25</td>
<td>0.0143</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>0.0585</td>
<td></td>
<td>25</td>
<td>0.0036</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>0.0314</td>
<td></td>
<td>25</td>
<td>0.0083</td>
</tr>
<tr>
<td>0.5</td>
<td>23</td>
<td>0.0699</td>
<td>1.0</td>
<td>25</td>
<td>0.0091</td>
</tr>
<tr>
<td></td>
<td>23.5</td>
<td>0.0528</td>
<td></td>
<td>25</td>
<td>0.0143</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>0.0143</td>
<td></td>
<td>26</td>
<td>0.0091</td>
</tr>
</tbody>
</table>

Table 4: Upper Tail Critical Values for the $G$ Statistic, $N = 20, n = 10$

<table>
<thead>
<tr>
<th>$p_c$</th>
<th>$G$</th>
<th>$P(G \geq g)$</th>
<th>$p_c$</th>
<th>$G$</th>
<th>$P(G \geq g)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>115</td>
<td>0.1098</td>
<td>0.6</td>
<td>126</td>
<td>0.0526</td>
</tr>
<tr>
<td></td>
<td>115.5</td>
<td>0.0418</td>
<td></td>
<td>126.5</td>
<td>0.0465</td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>0.0279</td>
<td></td>
<td>134</td>
<td>0.0111</td>
</tr>
<tr>
<td></td>
<td>121</td>
<td>0.0070</td>
<td></td>
<td>134.5</td>
<td>0.0094</td>
</tr>
<tr>
<td>0.2</td>
<td>120</td>
<td>0.0653</td>
<td>0.7</td>
<td>127</td>
<td>0.0503</td>
</tr>
<tr>
<td></td>
<td>120.5</td>
<td>0.0376</td>
<td></td>
<td>127</td>
<td>0.0483</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>0.0201</td>
<td></td>
<td>135</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td>125.5</td>
<td>0.0099</td>
<td></td>
<td>135.5</td>
<td>0.0086</td>
</tr>
<tr>
<td>0.3</td>
<td>122</td>
<td>0.0577</td>
<td>0.8</td>
<td>127</td>
<td>0.0504</td>
</tr>
<tr>
<td></td>
<td>122.5</td>
<td>0.0487</td>
<td></td>
<td>127</td>
<td>0.0445</td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>0.0109</td>
<td></td>
<td>135</td>
<td>0.0110</td>
</tr>
<tr>
<td></td>
<td>130.5</td>
<td>0.0067</td>
<td></td>
<td>135.5</td>
<td>0.0092</td>
</tr>
<tr>
<td>0.4</td>
<td>124</td>
<td>0.0538</td>
<td>0.9</td>
<td>127</td>
<td>0.0517</td>
</tr>
<tr>
<td></td>
<td>124.5</td>
<td>0.0481</td>
<td></td>
<td>127</td>
<td>0.0450</td>
</tr>
<tr>
<td></td>
<td>131.5</td>
<td>0.0104</td>
<td></td>
<td>135</td>
<td>0.0114</td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>0.0098</td>
<td></td>
<td>135.5</td>
<td>0.0093</td>
</tr>
<tr>
<td>0.5</td>
<td>125</td>
<td>0.0552</td>
<td>1.0</td>
<td>127</td>
<td>0.0526</td>
</tr>
<tr>
<td></td>
<td>125.5</td>
<td>0.0482</td>
<td></td>
<td>128</td>
<td>0.0446</td>
</tr>
<tr>
<td></td>
<td>133</td>
<td>0.0112</td>
<td></td>
<td>135</td>
<td>0.0116</td>
</tr>
<tr>
<td></td>
<td>133.5</td>
<td>0.0093</td>
<td></td>
<td>136</td>
<td>0.0093</td>
</tr>
</tbody>
</table>
B  Simulation Results—Time Savings

The following seven figures display additional simulation results demonstrating the timesaving for various combinations of \( p_c \) and \( p_v \).

Figure 5. Time saved with use of \( G \) test when \( p_c = 0.9 \) for various values of \( p_v \).

Figure 6. Time saved with use of \( G \) test when \( p_c = 0.7 \) for various values of \( p_v \).

Figure 7. Time saved with use of \( G \) test when \( p_c = 0.6 \) for various values of \( p_v \).

Figure 8. Time saved with use of \( G \) test when \( p_c = 0.5 \) for various values of \( p_v \).

Figure 9. Time saved with use of \( G \) test when \( p_c = 0.4 \) for various values of \( p_v \).

Figure 10. Time saved with use of \( G \) test when \( p_c = 0.3 \) for various values of \( p_v \).

Figure 11. Time saved with use of \( G \) test when \( p_c = 0.1 \) for various values of \( p_v \).
A Comparison of Graphical Methods for Assessing the Proportional Hazards Assumption in the Cox Model

Inger Persson, AstraZeneca R&D, Södertälje, SWEDEN
Harry Khamis, Statistical Consulting Center, Wright State University, Dayton, Ohio 45435, USA

ABSTRACT Six graphical procedures to check the assumption of proportional hazards for the Cox model are described and compared. A new way of comparing the graphical procedures using a Kolmogorov-Smirnov like maximum deviation criterion for rejection is derived for each procedure. The procedures are evaluated in a simulation study under proportional hazards and five different forms of nonproportional hazards: (1) increasing hazards, (2) decreasing hazards, (3) crossing hazards, (4) diverging hazards, and (5) nonmonotonic hazards. The procedures are compared in the two-sample case corresponding to two groups with different hazard functions. None of the procedures under consideration require partitioning of the survival time axis. Results indicate that the Arjas plot, a plot of estimated cumulative hazard versus number of failures, is superior to the other procedures under almost every form of nonproportional hazards, especially crossing and nonmonotonic hazards. For increasing hazards, the smoothed plot of the ratio of log cumulative baseline hazard rates versus time or the smoothed plot of scaled Schoenfeld residuals versus time perform the best. The Andersen plot performs very poorly for increasing, decreasing, and diverging hazards.

1. INTRODUCTION

The relation between the distribution of event times and time-invariant covariates or risk factors \( z \) (\( z \) is a \( p \times 1 \) vector) can be described in terms of a model according to Cox (1972), in which the hazard rate at time \( t \) for an individual is

\[
\lambda(t, z) = \lambda_0(t)e^{\beta'z},
\]

(1.1)

where \( \lambda_0(t) \) is the baseline hazard rate, an unknown (arbitrary) function giving the hazard rate for the standard set of conditions \( z = 0 \), and \( \beta \) is a \( p \times 1 \) vector of unknown parameters. The factor \( e^{\beta'z} \) describes the hazard for an individual with covariates \( z \) relative to the hazard at standard conditions \( z = 0 \).

The ratio of the hazard functions for two individuals with covariate values \( z \) and \( z' \) is

\[
\frac{\lambda(t \mid z)}{\lambda(t \mid z')} = e^{\beta'(z-z')},
\]

an expression that does not depend on \( t \). Thus, the Cox model in (1.1) is only valid for data consistent with the assumption of proportional hazards.

Since the validity of the Cox regression analysis based on the model in (1.1) relies on the assumption of proportionality of the hazard rates of individuals with distinct covariate values, it is important to be able to reliably determine if the assumption is plausible. This can be done graphically or numerically. A partial review of numerous graphical and analytical methods for checking the adequacy of Cox models was given by Lin and Wei (1991). Some authors recommend using numerical tests for such determinations (e.g., Hosmer and Lemeshow, 1999, p. 207). However, others recommend graphical procedures arguing that the proportional hazards assumption only approximates the correct model for a covariate, and that any formal statistical test, based on a large enough sample size, will reject the null hypothesis of proportionality (Klein and Moeschberger, 1997, p. 354). A comprehensive comparative study of numerical procedures is given elsewhere (Persson, 2002). This paper focuses on the effectiveness of graphical procedures. In section 2, six graphical methods for determining the plausibility of the proportional hazards assumption are described. In section 3, the results of a comparative simulation study are presented. A discussion is given in section 4, an example is presented in section 5, and conclusions are given in section 6.

2. GRAPHICAL METHODS COMPARED

Hess (1995) describes eight graphical methods for detecting violations of the proportional hazards
assumption and demonstrated each on three authentic data sets. Five of those methods are described in sections 2.1 – 2.5 below. The methods not included in this paper are (1) methods that require a partitioning of the time axis, which introduces a certain degree of arbitrariness into the procedure, leading to different conclusions depending on the partition used, or (2) methods that do not allow a comparison with other methods through the use of a maximum deviation criterion proposed in this paper. Section 2.6 describes an additional graphical method not included in the article by Hess (1995), the lesser used Arjas plot (Arjas, 1988). These six graphical methods are compared through a simulation study.

Comparing graphical methods can be somewhat arbitrary since there are no clear guidelines for how to interpret the plots. The conclusions are highly dependent on the subjectivity of the viewer. However, to make it possible to compare the results of the different methods, a criterion for rejection is derived for each method individually using measures described in sections 2.1 – 2.6. In each case, a Kolmogorov-Smirnov like maximum deviation criterion is used. See Lin et al. (1993) for an illustration of this approach.

2.1 Method 1: Plot of Survival Curves Based on the Cox Model and Kaplan-Meier Estimates for Each Group

The survival function, \( S(t) \), is related to the cumulative hazards function, \( H(t) \), as follows:

\[
S(t) = e^{-H(t)} = \exp\{-H_0(t)\exp(\beta z)\}
\]

\( = S_0(t)^{\exp(\beta z)}, \quad (2.1.1)\)

where \( H_0(t) \) is the cumulative baseline hazard and \( S_0(t) \) is the baseline survival function. Breslow (1974) gives an estimate for the cumulative baseline hazard based on the Cox proportional hazards model,

\[
\hat{H}_0(t) = \sum_{i, t_i \leq t} \delta_i \sum_{j \in R_i} e^{\hat{\beta}_j}, \quad (2.1.2)
\]

where \( \delta_i \) is the event indicator for the \( i \)th individual and \( R_i \) is the risk set at time \( t_i \), i.e., the set of individuals still under study at a time just prior to \( t_i \). Kalbfleish and Prentice (1980) and Link (1984) provide additional estimates for the cumulative baseline hazard. The baseline survival function can be written \( S_0(t) = e^{-H_0(t)} \). Thus, an estimate of the baseline survival function based on the Cox model is given by

\[
\hat{S}_0(t) = e^{-\hat{H}_0(t)}. \quad (2.1.3)
\]

It is possible to assess violations of the assumption of proportional hazards by comparing survival estimates based on the Cox model with estimates computed independently of the model, such as the Kaplan-Meier product-limit estimate for each group (Kaplan and Meier, 1958), defined by

\[
\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left[ 1 - \frac{d_i}{Y_i} \right] & \text{if } t \geq t_1 \end{cases} \quad (2.1.4)
\]

where \( d_i \) is the observed number of events at time \( t_i \) and \( Y_i \) is the number at risk at time \( t_i \) (i.e., the number of individuals who are alive at time \( t_i \) or experience the event of interest at time \( t_i \)). See Kleinbaum (1996), Chapter 3 for a discussion of the quantitative comparison of the Kaplan-Meier and Cox regression estimates.

Clear departures of the two estimates provide evidence against the assumption of proportional hazards. Figure 2.1.1 shows an example of plots of survival curves based on the Cox model along with Kaplan-Meier estimates for each of two groups of patients.

![Figure 2.1.1 Survival curves based on the Cox model along with Kaplan-Meier estimates for each of two groups of patients](image-url)
The maximum absolute difference between the curves is used to establish possible deviations from the assumption of proportional hazards. This criterion is used in the Kolmogorov-Smirnov test for goodness-of-fit of two cumulative distribution functions (see, e.g., Sokal and Rohlf, 1995). The larger the absolute difference between the curves, the stronger the indication of violations of the proportional hazards assumption. Let $\text{Diff}_{\text{max1}}$ denote the maximum absolute difference between the curves, then the hazards are proportional if $\text{Diff}_{\text{max1}} = 0$. The larger the value of $\text{Diff}_{\text{max1}}$, the stronger the evidence of nonproportionality. Figure 2.1.2 shows the distribution of 10,000 generated $\text{Diff}_{\text{max1}}$ values under proportional hazards.

This distribution can be used to establish a criterion for determining that the proportional hazards assumption is not plausible. We use the 95th percentile as such a criterion, namely, that value of $x$ for which $\Pr(\text{Diff}_{\text{max1}} > x) = 0.05$. So, to check the assumption of proportional hazards, $\text{Diff}_{\text{max1}}$ is calculated and it is concluded that the hazards are not proportional if $\text{Diff}_{\text{max1}}$ exceeds $x$.

### 2.2 Method 2: Plot of Cumulative Baseline Hazards in Different Groups

Another method to graphically check the assumption of proportional hazards is based on the estimated cumulative baseline hazard rate, namely, the Andersen (1982) plot.

Let $\hat{H}_g(t)$ be the estimated cumulative baseline hazard rate in stratum $g$, $g = 1, 2, \ldots, K$. Plot, for all $t$, $\hat{H}_g(t)$ for $g = 2, \ldots, K$. If the proportional hazards assumption is true, then these curves should be straight lines through the origin. Figure 2.2.1 shows an example of a plot of the cumulative baseline hazards in two groups of patients.

![Figure 2.2.1 Estimated cumulative baseline hazard rate in group 2 versus group 1](image)

To determine if this curve follows a straight line through the origin, estimation of a linear regression with no intercept of $\hat{H}_2(t)$ on $\hat{H}_1(t)$ is proposed. Let $\text{Diff}_{\text{max2}}$ denote the maximum absolute difference between $\hat{H}_2(t)$ and the estimated (fitted) values from the regression. Figure 2.2.2 shows the distribution of 10,000 generated $\text{Diff}_{\text{max2}}$ values under proportional hazards.

![Figure 2.2.2 10,000 $\text{Diff}_{\text{max2}}$ values generated under proportional hazards](image)

The assumption of proportional hazards is concluded to be implausible if the calculated value of $\text{Diff}_{\text{max2}}$ exceeds the 95th percentile of this distribution.
2.3 Method 3: Plot of the Difference of the Log Cumulative Baseline Hazard Versus Time

Schumacher (1990) suggested plotting \( \hat{\gamma}(t) \) versus \( t \), where

\[
\hat{\gamma}(t) = \log[H_i(t)] - \log[H_0(t)]. \tag{2.3.1}
\]

Under proportional hazards this plot is constant over \( t \), centered around the estimated log hazard ratio \( \hat{\beta} \). Figure 2.3.1 shows an example of a plot of the difference of the log cumulative baseline hazard versus time.

![Figure 2.3.1 Plot of \( \hat{\gamma}(t) \) versus time](image)

Let \( \text{Diff}_{\text{max}3} \) denote the maximum absolute difference between \( \hat{\gamma}(t) \) and \( \hat{\beta} \). Figure 2.3.2 shows the distribution of 10,000 generated \( \text{Diff}_{\text{max}3} \) values under proportional hazards.

![Figure 2.3.2 10,000 \( \text{Diff}_{\text{max}3} \) values generated under proportional hazards](image)

The hazards are concluded to be nonproportional if the calculated value of \( \text{Diff}_{\text{max}3} \) exceeds the 95\(^{th}\) percentile of this distribution.

2.4 Method 4: Smoothed Plot of the Ratio of Log Cumulative Baseline Hazard Rates Versus Time

Smoothing helps describe the pattern of dependence, thus making it easier to check the constancy of \( \hat{\gamma}(t) \) when plotting it against \( t \) as described in subsection 2.3. The choice of smoothing technique is usually not very important as long as the smoother (1) is sensitive to local rather than global features of the data and (2) has an appropriate number of degrees of freedom (Hastie and Tibshirani, 1990). For example, LOWESS (locally-weighted scatter plot smoothing) employs iterated weighted least squares with a robustness feature that identifies and down-weights outliers in successive smoothings. Figure 2.4.1 shows an example of a smoothed plot of the difference of the log cumulative baseline hazard versus time using LOWESS.

![Figure 2.4.1 Smoothed plot of \( \hat{\gamma}(t) \) versus time](image)

Let \( \text{Diff}_{\text{max}4} \) denote the maximum absolute difference between the smoothed values of \( \hat{\gamma}(t) \) and \( \hat{\beta} \). Figure 2.4.2 shows the distribution of 10,000 generated \( \text{Diff}_{\text{max}4} \) values under proportional hazards.

![Figure 2.4.2 \( \text{Diff}_{\text{max}4} \) values generated under proportional hazards](image)
Figure 2.4.2 10,000 $\text{Diff}_{\text{max}4}$ values generated under proportional hazards

The hazards are concluded to be nonproportional if the calculated value of $\text{Diff}_{\text{max}4}$ exceeds the 95\textsuperscript{th} percentile of this distribution.

2.5 Method 5: Smoothed Plot of Scaled Schoenfeld Residuals Versus Time

Schoenfeld (1980) defined partial residuals for the Cox model that do not depend on time, so that the \( j \)th residual can be plotted against \( t_j \) to detect violations of the proportional hazards assumption, where \( j \) indexes individuals (\( j = 1, 2, \ldots, n \)). The Schoenfeld residuals are defined as

\[
\begin{align*}
    r_i(\hat{\beta}) &= z_i - M(\hat{\beta}, t_i), \\
    \end{align*}
\]

where \( z_i \) is the covariate vector of the subject with an event at time \( t_i \), where \( i \) indexes event times (\( i = 1, 2, \ldots, D \)), and \( M(\hat{\beta}, t_i) \) is the conditional weighted mean of the covariate vector at time \( t_i \) as described in Persson (2002), section 2.2. Grambsch and Therneau (1994) describe a scale adjustment for Schoenfeld’s residuals,

\[
\begin{align*}
    \hat{r}_i^*(\hat{\beta}) &= \hat{\beta} + \hat{r}_i d \hat{V}(\hat{\beta})^{-1}, \\
    \end{align*}
\]

where \( \hat{\beta} \) is the maximum partial likelihood estimate under proportional hazards, \( \hat{V}(\hat{\beta}) \) is the estimated variance of \( \hat{\beta} \), and \( d \) is the total number of events where individuals from both groups remain at risk. For a binary covariate coded 0 or 1, a plot of \( r_i^* \) versus \( t_i \) yields two horizontal bands of residuals. If the proportional hazards assumption holds, then the residuals center around \( \hat{\beta} \). Smoothing improves the interpretability of the residual plots, so LOWESS is applied. Figure 2.5.1 shows an example of a smoothed plot of the scaled Schoenfeld residuals versus time.

Figure 2.5.1 Smoothed plot of scaled Schoenfeld residuals versus time

Let $\text{Diff}_{\text{max}5}$ denote the maximum absolute difference between the smoothed residuals and $\hat{\beta}$. Figure 2.5.2 shows the distribution of 10,000 generated $\text{Diff}_{\text{max}5}$ values under proportional hazards.

Figure 2.5.2 10,000 $\text{Diff}_{\text{max}5}$ values generated under proportional hazards

The hazards are concluded to be nonproportional if the calculated value of $\text{Diff}_{\text{max}5}$ exceeds the 95\textsuperscript{th} percentile of this distribution.

2.6 Method 6: Plot of the Estimated Cumulative Hazard Versus Number of Failures

Arjas (1988) suggested a plot of the estimated cumulative hazard versus the number of failures in each stratum (Arjas plot). See Section
VII.3.4 of Andersen et al. (1993) for the development, of which the following is a brief summary. Consider the differences

\[ N_h(t) - \int_0^t \sum_{h(t)=h} p_i(u, \beta_0) dN(u), h = 1, ..., k, \]

where \( N_h(t) = \sum_{h(t)=h} N_i(t) \), \( N \) is the process counting the observed failures, and \( \beta_0 = (\beta_0^1, ..., \beta_0^{p-1}, \beta_0^p, ..., \beta_0^{p+k-1}) \) is the true parameter vector. These differences are (local) martingales. Therefore, plots of

\[ \int_0^t \sum_{h(t)=h} p_i(u, \hat{\beta}) dN(u), m = 1, ..., N_h(\tau), h = 1, ..., k \]

versus \( m \), where \( X^h_m, m = 1, ..., N_h(\tau) \) are the ordered jump times in stratum \( h \), should be approximately straight lines with unit slope.

This plot is a Total Time on Test plot for the residuals \( \hat{p} \). Tests for the proportional hazards model based on these residuals were briefly discussed by Arjas (1988) and Arjas and Haara (1988).

Figure 2.6.1 shows an example of an Arjas plot of estimated cumulative hazard versus number of failures in each stratum (here the strata are the two groups).

If the hazards are proportional, then these curves should be approximately linear with slope close to one. However, even under proportional hazards the curves may differ from the 45 degree line, as seen in Figure 2.6.1, but they are still fairly linear. When the hazards are not proportional, then the curves are roughly as close to the 45 degree line as under proportionality, but the curves are not linear. To determine if these curves differ nonlinearly from the 45 degree line, one can estimate, for each stratum, a linear regression of \( \sum H_j(t_i) \) on \( \sum N_j(T_i) \), where \( H_j(t_i) \) is the cumulative hazard for the \( j \)th individual in the sample at time \( t_i \), \( i = 1, ..., D \). Let \( \text{Diff}_{max6} \) denote the maximum absolute difference between \( \sum H_j(t_i) \) and the estimated (fitted) values from the regression. Figure 2.6.2 shows the distribution of 10,000 generated \( \text{Diff}_{max6} \) values under proportional hazards.

![Estimated cumulative hazard](image)

*Figure 2.6.1 Arjas plot*

The hazards are concluded to be nonproportional if the calculated value of \( \text{Diff}_{max6} \) exceeds the 95th percentile of this distribution.

3. SIMULATION STUDY

The six graphical methods described in subsections 2.1 – 2.6 are evaluated under proportional hazards and five different forms of nonproportional hazards: (1) increasing hazards, (2) decreasing hazards, (3) crossing hazards, (4) diverging hazards, and (5) nonmonotonic hazards. The methods are compared in the two-sample case corresponding to two groups with different hazard functions. Equal sample sizes
of 30, 50, and 100 observations per group are used along with average censoring rates of 10, 25, 50, and 70 percent. Random (noninformative) censoring using an exponential censoring distribution is incorporated. The smallest sample size is not used at the highest censoring rate because of the small number (18) of events that would result. The number of repetitions used in each simulation is 10,000.

Random samples of survival times, $T_s$, are generated from the Weibull distribution in all cases except for the nonmonotonic hazards, where the lognormal distribution, having probability density function

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[ -\frac{1}{2} \left( \frac{\ln(t) - \mu}{\sigma} \right)^2 \right].$$

is used. The hazard of the Weibull distribution is defined as $h(t) = \alpha \gamma (\alpha t)^{\gamma-1}$, where $\alpha$ is the scale parameter and $\gamma$ is the shape parameter. Details about parameter values are described in each case below. Censoring times, $T_c$, are generated from the exponential distribution with hazard function $h(t) = \lambda$, where the value of the parameter $\lambda$ is adjusted to achieve the desired censoring rates. The time on study, $T$, is defined as $T = \min(T_s, T_c)$, where $T_s$ and $T_c$ are independent.

The criteria described in subsections 2.1 – 2.6 are used for rejection of the null hypothesis of proportional hazards for a test procedure that is adjusted for the appropriate sample size and censoring rate. A significance level of 5% is used. The results from the simulations are presented in the form of a plot and a numerical summary (table) presenting the proportion of times that the criterion, $D_{\text{max}k}$, $k = 1, 2, \ldots, 6$, exceeds the $95^{\text{th}}$ percentile of the corresponding reference distribution, thus indicating “strong” evidence that the hazards are not proportional.

In subsections 3.1 – 3.6 below, the parameter settings for the survival distributions and the figures are given as follows.

<table>
<thead>
<tr>
<th>Subsec</th>
<th>Hazards</th>
<th>Survival Dist.</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Fig</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Proportional</td>
<td>Weibull($\alpha, \gamma$)</td>
<td>$\alpha, \gamma$=(1,1)</td>
<td>$\alpha, \gamma$=(2,1)</td>
<td>3.1.1</td>
</tr>
<tr>
<td>3.2</td>
<td>Increasing</td>
<td>Weibull($\alpha, \gamma$)</td>
<td>$\alpha, \gamma$=(2,1.5)</td>
<td>$\alpha, \gamma$=(2,2)</td>
<td>3.2.1</td>
</tr>
<tr>
<td>3.3</td>
<td>Decreasing</td>
<td>Weibull($\alpha, \gamma$)</td>
<td>$\alpha, \gamma$=(2,3)</td>
<td>$\alpha, \gamma$=(2,5)</td>
<td>3.3.1</td>
</tr>
<tr>
<td>3.4</td>
<td>Crossing</td>
<td>Weibull($\alpha, \gamma$)</td>
<td>$\alpha, \gamma$=(2,1.5)</td>
<td>$\alpha, \gamma$=(5,1)</td>
<td>3.4.1</td>
</tr>
<tr>
<td>3.5</td>
<td>Diverging</td>
<td>Weibull($\alpha, \gamma$)</td>
<td>$\alpha, \gamma$=(1,95)</td>
<td>$\alpha, \gamma$=(1,1.5)</td>
<td>3.5.1</td>
</tr>
<tr>
<td>3.6</td>
<td>Nonmonotonic</td>
<td>Lognormal($\mu, \sigma$)</td>
<td>$\mu, \sigma$=(.3,1)</td>
<td>$\mu, \sigma$=(1,1)</td>
<td>3.6.1</td>
</tr>
</tbody>
</table>

### 3.1 Proportional Hazards

The proportion of times that $D_{\text{max}k}$ exceeds the $95^{\text{th}}$ percentile of the reference distribution is given for each censoring rate, sample size, and $k = 1, 2, \ldots, 6$ in Figure 3.1.2. The numerical values can be found in Appendix Table A1.

![Figure 3.1.1 Constant hazards.](image)
In order to compare the percentages, one can calculate the standard deviation of the proportion under the proportional hazards model:

\[
\sigma_p = \sqrt{\frac{p(1-p)}{n_{rep}}} = \sqrt{\frac{(0.05)(0.95)}{5000}} \approx 0.0031
\]

The standard deviation of the difference between two independent proportions is \(\sigma_{\text{diff}} \approx 0.0044\), so \(3\sigma_{\text{diff}} \approx 0.013\) can be used as an informal benchmark for a real difference between significance levels. However, due to the multiplicity of comparisons, this benchmark must be used cautiously.

All of the graphical procedures behave as expected, with the percentage of rejections of proportional hazards close to the significance level of 5%. In fact, all percentages fall between 0.045 and 0.055.

3.2 Increasing Hazards

The proportion of times that \(\text{Diff}_{\text{max}}\) exceeds the 95\(^{th}\) percentile of the reference distribution is given for each censoring rate, sample size, and \(k = 1, 2, \ldots, 6\) in Figure 3.2.2. The numerical values can be found in Appendix Table A2.

---

![Figure 3.1.2 Proportional hazards](image1)

![Figure 3.2.1 Increasing hazards](image2)
Figure 3.2.2 Increasing hazards

In order to compare the percentages, one can calculate the maximum standard deviation of the difference between two proportions:

$$\sigma_{\text{diff}}^{\text{max}} = \sqrt{2 \times \left(\frac{0.5}{5000}\right)^2} = 0.01,$$

and

$$3\sigma_{\text{diff}}^{\text{max}} = 0.03$$
serves as a conservative benchmark signifying a real difference. The same benchmark can be used in subsections 3.3 – 3.6. Again, because of the multiplicity of comparisons this benchmark must be used cautiously.

Methods 2 and 6 perform relatively poorly at the 10% and 25% censoring rates, while method 1 performs poorly at the 50% and 70% censoring rates.

3.3 Decreasing Hazards

The proportion of times that $\text{Diff}_{\text{max}}$ exceeds the 95\textsuperscript{th} percentile of the reference distribution is given for each censoring rate, sample size, and $k = 1, 2, \ldots, 6$ in Figure 3.3.2. The numerical values can be found in Appendix Table A3.

Figure 3.3.1 Decreasing hazards.
Method 2 is consistently inferior. Methods 1 and 3 perform well at the 10% and 25% censoring rates. Method 6 performs well at the 50% and 70% censoring rates.

3.4 Crossing hazards

The proportion of times that $\text{Diff}_k$ exceeds the 95\textsuperscript{th} percentile of the reference distribution is given for each censoring rate, sample size, and $k = 1, 2, ..., 6$ in Figure 3.4.2. The numerical values can be found in Appendix Table A4.
3.5 Diverging Hazards

The proportion of times that $\text{Diff}_{\text{maxk}}$ exceeds the 95th percentile of the reference distribution is given for each censoring rate, sample size, and $k = 1, 2, …, 6$ in Figure 3.5.2. The numerical values can be found in Appendix Table A5.
Method 2 performs consistently poorly, especially at the 10% and 25% censoring rates.

3.6 Nonmonotonic Hazards

The proportion of rejections is shown for each censoring rate, sample size, and $k = 1, 2, \ldots, 6$ in Figure 3.6.2. The numerical values can be found in Appendix Table A.6.

Figure 3.5.2 Diverging hazards

Figure 3.6.1 Nonmonotonic hazards.
Method 6 is consistently superior. Method 2 also performs consistently well.

4. DISCUSSION

In each of the five forms of nonproportional hazards, the proportion of rejections generally increases with sample size and decreases with censoring rate. The conclusions from the simulations is that Method 6, the Arjas plot, finds nonproportionality more often than the other methods, especially (1) for crossing and nonmonotonic hazards and (2) at the higher censoring rates for decreasing and diverging hazards. For decreasing hazards, Method 1, the Cox and Kaplan-Meier survival versus time plot, is superior at the low censoring rates. The Andersen plot, Method 2, performs poorly in all situations except for nonmonotonic hazards where it performs well.

These results are consistent with a viewing of the plots derived from data sets. Figure 4.1 shows examples of Method 6, which performed well for crossing and nonmonotonic hazards. The sample size is 100 and the censoring rate is 10% for both plots.
It is fairly easy to see that at least one of the curves differs nonlinearly from the 45 degree line. Under crossing hazards the distance between the curves and the 45 degree line is also larger than it is under proportional hazards (see Figure 2.6.1). Even though the Arjas plot did not perform as well under increasing hazards, it is still easy to see that the curves differ nonlinearly from the 45 degree line (Figure 4.1a; compare to Figure 2.6.1).

The other method that performed well in the simulations, especially at low censoring rates, except for crossing and nonmonotonic hazards, is Method 1. Figure 4.2 shows an example of that plot under decreasing hazards, sample size 100 and 10% censoring rate. A departure of the two estimates can be seen in the figure (compare to Figure 2.1.1).

The Andersen plots look similar to this under any of the nonproportional hazards cases; only in a few cases would it be possible to detect a deviation from a straight line with the naked eye.

5. APPLICATION

A multicenter study of the disease CML, chronic myeloid leukemia, was initiated in 1984 at the University Hospital in Uppsala, Sweden. CML is a cancer of the blood cells where the patient has a high number of on time od cells, granulocytes, in bone marrow and blood. The treatment of this disease aims to reduce the number of white blood cells. “Cell-restraining drugs” which reduce the production of these blood cells are used in treatment. The two treatments, busulphane and hydroxyurea, were widely used all over the world at the time of this study. In previous studies, these treatments were found to be equally effective at prolonging the lifetimes of the patients (Hehlmann et al., 1993 and Alan et al., 1995).

Patients were recruited from all hospitals in Sweden. All patients older than five years and willing to participate, diagnosed with CML from January 1, 1984 until December 31, 1988, were included in the study. The patients were randomized to one of the two treatments at the date of diagnosis. All patients younger than approximately 45 years of age with a compatible donor (only brothers or sisters) were offered bone marrow transplantation. The last patient was included in the study in May 1988, and all patients were followed until February 1998. A total of 63 patients were included in the study, 26 of which received bone marrow transplantation. Figure 5.1 shows the Kaplan-Meier survival curves for patients who received a transplant (transpl 1) and those who did not receive a transplant (transpl 0).
Figure 5.1 Kaplan-Meier survival for transplanted and not transplanted patients.

The censoring rate for these data is 16%. The transplantation covariate (1 = Yes, 0 = No) was believed to be time-dependent, so that the proportional hazards assumption for the Cox model was under question. Figure 5.2 shows the hazard rates for the two groups.

Figure 5.2 Hazard rates for transplanted and not transplanted patients.

The hazard rate $\lambda$ early times and then diverge. From the results of section 3, the Arjas plot (Method 6) should be an effective graphical method to assess the proportional hazards assumption.

Figures 5.3 – 5.8 show the six different graphical methods described in sections 2.1 – 2.6 applied to the CML data with transplantation as a single binary covariate.

Figure 5.3 Method 1, Survival curves based on the Cox model along with Kaplan-Meier estimates for transplanted and not transplanted patients.

Figure 5.4 Method 2, Estimated cumulative baseline hazard rate for transplanted patients versus not transplanted patients (Andersen plot).

Figure 5.5 Method 3, Plot of $\hat{\lambda}(t)$ versus time.
General conclusions from these plots are given as follows.

Method 1.

There is a clear difference between the two estimates, especially for transplanted patients (compare to Figure 2.1.1), which signals a violation of the proportional hazards assumption.

Method 2.

A deviation from linearity can be detected in the Andersen plot (compare to Figure 2.2.1).

Method 3.

This plot is not constant over time, as would be the case if the hazards were proportional (compare to Figure 2.3.1).

Method 4.

The smoothing helps the analyst to determine that the plot is not constant over time (compare to Figure 2.4.1).

Method 5.

The residuals do not tend to center around $\hat{\beta}$ (compare to Figure 2.5.1).

Method 6.

Both curves cross the 45 degree line and differ nonlinearly from it, especially the curve for the transplanted patients (compare to Figure 2.6.1).

CONCLUSION

Assessing graphs for the purpose of determining the severity of model assumption violations can be difficult because of the lack of objectivity involved. To the untrained eye, several of the plots in Figures 2.1.1 – 2.6.1 may appear to signal a violation of the proportional hazards assumption (e.g., Figures 2.2.1 and 2.4.1) even though they were generated from models having proportional hazards.

By using a Kolmogorov-Smirnov like maximum deviation criterion upon which to base comparisons of six different graphical procedures, this simulation study shows that the Arjas plot is generally the most effective at identifying nonproportional hazards, especially for (1) crossing and nonmonotonic hazards and (2) decreasing and diverging hazards where the censoring rate is high. It is interesting to note that the effectiveness of the Arjas plot at identifying nonproportional hazards remains relatively constant across the censoring rates while for most all of the other methods the
The proportion of rejections tends to decrease with censoring rate.

When the proportion of rejections is averaged over sample sizes and censoring rates, Method 2 performs the worst under increasing, decreasing, and diverging hazards while Method 6 performs the best under crossing and nonmonotonic hazards. The average rejection rates are given as follows:

<table>
<thead>
<tr>
<th>Method</th>
<th>Hazards</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increasing</td>
<td>.18</td>
<td>.08</td>
<td>.16</td>
<td>.20</td>
<td>.20</td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td>Decreasing</td>
<td>.39</td>
<td>.02</td>
<td>.35</td>
<td>.21</td>
<td>.26</td>
<td>.37</td>
<td></td>
</tr>
<tr>
<td>Crossing</td>
<td>.21</td>
<td>.17</td>
<td>.14</td>
<td>.33</td>
<td>.12</td>
<td>.56</td>
<td></td>
</tr>
<tr>
<td>Diverging</td>
<td>.71</td>
<td>.13</td>
<td>.79</td>
<td>.83</td>
<td>.65</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>Nonmonotonic</td>
<td>.11</td>
<td>.27</td>
<td>.08</td>
<td>.12</td>
<td>.13</td>
<td>.41</td>
<td></td>
</tr>
</tbody>
</table>

Method 6, the Arjas plot, has one of the top two average rejection rates for four of the five forms of nonproportional hazards. For increasing hazards, where Method 6 has the fifth highest average rejection rate, Methods 4 or 5 would be recommended.

The maximum absolute deviation criterion used in this study is consistent with the practical usage of plots to determine if model assumptions are plausible. That is, when one visually analyzes a plot, one is searching for the deviation between the observed plot and the plot one expects to see under the model assumption. This study merely formalizes this process.

It is recommended that in general the Arjas plot (Method 6) be used as the preferred graphical procedure for checking the proportional hazards assumption if the form of the nonproportional hazards is anything but increasing. For increasing hazards, Methods 4 and 5 are superior.

Generally, it is recommended that the proportional hazards assumption always be checked in the Cox model, and that while a plot such as the Arjas plot is a helpful tool, it should not be the only basis upon which to make a decision regarding the plausibility of the proportional hazards assumption.

REFERENCES


## APPENDICES

### Table A.1. Proportion of rejections for constant hazards.

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>Censoring</th>
<th>Sample size in each group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>25%</td>
</tr>
<tr>
<td>1. Cox and Kaplan-Meier Survival vs. time</td>
<td>.049</td>
<td>.048</td>
</tr>
<tr>
<td>2. Cumulative Baseline Hazard (Andersen plot)</td>
<td>.045</td>
<td>.048</td>
</tr>
<tr>
<td>5. Smoothed scaled Schoenfeld residuals vs. Time</td>
<td>.048</td>
<td>.049</td>
</tr>
<tr>
<td>6. Arjas plot of Cum. Hazard vs. number of failures</td>
<td>.050</td>
<td>.051</td>
</tr>
</tbody>
</table>

### Table A.2. Proportion of rejections for increasing hazards.

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>Censoring</th>
<th>Sample size in each group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>25%</td>
</tr>
<tr>
<td>1. Cox and Kaplan-Meier Survival vs. time</td>
<td>.207</td>
<td>.296</td>
</tr>
<tr>
<td>2. Cumulative Baseline Hazard (Andersen plot)</td>
<td>.029</td>
<td>.021</td>
</tr>
<tr>
<td>6. Arjas plot of Cum. Hazard vs. number of failures</td>
<td>.066</td>
<td>.081</td>
</tr>
</tbody>
</table>
### Table A.3. Proportion of rejections for decreasing hazards

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>Censoring</th>
<th>Sample size in each group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>30  50  100</td>
<td>30  50  100</td>
</tr>
<tr>
<td>2. Cumulative Baseline Hazard (Andersen plot)</td>
<td>.020 .011 .004</td>
<td>.021 .011 .006</td>
</tr>
<tr>
<td>5. Smoothed scaled Schoenfeld residuals vs. Time</td>
<td>.204 .313 .596</td>
<td>.210 .310 .481</td>
</tr>
<tr>
<td>6. Arjas plot of Cum. Haz. vs. number of failures</td>
<td>.198 .298 .578</td>
<td>.216 .322 .559</td>
</tr>
</tbody>
</table>

### Table A.4. Proportion of rejections for crossing hazards

<table>
<thead>
<tr>
<th>Test statistic</th>
<th>Censoring</th>
<th>Sample size in each group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>30  50  100</td>
<td>30  50  100</td>
</tr>
<tr>
<td>1. Cox and Kaplan-Meier Survival vs. time</td>
<td>.159 .296 .601</td>
<td>.127 .216 .484</td>
</tr>
<tr>
<td>2. Cumulative Baseline Hazard (Andersen plot)</td>
<td>.141 .183 .214</td>
<td>.166 .188 .232</td>
</tr>
<tr>
<td>5. Smoothed scaled Schoenfeld residuals vs. Time</td>
<td>.052 .148 .369</td>
<td>.040 .090 .196</td>
</tr>
</tbody>
</table>
Table A.5. Proportion of rejections for diverging hazards

<table>
<thead>
<tr>
<th>Censoring</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test statistic</strong></td>
<td><strong>Sample size in each group</strong></td>
<td><strong>30</strong></td>
<td><strong>50</strong></td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>1. Cox and Kaplan-Meier Survival vs. time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Cumulative Baseline Hazard (Andersen plot)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Smoothed scaled Schoenfeld residuals vs. Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Arjas plot of Cum. Hazard vs. number of failures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.6. Proportion of rejections for nonmonotonic hazards

<table>
<thead>
<tr>
<th>Censoring</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test statistic</strong></td>
<td><strong>Sample size in each group</strong></td>
<td><strong>30</strong></td>
<td><strong>50</strong></td>
<td><strong>100</strong></td>
</tr>
<tr>
<td>1. Cox and Kaplan-Meier Survival vs. time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Cumulative Baseline Hazard (Andersen plot)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Smoothed scaled Schoenfeld residuals vs. Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Arjas plot of Cum. Hazard vs. number of failures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A Multistage Model for Analyzing Repeated Observations on Depression in Elderly

M. Ataharul Islam*
Professor
School of Mathematical Sciences,
Universiti Sains Malaysia,
11800 Penang, Malaysia

Rafiqul I Chowdhury
Senior Lecturer,
Department of Health Information Administration,
Faculty of Allied Health Sciences,
Kuwait University, Kuwait

Shahariar Huda
Professor
Department of Statistics and OR,
College of Science, Kuwait University, Kuwait

*Corresponding Author: E-mail: mataharul@yahoo.com

Abstract

The problem of depression in elderly, in terms of short and long-term health and economic consequence, needs to be assessed more critically. This study uses the longitudinal data from the six waves of the Health and Retirement Survey to examine the transition to depression, reverse transition from depression to no depression and also repeated transition from no depression to depression after experiencing a reverse transition during a study period. Covariate-dependent Markov models are proposed using the logistic link function based on the Chapman-Kolmogorov equation. A simplified and more flexible approach of constructing the likelihood function is demonstrated in this paper to deal with the branching of a number of transition types starting from a group of subjects with no depression at the beginning of the study.

AMS (2000) subject classification: 62N01, 62N02

Key Word: Chapman-Kolmogorov Equation, Depression, Logistic Regression, Markov Chain, Multistate Model, and Transitions.

1. Introduction

During the past decades, many developed and developing countries have experienced a steady increase in their elderly population. With an increase in proportion of elderly population, the focus of public health also needs major adjustments in health policies in order to face challenges due to change in age composition. One major problem in the elderly population is to deal with the problem of depression. It is evident from several studies that there is a steady increase in the incidence and prevalence of depression and related problems among the elderly population (Doraiswamy, 2001, Birrer and Vemuri, 2004, American Psychiatric Association, 1994, Beekman et al. 2002, Kales and Valenstein, 2002, Schovers et al., 2000, Crystal et al., 2003, Fischer et al., 2003, Williams and Connolly, 1990; Lundquist et al., 1997, Gurland et al., 1980; Murrell et al., 1983).

It is important to understand how the transitions among the different states of depression take place and how the covariates influence these transitions. Markov chain is one class of statistical models, which is used to model these types of transitions. The Markov chain models for discrete variate time series appear to be restricted due to over-parameterization and several attempts have been made to simplify the application of Markov chain models. There are different approaches of employing Markov chain models. Raftery (1985), Raftery and Tavare
(1994), and Berchtold and Raftery (2002) addressed one such area of problems in estimating transition probabilities following the work of Pegram (1980). This area of research is popularly known as the mixture transition distribution (MTD), deals with modeling of high-order Markov chains for a finite state space. In another development, Albert (1994) proposed a finite Markov chain model for analyzing sequences of ordinal data from a relapsing remitting of a disease. In addition, Albert and Waclawiw (1998) developed a class of quasi-likelihood models for a two state Markov chain with stationary transition probabilities for heterogeneous transitional data.

Regier (1968) proposed a model for estimating odds ratio from a two state transition matrix. A grouped data version of the proportional hazards regression model for estimating computationally feasible estimators of the relative risk function was proposed by Prentice and Gloeckler (1978). The role of previous state as a covariate was examined by Korn and Whittemore (1979). Wu and Ware (1979) proposed a model which included accumulation of covariate information as time passes before the event and considered occurrence or non-occurrence of the event under study during each interval of follow up as the dependent variable. The method could be used with any regression function such as the multiple logistic regression model and also the one suggested by Prentice and Gloeckler (1978). Kalbfleisch and Lawless proposed the models for continuous time (1985). They presented procedures for obtaining estimates for transition intensity parameters in homogeneous models. For a first order Markov model, they introduced a model for covariate dependence of log-linear type.

Another class of models has emerged for analyzing transition models with serial dependence of the first or higher orders on the basis of the marginal mean regression structure models. Azzalini (1994) introduced a stochastic model, more specifically, first order Markov model, to examine the influence of time-dependent covariates on the marginal distribution of the of the binary outcome variables in serially correlated binary data. Markov chains are expressed in transitional form rather than marginally and the solutions are obtained such that covariates relate only to the mean value of the process, independent of association parameters. Following Azzalini (1994), Heagerty and Zeger (2000) presented a class of marginalized transition models (MTM) and Heagerty (2002) proposed a class of generalized MTMs to allow serial dependence of first or higher order. These models are computationally tedious and the form of serial dependence is quite restricted. If the regression parameters are strongly influenced by inaccurate modeling for serial correlation then the MTMs can result in misleading conclusions. Heagerty (2002) provided derivatives for score and information computations.

In recent years, there is a great deal of interest in the development of multivariate models based on the Markov Chains. These models have wide range of applications in the fields of reliability, economics, survival analysis, engineering, social sciences, environmental studies, biological sciences, etc. Muenz and Rubenstein (1985) employed logistic regression models to analyze the transition probabilities from one state to another but still there is serious lack of a general methodology for analyzing transition probabilities of higher order Markov models. In a higher order Markov model, we can examine some inevitable characteristics that may be revealed from the analysis of transitions, reverse transitions and repeated transitions. Islam and Chowdury (2006) extended the model for higher order Markov model with covariate dependence for binary outcomes. It is noteworthy that the covariate dependent higher order Markov models can be used to identify the underlying factors associated with such transitions. In this study, it is aimed to provide a comprehensive covariate-dependent Markov Model for higher order. A general procedure is developed comprehensively in this paper to propose the estimation procedure for Markov models for any order. The proposed model and inference procedures are simple and the covariate dependence of the transition probabilities of any order can be examined without making the underlying model complex. Another advantage of the model lies in the fact that the estimation and test procedures for both the specific parameter of interest and the overall model remain easy for practical applications for any longitudinal data.

In a recent study, Harezlak, Gao and Hui (2003), proposed a Markov model consisting of three states, non-diseased, diseased and dead, and estimated the transition hazard parameters using the maximum likelihood approach for analyzing dementia data. Their model does not include any reverse transition or repeated transition. However, instead of considering death as missing, they considered it as a separate state and thus the results became more meaningful. Islam and Singh (1992), Islam (1994), and Islam et al. (2004) demonstrated the analysis of transitions, reverse transitions and repeated transitions for longitudinal data. They employed hazards model for continuous time data with discrete state outcomes. Islam et al. (2004) indicated the possibility of using logistic link functions for
similar analysis. In the previous works, the transitions were considered to follow hidden Markov properties of first or higher orders. In other words, we needed to strictly observe consecutive follow-ups to obtain the history of events. In that case, the branching of possible occurrence of events would produce a large number of models based on past history. This paper introduces a simplified approach based on Chapman-Kolmogorov equations, to formulate the models for transitions, reverse transitions and repeated transitions. The likelihood function proposed in this paper makes the longitudinal problems for transitions among a number of states much flexible and thus provides a more useful generalization of the estimation procedure for repeated measures. In addition, the confirmed deaths during consecutive follow-ups are considered from both non-depressed or depressed states so that the deaths are not included as censored because there might be both qualitative and quantitative differences between other types of censoring and death, particularly in analyzing depression.

The emerging issues concerning the analysis of depression among elderly include the identification of risk factors from longitudinal studies. The Health and Retirement Study (HRS) conducted in the USA provides an opportunity to examine the follow-up data on elderly population covering six consecutive follow-ups over a period of 10 years, each 2 years apart. In this paper, we have proposed a covariate dependent Markov model for three states, no depression, depression and death. We have considered death as a separate state in order to provide a more useful procedure for analyzing the transition from no depression to depression as well as for analyzing reverse transition from depression to no depression as well as repeated transition for those who made a reverse transition at a previous stage.

2. The Covariate Dependent Markov Model for Transition, Reverse Transition and Repeated Transition

Let us consider that a stationary process \((Y_{ij1}, Y_{ij2}, \ldots, Y_{ij})\) represents the past and present responses for subject \(i\) (\(i=1, 2, \ldots, n\)) at follow-up \(j\) (\(j=1, 2, \ldots, J_i\)). Here, \(y_{ij}\) is the response at time \(t_{ij}\). We can think of \(y_{ij}\) as an explicit function of past history of subject \(i\) at follow-up \(j\) denoted by \(H_{ij} = \{y_{ik}, k=1, 2, \ldots, j-1\}\). The transition models for which the conditional distribution of \(y_{ij}\) given \(H_{ij}\) depends on \(q\) prior observations \(y_{ij-1}, \ldots, y_{ij-q}\), is considered as the model of order \(q\).

The multiple outcomes defined by \(y_{ij} = s, s=0, 1, 2, \ldots, m-1\) if an event of level \(s\) occurs for the \(i\)th subject at the \(j\)th follow-up where \(y_{ij}=0\) indicates that no event occurs. Then the first order Markov model can be expressed as

\[
P(y_{ij}|y_{ij-q}, \ldots, y_{ij-1}) = P(y_{ij}|y_{ij-1}). \tag{2.1}
\]

Here, \(0, 1, \ldots, m-1\) are the \(m\) possible outcomes of a dependent variable, \(Y\). The probability of a transition from \(u (u=0, \ldots, m-1)\) at time \(t_{j-1}\) to \(v (v=0, \ldots, m-1)\) at time \(t_j\) is

\[
\pi_{uv} = P(Y_j = v | Y_{j-1} = u). \quad \text{For any } u, \sum_{v=0}^{m-1} \pi_{uv} = 1, \quad u=0, \ldots, m-1. \tag{2.2}
\]

Now let us consider \(m=3\). The states are: no depression (0), depression (1) and death (2). Figure 1 displays the flow of transitions, reverse transitions and repeated transitions, starting at the baseline from the state of no depression. We can define the following probabilities for \(m=3\), using the Chapman-Kolmogorov equations and also equation (2.1). The probability of a transition from \(u (u=0, \ldots, m-1)\) at time \(t_{j-1}\) to \(v (v=0, \ldots, m-1)\) at time \(t_j\) is, where \(t_{j-1}\) is the time of follow-up just prior to \(t_j\):

\[
\pi_{uv} = P(Y_j = v | Y_{j-1} = u).P(Y_{j-1} = u) \tag{2.3}
\]

The probability of a transition from \(u (u=0, \ldots, m-1)\) at time \(t_{j-1}\) (just prior to the follow-up at time \(t_j\)) to \(v (v=0, \ldots, m-1)\) at time \(t_{j-1}\) (just prior to the follow-up at time \(t_{j-1}\)) and \(w\) at time \(t_{j-1}\) is (\(j_2 > j_1\)):

\[
\pi_{uvw} = P(Y_{j_2} = w | Y_{j_2-1} = u, Y_{j_2-1} = v) = P(Y_{j_2} = w | Y_{j_2-1} = v)P(Y_{j_2-1} = v | Y_{j_2-1} = u)P(Y_{j_2-1} = u) \tag{2.4}
\]

The probability of a transition from \(u (u=0, \ldots, m-1)\) at time \(t_{j-1}\) (just prior to the follow-up at time \(t_{j-1}\)) to \(v (v=0, \ldots, m-1)\) at time \(t_{j-1}\) (just prior to the follow-up at time \(t_{j-1}\)) and \(w\) at time \(t_{j-1}\) (just prior to the follow-up at time \(t_{j-1}\)) is:

\[
\pi_{uvw} = P(Y_{j_2} = w | Y_{j_2-1} = u, Y_{j_2-1} = v) = P(Y_{j_2} = w | Y_{j_2-1} = v)P(Y_{j_2-1} = v | Y_{j_2-1} = u)P(Y_{j_2-1} = u) \tag{2.4}
\]
\[ \pi_{uvw} = P(Y_{j3} = s \mid Y_{j1} = u, Y_{j2} = v, Y_{j3-1} = w) = P(Y_{j3} = s \mid Y_{j1} = u, Y_{j2} = w) \]
\[ \pi_{uv} = P(Y_{j2} = w \mid Y_{j1} = u, Y_{j2-1} = v) \]
\[ \pi_{uvw} = P(Y_{j3} = s \mid Y_{j1} = u, Y_{j2} = v, Y_{j3-1} = w) = P(Y_{j3} = s \mid Y_{j1} = u, Y_{j2} = w) \]
\[ \pi_{uvw} = P(Y_{j2} = w \mid Y_{j1} = u, Y_{j2-1} = v) \]
\[ \pi_{uvw} = P(Y_{j3} = s \mid Y_{j1} = u, Y_{j2} = v, Y_{j3-1} = w) = P(Y_{j3} = s \mid Y_{j1} = u, Y_{j2} = w) \]
\[ \pi_{uvws} = P(Y_{j3} = s \mid Y_{j1} = u, Y_{j2} = v, Y_{j3-1} = w, Y_{j4} = s) \]
\[ \pi_{uvws} = P(Y_{j3} = s \mid Y_{j1} = u, Y_{j2} = v, Y_{j3-1} = w) \]

It is observed that \( \pi_{uv}, \pi_{uvw}, \pi_{uvws} \) are initially first, second and third order joint probabilities respectively but later these are expressed in terms of first order Markov probabilities using equation (2.1) as follows:

\[ \pi_{uv} = \pi_{uv}^{(1)} \]
\[ \pi_{uvw} = \pi_{uvw}^{(2)} \]
\[ \pi_{uvws} = \pi_{uvws}^{(3)} \]

In the above conditional probabilities, we have considered that once a transition is made from \( u \) to \( v \), then the time of event \( u \) will remain fixed for all other subsequent transitions. Here a transition from \( u \) to \( v \) can happen in the second follow-up or the process can remain in the same state for the first observed time to transition to \( v \), will remain fixed for all other subsequent transitions. Thus a transition from \( u \) to \( v \) can happen in the second follow-up or the process can remain in the same state for the first observed time to transition to \( v \), will remain fixed for all other subsequent transitions. Similarly, in case of a transition from \( v \) to \( w \), the last observed time in state \( v \), before making a transition to \( w \), will remain fixed for any subsequent transition. In other words, we will consider that once a transition is made from \( u \) to \( v \), the last observed time to transition to \( v \) will remain fixed for the first time as found in expressions for \( \pi_{uvw} \) (for the first observed time to transition to \( w \) and last observed times for \( u \) and \( v \)) and \( \pi_{uvws} \) (for the first observed time to transition to \( s \) and last observed times for \( u, v \) and \( w \)).

Let us define the following notations:

\[ X_i = [X_{i1}, X_{i2}, ..., X_{ip}] \]

vector of covariates for the \( i \)th person;

\[ \beta_{uv} = [\beta_{uv0}, \beta_{uv1}, ..., \beta_{uvp}] \]

vector of parameters for the transition from \( u \) to \( v \).

\[ \beta_{uw} = [\beta_{uw0}, \beta_{uw1}, ..., \beta_{uwp}] \]

vector of parameters for the transition from \( u \) to \( w \), and

\[ \beta_{ws} = [\beta_{ws0}, \beta_{ws1}, ..., \beta_{wsp}] \]

vector of parameters for the transition from \( w \) to \( s \).

The transition probabilities can be expressed from state \( u \) to state \( v \), state \( v \) to state \( w \), and state \( w \) to state \( s \) in terms of conditional probabilities as function of covariates as follows:

\[ \pi_{uv}(X) = P(Y_{j1} = v \mid Y_{j1} = u, X) = \frac{e^{g_{uv}(X)}}{\sum_{k=0}^{\infty} e^{g_{lk}(X)}} \]

where \( v = 0, v = 0,1,2 \); \( u = 0, v = 0,1,2 \); \( s = 0,1,2 \); \( w = 0,1,2 \); \( k = 0 \); \( \beta_{uv0}, \beta_{uv1}, ..., \beta_{uvp} \);

Similarly,

\[ \pi_{vw}(X) = P(Y_{j2} = w \mid Y_{j1} = v, X) = \frac{e^{g_{vw}(X)}}{\sum_{k=0}^{\infty} e^{g_{vk}(X)}} \]

where \( v = 0, v = 0,1,2 \); \( w = 0,1,2 \); \( k = 0 \); \( \beta_{vw0}, \beta_{vw1}, ..., \beta_{vwp} \);

and

\[ \pi_{ws}(X) = P(Y_{j3} = s \mid Y_{j1} = w, X) = \frac{e^{g_{ws}(X)}}{\sum_{k=0}^{\infty} e^{g_{sk}(X)}} \]

where \( v = 0, v = 0,1,2 \); \( w = 0,1,2 \); \( k = 0 \); \( \beta_{vw0}, \beta_{vw1}, ..., \beta_{wsp} \);

3. Estimation
The likelihood function for \( n \) individuals with \( i \)th individual having \( J_i \) (\( i = 1,2, ..., n \)) follow-ups can be expressed as
The log likelihood function is given by

\[
\ln L = \sum_{i=1}^{n} \{ \sum_{j_1=1}^{J_1} \{ \sum_{j_2=1}^{J_2} \{ \sum_{j_3=1}^{J_3} \{ \sum_{u=0}^{v=0} \{ \sum_{w=0}^{s=0} (\pi_{j_1j_2j_3uwv}(X_i)) \delta_{ij_1j_2j_3uwv} \} \} \} \} \}
\]

(3.1)

The above likelihood function can be expressed as follows:

\[
L = \prod_{i=1}^{n} \left\{ \prod_{j_1=1}^{J_1} \left\{ \prod_{j_2=1}^{J_2} \left\{ \prod_{j_3=1}^{J_3} \left\{ \prod_{u=0}^{v=0} \left\{ \prod_{w=0}^{s=0} (\pi_{j_1j_2j_3uwv}(X_i)) \delta_{ij_1j_2j_3uwv} \right\} \right\} \right\} \right\} \}
\]

(3.2)

where \( \delta_{ij_1j_2j_3uwv} = 1 \) if a transition type \( u \rightarrow v \) (\( u=0, v=1,2 \)) is observed at the \( j_1 \)th follow-up for the \( i \)th individual, \( \delta_{ij_1j_2j_3uwv} = 0 \), otherwise; \( \delta_{ij_1j_2j_3uwv} = 1 \), if a transition type \( u \rightarrow v \) (\( u=0, v=1,2 \)) is observed at the \( j_1 \)th follow-up and a transition type \( v \rightarrow w \) (\( v=1, w=0,2 \)) is observed at the \( j_2 \)th follow-up, \( \delta_{ij_1j_2j_3uwv} = 0 \), if a transition type \( u \rightarrow v \) (\( u=0, v=1,2 \)) is observed at the \( j_1 \)th follow-up and a transition type \( v \rightarrow w \) (\( v=1, w=0,2 \)) does not occur at the \( j_2 \)th follow-up; \( \delta_{ij_1j_2j_3uwv} = 1 \) if a transition type \( u \rightarrow v \) (\( u=0, v=1,2 \)) is observed at the \( j_1 \)th follow-up, a transition type \( v \rightarrow w \) (\( v=1, w=0,2 \)) is observed at the \( j_2 \)th follow-up, and a transition type \( w \rightarrow s \) (\( w=0, s=1,2 \)) does not occur at the \( j_3 \)th follow-up.

The log likelihood function is given by

\[
\ln L = \sum_{i=1}^{n} \{ \sum_{j_1=1}^{J_1} \{ \sum_{j_2=1}^{J_2} \{ \sum_{j_3=1}^{J_3} \{ \sum_{u=0}^{v=0} \{ \sum_{w=0}^{s=0} \delta_{ij_1j_2j_3uwv} \ln \pi_{j_1j_2j_3uwv}(X_i) \} \} \} \} \}
\]

(3.3)

Differentiating (3.3) with respect to the parameters and solving the following equations we obtain the likelihood estimates for:

\[
\frac{\partial \ln L}{\partial \beta_{uvq}} = \sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) = 0
\]

(3.4)

\[
\frac{\partial \ln L}{\partial \beta_{uvq}} = \sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) = 0
\]

(3.5)

\[
\frac{\partial \ln L}{\partial \beta_{uvq}} = \sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) = 0
\]

(3.6)

The observed information matrix can be obtained from the following second derivatives:

\[
\frac{\partial^2 \ln L}{\partial \beta_{uvq} \partial \beta_{uwq}} = -\sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) \ln \pi_{j_1j_2j_3uwv}(X_i)
\]

(3.7)

\[
\frac{\partial^2 \ln L}{\partial \beta_{uvq} \partial \beta_{uwq}} = -\sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) \ln \pi_{j_1j_2j_3uwv}(X_i)
\]

(3.8)

\[
\frac{\partial^2 \ln L}{\partial \beta_{uvq} \partial \beta_{uwq}} = -\sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) \pi_{j_1j_2j_3uwv}(X_i)
\]

(3.9)

Similarly,

\[
\frac{\partial^2 \ln L}{\partial \beta_{uwq} \partial \beta_{uwq}} = -\sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) \pi_{j_1j_2j_3uwv}(X_i)
\]

(3.10)

\[
\frac{\partial^2 \ln L}{\partial \beta_{uwq} \partial \beta_{uwq}} = -\sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) \pi_{j_1j_2j_3uwv}(X_i)
\]

(3.11)

\[
\frac{\partial^2 \ln L}{\partial \beta_{uwq} \partial \beta_{uwq}} = -\sum_{i=1}^{n} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \delta_{ij_1j_2j_3uwv} X_{qi}(1 - \pi_{j_1j_2j_3uwv}(X_i)) \pi_{j_1j_2j_3uwv}(X_i)
\]

(3.12)

4. Application to Depression Data

For this study, an application is showed in this section from the Health and Retirement Study (HRS) data. The HRS is sponsored by the National Institute of Aging (grant number NIA U01AG09740) and conducted by the University
of Michigan (2002). This study is conducted nationwide for individuals over age 50 and their spouses. We have used the panel data from the six rounds of the study conducted on individuals over age 50 years in 1992, 1994, 1996, 1998, 2000 and 2002. This study uses data documented by RAND. We have used the panel data on depression for the period, 1992-2002. The depression index is based on the score on the basis of the scale proposed by the Center for Epidemiologic Studies Depression (CESD). As indicated in the documentation of the RAND, the CESD score is computed on the basis of eight indicators attributing depression problem. The indicators of depression problem are based on six negative (all or most of the time: depressed, everything is an effort, sleep is restless, felt alone, felt sad, and could not get going) and two positive indicators (felt happy, enjoyed life). These indicators are yes/no responses of the respondent’s feelings much of the time over the week prior to the interview. The CESD score is the sum of six negative indicators minus two positive indicators. Hence, severity of the emotional health can be measured from the CESD score. From the panels of data, we have used 9761 respondents for analyzing depression among the elderly in the USA during 1992-2002.

We considered the following dependent and explanatory variables: no depression (CESD score ≤ 0) = 0, depression (CESD score > 0) = 1, death (=2 if a respondent died during any two consecutive waves; age (≤ 60 years=0, 61+ =1), gender (male=1, female=0), education (high school or lower=0, some college or higher=1), ethnic group (white=1, else 0; black=1, else 0; others= reference category).

Table 1 shows counts based on different transition types. Table 2 summarizes the counts in terms of the proposed models for transitions, reverse transitions and repeated transitions. We considered only subjects with no depression at the baseline and observed that 61 percent of all the subjects remain as depression free during the study period. Out of 8318 depression free subjects, 32 percent moved to the state of depression and 7 percent made transition to death state. Out of the 2660 transitions, 21.1 percent remained in the state of depression during the remaining period (in addition to those who are lost to follow-up during the study period), 72.3 percent made reverse transition to no depression from the immediate prior state of depression and 6.7 percent died from the state of depression. Out of 1922 reverse transitions, 66.4 percent remained in the same state of no depression, 29.3 percent experienced transition to depression once again (repeated transition) and 4.3 percent died from the last prior state of no depression.

Table 3 displays the estimates of the parameters for models on transitions, reverse transition and repeated transitions as well as for models on transition to death states from no depression, depression (transition), and no depression (reverse). It is surprising that age is negatively associated with all these transitions. Similarly, gender is negatively associated with transition and reverse transition (females are more likely to experience both transitions and reverse transitions), but after reverse transition, males are more likely to make transition to death state. Education is negatively associated with transition and repeated transition to depression but education does not show any association with reverse transition. Similarly, it is evident that there is lower transition from no depression, as well as from no depression (reverse transition) state, for higher education group. There is no association between education and transition to death from no depression but there is no association between education and transition to death after reverse transition to depression. Black subjects have higher transition from no depression to death compared to other ethnic groups but White subjects have lower transition to death after reverse transition to no depression as compared to other ethnic groups. Interaction between age and gender exert positive impact on transition to depression as well as transition to death states.

5. Conclusions

The problem of depression has emerged as a major health concern during the past decades. This problem can have long-term impact in terms of health risk as well as major socio-economic consequences. In some cases, the transitions, reverse transitions as well as repeated transitions to or from the state of depression can be of interest both to the researchers and the health professionals. This paper simplifies the discrete-time and discrete state Markov model with covariate dependence using Chapman-Kolmogorov equation and shows a more generalized form of the likelihood function in order to cover transitions, reverse transitions and repeated transitions for longitudinal data. In this paper, we have also made the model more realistic defining deaths from no depression or depression states as a separate state instead of considering it as a potential source of censoring. We have considered polytomous logistic link functions for analyzing the different types of transitions. The proposed model can be extended easily for continuous-time and discrete-state processes with different link functions. The application of the proposed model to the HRS data on depression for six waves during 1992-2002 reveals some interesting results that reflect
the underlying differences in making transitions, reverse transitions and repeated transitions.

6. Acknowledgements
The HRS (Health and Retirement Study) is sponsored by the National Institute of Aging (grant number NIA U01AG09740) and conducted by the University of Michigan.

7. References


Health and Retirement Study, (Wave [1-6]/Year [1992-2002]) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG09740). Ann Arbor, MI.


Table 1. Counts based on transitions, reverse transitions and repeated transition.

<table>
<thead>
<tr>
<th>Transition Types</th>
<th>N</th>
<th>Transition Types</th>
<th>N</th>
<th>Transition Types</th>
<th>N</th>
<th>Transition Types</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transitions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No transition</td>
<td>0-0-0-0-0-3</td>
<td>247</td>
<td>0-0-0-0-1</td>
<td>504</td>
<td>0-0-2</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>0-3</td>
<td>176</td>
<td>0-0-0-0-0-0</td>
<td>3183</td>
<td>0-0-0-0-1</td>
<td>307</td>
<td>0-0-0-2</td>
<td>126</td>
</tr>
<tr>
<td>0-0-3</td>
<td>672</td>
<td>Transition 0-1</td>
<td>219</td>
<td>0-0-0-0-0-1</td>
<td></td>
<td>0-0-0-0-2</td>
<td>130</td>
</tr>
<tr>
<td>0-0-0-3</td>
<td>463</td>
<td>0-1</td>
<td>1099</td>
<td>Transition 0-2</td>
<td>219</td>
<td>0-0-0-0-0-2</td>
<td>111</td>
</tr>
<tr>
<td>0-0-0-0-3</td>
<td>332</td>
<td>0-0-1</td>
<td>531</td>
<td>0-2</td>
<td>115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reverse Transitions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No transition</td>
<td>0-0-0-0-1-3</td>
<td>5</td>
<td>0-0-0-0-1-0</td>
<td>208</td>
<td>Reverse Tran. 0-1-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-1-1-1-1-1</td>
<td>113</td>
<td>0-0-0-1-1-3</td>
<td>3</td>
<td>0-0-0-1-1-0</td>
<td>66</td>
<td>0-1-2</td>
<td>50</td>
</tr>
<tr>
<td>0-0-1-1-1-1</td>
<td>48</td>
<td>0-0-1-3</td>
<td>7</td>
<td>0-1-1-1-1-0</td>
<td>27</td>
<td>0-0-1-2</td>
<td>17</td>
</tr>
<tr>
<td>0-0-0-1-1-1</td>
<td>53</td>
<td>Reverse Tran. 0-1-0</td>
<td>0-1-1-1-1-0</td>
<td>50</td>
<td>0-1-1-2</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>0-0-0-0-1-1</td>
<td>76</td>
<td>0-1-0</td>
<td>246</td>
<td>0-1-0-0-0-0</td>
<td>160</td>
<td>0-0-0-1-2</td>
<td>14</td>
</tr>
<tr>
<td>0-0-0-0-0-1</td>
<td>219</td>
<td>0-0-1-0</td>
<td>161</td>
<td>0-0-1-0-0-0</td>
<td>241</td>
<td>0-0-1-1-2</td>
<td>12</td>
</tr>
<tr>
<td>0-1-3</td>
<td>16</td>
<td>0-1-0-0</td>
<td>181</td>
<td>0-1-0-0-0-0</td>
<td>162</td>
<td>0-1-1-1-2</td>
<td>15</td>
</tr>
<tr>
<td>0-1-1-3</td>
<td>8</td>
<td>0-1-1-0</td>
<td>93</td>
<td>0-1-1-0-0-0</td>
<td>47</td>
<td>0-0-0-0-1-2</td>
<td>18</td>
</tr>
<tr>
<td>0-1-1-1-3</td>
<td>5</td>
<td>0-0-0-1-0-0</td>
<td>109</td>
<td>0-1-0-0-0-0</td>
<td>49</td>
<td>0-0-0-1-1-2</td>
<td>12</td>
</tr>
<tr>
<td>0-1-1-1-1-3</td>
<td>2</td>
<td>0-0-1-1-0</td>
<td>41</td>
<td>0-1-1-1-0-0</td>
<td>33</td>
<td>0-0-1-1-1-2</td>
<td>9</td>
</tr>
<tr>
<td>0-0-0-1-3</td>
<td>6</td>
<td>0-1-1-1-0</td>
<td>48</td>
<td>0-1-1-1-1-0</td>
<td></td>
<td>0-1-1-1-1-2</td>
<td>16</td>
</tr>
<tr>
<td>Repeated Transitions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No transition</td>
<td>0-1-0-3</td>
<td>13</td>
<td>0-1-0-1-0-0</td>
<td>16</td>
<td>Rep. Tran. 0-1-0-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-0-1-0-0-0</td>
<td>47</td>
<td>0-1-0-0-3</td>
<td>31</td>
<td>0-1-0-0-1-0</td>
<td>38</td>
<td>0-1-0-2</td>
<td>12</td>
</tr>
<tr>
<td>0-0-1-0-0-0</td>
<td>162</td>
<td>0-1-0-1-0-0</td>
<td>32</td>
<td>0-1-0-1</td>
<td>127</td>
<td>0-0-1-0-2</td>
<td>8</td>
</tr>
<tr>
<td>0-1-1-0-0-0</td>
<td>33</td>
<td>0-0-1-0-1-0</td>
<td>30</td>
<td>0-0-1-0-1</td>
<td>20</td>
<td>0-1-1-0-2</td>
<td>7</td>
</tr>
<tr>
<td>0-0-0-1-0-0</td>
<td>241</td>
<td>0-1-1-1-0-0</td>
<td>10</td>
<td>0-1-1-0-0-0</td>
<td>23</td>
<td>0-1-0-0-2</td>
<td>17</td>
</tr>
<tr>
<td>0-0-0-0-1-0</td>
<td>208</td>
<td>0-1-0-0-0-0</td>
<td>17</td>
<td>0-1-0-0-1</td>
<td>26</td>
<td>0-0-0-1-0-2</td>
<td>10</td>
</tr>
<tr>
<td>0-0-0-1-1-0</td>
<td>66</td>
<td>0-1-0-0-0-0-0</td>
<td>7</td>
<td>0-0-0-1-0-1</td>
<td>69</td>
<td>0-0-1-0-0-2</td>
<td>8</td>
</tr>
<tr>
<td>0-1-0-0-0-0</td>
<td>160</td>
<td>0-1-0-0-0-0-0</td>
<td>18</td>
<td>0-0-1-0-1-0-0</td>
<td>28</td>
<td>0-1-1-0-2</td>
<td>3</td>
</tr>
<tr>
<td>0-1-1-0-0-0</td>
<td>49</td>
<td>0-0-1-1-0-0-0</td>
<td>13</td>
<td>0-1-1-0-0-0-0</td>
<td>21</td>
<td>0-1-0-0-0-2</td>
<td>11</td>
</tr>
<tr>
<td>0-1-1-1-0-0</td>
<td>50</td>
<td>Rep. Tran. 0-1-0-1</td>
<td>0-1-0-0-0-1</td>
<td>37</td>
<td>0-1-0-0-0-2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>0-0-1-1-1-0</td>
<td>27</td>
<td>0-1-0-1-0-0-0</td>
<td>45</td>
<td>0-1-1-1-0-1</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-1-0-3</td>
<td>62</td>
<td>0-0-1-0-1-0</td>
<td>39</td>
<td>0-1-0-0-0-1</td>
<td>40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: No depression = 0; Depression = 1; Death = 2; and Missing = 3.
Table 2 Distribution of respondents in terms of the proposed models for transitions, reverse transitions and repeated transitions.

<table>
<thead>
<tr>
<th>Transition Types</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transitions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No transition from no depression</td>
<td>5073</td>
<td>61.0</td>
</tr>
<tr>
<td>No depression to depression (0→1)</td>
<td>2660</td>
<td>32.0</td>
</tr>
<tr>
<td>No depression to death (0→2)</td>
<td>585</td>
<td>7.0</td>
</tr>
<tr>
<td>Total</td>
<td>8318</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Reverse Transitions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No transition after transition to depression</td>
<td>561</td>
<td>21.1</td>
</tr>
<tr>
<td>No depression to depression to no depression (0→1→0)</td>
<td>1922</td>
<td>72.3</td>
</tr>
<tr>
<td>No depression to depression to death (0→1→2)</td>
<td>177</td>
<td>6.7</td>
</tr>
<tr>
<td>Total</td>
<td>2660</td>
<td>100.0</td>
</tr>
<tr>
<td><strong>Repeated Transitions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No transition after reverse transition</td>
<td>1276</td>
<td>66.4</td>
</tr>
<tr>
<td>No depression to depression to no depression to depression (0→1→0→1)</td>
<td>564</td>
<td>29.3</td>
</tr>
<tr>
<td>No depression to depression to no Depression to death (0→1→0→2)</td>
<td>82</td>
<td>4.3</td>
</tr>
<tr>
<td>Total</td>
<td>1922</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Table 3 Estimates of the parameters of the proposed models for transitions, reverse transitions and repeated transitions

| Variables | Transitions | 0→1 | | 0→2 | | 95% CI | | 95% CI |
|-----------|-------------|-----|---|-----|---|---|---|
| B | SE | OR | B | SE | OR | B | SE | OR |
| Intercept | 0.87** | 0.14 | -1.53** | 0.25 | | | | |
| Age | -2.15** | 0.07 | 0.12 | 0.10 | 0.13 | -1.34** | 0.14 | 0.26 | 0.20 | 0.35 |
| Gender | -0.83** | 0.07 | 0.44 | 0.38 | 0.51 | 0.00 | 0.13 | 1.00 | 0.78 | 1.29 |
| Education | -0.41** | 0.06 | 0.66 | 0.59 | 0.74 | -0.36** | 0.09 | 0.70 | 0.58 | 0.84 |
| White | -0.04 | 0.14 | 0.96 | 0.73 | 1.26 | 0.004 | 0.24 | 1.00 | 0.62 | 1.61 |
| Black | 0.20 | 0.15 | 1.22 | 0.91 | 1.63 | 0.61* | 0.26 | 1.83 | 1.11 | 3.03 |
| Age*Gender | 0.41* | 0.11 | 1.50 | 1.21 | 1.87 | 0.63* | 0.18 | 1.87 | 1.31 | 2.67 |

| Variables | Reverse Transitions | 0→1→0 | | 0→1→2 | | 95% CI | | 95% CI |
|-----------|---------------------|-----|---|-----|---|---|---|
| B | SE | OR | B | SE | OR | B | SE | OR |
| Intercept | 3.86** | 0.36 | 0.80 | 0.52 | | | | |
| Age | -2.75** | 0.21 | 0.06 | 0.04 | 0.10 | -2.03* | 0.30 | 0.13 | 0.07 | 0.24 |
| Gender | -0.60* | 0.28 | 0.55 | 0.32 | 0.96 | 0.264 | 0.36 | 1.30 | 0.64 | 2.65 |
| Education | -0.06 | 0.12 | 0.95 | 0.75 | 1.19 | -0.28 | 0.20 | 0.75 | 0.51 | 1.12 |
| White | -0.51 | 0.31 | 0.60 | 0.33 | 1.10 | -0.63 | 0.46 | 0.53 | 0.21 | 1.31 |
| Black | -0.86** | 0.32 | 0.42 | 0.22 | 0.80 | -0.36 | 0.48 | 0.70 | 0.27 | 1.79 |
| Age*Gender | 0.54 | 0.31 | 1.71 | 0.94 | 3.12 | 0.35 | 0.42 | 1.42 | 0.62 | 3.24 |

| Variables | Repeated Transitions | 0→1→0→1 | | 0→1→0→2 | | 95% CI | | 95% CI |
|-----------|---------------------|-----|---|-----|---|---|---|
| B | SE | OR | B | SE | OR | B | SE | OR |
| Intercept | 0.41 | 0.28 | -1.43** | 0.49 | | | | |
| Age | -1.50** | 0.15 | 0.22 | 0.17 | 0.30 | -0.86* | 0.40 | 0.42 | 0.19 | 0.93 |
| Gender | -0.24 | 0.22 | 0.79 | 0.51 | 1.21 | 0.962* | 0.48 | 2.62 | 1.03 | 6.64 |
| Education | -0.35** | 0.12 | 0.71 | 0.56 | 0.89 | -1.06** | 0.32 | 0.35 | 0.18 | 0.65 |
| White | 0.05 | 0.26 | 1.05 | 0.63 | 1.75 | -1.04** | 0.40 | 0.35 | 0.16 | 0.77 |
| Black | 0.22 | 0.28 | 1.24 | 0.72 | 2.15 | -0.41 | 0.44 | 0.66 | 0.28 | 1.56 |
| Age*Gender | 0.04 | 0.26 | 1.04 | 0.63 | 1.73 | 0.04 | 0.55 | 1.04 | 0.36 | 3.02 |

Model Chi-square | 2556.704 | (d.f =36, p-value=0.000) |

** Significant at 1% level; * Significant at 5% level.
The Stochastic Analysis of Minimum $x^2$ for Various Clumping Models

M.I. A. Ageel
King Khalid University, Department of Mathematics, Abha, P.O. Box 9013, Saudi Arabia

Abstract. The paper considers three theoretical and stochastic approaches to approximate $E(x^2)$ for various clumping models.

1. Introduction

A flexible system of nonlinear statistical models has been developed to describe aspects of animal growth and development in quantitative terms. Models from this system have been applied to data concerning the generation and growth of ovarian follicles in infant rats and sheep. A review of this ovarian folliculogenesis model and the results of fitting it, using minimum $x^2$ statistics, to independent individuals see Read et al. (1979, 1981, 1989), Berry (1983) and Ageel (1992). In the folliculogenesis model, it is assumed that within each animal individual follicles develop independently. Ageel (1992) assessed the validity of this basic assumption, for the real sheep data. The results showed that there is a possible contradiction of the assumption that individual follicles are generated independently, and it may be that some follicles, of the same or similar recorded size, start to grow in clusters rather than individually. In view of the possibility of serial or spatial dependence due to natural hormonal activity in the ovary, Ageel repeated his modeling analysis counting only one individual from each set of follicles with the same size. The results of this modified analysis are shown in table1 and demonstrated the model to be very much more successful, giving an acceptable fit in all but three cases. This improvement could be taken to suggest that the above folliculogenesis model describes the development of clusters or clumps of follicles rather than of separate individuals. It would be therefore be of interest to examine the behavior of the goodness of fit under clumping.

It is hard to calculate the exact effect of the clumping because the chances, that some or all of the members of a clump are in a particular size group when the animal is observed, depend on the delay time, $t_j$ say, at which the clump started to grow. This dependence makes the variance and higher moments of the $x^2$ statistics, but not the expectation, very hard to calculate. To address this problem we consider the following three theoretical and stochastic approaches:

(i) to assume that all clumps have the same probabilities of contributing to any given size group regardless of the time at which they began to grow, and that the members of a clump are either all alive or else all dead at the time of observation

(ii) to assume as in (i) that all clumps have the same probabilities of contributing to any given size group regardless of the time at which they began to grow, but that the members of a clump survive independently of each other.

(iii) To develop exact expressions for $E(x^2)$ in integral form and calculate stochastic approximations to this quantity.

The reason for considering these approaches is that they avoid the calculation of the probabilities of being observed in different size groups conditional on $t_j$

2. Approach 1

Assume that there are $N$ clumps. For an animal examined at age $t$, in which follicular growth was initiated at time $t_0$, suppose that each clump has probability $p_j$, say of being in the $i^{th}$ group of size, $i = 1, \ldots, k$ and $(k+1)^{th}$ group of unobservable clumps has probability $1 - \sum_{i=1}^k p_i$, or $1 - p$.

Let $X$ be the number in clumps which are distributed as $1 + Poisson(\lambda)$ with probability
generating function (pgf) $F(s) = se^{4(s-1)}$. Let $R$ be the number of clumps in the $i^{th}$ size group which are distributed as a binomial with parameters $N$ and $p_j$, and with probability generating function (pgf) $G_j(s) = (p_j se^{4(s-1)} + 1 - p_j)^N$. Here, we assume that members of a clump are either all alive or all dead. Then the pgf of the number of follicles in the $i^{th}$ group, $Y$ say, is

$$G_i(F(s)) = H_i(s) = [p_j se^{4(s-1)} + 1 - p_j]^N$$

It is desirable to present some of the general formulae which will be used in this theory. For simplicity we omit the suffix $i$ in the pgfs $G_i(s)$ and $H_i(s)$.

$$H'(s) = G'(F(s)) \times F'(s); s = 1 \Rightarrow \mu_e = \mu_j \mu_e$$

$$H''(s) = G''(F(s)) \times (F'(s))^2 + G'(F(s)) \times F''(s)$$

$$H''(l) = G''(F(l)) \times (F'(l))^2 + G'(F'(l)) \times F''(l)$$

We have

$$\mu^2_e - \mu = \mu^2_j + \sigma^2_e - \mu_e + \mu_j \sigma^2_e - \mu_e$$

Hence

$$\sigma^2_e = \mu^2_j + \sigma^2_e - \mu_e + \mu_j \sigma^2_e - \mu_e$$

Taking the above distributions, we have for the $i^{th}$ size group $\mu_j = 1 + \lambda, \; \sigma^2_j = \lambda, \; \mu_e = Np_i, \; \sigma^2_e = Np_i (1 - p_j)$

Hence

$$\sigma^2_i = (1 + \lambda)^2 Np_i (1 - p_j) + Np_i \lambda = V_i \text{ say}$$

$$\mu_i = Np_i (1 + \lambda) = E_i \text{ for groups } i = 1, \ldots, k$$

For the $(k+1)^{th}$ (unobservable) group the expected number of clumps is $N(1 - p_j)$ and each clump has expected size $(1 + \lambda)$, hence the expected number unobserved is $N(1 + \lambda)(1 - p_j) = E_{k+1}$ say. The variance of the number unobserved is given by $\sigma^2_i$ with $p_i$ replaced by $(1 - p_j)$. Hence,

$$V_{k+1} = N(1 + \lambda)^2 (1 - p_j) p + N\lambda(1 - p_j).$$

For $E(x^2)$ we therefore have

$$E \left( \sum_{i=1}^{k+1} (O_i - E_i)^2 \right) = \sum_{i=1}^{k+1} E_i - \left( \sum_{i=1}^{k+1} E_i = N(1 + \lambda) \right)$$

$$E(\chi^2) = \sum_{i=1}^{k+1} \frac{(E_i + \lambda)}{E_i} - N(1 + \lambda)$$

$$= \sum_{i=1}^{k+1} \frac{1}{E_i} - \frac{N(1 + \lambda) p + N\lambda(1 - p_j)}{N(1 + \lambda)}$$

$$= \sum_{i=1}^{k+1} \frac{(1 + \lambda)(1 - p_j) + \frac{k\lambda}{1 + \lambda} + p(1 + \lambda) + \frac{\lambda}{1 + \lambda}}{1 + \lambda}$$

$$= k(1 + \lambda) + \frac{(k + 1)\lambda}{1 + \lambda}$$

3. Approach 2

Suppose that $N$ clumps are generated as in approach one, and assume that follicles in a clump start to grow all at the same time but die or live independently. Then for group $i, \; i = 1, \ldots, k$

$$O_i = \sum_{j=1}^{N} \text{ number in clump } j \text{ which are in size group } i \text{ say } (N \text{ fixed})$$

where any $Z_{ij}$ is the sum of $1 + P(\lambda)$ random variable number of Bernoulli trials with chance of success $p_j$; let $R_j$ be the size of the $j^{th}$ clump in size group $i$. For $O_{k+1}$ the chance of ‘success’ is $1 - p_j$, we can write

$$Z_j = \sum_{i=1}^{N} W_{ij},$$

where $R_j$ is distributed as $1 + P(\lambda)$ for all $j$ and the pgf of $R_j$ is $G(s) = e^{[1(p_j + s)\lambda]}$ and the pgf of
\[ W_{ij} = F(s) = p_i s + q_i \text{ for all } i, j, k, \text{ hence the pgf of } Z_{ij} \] is

\[ G(F(s)) = (p_i s + q_i) e^{\lambda i (s + q_i)} , \]

and then \( \mu_i = p_i (1 + \lambda) \).

\( \sigma_i^2 = p_i^2 \lambda + (1 + \lambda) p_i (1 - p_j) \), and since for \( i = 1, \ldots, k \), \( O_i \) is the sum of \( N \) iid random variables distributed as in \( Z_{ij} \) we have

\[ E(O_i) = N p_i (1 + \lambda) = E_i \]

\[ V(O_i) = N p_i [\lambda p_i + (1 + \lambda) (1 - p_j)] = V_i \text{ say.} \]

For \( i = k + 1 \) we have

\[ E(O_{i+1}) = N (1 - p) (1 + \lambda) \]

\[ V(O_{i+1}) = N (1 - p) [\lambda (1 - p) + p (1 + \lambda)] \]

As in Approach one,

\[ E(x^2) = \sum_{i=1}^{k+1} \frac{V_i}{E_i} \]

\[ = \sum_{i=1}^{k+1} \frac{N p_i [\lambda p_i + (1 + \lambda) (1 - p_j)]}{N p_i (1 + \lambda)} \cdot \frac{\lambda (1 - p)}{(1 + \lambda) + p} \]

\[ = \frac{1}{(1 + \lambda)} [4 p + (1 + \lambda) (k - p) + \lambda (1 - p) + p (1 + \lambda)] \]

\[ = k + \frac{\lambda}{1 + \lambda} \]

4. Calculations for Approaches 1 and 2

A Pascal program has been written to simulate a set of data of 2000 clumps of size 1 + Poisson(\( \lambda \)) random variable, where Poisson (\( \lambda = 1 \)) . Thus, the average clump size is 2 and the expected total number of follicles generated is \( 2000 \times 2 = 4000 \). Then we apply the two preceding approaches (1 and 2) to approximate \( E(x^2) \). It is worth noting that the reason for changing to 2000 clumps is to limit the burden of calculation in the stochastic approximation method which is found to be necessary. Having made this decision we start by comparing \( E(x^2) \) from approaches 1 and 2, which are theoretically but not stochastically approximate, for an animal of 2000 clumps. These non-stochastic results are given in table 2. From table 2, it appears that the approximation \( E(x^2) \) produced by approach 1 is larger compared to the observed \( x^2 \) and the \( E(x^2) \) approximated by approach 2 is less. One possible interpretation is that since the members of each clump have in fact been simulated independently conditional on their common time of starting growth, it can be expected that approach 1 will overestimate \( x^2 \). This is because, under approach 1, the observed numbers are represented as sum of sets of perfectly correlated Bernoulli trials whereas the individuals in those sets or clumps are simulated as conditionally independent. On the other hand, approach 2 is likely to underestimate \( x^2 \) because it ignores the fact that the actual probabilities of a follicle being observed in any size group vary according to the time \( j_t \) at which its particular clump started to grow. Thus, these results suggest that neither of these two arguments is correct. This also shows the need for a more accurate theoretical calculation and we consider an exact approach. Because of the significant greater algebraic complexity, this third approach is carried out by stochastic approximation and is described in the next section

5. Approach 3

As expected, approaches 1 and 2 fail to produce an accurate approximation for \( E(x^2) \). We now consider the following exact approach.

Assume that \( N \) clumps are generated as in (1) and (2), so we have \( N \) independent starting points such that for \( j = 1, \ldots, N \) all members of clump \( j \) start to grow at time \( t_j \). The delay times \( t_j \) are distributed exponentially with probability density function:

\[ \lambda_0 e^{-\lambda_0 (t-t_0)}, t \geq t_0 \]

of course, when the animal is observed (sacrificed) at age \( t \) after follicular growth for at most \( t' = t - t_0 \), there will be a residue of primordial follicles which have not yet begun to grow and all observed growing follicles must have had delay times of \( t' \) or less. This analysis taken account of the fact that the chance that a follicle finishes in size group \( i, i = 1, \ldots, k + 1, \) depends on the time
When its clump started to grow. We therefore write these probabilities as \( p_i(t_j) \). Number in each clump \( 1 + \text{Poisson}(\lambda) \) (we use \( \lambda_0 \) for exponential parameter to distinguish it from the clump size parameter).

As in approach 2, we have \( N \) clumps. Thus

\[
O_i = \sum_{j=1}^{N} \text{number in clump } j \text{ which are in size group } i
\]

\[
= \sum_{j=1}^{N} Z_{ij}
\]

\[
Z_{ij} = \sum_{j=1}^{W_{ij}}, j = 1, ..., N
\]

where for all \( j \) pgf of the size \( R_j \) of clump \( j \) is

\[
G(s) = se^{\lambda(x-1)}. \quad \text{The } W_{ij} \text{ are independent Bernoulli trial random variables, each with chance of success } p_i(t_j) \text{ and pgf given by:}
\]

\[
F_g(s) = p_i(t_j)s + 1 - p_i(t_j), l = 1, ..., R_g
\]

Thus conditional on a clump starting to develop at time \( t_j \), the pgf of the number in the \( i^{th} \) size group is

\[
\left[ \frac{p_i(t_j)s + 1 - p_i(t_j)}{e^{\lambda_i(t_j)(1-1)}} \right]
\]

The unconditional pgf of \( Z_j \) is given by

\[
H_i(s) = \int_0^t \lambda e^{-\lambda s + \lambda s_p} p_i(u) s + 1 - p(u) du,
\]

where \( t = t - t_0 \) as in earlier formulae and \( u = t_j - t_0 \). Then, taking \( \frac{d}{ds} \) under the integral sign,

\[
H_i'(s) = \int_0^t \lambda e^{-\lambda s + \lambda s_p} p_i(u) [1 + \lambda p_i(u)] s + 1 - p(u) du
\]

and

\[
H_i'(1) = \int_0^t \lambda e^{-\lambda u + \lambda s_p} p_i(u) du = p_i(1 + \lambda),
\]

where \( p_i = \int_0^t p_i(u) \lambda e^{-\lambda u} du \) is the overall probability of an individual follicle being observed in size group \( i \) in an animal of age \( t \).

Differentiating again, we have

\[
H_i''(s) = \int_0^t \lambda e^{-\lambda s + \lambda s_p} p_i(u) [1 + \lambda p_i(u)] s + 1 - p(u) du
\]

\[
= \int_0^t \lambda e^{-\lambda s + \lambda s_p} p_i(u) [2 + \lambda (p_i(u) s + 1 - p_i(u))] du
\]

So,

\[
H_i''(1) = \int_0^t \lambda e^{-\lambda u + \lambda s_p} p_i(u) du
\]

\[
= \lambda(1 + \lambda) \left[ 2 + \lambda (\lambda p_i(u) s + 1 - p_i(u)) du \right] = \lambda(1 + \lambda) (\lambda + p_i(u))
\]

say, where \( v_i = \text{variance} \left( \frac{p_i(u)}{u} \right) \) and

\[
p_i = \text{Expectation} \left[ \frac{p_i(u)}{u} \right].
\]

Now from above , for \( i = 1, ..., k \), we have

\[
E_i = E(O_i) = N H_i'(1) = N p_i(1 + \lambda)
\]

\[
E(O_i^*) = N^2 \left( H_i'(i) - H_i'(1) \right)^2 + N \sum_{j=0}^{k} V_j = V(O_i) = N \left( H_i'(i) - (H_i'(1)) \right)^2
\]

\[
= N \left( \lambda(1 + \lambda + \lambda) (\lambda + p_i(u)) \right)
\]

For the \((k+1)^{th}\) (unobservable) group the above arguments follow through with \( p_i \) replaced by the residual probability \( 1 - p \), hence

\[
V(O_{k+1}) = N \left( \lambda(1 + \lambda + \lambda) (\lambda + p_{k+1}) \right)
\]

and so finally

\[
E(x^2) = \sum_{i=1}^{k+1} \frac{V_i}{E_i}
\]

\[
= \frac{1}{1 + \lambda} \left( \sum_{i=1}^{k+1} \lambda(1 + \lambda + \lambda) (\lambda + p_{k+1}) \right)
\]

\[
= \frac{1}{1 + \lambda} \left( \sum_{i=1}^{k+1} \lambda(1 + \lambda) (\lambda + p_{k+1}) \right)
\]

\[
= \frac{1}{1 + \lambda} \left( \sum_{i=1}^{k+1} \lambda(1 + \lambda) (\lambda + p_{k+1}) \right)
\]

\[
= k + \lambda \left( 1 + \lambda \right) \left( \sum_{i=1}^{k+1} \lambda(1 + \lambda) (\lambda + p_{k+1}) \right)
\]

6. Calculations for Approach 3

In the expression for \( E(x^2) \), the quantities \( V_1, \ldots, V_{k+1} \) are extremely tedious to evaluate algebraically. An alternative approach is to approximate them by simulation. The calculations of approximating \( E(x^2) \), for approach 3, show
that the approximating $E(x^2)$ values (=65.446) falls a little short of the observed value (71.486), based on the estimated parameter values, by 6.04. Assuming that the true model parameter values are 1.04. This based on the fact that in the usual calculations of the goodness-of-fit $x^2$ statistics from a total of $k+1$ groups, the effect of using estimated values is approximately to reduce $E(x^2)$ by at most the number ($s$, say) of estimated probability parameters, from $k$ to a minimum limit of $k - s$.

7. Conclusion

We have studied the effect of clumping on the $x^2$ statistics. Three approaches were considered. Approaches 1 and 2 are used to approximate the effect of clumping on the minimum $x^2$ statistics theoretically. The calculations for these two approaches showed that neither of these two arguments is right. Accordingly, a third approach was carried out using stochastic approximation. The results for approach 3 show that the approximating $E(x^2)$ value is better and it falls a little short of the observed value. This may suggest the reliability of this approach in the assessment of goodness of fit under clumping in folliculogenesis model. To examine the importance of this little discrepancy in approach 3, we can carry out a similar approximation for variance of $x^2$, $V(x^2)$, and which we hope to discuss in some other publications.

used in the calculation of $x^2$, the approximating $E(x^2)$ values (= 70.446) is less than the observed $x^2$

8. References


Table 1: Estimates of Parameters for Clumping Follicle Data.

<table>
<thead>
<tr>
<th>Animal</th>
<th>Follicle</th>
<th>$a$</th>
<th>$\beta_1$</th>
<th>$\lambda$</th>
<th>$\mu$</th>
<th>$r^*$</th>
<th>$\beta_2$</th>
<th>$t_0$</th>
<th>$c$</th>
<th>$N_g$</th>
<th>$x^2$</th>
<th>D</th>
<th>FS</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>D\textsuperscript{1}O</td>
<td>68</td>
<td>0.0004</td>
<td>0.0343</td>
<td>0.0320</td>
<td>0.2176</td>
<td>20.001</td>
<td>75.45</td>
<td>0.0808</td>
<td>0.8723</td>
<td>624</td>
<td>0.42</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>0.2535</td>
<td>0.0737</td>
<td>0.0283</td>
<td>0.2043</td>
<td>8.64</td>
<td>-17.13</td>
<td>0.2201</td>
<td>0.2988</td>
<td>400</td>
<td>4.49</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>0.0557</td>
<td>0.0329</td>
<td>0.0257</td>
<td>0.1450</td>
<td>3.24</td>
<td>-88.80</td>
<td>0.0666</td>
<td>0.7861</td>
<td>734</td>
<td>0.26</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>0.1737</td>
<td>0.0392</td>
<td>0.0274</td>
<td>0.2167</td>
<td>20.0</td>
<td>-52.80</td>
<td>0.1936</td>
<td>0.9749</td>
<td>400</td>
<td>0.51</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>163</td>
<td>0.1631</td>
<td>0.0925</td>
<td>0.0377</td>
<td>0.0201</td>
<td>1.71</td>
<td>-14.17</td>
<td>0.1060</td>
<td>0.3000</td>
<td>200</td>
<td>2.75</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>0.0522</td>
<td>0.5828</td>
<td>0.0148</td>
<td>0.1837</td>
<td>8.91</td>
<td>-15.42</td>
<td>0.2379</td>
<td>0.9005</td>
<td>400</td>
<td>0.33</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>0.1123</td>
<td>0.0538</td>
<td>0.0137</td>
<td>0.2140</td>
<td>9.85</td>
<td>-22.96</td>
<td>0.2511</td>
<td>1.2942</td>
<td>400</td>
<td>0.65</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F\textsuperscript{1}O</td>
<td>137</td>
<td>0.0974</td>
<td>0.1089</td>
<td>0.0070</td>
<td>0.2149</td>
<td>20.0</td>
<td>-39.18</td>
<td>0.1939</td>
<td>0.3980</td>
<td>2401</td>
<td>16.05</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>136</td>
<td>0.1849</td>
<td>0.1242</td>
<td>0.0321</td>
<td>0.2159</td>
<td>9.08</td>
<td>-14.51</td>
<td>0.0814</td>
<td>0.9378</td>
<td>447</td>
<td>1.39</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>209</td>
<td>0.1883</td>
<td>0.0640</td>
<td>0.0351</td>
<td>0.2143</td>
<td>20.0</td>
<td>-119.79</td>
<td>0.0250</td>
<td>1.4597</td>
<td>1000</td>
<td>4.31</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>244</td>
<td>0.1235</td>
<td>0.0968</td>
<td>0.0058</td>
<td>0.2117</td>
<td>20.0</td>
<td>-40.63</td>
<td>0.0381</td>
<td>0.8178</td>
<td>2097</td>
<td>11.27</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>0.1143</td>
<td>0.0478</td>
<td>0.0378</td>
<td>0.1741</td>
<td>20.0</td>
<td>-49.18</td>
<td>0.0239</td>
<td>1.1525</td>
<td>474</td>
<td>0.72</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>0.1799</td>
<td>0.0608</td>
<td>0.0372</td>
<td>0.2179</td>
<td>20.0</td>
<td>-34.84</td>
<td>0.0429</td>
<td>0.9665</td>
<td>400</td>
<td>2.70</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>0.1311</td>
<td>0.1015</td>
<td>0.0325</td>
<td>0.2110</td>
<td>9.93</td>
<td>-11.09</td>
<td>0.4340</td>
<td>0.9944</td>
<td>400</td>
<td>1.64</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D\textsuperscript{1}L</td>
<td>39</td>
<td>0.4325</td>
<td>0.0445</td>
<td>0.0325</td>
<td>0.1872</td>
<td>8.44</td>
<td>-27.57</td>
<td>0.0138</td>
<td>1.9026</td>
<td>465</td>
<td>0.39</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T\textsuperscript{1}L</td>
<td>309</td>
<td>0.0780</td>
<td>0.0844</td>
<td>0.1093</td>
<td>0.1888</td>
<td>4.87</td>
<td>-7.33</td>
<td>0.2670</td>
<td>1.5451</td>
<td>499</td>
<td>24.25</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>309</td>
<td>0.1324</td>
<td>0.0699</td>
<td>0.0652</td>
<td>0.2050</td>
<td>20.0</td>
<td>-49.20</td>
<td>0.0407</td>
<td>0.9978</td>
<td>5596</td>
<td>7.51</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>156</td>
<td>0.2778</td>
<td>0.0831</td>
<td>0.0053</td>
<td>0.0424</td>
<td>9.84</td>
<td>-1.26</td>
<td>0.1340</td>
<td>1.4444</td>
<td>2019</td>
<td>6.75</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>154</td>
<td>0.1026</td>
<td>0.0798</td>
<td>0.0297</td>
<td>0.1172</td>
<td>9.22</td>
<td>-1.60</td>
<td>0.1147</td>
<td>1.4944</td>
<td>400</td>
<td>2.75</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D\textsuperscript{1}L</td>
<td>326</td>
<td>0.0872</td>
<td>0.0696</td>
<td>0.0874</td>
<td>0.2151</td>
<td>8.89</td>
<td>-19.53</td>
<td>0.1039</td>
<td>1.7554</td>
<td>2584</td>
<td>10.61</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>236</td>
<td>0.2036</td>
<td>0.0580</td>
<td>0.2588</td>
<td>0.2039</td>
<td>20.0</td>
<td>-3.72</td>
<td>0.2031</td>
<td>1.6934</td>
<td>10000</td>
<td>20.29</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T\textsuperscript{1}L</td>
<td>315</td>
<td>0.2101</td>
<td>0.0612</td>
<td>0.0920</td>
<td>0.2034</td>
<td>20.0</td>
<td>-31.90</td>
<td>0.0595</td>
<td>0.0725</td>
<td>1000</td>
<td>10.58</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>435</td>
<td>0.0274</td>
<td>0.0825</td>
<td>0.1070</td>
<td>0.1980</td>
<td>9.77</td>
<td>-17.86</td>
<td>0.0617</td>
<td>0.9325</td>
<td>5104</td>
<td>27.60</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Approximating $E(x^2)$ values for Approaches 1, 2, and 3

<table>
<thead>
<tr>
<th>Minimum $x^2$</th>
<th>Approximating $E(x^2)$ using Approach 1</th>
<th>Approximating $E(x^2)$ using Approach 2</th>
<th>Approximating $E(x^2)$ using Approach 3 (based on estimated parameter values)</th>
<th>Approximating $E(x^2)$ using Approach 3 (based on true parameter values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>71.468</td>
<td>108.5</td>
<td>39.5</td>
<td>65.446</td>
<td>70.446</td>
</tr>
</tbody>
</table>

60
Two Nonlinear Models for Time Series

David A. Dickey  
Sangil Hwang  
North Carolina State University  
Bank of Korea

Abstract.

We discuss two applications of the logistic function in time series. In the first, we look at a transfer function model for some stream flow data. Using a logistic function of lagged flow, we allow the lag relationship between the upstream and downstream stations to vary over time.

In the second application, we look at properties of the autoregressive order one process with a hyperbolic tangent function of the lagged response replacing the usual fixed lag 1 coefficient. This function is just a logistic function linearly stretched to span the interval (-1,1), the range of stationary values. Because the estimate of the autoregressive order 1 coefficient has quite a different distribution, in the fixed coefficient case, for coefficients at or near 1 than it does for coefficients well inside the stationarity region.

1. Introduction.

ARIMA models in which a current time series value is related to past values and current and past errors are often used in practice as are transfer function models that add input time series to the list of explanatory variables. All of these typically assume fixed parameters. Our first application shows an extension of the transfer function model that allows time varying coefficients that depend, through a logistic function, on lags of the predictor variables.

It becomes clear to the practitioner that not all series observed in practice satisfy the stationarity condition. This condition guarantees limiting normal distributions of least squares and maximum likelihood estimates. The suggested solution is to analyze the series of differences. Visual inspection of the series and its autocorrelations is a way to guess which series need this differencing.

Dickey and Fuller (1979) showed that least squares estimators and their studentized statistics can have limit distributions that differ substantially from the normal when differencing is appropriate but is not done. Taking the lag 1 case in particular, they look at these distributions for data generated from $Y_t = \rho Y_{t-1} + e_t$ where $e_t$ is white noise and, $Y_0$ is assumed to be 0. The stationarity condition here is $|\rho| < 1$. They studied distributions for least squares estimators in the $\rho=1$ case. Gonzalez-Farias and Dickey (1999) show yet another nonstandard limit distribution for the case in which exact maximum likelihood estimates are used.

An interesting feature of this case is that the sum of squares, $\sum_{t=1}^{n} Y_{t}^2$ is $O_p(n^2)$ rather than $O_p(n)$ as in the stationary case and when thus normalized, converges to a random variable rather than a constant. In our second application we replace the lag 1 coefficient $\rho$ with the hyperbolic tangent function $(e^{L_t} - 1)/(e^{L_t} + 1)$ where $L_t = \alpha + \beta Y_{t-1}$. A logistic function could be used as well, and the argument could be taken as $|Y_{t-1}|$.

Tong (1990) summarized many results for time series models with time varying coefficients. Theoretical results in this area are quite hard to come by, and those that were available at the time of publication are nicely described by Tong. In the context of nonlinear analogues to the AR(1) model, which we study herein, results in Tong require that the time varying lag 1 parameter be restricted to lie in an interval $[-C,C]$ with $0<C<1$. This requirement forces a buffer zone between the parameter range and the
nonstationary boundary. Thus, although the model is more flexible than a fixed coefficient AR(1), the range is restricted to be strictly inside the stationarity region (−1,1). We use the term “stationarity region” to refer to that region in which a fixed coefficient model would be stationary.

In this paper, we investigate what happens if the time varying coefficient in these models is allowed to span the range (−1,1) or (0,1). Thus, although strictly speaking our coefficient stays in the stationarity region, it can be arbitrarily close to 1. The buffer zone (C,1] is removed.

In section 2 we present the transfer function example. Nonstandard distributions do not come into play there. In section 3 we briefly review some specifics of stationary and nonstationary fixed coefficient AR(1) models. In section 4 we introduce time varying coefficient models analogous to the AR(1) that allow coefficients arbitrarily close to the boundary of the stationarity region, and we give some examples of the behavior of data generated by these. In section 5, we prove some distributional results for a special case and suggest reasons that such analytic proofs will not generalize to the entire class of such models. In section 6 we review some results suggesting that estimation works reasonably well when the true model parameters are in a particular region, while outside that region, even numerical convergence of our nonlinear estimators is hard to come by. Examples are given in section 7.

2. A transfer function model.

In this example, we model 400 daily log transformed flow rates of the Neuse River between Goldsboro N.C. (Gt, upstream) and Kinston, N.C. (Kt, about 30 miles downstream) as given in Brocklebank and Dickey (2003). Those authors fit the model

\[ \Delta K_t = 0.50(\Delta G_{t-1}^*) + 0.55\Delta G_{t-2}^* + e_t \]

where \( e_t \) is a stationary time series, \( \rho(G_t^*) = \exp(\delta + \gamma G_t^*)/(1 + \exp(\delta + \gamma G_t^*)) \), and \( G_t^* \) is the deviation of \( G_t \) from its mean. The use of these deviations avoids convergence problems encountered when the Goldsboro mean was estimated as another nonlinear parameter. Figure 1 shows a graph of the two series. Goldsboro is the lower solid line and Kinston the higher (red) dashed line.

\[ K_t = \beta_0 + \beta_1 \rho(G_{t-1}^*) G_{t-1}^* + \beta_2 (1 - \rho(G_{t-2}^*))G_{t-2}^* + Z_t \]

Notice how \( \rho( ) \) allocates different weights to lags 1 and 2 depending on lagged upstream flow. It is not clear whether higher flows, with their greater volume at G, would induce longer lasting effects so that more weight would be given to lag 2, or whether the fact that the river is carrying parcels of water downstream faster would induce more weight on the closer lag. We will fit our model and let the sign of \( \gamma \) inform us on this issue. An initial fit of the model indicates that the errors are autocorrelated. A stationary lag 4 model for \( Z_t \), \( Z_t = \alpha_1 Z_{t-1} + \ldots + \alpha_4 Z_{t-4} + e_t \), seemed sufficient and with that feature, the problem becomes a fairly standard nonlinear regression problem. Using the SAS\(^1\) procedure NLIN we get Table 1.

To address these concerns, we propose fitting the following random coefficients model:

\[ K_t = \beta_0 + \beta_1 \rho(G_{t-1}^*) G_{t-1}^* + \beta_2 (1 - \rho(G_{t-2}^*))G_{t-2}^* + Z_t \]

\( ^1\) SAS is the registered trademark of SAS Institute, Cary N.C.
Table 1. Nonlinear Regression Output.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>22666.5</td>
<td>2518.5</td>
</tr>
<tr>
<td>Error</td>
<td>385</td>
<td>2.2575</td>
<td>0.00586</td>
</tr>
<tr>
<td>Uncor Tot</td>
<td>394</td>
<td>22668.7</td>
<td></td>
</tr>
</tbody>
</table>

Parameter | Estimate | Approx 95% CI |
-----------|----------|---------------|
\(\beta_0\) | 7.5861   | 7.4143 7.7578 |
\(\beta_1\) | 0.8933   | 0.6938 1.0929 |
\(\beta_2\) | 0.5500   | 0.4211 0.6789 |
\(\delta\)  | -0.1643  | -0.5723 0.2437 |
\(\gamma\)  | -0.4081  | -0.6105 -0.2058 |
\(\alpha_1\) | 1.3288   | 1.2291 1.4286 |
\(\alpha_2\) | -0.5278  | -0.6945 -0.3610 |
\(\alpha_3\) | 0.0320   | -0.1366 0.2007 |
\(\alpha_4\) | 0.1212   | 0.0182 0.2243 |

From the estimate of  \(\gamma\), it seems that the higher the flow, the more weight is applied to the lag two upstream value.

3. Fixed coefficient models.

Our focus here is on the autoregressive order 1 (AR(1)) model. Such a model with fixed coefficients is written:

\((Y_t - \mu) = \rho (Y_{t-1} - \mu) + e_t\)

where \(\mu\) is the long run mean of the series, and \(e_t\) is white noise, that is, an uncorrelated sequence of mean 0 shocks with constant variance. The least squares estimator of \(\rho\) is

\[\hat{\rho} = \frac{\sum_{t=2}^{n} (Y_t - \mu)(Y_{t-1} - \mu)}{\sum_{t=2}^{n} (Y_{t-1} - \mu)^2}\]

where the sample mean of \(Y\) can be substituted for an unknown \(\mu\). This statistic can be computed with any standard regression program by regressing \((Y_t - \bar{Y})\) on \((Y_{t-1} - \bar{Y})\) with no intercept and an asymptotically equivalent statistic is given by just regressing \(Y_t\) on \(Y_{t-1}\) with an intercept. When \(|\rho| < 1\), \(\sqrt{n}(\hat{\rho} - \rho)\) converges to a N(0, 1-\(\rho^2\)) and the t test statistic \(t = (\hat{\rho} - \rho)/se\) converges to N(0,1) where se is the standard error of the estimate computed by the least squares formulas and output by most regression programs. When \(\rho\) is close to 1, distributions are closer to those for \(\rho=1\) than to their limit even for large samples, that is, convergence is quite slow.

4. The Random Coefficient Model

We mimic the idea of ARCH models by letting the lag 1 coefficient be a function of the most recently observed series value and write our model as \(Y_t = [\rho(Y_{t-1})]Y_{t-1} + e_t\) where \(\rho(x)\) is the hyperbolic tangent function. It is just a rescaled logistic function defined by

\[\rho(x) = \frac{\exp(\alpha + \beta x) - 1}{\exp(\alpha + \beta x) + 1}\]

which monotonically spans the range (-1,1). A logistic function can replace the hyperbolic tangent and in either case absolute values \([\rho(|Y_{t-1}|)]\) can be used.

Generating data from this model is straightforward. Figure 2A presents a realization of the process with \(\alpha = 0, \beta = 1\) in dark red. The light gray series is a random walk constructed from the same shocks. In some areas our model series behaves quite like the random walk and in others our series is quite stationary looking. Aligned below this is Figure 2B, a plot of \([\rho(Y_{t-1})]\) versus \(t\) for that series. This shows where \([\rho(Y_{t-1})]\) ventures close to 1. Figure 2C shows that \(\rho(t)\) covers an informative portion of the logistic curve. Figure 2D is a similar graph for \(\alpha = 3, \beta = 3\). Here our series is almost exactly like the random walk.

For \(\alpha = 3, \beta = 3\), the plot corresponding to Figure 2C would be almost a horizontal line at 1, that is, it would not trace out much of the hyperbolic tangent function. This suggests that nonlinear
estimation of $\alpha$ and/or $\beta$ will be difficult or impossible for such cases.

Figure 2A. Data, $\alpha=0$ and $\beta=1$.

Figure 2B. $\rho(t)$ when $\alpha=0$ and $\beta=1$.

Figure 2C. $\rho(t)$ versus $Y(t-1)$.

Figure 2D. Data, $\alpha=3$ and $\beta=3$.

Tong defines the skeleton of a process as the recursive equation without the shocks. For our case this is $y_t = [\rho(y_{t-1})]y_{t-1}$. Generating this skeleton from various starting values $y_0$ gives, for $\alpha=0$ and $\beta=1$, the skeletons in Figure 3. Because $|\rho(y)| < |y|$ the skeleton plot will approach 0 from any starting point, and the plot gives some idea of persistence for given initializations. Also $\rho(y)$ is an even function so that paths starting at $-y$ and those starting at $y$ will be the same after the first step.

Figure 3. Skeletons, $\alpha=0$, $\beta=1$. 
5. Distributions.

We investigate the properties of least squares estimates in the model above. The model is nonlinear so iterative methods are required for calculating the estimates. For example, this output from PROC NLIN in SAS was obtained for the Figure 2A. In the model, $Y_t - \mu$ was substituted for $Y_t$ and $\mu$ treated as a parameter to be estimated. Estimates converged in 9 iterations, and the error mean square was 1.04, estimating the innovation variance 1. Parameter estimates are shown in Table 2. The estimates, A and B, are estimates of 0 and 1 and MU an estimate of 0 so it appears the nonlinear procedure worked reasonably well on this generated series. As anticipated, estimates for $\alpha=3$, $\beta=3$ did not even converge.

Table 2. Nonlinear Estimates.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>4039.2</td>
<td>1299.72</td>
</tr>
<tr>
<td>Error</td>
<td>597</td>
<td>618.4</td>
<td></td>
</tr>
<tr>
<td>Uncor. Tot.</td>
<td>600</td>
<td>4657.7</td>
<td></td>
</tr>
</tbody>
</table>

Approx 95%

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.4188</td>
<td>-0.8727 0.0351</td>
</tr>
<tr>
<td>B</td>
<td>1.0695</td>
<td>0.8091 1.3299</td>
</tr>
<tr>
<td>MU</td>
<td>0.0437</td>
<td>-0.0555 0.1429</td>
</tr>
</tbody>
</table>

Results of Tong imply that when $[\rho(Y_{t+1})]$ is replaced by $C[\rho(Y_{t+1})]$ with $0<C<1$, the series is ergodic. Tong also provides some nice theorems for showing ergodicity and for showing that typical parameter estimates in stationary ergodic cases are asymptotically normal. Our simulations indicate that normality is not a good approximation, even for large $n$, for certain $(\alpha, \beta)$ combinations when the logistic or hyperbolic tangent function is used. One of the conditions Tong uses to show ergodicity is that there exist $K>0$ and $0<\gamma<1$ such that, from any starting point $y_0$, the skeleton value of $Y_t$ (see Figure 3) is bounded by $|y_0|K^T$. This is satisfied when $\beta=0$ but when $\beta$ is not 0, it can’t hold.

To show this, suppose such $(K, \gamma)$ did exist. At some time $T$ we must have $K^{T} < 1$. Let $M=K^{T} < 1$ and $B = M^{1/T}$. Note that $M \leq B = M^{1/T} < 1$. Let $\rho(\ )$ be a strictly increasing hyperbolic tangent function ($\beta > 0$). Pick $y_1 > 0$ with $\rho(y_1) > B$ and take $y_0 = y_1B^{-T} > y_1$. For any $y > y_1$, $\rho(y) > B$ by monotonicity. Notice that $B^T y_0 = y_1B^{-T}y_1$ for $t \leq T$, so $\rho(B^T y_0) > B$ for $t \leq T$. Thus $\rho(y_0) > B$ and $y_1 = y_0 \rho(y_0) > By_0 > y_1$ so again $\rho(y_1) > \rho(y_1) > B$. Proceeding recursively, $y_2 = y_1\rho(Y_1) > By_1 > B^2 y_0$ for a sequence $y_1$ from the skeleton. By induction, $y_1 > B^Ty_0$ for $t=1,...,T$ and in particular $y_T > M y_0 = K^T y_0$, a contradiction. Thus a standard condition used to show ergodicity does not hold here.

The development above can be visualized in Figure 4 below which shows the ratio of time $T$ skeleton value to the start value, $y_T/y_0$ plotted against $y_0$, $-10 < y_0 < 10$, for $t=1$ to $T$ using the hyperbolic tangent with $\beta=1$ and $\alpha=0$. Horizontal lines at -.95 and .95 illustrate that if we start far enough to the right, the skeleton value $y_T$ will remain above 0.95$y_0$ after $T=12$ steps.

Figure 5. Skeleton ratios for other parameter settings.

Figure 4 A. Skeleton ratios, $\alpha=0$, $\beta=1$.  "Skeleton ratios" means the ratio of the last value to the first value for simulations of $Y_T/y_0$.
Similar plots for other parameter combinations, \((\alpha, \beta) = (2.5)\) and \((-1, -0.8)\), are shown above as Figure 5. A variety of behaviors is possible, but in all cases given \(M = 0.95\), if one starts far enough to the right or left, the skeleton will remain above \(M|Y|\) after \(T = 12\) steps.

There is one special case in which straightforward analytic results follow. Note that if \(\beta = 0\), then \(\rho(Y_{t+1})\) is a constant less than 1 in magnitude so that the series is just a standard AR(1) model whose sums of squares and cross products for \((Y_t, Y_{t+1})\) have well known properties. The one step Gauss Newton estimator, for the nonlinear model and data generated under the null hypothesis that \(\beta = 0\), will produce a hypothesis test for linearity versus the particular type of nonlinearity imposed by our model.

Letting \(\rho_0 = (e^{\alpha} - 1)/(e^{\alpha} + 1)\), the AR(1) coefficient under the null hypothesis, we find the derivatives of \((Y_t - Y_{t+1})p(Y_{t+1})\). The logistic case, \(\rho_0 = e^{\alpha}/(e^{\alpha} + 1)\), follows in the same manner.

First, the derivatives of \(Y_{t+1}p(Y_{t+1}) = Y_{t+1}(e^{\alpha(Y_{t+1} - 1)}/(e^{\alpha(Y_{t+1} - 1)} + 1)\) with respect to \(\alpha\) and \(\beta\) evaluated at \(\beta = 0\), are

\[C_0(Y_{t+1}, Y_{t+1})\] where \(C_0\) is evaluated at \(\beta = 0\). Consistent starting values under the null hypothesis are provided by setting \(\beta_0 = 0\), fitting an AR(1) model with OLS, then solving for \(\alpha_0\). The one step Gauss Newton estimator consists of computing residuals, in this case the OLS residuals from the AR(1) fit, then regressing the vector of residuals, \(R_n\), on matrix \(F_n\) of derivatives to get the changes in the estimates. Assume for simplicity a normal \(N(0, \sigma^2)\) distribution for the errors. The sum of squares and cross products matrix for this regression, divided by \((n-1)\), is seen from the derivative expression above to be

\[
(n-1)^{-1}F_n'F_n =
\begin{pmatrix}
\sum_{t=1}^{n} Y_t^2/(n-1) & \sum_{t=1}^{n} Y_t Y_{t-1}/(n-1) \\
\sum_{t=1}^{n} Y_t Y_{t-1}/(n-1) & \sum_{t=1}^{n} Y_{t-1}^2/(n-1)
\end{pmatrix}
\rightarrow
C_0^2 \begin{pmatrix}
\sigma^2/(1-\rho_0^2) & 0 \\
0 & 3\sigma^4/(1-\rho_0^2)^2
\end{pmatrix} = G.
\]

The sum of cross products matrix \((n-1)^{-1/2}F_n'R\) converges to a multivariate normal distribution with mean 0 and covariance matrix \(G\) so that the parameter estimates are asymptotically normal with limit variance matrix \(G^{-1}\sigma^2\). As in the usual AR(1) case, the variance of \(\hat{\alpha}\) is thus unrelated to \(\sigma^2\), but in contrast, the variance of \(\hat{\beta}\) is inversely related to the error variance, indicating that in this (true \(\alpha = 0\)) case, the more spread out the data, the smaller the variance of \(\hat{\beta}\).


Should a series venture into a region where the logistic or hyperbolic tangent is close to 1, it will begin behaving like a random walk and thus may stay in this region for a prolonged period of time. In this region, the functions are quite flat, implying that it will be very hard to estimate \(\alpha\) and \(\beta\). Sigmoidal functions are most accurately estimated when the data surround the point of steepest slope, not when they mostly lie far to one side of this point. Our simulations reinforce this intuition. For certain \(\alpha\) and \(\beta\) combinations, the series move into a region in which the lag coefficient \(p(Y_{t+1})\) is very near 1. Once there, sequential values stay close to their immediate predecessors thus perpetuating \(\hat{\alpha}\) values near 1. In these situations, the nonlinear estimation of \(\alpha\) and \(\beta\) using nonlinear least squares, not surprisingly, can fail to converge. Data within the flat regions of the logistic function contain little information on these parameter values.

Simulations in Hwang (2004) suggest that distribution and convergence problems occur
infrequently for series in which a unit root test rejects the (unit root) null hypothesis. If \( p(Y_{t+1}) \) remains very near 1, we suggest abandoning the nonlinear model in favor of an ARIMA\((p,1,q)\).

The parameter estimate of \( \beta \) depends on the innovations variance \( \sigma^2 \), unlike what happens in the fixed coefficients case. Notice that \( Y_t = p(Y_{t-1}) Y_{t-1} + e_i \) is equivalent to

\[
Y_t / \sigma = p(Y_{t-1}) Y_{t-1} / \sigma + e_i / \sigma
\]

where \( p(Y_{t-1}) = \exp(\alpha + \beta \sigma(Y_{t-1}) - 1) / (1 + \exp(\alpha + \beta \sigma(Y_{t-1})/\sigma)) \) and similarly for the logistic case. Thus one need only study the distribution of estimates using an innovation variance 1 as long as the slope parameter is expressed as \( \beta \sigma \), that is, as a multiple of \( \sigma \).

There are certain symmetries in this estimation problem if the distribution of \( e \) is symmetric. First, note that \( -Y_t = p(Y_{t-1}) (Y_{t-1}) - e_i \), where \( p(Y_{t-1}) = (\exp(\alpha - \beta (Y_{t-1}) - 1) / \exp(\alpha - \beta (Y_{t-1}) + 1) \) so any data generated from our original model is precisely equivalent to data generated with the negatives of those shocks and using the negative of the original slope. Since the negatives of the shocks have the same distribution as the original shocks, this implies that the distribution of estimates for \( -\beta \) is the mirror image of that for \( \beta \). Simulating both series one observes that, in fact the \( \alpha \) estimates are identical for both while the \( \beta \) estimates are exactly the negatives of each other. Thus the region of well behaved estimators must be symmetric in \( \beta \) when \( Y_{t-1} \) is used as the argument of the hyperbolic tangent function.

If we use \( |Y_{t-1}| \) as the argument of the hyperbolic tangent, another symmetry results from the fact that for any \( x \), \((e^{2x} - 1)/(e^{2x} + 1) = -(e^{x} - 1)/(e^{x} + 1)\). Generate a sequence \( Y_t = (\exp(\alpha + \beta |Y_{t-1}| - 1) / \exp(\alpha + \beta |Y_{t-1}| + 1) \)\( Y_{t-1} + e_i \). Now let \( X_i = (-1)^{t-1} Y_t \) and notice that \((-1)^{t-1} Y_t = -(\exp(\alpha + \beta |Y_{t-1}| - 1) / \exp(\alpha + \beta |Y_{t-1}| + 1) \)\( (-1)^{t-2} Y_{t-1} + (-1)^{t-3} e_i \), which shows that \( X_i = (\exp(-\alpha - \beta |X_{t-1}| - 1) / \exp(-\alpha - \beta |X_{t-1}| + 1) X_{t-1} + (-1)^{t-3} e_i \). Indeed, simulations verify that the nonlinear estimates using \( X \) are exactly the negatives of those using \( Y \). If we have well behaved estimators at some \( (\alpha, \beta) \) we would have well behaved estimators at \( (-\alpha, -\beta) \).

Hwang’s simulations of series with various true values of \( \alpha \) and \( \beta \) indicated that the region in which convergence problems are minimal is about the same as the region in which unit root tests reject. He looked at a range of \( (\alpha, \beta) \) parameters with values from -4 to 4 for both. He looked for cases with a high rate of numerical convergence (99.5%) and close to the claimed 95% coverage rate for confidence intervals on the parameters. The hyperbolic tangent using \( Y_{t-1} \) and that using \( |Y_{t-1}| \) both satisfied his criteria in a triangle given approximately by \( 0 < \beta \sigma < 3 \) and \(-4 < \alpha < 3 - 7\beta \sigma/3 \). His results verify our expectations above: for \( Y_{t-1} \), the criteria were also satisfied at the corresponding \( (\alpha, -\beta) \) and for \( |Y_{t-1}| \) at \((-\alpha, -\beta) \).

7. Examples

Returning to the Kinston flow data, we fit our nonlinear (hyperbolic tangent) autoregressive model to deviations from a yearly sinusoid. Using a sine-cosine pair, we add their coefficients \( S_1b \) and \( C_1b \) to the list of parameters in the nonlinear estimation. The results appear in Table 3. \( A \) and \( B \) are the logistic constant and trend coefficients.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>22857.3</td>
<td>199935</td>
</tr>
<tr>
<td>Error</td>
<td>28666.3</td>
<td></td>
</tr>
<tr>
<td>Uncor Tot</td>
<td>22857.3</td>
<td>199935</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parm</th>
<th>Est</th>
<th>Approx 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3.2817</td>
<td>2.0700</td>
</tr>
<tr>
<td>B</td>
<td>-1.3314</td>
<td>-2.4949</td>
</tr>
<tr>
<td>Mu</td>
<td>8.7117</td>
<td>8.1481</td>
</tr>
<tr>
<td>S1b</td>
<td>0.9541</td>
<td>0.4813</td>
</tr>
<tr>
<td>C1b</td>
<td>1.8079</td>
<td>1.3092</td>
</tr>
</tbody>
</table>

This \( B<0 \) suggests a larger lag 1 coefficient for low than for high flows, indicating more stability during low flows. However, an improved fit can be obtained using more lags.

The stationarity region for the autoregressive order 2 model, \( Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + e_i \) is the region where the roots of the characteristic equation, \( m^2 - \alpha_1 m - \alpha_2 = 0 \), are both less than 1 in magnitude with unit root processes resulting when one or both are 1. Because \( -\alpha_2 \) is the
product of these roots, replacing it by a hyperbolic tangent \( \rho(Y_{t-2}) \) allows it to cover the entire region implied by the stationarity condition. We do so, letting the lag 1 coefficient be constant, and present the results in Table 4. Starting values are obtained by fitting a fixed coefficient AR(2) and sinusoidal trend, setting the initial \( \beta \) to 0, then solving for the other parameters to match the fixed coefficient model.

**Table 4. Flow Analysis.**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>F Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>22818.9</td>
<td>241190</td>
</tr>
<tr>
<td>Error</td>
<td>392</td>
<td>6.1812</td>
<td></td>
</tr>
<tr>
<td>Uncor Tot</td>
<td>398</td>
<td>22825.1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Approx 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-1.3267</td>
<td>-1.5837 -1.0698</td>
</tr>
<tr>
<td>B</td>
<td>0.0646</td>
<td>0.0120 0.1173</td>
</tr>
<tr>
<td>Mu</td>
<td>7.3838</td>
<td>7.1136 7.6540</td>
</tr>
<tr>
<td>S1b</td>
<td>0.9528</td>
<td>0.5856 1.3200</td>
</tr>
<tr>
<td>C1b</td>
<td>0.2498</td>
<td>-0.0606 0.5602</td>
</tr>
<tr>
<td>D</td>
<td>1.5278</td>
<td>1.4459 1.6097</td>
</tr>
</tbody>
</table>

In Figures 5A and 5B we see the series and forecasts from two models.

**Figure 5A. Flow predictions AR(2).**

**Figure 5B. AR(2) with nonlinear lag 2 coefficient.**

The gray dotted lines are the data. Figure 5A is the sinusoid plus fixed AR(2) model and 5B is the nonlinear fit just described. The sinusoid in 5A is higher on the left, overpredicting the early part of the series and underpredicting the later part. The forecast returns quickly to that sinusoid. In 5B, the sinusoid fits the early data much better and thus lies even farther below the later data. This along with the seemingly strong autocorrelation near the end of the series suggests that the dominant root of the characteristic equation is higher toward the end.

If we compute the quadratic characteristic polynomial as in the fixed coefficient case, we get a different equation and pair of roots at each time t. These are plotted against time in Figure 6. When the roots are complex, their magnitude is plotted as a single point at those times.

**Figure 6. Roots of \( m^2 - 1.53m - \rho(Y_{t-2}) \).**
The dominant root is near 1 at the end of the series and at historic peak flows. This might be interpreted as smoother behavior near peaks than near troughs. The large variance associated with near unit root episodes allows the fitted sinusoidal trend to discount the high flows toward the end of the series. Roots associated with the forecast approach limits well within the stationarity region as we forecast out into the future.

Recall that the requirement that the second lag coefficient be between -1 and 1 is a necessary, but not sufficient condition for the parameters to lie in the AR(2) stationarity region. If we take 1.5278, the estimate, as known and look at $m^2 - 1.5278m - C$, we see that both roots are complex with squared magnitude $|C|$ or the largest is positive. Requiring the largest to be less than 1 requires $1.5278 + (1.5278^2 + 4C)^{1/2} < 2$, that is, $C < -0.5278$. As a check, we use the estimated parameters to generate 100 future values of the time series conditional on the observed data. We repeat this 15,000 times using independent generated errors then compute, at each time $t$, the mean and the 5th and 95th percentiles of the 15000 future values. These (light dotted lines) are overlaid on the plug in forecasts and sinusoidal trend in Figure 7. In the simulated continuations of the series, $\max(\rho(Y_{t-2}), -0.5278)$ was used instead of $\rho(Y_{t-2})$. Without this bound the generated series often ventured into the explosive region, producing extremely high values and unreasonable averages.

**Figure 7. Two Forecasts, Nonlinear Model.**

The second example consists of weekly soybean prices in N.Carolina, data kindly provided by Prof. Nick Piggott. A graph of the data and below it the $\rho(Y_{t-1})$ estimates, with some reference lines, appear as Figure 8. An initial unit root test with 9 augmenting lags rejected unit roots. The nonlinear lag 1 model fit to these data gave Table 5.

**Figure 8. Soybean Prices**

**Table 5. N.C. Soybean Price Data.**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>37652.0</td>
<td>12550.7</td>
</tr>
<tr>
<td>Error</td>
<td>895</td>
<td>37.2886</td>
<td>0.0417</td>
</tr>
<tr>
<td>Uncor Tot</td>
<td>898</td>
<td>37689.3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parm</th>
<th>Estimate</th>
<th>Approximate 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4.3006</td>
<td>2.1949</td>
</tr>
<tr>
<td>mu</td>
<td>7.8189</td>
<td>6.8801</td>
</tr>
<tr>
<td>b</td>
<td>-1.2435</td>
<td>-1.9299</td>
</tr>
</tbody>
</table>
The first order autocorrelation of the residuals was -0.01 and no lag through 24 had autocorrelation exceeding 0.1 in magnitude and we ignore these, though with almost 900 observations some are statistically significant.

The b<0 estimate suggests that a smaller lag 1 coefficient, implying less persistence and more mean reversion tendency, is associated with higher prices. When prices are high they depend less on their predecessors than when they are low.

**Summary**

A transfer function has been introduced that allows different lag coefficients depending on the level of the input series. River flows at two stations are used to motivate and illustrate the model. For the example series, larger upstream flows seem to have longer lasting effects on the downstream flows in the sense that more weight is given to further back lags in that case.

A nonlinear autoregressive model that allows a series to exhibit both locally stationary and nonstationary features has been introduced. Analytic results appear to be quite hard to prove. Some symmetry properties of the estimators, limiting normality for a very particular case, and failure of a property commonly found in ergodic processes were shown. Simulations suggest a region of parameter values for which reasonable behavior of estimators is to be expected. An illustration of this model on detrended river flow data is given. A second model is fit using more lags, and its forecasts compared to a standard fixed coefficient model. A final example using soybean prices is given.

**References**


On A- and D-rotatability of two-dimensional third-order designs

S. Huda*, L. Benkherouf and F. Alqallaf
Department of Statistics, Faculty of Science, Kuwait University
P.O.Box-5969, Safat-13060, Kuwait
*E-mail: shuda@kuc01.kuniv.edu.kw

Abstract

The expression for variance-covariance matrix of the estimated axial slopes at a point in the factor space is obtained for a symmetric balanced two dimensional third-order design. The trace and determinant of the matrix are derived in order to show that symmetry and balance are not sufficient for either A-rotatability or D-rotatability of the design. A necessary condition for A-rotatability is provided. It is further shown that rotatability in Box and Hunter (1957) sense is sufficient for both A- and D-rotatability. The expression for the determinant provided may be useful for obtaining the D-minimax design in two dimensions.

Key words: A-rotatability, D-rotatability, Response surface designs, Third-order designs

AMS Subject Classification: 62K05, 62K99

1. Introduction

Classical experimental designs were concerned with comparative experiments, that is, experiments in which the primary objective is to compare the effects of various treatments and, especially, to estimate treatment contrasts. An exception is the more recently developed field of response surface designs in which treatments are various combinations of different levels of the factors that are quantitative. Here the main objective of the experimenter is usually to estimate the absolute response or the parameters of a model providing the relationship between the response and the factors. In this context, rotatable designs were introduced by Box and Hunter (1957) in order to explore the response surface. Rotatable designs have the nice property that the variance of the estimated response is constant at points equidistant from the centre of the design, conventionally taken to be the origin of the factor space, after transformations if necessary. Rotatable designs generate information about the response surface equally in all directions and are therefore useful when no or little prior knowledge is available about the nature of the response surface. The class of rotatable designs is also very rich in the sense that under many commonly used criteria, such as D-optimality, the optimal designs for polynomial regression models over hyperspherical regions may be found within this class (Kiefer (1960)). Because of the above reasons, a large volume of literature in experimental design is devoted to the investigation of properties and constructional problems of rotatable designs.

It has been recognized in recent years that even in response surface designs the main interest of the experimenter may not always be in the response at individual locations. Sometimes, the differences between responses at various locations may be of greater interest (Herzberg (1967), Box and Draper (1980) and Huda and Mukerjee (1984)). For more recent work dealing with such situations the reader is requested to see Huda (2006 a) and the references therein. If interest is in difference between responses at points close together in the factor space, the estimation of local slopes of the response surface becomes important. Estimation of slopes is particularly relevant in situations where the experimenter wishes to determine optimal settings of the factors in order to produce the maximum (minimum) value of the response. Atkinson (1970) initiated research into designs for estimating slopes. Subsequently, Ott and Mendenhall (1972), Murty and Studden (1972),
2. Background material

Suppose that the scalar response variable $y$ depends on $k$ quantitative factors $x_1, \ldots, x_k$ through a smooth functional relationship $y(x) = \phi(x, \theta)$ where $x = (x_1, \ldots, x_k)^t$ and $\theta = (\theta_1, \ldots, \theta_p)^t$ is a column vector of $p$ unknown parameters. Any design $\xi$ is a probability measure on the experimental/design region $\chi$, the part of the factor space allowing experimentation. Let $y_i$ be the response observed at point $x_i = (x_{i1}, \ldots, x_{ik})^t$ ($i = 1, \ldots, N$) chosen according to the design $\xi$. We assume that $y_i = \phi(x_i, \theta) + \epsilon_i$ where the $\epsilon_i$’s are uncorrelated zero-mean random errors with a constant variance $\sigma^2$.

Let $\hat{\theta}$ be the least squares estimate (l.s.e.) of $\theta$. Then $\hat{y}(x) = \phi(x, \hat{\theta})$ is the l.s.e. of $y(x)$. The column vector of estimated slopes along the factor axes at a point $z$ is given by $d\hat{y}/dz = (\partial\hat{y}(z)/\partial z_1, \ldots, \partial\hat{y}(z)/\partial z_k)^t$. Let $V(\bar{\xi}, z) = (N/\sigma^2) \text{ cov } (d\hat{y}/dz)$ be the normalized variance-covariance matrix of the estimated slopes at the point $z$. Clearly $V(\bar{\xi}, z)$ depends on the design $\bar{\xi}$ used as well as the point $z$ at which the slopes are being estimated. The estimated directional derivative at point $z$ in the direction specified by the vector of direction cosines $c = (c_1, \ldots, c_k)^t$ is $c^t d\hat{y}/dz$.

Further, $\{(d\hat{y}/dz)^t\}^{-1/2} d\hat{y}/dz$ is the estimate of the direction in which this derivative is largest.

In the linear model setting $\phi(x, \theta) = f^t(x) \theta$ where $f^t(x) = (f_1(x), \ldots, f_p(x))$ contains $p$ linearly independent functions of $x$. In this case, $(N/\sigma^2) \text{ cov } \hat{\theta} = M^{-1}(\bar{\xi})$ where $M(\bar{\xi}) =
\[ \int f(x) f^t(x) \xi(dx) \] is the information matrix of the design \( \xi \). Further, \( V(\xi, z) = H(z)M^{-1}(\xi)H^t(z) \) where \( H(z) \) is the \( k \times p \) matrix whose \( i \)-th row is \( \frac{\partial f^i(z)}{\partial z_i} = (\frac{\partial f_1(z)}{\partial z_1}, \ldots, \frac{\partial f_p(z)}{\partial z_p}) \) (\( i = 1, \ldots, k \)).

Among the linear models, polynomial models appear to be the most commonly employed in experimental design. The model is called a polynomial model of degree (or order) \( d \) if \( f(x) \) contains the powers and products of \( x_i \)'s up to and including degree \( d \). A full \( d \)-th order model contains \( k + 2d \) parameters. A design \( \xi \) is called a \( d \)-th order design if it is non-singular for a \( d \)-th order model, that is, if it allows estimation of all the parameters in a \( d \)-th order model.

A design \( \xi \) of order \( d \) is said to be symmetric if all the \( " \text{odd moments} \)" up to and including order \( 2d \) are zero, that is, if \( \int \chi_1 \ldots \chi_d \xi(dx) = 0 \) whenever one or more of the \( \chi_i \)'s are odd integers subject to \( \sum_{i=1}^d d_i \leq 2d \). A design is said to be balanced (permutation invariant) if the moments are invariant with respect to permutation of the factors \( x_1, \ldots, x_k \).

A \( d \)-th order design is said to be A- (D-) rotatable if the trace (determinant) of \( V(\xi, z) \) depends on the point \( z \) only through \( \rho = (z^t z)^{1/2} \). The moment requirements of a second-order design to be A- rotatable have been investigated in the past by several researchers. The moment requirements of a second-order design to be D-rotatable have been investigated in Huda (2006a) and Huda, Benkherouf and Alqallaf (2006). In what follows we consider third-order designs.

3. Third-order designs

For the full third-order model it is most convenient to express \( f^t(x) \) as
\[
\begin{pmatrix}
  f_1^t(x), & f_2^t(x), & f_3^t(x), & f_4^t(x)
\end{pmatrix}
\]
(1)

where
\[
\begin{align*}
f_1^t(x) &= (1, x_1^2, \ldots, x_k^2), \\
f_2^t(x) &= (x_1 x_2, \ldots, x_{k-1} x_k), \\
f_3^t(x) &= (x_1 x_2 x_3, x_1 x_2 x_4, \ldots, x_k x_{k-2} x_{k-3}), \\
f_4^t(x) &= (g_1^t(x), \ldots, g_k^t(x)),
\end{align*}
\]
and
\[
\begin{align*}
g_i^t(x) &= (x_1, x_i^2, x_i x_1^2, \ldots, x_i x_{i-1}^2, \ldots, x_k x_{k-1}^2, \ldots, x_k x_1^2) \ (i = 1, \ldots, k).
\end{align*}
\]
Thus, for a third-order design \( \xi \), the information matrix \( M(\xi) \) is a partitioned matrix given by
\[
M(\xi) =
\begin{pmatrix}
  M_{11}(\xi), & M_{12}(\xi), & M_{13}(\xi), & M_{14}(\xi) \\
  M_{21}(\xi), & M_{22}(\xi), & M_{23}(\xi), & M_{24}(\xi) \\
  M_{31}(\xi), & M_{32}(\xi), & M_{33}(\xi), & M_{34}(\xi) \\
  M_{41}(\xi), & M_{42}(\xi), & M_{43}(\xi), & M_{44}(\xi)
\end{pmatrix}
\]
(2)
where \( M_{ij}(\xi) = \int f_i(x) f_j^t(x) \xi(dx) \) (\( i,j = 1, \ldots, 4 \)) correspond to the partitioning of \( f^t(x) \) given in (1).

Without some assumptions about the moments of the design \( \xi \), it is not possible to obtain the inverse of \( M(\xi) \) algebraically, except perhaps in the simplest case of \( k = 2 \). With \( k = 2 \), \( M(\xi) \) is a 10 x 10 matrix and it is extremely difficult, if not impossible, to obtain the inverse of an arbitrary (although symmetric and positive definite) 10 x 10 matrix. Hence, in what follows we shall consider only symmetric designs.

For a symmetric design \( \xi \), \( M_{ij}(\xi) \) (\( i \neq j \)) are all null matrices and \( M(\xi) \) becomes a block diagonal matrix, as (2) is reduced to
\[
M(\xi) = \text{Diag} \left( M_{11}(\xi), M_{22}(\xi), M_{33}(\xi), M_{44}(\xi) \right).
\]
(3)
Corresponding to the partitioning of \( f^t(x) \) given in (1), the matrix \( H(x) \) may be expressed as
\[
H(x) = \{ H_1(x), H_2(x), H_3(x), H_4(x) \}
\]
(4)
where the $i$-th row of $H_j(x)$ is given by $\partial f_i(x)/\partial x_j$ (j = 1, 2, 3, 4 and i = 1, ..., k). Then for a symmetric design $\xi$, using (3) and (4), it is seen that

$$V(\xi, z) = \sum_{i=1}^{4} V_i(\xi, z)$$  \hspace{1cm} (5)

where $V_i(\xi, z) = H_i(z) M_j^{-1}(\xi) H_i'(z)$ (i = 1, 2, 3, 4).

Note that for a symmetric design $\xi$, $M_{22}(\xi)$, and $M_{33}(\xi)$ are diagonal matrices and further $M_{44}(\xi)$ itself is a block diagonal matrix given by

$$M_{44}(\xi) = \text{Diag}\{M_1^*(\xi), ..., M_k^*(\xi)\}$$  \hspace{1cm} (6)

where $M_i^*(\xi) = \int_{x} g_i(x) g_i'(x) \xi(dx)$ (i = 1, ..., k).

Even with the simplification provided by (6) it would not be possible to obtain $M^{-1}(\xi)$ for large $k$ unless further assumptions are made. Hence, we consider only symmetric balanced designs. This is also justified by the fact that under almost every conceivable design criterion an optimal design can always be found within this class. For a symmetric balanced third-order design $\xi$,

$$M_{11}(\xi) = \begin{bmatrix} 1 & \alpha_1 & \alpha_1^' \\ \alpha_1 & \alpha_1 & \alpha_1^' \end{bmatrix}$$

$$M_{22}(\xi) = \alpha_2 I_{k^*},$$

$$M_{33}(\xi) = \alpha_3 I_{k^*}. $$

$$M_{44}(\xi) = \begin{bmatrix} \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 \end{bmatrix}$$  \hspace{1cm} (7)

where $\alpha_i = \int_{x} x_i \xi(dx)$ (i = 1, ..., k) (j = 2, 4, 6);

$$\alpha_{22} = \int_{x} x_1^2 x_1^2 \xi(dx), \quad \alpha_{22} = \int_{x} x_1^2 x_1^2 \xi(dx), \quad \alpha_{22} = \int_{x} x_1^2 x_1^2 \xi(dx)$$  \hspace{1cm} (8)

where $\alpha_i = \int_{x} x_i \xi(dx)$ (i = 1, ..., k), k = (k-1)/2, k^* = k (k-1) (k-2)/6. $I_{k}$ is the identity matrix of order k, $I_{k}$ is the k-component column vector of 1’s and $E_{k^*} = 1_{k^*}$.

4. Two-dimensional third-order designs

For a third-order model in two variables $x_1$ and $x_2$, $f_1(x) = (1, x_1, x_2), f_2(x) = x_1 x_2$.

If $f_3(x)$ does not exist and $g_1(x) = (x_1, x_1^3, x_1 x_2^2)$, $g_2(x) = (x_2, x_2^3, x_1 x_2^2)$. Hence,

$$H_3(x) = \begin{bmatrix} 0 & 2x_1 & 0 \\ 0 & 0 & 2x_2 \end{bmatrix},$$

$$H_4(x) = \begin{bmatrix} 1 & 3x_1^2 & x_2^2 \\ 0 & 0 & 2x_1 x_2 \end{bmatrix}.$$  \hspace{1cm} (9)

If the two-dimensional third-order design $\xi$ is symmetric and balanced then from (7) we get

$$M_{11}(\xi) = \begin{bmatrix} 1 & \alpha_2 & \alpha_2 \\ \alpha_2 & \alpha_2 & \alpha_2 \\ \alpha_2 & \alpha_2 & \alpha_2 \end{bmatrix}$$

$$M_{22}(\xi) = \alpha_2 I_{k^*},$$

$$M_{44}(\xi) = \begin{bmatrix} \alpha_2 & \alpha_2 \\ \alpha_2 & \alpha_2 \\ \alpha_2 & \alpha_2 \end{bmatrix}$$  \hspace{1cm} (10)

Therefore, using (8) and (9) we obtain

$$v(\xi, z) = \sum_{i=1}^{4} V_i(\xi, z)$$

\text{Composition}
\[
V_4(\xi, z) = \frac{1}{\alpha_{22}} \begin{bmatrix}
    z_1^2 & z_1 z_2 \\
    z_1 z_2 & z_2^2
\end{bmatrix},
\]

(10)

\[V_4(\xi, z) = H_{40}(z)M^{-1}(\xi)H'_{40}(z) + H_{421}(z)M^{-1}(\xi)H'_{421}(z),\]

where

\[H_{40}(z)M^{-1}(\xi)H'_{40}(z) = \begin{bmatrix}
    v_{41(2,1)} & v_{41(2,2)} \\
    v_{41(2,1)} & v_{41(2,2)}
\end{bmatrix},\]

\[H_{421}(z)M^{-1}(\xi)H'_{421}(z) = \begin{bmatrix}
    v_{42(2,1)} & v_{42(2,2)} \\
    v_{42(2,1)} & v_{42(2,2)}
\end{bmatrix},\]

with

\[v_{41(2,1)} = A[\alpha_6 \alpha_{24} - \alpha_4^2] z_1^2 z_2,\]

\[v_{41(2,2)} = A[\alpha_6 \alpha_{24} - \alpha_4^2] z_1^2 z_2,\]

\[v_{42(2,1)} = 2A[(\alpha_4 \alpha_{24} - \alpha_2 \alpha_6) z_1 z_2 + 3(\alpha_4 \alpha_{24} - \alpha_2 \alpha_6) z_1^3 z_2 + (\alpha_6 \alpha_6 - \alpha_4^2) z_1^2],\]

\[v_{42(2,2)} = A[\alpha_6 \alpha_{24} - \alpha_4^2] + 6(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2 + 2(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2,\]

\[+ 9(\alpha_4 \alpha_{24} - \alpha_6 \alpha_{24}) z_1^4 + 6(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2,\]

\[+ 4(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2],\]

\[v_{21} = 4(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) + 6(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2 + 2(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2,\]

\[+ 9(\alpha_4 \alpha_{24} - \alpha_6 \alpha_{24}) z_1^4 + 6(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2,\]

\[+ 4(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2],\]

\[v_{22} = 4(\alpha_4 \alpha_{24} - \alpha_6 \alpha_{24}) + 6(\alpha_4 \alpha_{24} - \alpha_6 \alpha_{24}) z_1^2 + 2(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2,\]

\[+ 9(\alpha_4 \alpha_{24} - \alpha_6 \alpha_{24}) z_1^4 + 6(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2,\]

\[+ 4(\alpha_2 \alpha_{24} - \alpha_4 \alpha_{24}) z_1^2].\]

5. A-rotatability of two-dimensional third-order designs

Here we present some interesting results concerning A-rotatability of two-dimensional third-order designs.

**Lemma 1.** Symmetry and balance are not sufficient for A-rotatability of two-dimensional third-order designs.

**Proof:** For a design \( \xi \) to be rotatable, \( \text{tr} V(\xi, z) \) must be a function of \( z \) only through \( \rho = (z' z)^{1/2} \). Now from (11) it can be seen that for a symmetric balanced two-dimensional third-order design

\[\text{tr} V(\xi, z) = v_{11} + v_{22} = a + b \rho^2 + c(z_1^2 + z_2^2) + d z_1^2 z_2^2\]

(12)

where

\[a = 2 A (\alpha_6 \alpha_{24} - \alpha_4^2),\]

\[b = 0,\]

\[c = 6(\alpha_3 \alpha_{24} - \alpha_2 \alpha_6),\]

\[d = 2 A (6(\alpha_4 \alpha_{24} - \alpha_6 \alpha_{24}) + 4(\alpha_6 \alpha_6 - \alpha_4^2)).\]

and \( A \) as defined earlier. For (12) to be a function of \( \rho \) it is necessary to have \( d = 2c \). Since \( d \neq 2c \) for an arbitrary symmetric balanced design the lemma follows.

**Theorem 1.** A two-dimensional third-order symmetric balanced design \( \xi \) is A-rotatable if and only if its moments satisfy

\[\alpha_6 \alpha_{24} - \alpha_4^2 + 2\alpha_2 \alpha_{24} - 5\alpha_4 \alpha_{24} + 3\alpha_6^2 = 0.\]

(13)

**Proof:** The theorem follows from the preceding lemma since condition (13) is simply a restatement of \( d = 2c \).
Corollary 1. A two-dimensional third-order rotatable design is A-rotatable.

Proof: For a rotatable design
\[ \alpha_4 = 3\alpha_2, \alpha_6 = 5\alpha_2, \] and hence (13) is satisfied.

Corollary 1 above shows that rotatability is a sufficient condition for A-rotatability. However, we do not know if it is also a necessary condition. We do not know if there exist symmetric and balanced designs outside the rotatable subclass which are A-rotatable. The matter requires further investigation.

6. D-rotatability of two-dimensional third-order designs

For the two-dimensional third-order design to be D-rotatable, \( | V(\xi, z) | \) must be a function of \( z \) only through \( \rho \). Now from (11), we have \( | V(\xi, z) | = v_{11}v_{22} - v_{12}^2 \) which is a polynomial of order (degree) 8 in the two variables \( z_1, z_2 \) and the expression is quite nasty even for the symmetric balanced design. Obviously, the moment requirements of a design to be D-rotatable are going to be at least as stringent as the moment requirements of a design to be A-rotatable since D-rotatability involves all the elements of \( V(\xi, z) \) whereas A-rotatability involves only the diagonal elements. In view of this we consider only rotatable designs here and present the following result. We use the notation \( \alpha_1 = \lambda_1, \alpha_4 = 3\lambda_2, \alpha_6 = 5\lambda_4, \alpha_8 = 15\lambda_s = 15\lambda_8 \), as traditional for rotatable designs.

Theorem 2: For a two-dimensional third-order design rotatability is a sufficient condition for D-rotatability.

Proof: A direct substitution of \( \alpha_1 = \lambda_1, \alpha_4 = 3\lambda_2, \alpha_6 = 5\lambda_4, \alpha_8 = 15\lambda_s = 15\lambda_8 \) in the expression \( | V(\xi, z) | = v_{11}v_{22} - v_{12}^2 \) gives
\[
| V(\xi, z) | = B(\beta_0 + \beta_1\rho^2 + \beta_2\rho^4 + \beta_3\rho^6 + \beta_4\rho^8) \quad (14)
\]
where
\[
\beta_0 = 288\lambda_1^4\lambda_2^2 - 144\lambda_1^2\lambda_2^4, \\
\beta_1 = 720\lambda_1^2\lambda_2^2 + 388\lambda_1^4\lambda_2^2 - 288\lambda_1^2\lambda_2^4 - 864\lambda_1^4\lambda_2^4, \\
\beta_2 = 216\lambda_1^4\lambda_2^2 + 432\lambda_1^2\lambda_2^4 + 432\lambda_1^4\lambda_2^4 - 154\lambda_1^2\lambda_2^4 - 672\lambda_1^2\lambda_2^4 - 168\lambda_1^4\lambda_2^4, \\
\beta_3 = 280\lambda_1^4\lambda_2^2 + 216\lambda_1^2\lambda_2^4 + 396\lambda_1^4\lambda_2^4 - 168\lambda_1^2\lambda_2^4 - 684\lambda_1^4\lambda_2^4 - 9\lambda_2^4, \\
\beta_4 = 42\lambda_1^4\lambda_2^2 + 216\lambda_1^2\lambda_2^4 + 396\lambda_1^4\lambda_2^4 - 168\lambda_1^2\lambda_2^4 - 684\lambda_1^4\lambda_2^4 - 9\lambda_2^4, \\
B = 1/(16(2\lambda_1^2 - 3\lambda_2^2)\lambda_1^2\lambda_2^4(3\lambda_s^2 - 2\lambda_1^2)^2). \]
Expression (14) shows that rotatability is a sufficient condition for D-rotatability.

7. Some remarks

D-rotatability is a fairly recent concept and not much is known about it even for designs of low order. In particular, it is not known if a D-rotatable design may be found outside the class of symmetric balanced designs.

It was shown in Huda (2006b) that a second-order symmetric balanced two-dimensional design is D-rotatable if and only if it is rotatable. Huda, Benkerouf and Alqallaf (2006) extended the result and showed that any second-order symmetric balanced design in \( k \geq 2 \) dimensions is D-rotatable if and only if it is rotatable. The result presented here in Section 6 for two-dimensional third-order designs encourages us to conjecture a similar situation with \( k \)-dimensional third-order designs for any \( k \geq 2 \). However, even if the conjecture is correct, there may be, although highly unlikely, D-rotatable designs that are not symmetric or balanced. The matter remains to be investigated.

In comparison with A-rotatability, the D-rotatability is more difficult to deal with. None the less it may be an useful concept and deserves further investigations since it too has a statistical interpretation, namely, rotatability in terms of generalized variance of the estimated slopes.

Under D-minimax criterion the objective is to derive the design minimizing the value of \( | V(\xi, z) | \) maximized with respect to \( z \) over the region \( R \) of interest in the factor space (Huda and Al-Shiha (2000)). The formula for the determinant \( | V(\xi, z) | \) provided in (14) may be helpful in situations where derivation of optimal designs under the D-minimax criterion is needed.
References


Weibull-Based Approaches to Survival Analysis: An Application to a Breast Cancer Data Set

Abdus S. Wahed, The Minh Luong and Jong-Hyeon Jeong

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15217, USA.

SUMMARY

This article discusses various generalizations of the standard two-parameter Weibull distribution as applied to the survival analysis. It introduces a new generalization that incorporates the exponentiated Weibull model introduced by Mudholkar and Srivastava (1993). The generalized models along with the standard two-parameter Weibull model are fitted to a survival data set with competing risks. Comparison of cumulative incidence and survival among two treatment groups using all models resulted in similar conclusions.

1. Introduction

A suitable parametric model is often of interest in the analysis of survival data, as it provides insight into characteristics of failure times and hazard functions that may not be available with non-parametric methods. The Weibull distribution (Weibull, 1951) is one of the most commonly used families for modeling such data. However, only monotone increasing and decreasing hazard functions can be generated from the classic two-parameter Weibull distribution; as such this two-parameter model is inadequate when the true hazard shape is of unimodal or bathtub nature.

The widely used Kaplan-Meier product-limit estimator is a flexible method to model survival data but is noted to often be inefficient (Miller, 1983). Other semi-parametric methods such as proportional hazards require assumptions that may not be plausible in many situations (Cox and Oakes, 1984). Meanwhile, a vast array of parametric techniques have been developed to incorporate the wide variety of behaviors in survival data. Some of the proposed parametric models have incorporated a shape parameter into the classic Weibull distribution to account for additional possible hazard shapes. One of the first models of this type was proposed by Kalbfleisch and Prentice (1980), but this model may be impractical in the presence of censored data as it often requires the evaluation of an incomplete gamma integral or beta ratio.

An exponentiated version of the Weibull model was introduced by Mudholkar and Srivastava (1993) that incorporated one additional shape parameter. The distribution has a closed form of probability density,
survival, and hazard functions, that are flexible and able to generate a wide variety of hazard shapes, including the frequently observed unimodal and bathtub shapes. It can be shown that this extension of the Weibull can be achieved through a simple application of the probability integral transform, using the densities of Beta and two-parameter Weibull distributions. Mudholkar, Srivastava and Kollia (1996) proposed another generalization of the Weibull model, that incorporates proportional hazards and is able to generate the same types of hazard shapes as the exponentiated model; however irregularities may arise when the shape parameter $\lambda$ is positive, as the support becomes dependent on the parameters. A four-parameter generalization of the Weibull distribution was introduced by Jeong (2006), based on stable distributions proposed by Hougaard (1986) induced under the semiparametric frailty model. This model can be seen as a generalization of the model of Mudholkar et al (1996).

A decade has passed since the exponentiated Weibull and other extensions of the classic Weibull model has been introduced; however a limited amount of literature is available on their applications to survival modeling. This article presents a generalization of Weibull model, the beta-Weibull, that incorporates the two-parameter Weibull and the exponentiated Weibull distribution as special cases. It also examines the potential use of the beta-Weibull distribution as a model for survival data in comparison to other generalizations of the Weibull model. We demonstrate how the cumulative incidence functions can be estimated and tested based on these generalized Weibull models through an application to a breast cancer data set.

The article is organized as follows. In Section 2, the Weibull-based families are summarized along with an overview of its properties with respect to hazard functions, with particular emphasis on the beta-Weibull family. In Section 3, maximum likelihood estimation and their corresponding inferential procedures are described for the generalized Weibull models. In Section 4, these methods are applied to a previously analyzed breast cancer data set, and the performance of the exponentiated Weibull relative to other Weibull-based approaches are compared using conventional model assessment methods. A conclusion follows in Section 5.

2. The Weibull distribution and its extensions

A regular two-parameter Weibull distribution has the probability density function (pdf)

$$g_2(x; \gamma, \beta) = \frac{\gamma}{\beta} (x/\beta)^{\gamma-1} e^{-(x/\beta)^\gamma}, \quad 0 < x < \infty, \ 0 < \gamma, \beta < \infty$$

(1)

with its cumulative distribution function being

$$G_2(x; \gamma, \beta) = 1 - e^{-(x/\beta)^\gamma}, \quad 0 < x < \infty.$$  

(2)

While the distribution has been used as a practical tool for survival analysis such as proportional hazards and accelerated failure time models, and maximum likelihood estimates under this distribution are fairly easy to obtain using iterative methods, it can only be used for modeling data with monotonically increasing or decreasing hazard shapes. To overcome this limitation of the Weibull model, several generalized versions of
the classic two-parameter Weibull model have been introduced in order to increase flexibility while remaining practical with respect to parameter estimation and model assessment techniques. In this section we focus on two such models that required at most two extra parameters to be estimated.

2.1 A three-parameter extension

A three-parameter generalization (Mudholkar, Srivastava and Kollia, 1996) of the Weibull distribution is provided by the survival function

\[ S(x; \gamma, \beta, \lambda) = \{1 + (x/\beta)^\gamma/\lambda\}^{-\lambda}, \quad 0 < \gamma, \beta < \infty \]

(3)

with corresponding hazard function

\[ h(x; \gamma, \beta, \lambda) = \frac{\gamma}{\beta} \frac{(x/\beta)^{\gamma-1}}{1 + (x/\beta)^\gamma/\lambda}. \]

The regular case of this generalized Weibull distribution occurs when \( \lambda \geq 0 \) with the density having support \((0, \infty)\), and generates monotonically decreasing or unimodal hazard functions. In particular, the distribution approaches the two-parameter Weibull as \( \lambda \to \infty \). When \( \lambda < 0 \), monotonically increasing and bathtub hazard shapes are generated, but the support of the density becomes parameter-dependent and within the range \((0, \beta(-\lambda)^\gamma)\). Similar to other non-regular densities, in these situations the likelihood may become unbounded and thus maximum likelihood estimates may not exist, or alternately, if the maximum likelihood estimates exist their asymptotic properties might not follow classic asymptotic theory (Smith, 1985).

2.2 A four-parameter extension

A four-parameter Weibull model was proposed by Jeong (2006). This generalization of the Weibull distribution incorporates an additional parameter \( \tau \) to the three-parameter version proposed by Mudholkar et al. (1996) that was previously discussed. This generalization can be characterized by the survival function

\[ S(x; \gamma, \beta, \lambda, \tau) = \exp \left[ -\lambda^{1-\tau} \frac{(x/\beta)^\gamma + \lambda^{\tau} - \lambda}{\tau} \right], \quad 0 < x < \infty, \quad 0 < \gamma, \beta, \lambda < \infty, \quad -\infty < \tau < \infty \]

with the corresponding hazard function

\[ h(x; \gamma, \beta, \lambda, \tau) = \frac{\gamma(x/\beta)^\gamma \lambda^{1-\tau}}{x [(x/\beta)^\gamma + \lambda^{1-\tau}].} \]

This distribution reduces to Mudholkar’s generalized Weibull distribution as \( \tau \) approaches to zero. As one would expect, it covers a wider range of shapes than the three-parameter extension and is regular regardless of the parameter values.

2.3 The beta-Weibull distribution

Wahed (2006) showed that any parametric family of distribution can be incorporated into larger families through an application of the probability integral transform. Specifically, given two valid probability density
functions (pdf) $g_1(.)$ and $g_2(.)$ with the latter having support on the unit interval, a third valid density function $f(.)$ may be obtained by applying the equation

$$f(x) = g_2(G_1(x))g_1(x),$$

where $G_1(.)$ is the cumulative distribution function (cdf) corresponding to $g_1(.)$. Using this technique, we obtain yet another generalization of the two-parameter Weibull distribution.

To obtain the appropriate extension of the Weibull distribution, we considered $g_1(.)$ as a regular two-parameter Weibull pdf given by Equation (1) and $g_2(.)$ as a two-parameter beta distribution with pdf

$$g_2(u; \alpha_1, \alpha_2) \propto u^{\alpha_1-1}(1-u)^{\alpha_2-1}, \quad 0 < u < 1, \quad 0 < \alpha_1, \alpha_2 < \infty.$$  \hspace{1cm} (5)

Using (1) and (5) in Equation (4) we obtain the probability density function of the beta-Weibull distribution:

$$f(x; \alpha_1, \alpha_2, \beta, \gamma) = \frac{1}{B(\alpha_1, \alpha_2)} \frac{\gamma^\frac{\gamma}{\beta}}{1 - e^{-x/\beta}} \left(1 - e^{-x/\beta}\right)^{\alpha_1-1} e^{-\alpha_2 x/\beta} \left(x/\beta\right)^{\gamma-1},$$

where $0 < x < \infty$, $0 < \alpha_1, \alpha_2, \gamma, \beta < \infty$ and $B(\alpha_1, \alpha_2) = \Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1 + \alpha_2)$. We refer to this distribution as the beta-Weibull distribution. The density given by (6) is a proper density and when $\alpha_1 = \alpha_2 = 1$, it reduces to the classical two-parameter Weibull density with scale parameter $\beta$ and the shape parameter $\gamma$. This model is generated entirely based on two commonly used probability distributions, namely beta and Weibull, and therefore can be easily implemented using most conventional statistical software packages. The survival function of the beta-Weibull distribution cannot be expressed in a closed form; however it may be estimated by evaluating the regularized beta function at the cumulative distribution function of the Weibull distribution. More specifically, the survival function of the beta-Weibull distribution
Figure 2. Unimodal (left; \( \alpha_1 = 4, \gamma = 0.5, \beta = 2.0 \)) and v-shaped (right; \( \alpha_1 = 0.4, \gamma = 2.0, \beta = 2.0 \)) hazard shapes with constant \( \alpha_1 \) and varying \( \alpha_2 \) (left), and \( \alpha_2 \) and varying \( \alpha_1 \) (right), for the beta-Weibull distribution.

(6) is given by

\[
S(x; \alpha_1, \alpha_2, \gamma, \beta) = \frac{1}{B(\alpha_1, \alpha_2)} \int_{-e^{-(x/\beta)\gamma}}^{1} u^{\alpha_1-1} (1-u)^{\alpha_2-1} \, du \tag{7}
\]

\[
= 1 - I(1 - e^{-(x/\beta)\gamma}; \alpha_1, \alpha_2), \tag{8}
\]

where

\[
I(u; \alpha_1, \alpha_2) = \frac{\int_{0}^{u} u^{\alpha_1-1} (1-u)^{\alpha_2-1} \, du}{\int_{0}^{1} u^{\alpha_1-1} (1-u)^{\alpha_2-1} \, du} \tag{9}
\]

is the incomplete beta function. The corresponding hazard function is obtained as

\[
h(x; \alpha_1, \alpha_2, \beta, \gamma) = \frac{\gamma}{\beta} \frac{(1 - e^{-(x/\beta)\gamma})^{\alpha_1-1} e^{-\alpha_2(x/\beta)^{\gamma}} (x/\beta)^{\gamma-1}}{1 - I(1 - e^{-(x/\beta)\gamma}; \alpha_1, \alpha_2)}. \tag{10}
\]

The survival and hazard functions are plotted in Figures 1 and 2, respectively. The hazard plots show that this model can generate unimodal, v-shaped and bathtub shapes. The extra shape parameters, \( \alpha_1 \) and \( \alpha_2 \), are related to the beta-component of the distribution; as a result, the mean and variance of the distribution increases as \( \alpha_1 \) increases as they do for the beta distribution. Similarly, survival rate is higher for larger \( \alpha_1 \). The shape of the hazard function is influenced by the parameter \( \alpha_1 \) and \( \gamma \), while \( \beta \) remains a scale parameter. The second parameter from the beta distribution, \( \alpha_2 \), also primarily acts as a scale parameter as it is absorbed into the Weibull function through the exponent term \( e^{-(x/\beta)\gamma} \).

2.4 Exponentiated Weibull distribution

Notice that when \( \alpha_2 = 1 \), the beta-Weibull distribution defined by (6) reduces to

\[
f(x; \alpha_1, \beta, \gamma) = \frac{\alpha_1 \gamma}{\beta} (1 - e^{-(x/\beta)\gamma})^{\alpha_1-1} e^{-(x/\beta)\gamma} (x/\beta)^{\gamma-1}, \; 0 < x < \infty, \; 0 < \gamma, \beta, \alpha_1 < \infty, \tag{11}
\]
Figure 3. Exponentiated Weibull hazard functions, with varying \(\alpha\) and fixed \(\beta = 2.0\), for \(\gamma = 0.5\) (left), and \(\gamma = 2.0\) (right).

which is the so-called exponentiated Weibull distribution (Mudholkar and Srivastava, 1993; Mudholkar, Srivastava and Freimer, 1995). For this specific case, both survival function and hazard function are evaluable in closed form and are given by

\[
S(x; \alpha_1, \beta, \gamma) = 1 - \left[1 - e^{-\left(\frac{x}{\beta}\right)^\gamma}\right]^{\alpha_1} \tag{12}
\]

and

\[
h(x; \alpha_1, \beta, \gamma) = \frac{\alpha_1 \gamma}{\beta} \frac{\left(1 - e^{-\left(\frac{x}{\beta}\right)^\gamma}\right)^{\alpha_1-1} e^{-\left(\frac{x}{\beta}\right)^\gamma (x/\beta)^{\gamma-1}}}{1 - \left(1 - e^{-\left(\frac{x}{\beta}\right)^\gamma}\right)} \tag{13}
\]

respectively.

Defining \(\alpha = \alpha_1\) for notational convenience, the derivative of the hazard function with respect to \(x\) can be expressed as

\[
h'(x; \alpha, \beta, \gamma) = m(x; \alpha, \beta, \gamma) \left\{\gamma (x/\beta)^\gamma \left[\alpha e^{-\left(\frac{x}{\beta}\right)^\gamma} - 1 + (1 - e^{-\left(\frac{x}{\beta}\right)^\gamma})^\alpha\right] + (\gamma - 1)(1 - e^{-\left(\frac{x}{\beta}\right)^\gamma})[1 - (1 - e^{-\left(\frac{x}{\beta}\right)^\gamma})^\alpha]\right\} \tag{14}
\]

where

\[
m(x; \alpha, \beta, \gamma) = \frac{\alpha \gamma}{\beta^2} \frac{\left(1 - e^{-\left(\frac{x}{\beta}\right)^\gamma}\right)^{\alpha-2} e^{-\left(\frac{x}{\beta}\right)^\gamma (x/\beta)^{\gamma-2}}}{\left[1 - \left(1 - e^{-\left(\frac{x}{\beta}\right)^\gamma}\right)^\alpha\right]^2} \tag{15}
\]

Notice that \(m(x; \alpha, \beta, \gamma)\) is non-negative regardless of any values of \(x\) or parameters.

For \(\alpha < 1\), it is apparent that hazard shapes are unimodal if \(\alpha < 1/\gamma\) and monotonically decreasing otherwise; likewise when \(\alpha > 1\) bathtub shapes are generated for \(\alpha > 1/\gamma\), with monotonically increasing shapes otherwise. The hazard function is monotonically increasing when \(\alpha \geq 1, \gamma > 1\), while \(\alpha \leq 1, \gamma < 1\) implies a monotonically decreasing function; when both parameters are 1 the distribution reduces to the
Figure 4. Hazard functions, with varying $\gamma$ and fixed $\beta = 2.0$, for $\alpha = 0.2$ (left), and $\alpha = 2.0$ (right). Three types of hazard shapes are generated in each of the graph, with respect to $\gamma$ only.

exponential distribution (constant hazard). However, when one of the parameters is $> 1$ and the other $< 1$ more shapes can be generated. If $\alpha > 1$ and $\gamma < 1$ a unimodal hazard shape is generated if $\alpha > 1/\gamma$ and is monotonically decreasing otherwise; if $\alpha < 1$ and $\gamma > 1$ bathtub shapes are generated when $\alpha < 1/\gamma < 1$ and are monotonically increasing otherwise. Behavior of the hazard function with respect to the parameters is depicted in Figure 3 for values of $\alpha$ less than 1 (left) and greater than 1 (right). Similarly, the hazards for different values of the shape parameter $\gamma$ for $\alpha < 1$ (left) and $\alpha > 1$ (right) are plotted in Figure 4. Different types of hazard shapes may be obtained by altering this parameter while keeping the other two parameters fixed.

While both exponentiated Weibull distribution and three-parameter generalized Weibull (Section 2.1) reduce to the two-parameter Weibull distribution when $\lambda$ approaches $\infty$ and $\alpha$ approaches 1, there is no clear link between the parameters used in these distribution functions. This is apparent from the behavior of respective hazard functions; $\lambda < 0$ implies either a monotonically increasing or bathtub shape for the generalized Weibull, while $\alpha > 1$ may result in a monotone hazard function that is increasing or decreasing, or a unimodal function. Similarly, $\lambda > 0$ produces a monotonically decreasing or unimodal shape for the generalized Weibull, whereas if $0 < \alpha < 1$ for the exponentiated Weibull, a bathtub shape is possible, along with monotonically decreasing and increasing shapes. The exponentiated Weibull is a generalization of the Burr Type X distribution $S(x) = \left[1 - e^{-x/(\beta/\alpha)}\right]^\alpha$, while the family of the three-parameter generalized Weibull includes that of the Burr Type XII family of distributions as a special case.
3. Inference for competing risks model

In this section, we focus on the exponentiated Weibull distribution and its use in modelling censored survival data with competing risks (Kalbfleisch and Prentice, 1980).

The set of parameters $\phi = (\alpha, \gamma, \beta)$ of the exponentiated Weibull model for a given data set can be estimated through maximum likelihood method. For a given cause-specific event of type $k$, the observed data for the $i^{th}$ individual is denoted by $(x_i, \delta_{ki})$, where $x_i$ represents the event time (failure/censoring) and $\delta_{ki}$ is an indicator variable, taking value 1 when the $k^{th}$ cause-specific event occurs. The log-likelihood of the exponentiated Weibull model is then given by

$$
\ln \{l_k(\phi)\} = \sum_{i=1}^{n} \delta_{ki} \ln f_k(x_i, \phi_k) + (1 - \delta_{ki}) \ln S_k(x_i, \phi_k)
$$

$$
= \sum_{i=1}^{n} \delta_{ki} \left\{ \ln \frac{\alpha_k \gamma_k x_i^{\alpha_k-1} \exp(-\gamma_k x_i^{\beta_k})}{\beta_k} - (\alpha_k - 1) \ln \left( 1 - \frac{1 - e^{-(x_i/\beta_k)^{\gamma_k}}}{1 - e^{-(x_i/\beta_k)^{\gamma_k}}} \right) \right\}
$$

The first and second derivatives of the above expression can be used to obtain the score functions and observed information matrix for each type of events. Thus, an iterative procedure such as the Newton-Raphson method can be used to obtain maximum likelihood estimates for the parameter vector $\phi_k$ of a cause-specific event.

The score functions can be expressed as:

$$
S(\alpha_k) = \sum_{i=1}^{n} \delta_{ki} \left\{ \frac{1}{\alpha_k} + \ln \left( 1 - e^{-(x_i/\beta_k)^{\gamma_k}} \right) \right\} - \delta_{ki} \left\{ \frac{1 - e^{-(x_i/\beta_k)^{\gamma_k}}}{1 - e^{-(x_i/\beta_k)^{\gamma_k}}} \right\} \frac{\alpha_k}{\beta_k} \ln \left( 1 - e^{-(x_i/\beta_k)^{\gamma_k}} \right)
$$

$$
S(\gamma_k) = \sum_{i=1}^{n} \delta_{ki} \left\{ \frac{1}{\gamma_k} + \frac{(\alpha_k - 1) e^{-(x_i/\beta_k)^{\gamma_k}} (x_i/\beta_k)^{\gamma_k-1} \ln(x_i/\beta_k)}{1 - e^{-(x_i/\beta_k)^{\gamma_k}}} + \ln(x_i/\beta_k) - (x_i/\beta_k)^{\gamma_k-1} \ln(x_i/\beta_k) \right\}
$$

$$
S(\beta_k) = \sum_{i=1}^{n} \delta_{ki} \left\{ \frac{1}{\beta_k} - \frac{\gamma_k (\alpha_k - 1) e^{-(x_i/\beta_k)^{\gamma_k}} (x_i/\beta_k)^{\gamma_k}}{1 - e^{-(x_i/\beta_k)^{\gamma_k}}} - \frac{\gamma_k - 1}{\beta_k} + \frac{\gamma_k (x_i/\beta_k)^{\gamma_k}}{1 - e^{-(x_i/\beta_k)^{\gamma_k}}} \right\}
$$

where $\delta_{ki}^* = 1 - \delta_{ki}$. To solve the score equations and thus obtain maximum likelihood estimates, the S-Plus function *nlinmod* was used to implement the above non-linear minimization procedures. The initial values of the parameters $\gamma$ and $\beta$ were chosen from the results of fitting a two-parameter Weibull model to the data, while the initial value of $\alpha$ can be set close to zero or one. When using the Newton-Raphson procedure to obtain the MLEs, in order to facilitate convergence it is suggested to set the options of the minimization procedure such that the iterative step length is of similar magnitude between each parameter (Thisted, 1988).
Once the MLE of $\phi = (\alpha, \gamma, \beta)$ is obtained, the variance of these estimates are calculated by inverting the observed information matrix.

The corresponding estimator for the survival function is obtained from (12) through the invariance property of the maximum likelihood estimate, namely,

$$\hat{S}(x; \alpha, \beta, \gamma) = S(x; \hat{\alpha}, \hat{\beta}, \hat{\gamma}) = 1 - \left\{1 - e^{-\left(\frac{x}{\hat{\beta}}\right)^\alpha}\right\}.$$

The standard error of (16) is then obtained through an application of multivariate delta method.

When multiple cause-specific events, or competing risks, are present, the cumulative incidence function assesses the cumulative probability of a particular cause-specific event while taking into account other types of events. This prevents overestimation of survival probability of a distribution of cause specific events of interest, which would occur if failures due to other causes are considered as censored (Gooley, 1999). For data consisting of only two types of cause-specific events, the cumulative incidence function for a cause-specific event of type 1 is estimated by:

$$\hat{F}_1(t; \phi) = \int_0^t S_2(u; \hat{\phi}_2) f_1(u; \hat{\phi}_1) du$$

with an analogous expression for the cumulative incidence function for the other competing risk.

The multivariate delta method can be used to approximate the variance of the cumulative incidence or survival function (see Jeong (2006) for derivation):

$$\text{var}(\hat{F}_1(t; \phi)) = \left. \left( \frac{\partial F_1(t; \phi)}{\partial \phi} \right) \right|_{\phi = \hat{\phi}} \text{var}(\hat{\phi}) \left( \frac{\partial F_1(t; \phi)}{\partial \phi} \right)^T \left. \text{var}(\hat{\phi}) \right|_{\phi = \hat{\phi}}.$$

Equality of survival curves or incidence functions between independent treatment groups can then be tested using one of the three widely known methods of hypothesis testing: the likelihood ratio test, Wald test and score test. Here we will concentrate on the Wald test; for testing the equality of two parametric functions $\theta^{(1)}(t)$ and $\theta^{(2)}(t)$ from two different treatments at a specific time point $t$, the appropriate test statistic is

$$Z = \frac{\hat{\theta}^{(1)}(t) - \hat{\theta}^{(2)}(t)}{\sqrt{\text{var}(\hat{\theta}^{(1)}(t)) + \text{var}(\hat{\theta}^{(2)}(t))}}.$$  

where $Z \sim N(0, 1)$. For instance, the appropriate Wald test from (19) is applied to test for differences between cumulative incidence functions by comparing the test statistic

$$Z = \frac{\hat{F}_1^{(2)}(t; \phi) - \hat{F}_1^{(1)}(t; \phi)}{\sqrt{\text{var}(\hat{F}_1^{(1)}(t; \phi)) + \text{var}(\hat{F}_1^{(2)}(t; \phi))}}$$

(20)

to the critical values from the standard normal distribution.

4. Application of Weibull-based models to breast cancer data

To investigate the practicality of the Weibull-based families, we fitted the models to a data set from a breast cancer study performed by the National Surgical Adjuvant Breast and Bowel Project Study (NSABP). The
study (Protocol B-20) was designed to test if the addition of chemotherapy to tamoxifen would result in a reduction in disease-free survival compared to tamoxifen alone. Here the disease-free survival events include breast cancer recurrences in local, regional, or distant sites, second primary cancers other than breast, or death prior to previously mentioned events, whichever occurs first. In this section, the cause-specific events of interest will be the local or regional recurrences, and the other competing events will include second primaries and death without evidence of disease. A total of 2,363 patients were randomized to tamoxifen, CMF (a chemotherapy containing the alkylating agent cyclophosphamide) or CMFT (CMF plus tamoxifen), but our example only consists of 770 patients in the tamoxifen group and 766 in the CMFT group. We present our data only up to first 10 years because the data get sparse later on.

4.1 Maximum likelihood estimates

The maximum likelihood estimates using the aforementioned non-linear minimization procedures are displayed in Table 1. The standard errors were obtained from inverting the matrix of minus the second-derivative of the log-likelihoods, evaluated at the MLEs. While the two-parameter Weibull model produced maximum likelihood estimates with relatively small estimated standard errors, the corresponding three extensions of the Weibull often produced very large estimate of the standard errors. This is due to the fact that near the boundary conditions, the MLE’s are often dependent on the initial parameters of the iterative procedures (Thisted, 1988), and the variance estimates are often unreliable (Smith, 1986).

Within the higher-risk tamoxifen group, the -2 x log-likelihood decreased significantly, according to the likelihood ratio test, with each parameter added to the Weibull model. The log-likelihoods of the two three-parameter models were comparable, though slightly lower for the exponentiated Weibull fit. In contrast, in the lower-risk CMFT group there were no significant differences between the log-likelihoods.

The hazard shapes for maximum likelihood estimates for each of the four model fits is displayed in Figure 5 for each type of the competing risks and in Figure 6 for all disease-free events. The non-parametric hazard shape was determined from life table estimates of the hazard at the midpoint of each year. Model fits were generally better for the other events, as local or regional recurrences were relatively infrequent and had more unpredictable hazard shapes.

The generalized Weibull models were able to produce both the basic monotone shapes (other events), as well as capture more complex unimodal hazard shapes (local or regional recurrences). Within the tamoxifen group, according to the non-parametric fit there was an apparent peak at 2-3 months which leveled off by the end of the study. This peak was inadequately captured by the two-parameter model which only produced monotonic shapes; the Weibull extensions produced unimodal shapes that had this peak. The curves from the exponentiated model diverged the least from the two-parameter model, tending to produce a more conservative peak at 2-3 months than those from the other two extensions of the Weibull. It is interesting to observe that, as the number of parameters increases, a model tends to pick up the peak better,
### Table 1

*Maximum likelihood estimates for breast cancer data*

<table>
<thead>
<tr>
<th>Distribution</th>
<th>par</th>
<th>specific</th>
<th>mle(se)</th>
<th>other</th>
<th>mle(se)</th>
<th>all events</th>
<th>mle(se)</th>
<th>-2LL</th>
<th>mle(se)</th>
<th>-2LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-parameter Weibull</td>
<td>γ</td>
<td>0.983</td>
<td>(0.130)</td>
<td>1.099</td>
<td>(0.084)</td>
<td>2084.06</td>
<td>1.067</td>
<td>(0.070)</td>
<td>1847.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>126.53</td>
<td>(47.39)</td>
<td>35.84</td>
<td>(4.57)</td>
<td>28.39</td>
<td>(2.82)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-parameter Weibull</td>
<td>γ</td>
<td>2.434</td>
<td>(0.819)</td>
<td>1.656</td>
<td>(0.305)</td>
<td>2070.84</td>
<td>1.831</td>
<td>(0.325)</td>
<td>1836.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>8.039</td>
<td>(5.217)</td>
<td>12.19</td>
<td>(4.76)</td>
<td>8.009</td>
<td>(2.566)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>λ</td>
<td>0.017</td>
<td>(0.009)</td>
<td>0.128</td>
<td>(0.063)</td>
<td>0.126</td>
<td>(0.048)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-parameter Weibull</td>
<td>γ</td>
<td>4.763</td>
<td>(4.114)</td>
<td>1.187</td>
<td>(4.300)</td>
<td>2093.16</td>
<td>1.500</td>
<td>(0.390)</td>
<td>1820.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>3.083</td>
<td>(1.464)</td>
<td>10.50</td>
<td>(4.43)</td>
<td>6.046</td>
<td>(1.921)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>λ</td>
<td>0.007</td>
<td>(0.007)</td>
<td>0.003</td>
<td>(0.001)</td>
<td>0.007</td>
<td>(0.003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>τ</td>
<td>0.133</td>
<td>(0.102)</td>
<td>0.070</td>
<td>(0.032)</td>
<td>0.114</td>
<td>(0.040)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponentiated Weibull</td>
<td>γ</td>
<td>0.130</td>
<td>(0.022)</td>
<td>0.130</td>
<td>(0.015)</td>
<td>2070.66</td>
<td>0.145</td>
<td>(0.018)</td>
<td>1832.5</td>
<td></td>
</tr>
<tr>
<td>(Beta-Weibull with fixed α)</td>
<td>β</td>
<td>0.012</td>
<td>(0.027)</td>
<td>0.0003</td>
<td>(0.0006)</td>
<td>0.001</td>
<td>(0.002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>α</td>
<td>26.41</td>
<td>(10.95)</td>
<td>71.23</td>
<td>(29.97)</td>
<td>57.67</td>
<td>(24.87)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution</th>
<th>par</th>
<th>specific</th>
<th>mle(se)</th>
<th>other</th>
<th>mle(se)</th>
<th>all events</th>
<th>mle(se)</th>
<th>-2LL</th>
<th>mle(se)</th>
<th>-2LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-parameter Weibull</td>
<td>γ</td>
<td>1.738</td>
<td>(0.445)</td>
<td>1.239</td>
<td>(0.100)</td>
<td>1581.49</td>
<td>1.272</td>
<td>(0.098)</td>
<td>1488.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>89.43</td>
<td>(53.21)</td>
<td>33.80</td>
<td>(4.18)</td>
<td>30.40</td>
<td>(3.35)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-parameter Weibull</td>
<td>γ</td>
<td>1.545</td>
<td>(0.775)</td>
<td>1.295</td>
<td>(0.213)</td>
<td>1581.32</td>
<td>20.23</td>
<td>(20.89)</td>
<td>1488.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>137.61</td>
<td>(245.71)</td>
<td>29.86</td>
<td>(12.47)</td>
<td>28.71</td>
<td>(11.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>λ</td>
<td>-0.043</td>
<td>(0.147)</td>
<td>1.179</td>
<td>(3.947)</td>
<td>2.70</td>
<td>(17.92)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-parameter Weibull</td>
<td>γ</td>
<td>120.70</td>
<td>(206.07)</td>
<td>20.06</td>
<td>(787.74)</td>
<td>1581.18</td>
<td>20.23</td>
<td>(20.89)</td>
<td>1488.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>1.602</td>
<td>(0.776)</td>
<td>1.442</td>
<td>(0.439)</td>
<td>1.440</td>
<td>(0.433)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>λ</td>
<td>2.77</td>
<td>(137.42)</td>
<td>0.011</td>
<td>(0.053)</td>
<td>0.014</td>
<td>(0.077)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>τ</td>
<td>54.95</td>
<td>(2672.23)</td>
<td>0.804</td>
<td>(0.314)</td>
<td>0.830</td>
<td>(0.321)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponentiated Weibull</td>
<td>γ</td>
<td>0.748</td>
<td>(1.668)</td>
<td>0.756</td>
<td>(0.720)</td>
<td>1581.30</td>
<td>0.875</td>
<td>(0.839)</td>
<td>1488.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>64.88</td>
<td>(133.31)</td>
<td>23.00</td>
<td>(26.53)</td>
<td>23.53</td>
<td>(21.92)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>α</td>
<td>2.513</td>
<td>(6.700)</td>
<td>1.833</td>
<td>(2.308)</td>
<td>1.573</td>
<td>(1.930)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

88
Figure 5. Estimated hazard functions, from a non-parametric fit (straight lines) and fitted curves from each of the 4 Weibull models, for local or regional recurrences (left) and other events (right), and the tamoxifen (top) and CMFT (bottom) groups.
Figure 6. Estimated hazard functions for all disease-free events, from a non-parametric fit (straight lines) and fitted curves from each of the four Weibull models, for the tamoxifen (left) and CMFT (right) groups, although the 4-parameter model does it a bit earlier than others.

There did not appear to be a similar single peak for the adverse events observed in the CMFT group. As such, monotonically increasing hazard shapes from the two-parameter Weibull model provided an adequate fit to the observed non-parametric curves. The three and four-parameter models very closely followed the two-parameter fit.

4.2 Cumulative incidence and survival

The cumulative incidence function was estimated using all four Weibull models for each of the two types of competing risk, i.e. local or regional recurrences and local events. The non-parametric cumulative incidence was estimated through Gray’s method (Gray, 1988), while the parametric cumulative incidence functions were obtained through the invariance property using the above maximum likelihood estimates from all four Weibull models. The cumulative incidence function is plotted using the exponentiated Weibull model fit in Figure 7 for each event type. The disease-free survival curves are plotted in Figure 8 using the exponentiated Weibull model. The fits from the model tend to follow the non-parametric cumulative incidence curves closely. There is a significant reduction in local or regional recurrences in the CMFT group, with no apparent difference between the two treatment groups for other types of events (Table 2). A formal Wald test also concluded that there was also a significant improvement in disease-free survival in the CMFT group for all models.
Figure 7. Cumulative incidence function estimates and 95% confidence intervals, between tamoxifen (solid) and CMFT (dashed) groups, for local or regional recurrences (left) and other events (right) estimated using the exponentiated Weibull model along with 95% confidence intervals.

Figure 8. Estimated survival in tamoxifen (solid) and CMFT (dashed) groups, for all disease-free events using the exponentiated Weibull model, along with the corresponding 95% confidence intervals. Non-parametric estimate (non-smooth curve) of survival was computed using the Kaplan-Meier method.
Table 2
Treatment group comparisons of estimates of the cumulative incidence function under competing risks and survival probabilities for all disease-free events

<table>
<thead>
<tr>
<th>Years</th>
<th>Treatment group</th>
<th>Local or regional</th>
<th>All events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cumulative incidence function</td>
<td>Survival function</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F_{ct}$ $\pm$ se($F_{ct}$) p-value</td>
<td>$F_{ct}$ $\pm$ se($F_{ct}$) p-value</td>
<td>$F_{surv}$ $\pm$ se($F_{surv}$) p-value</td>
</tr>
<tr>
<td>2</td>
<td>tamoxifen 0.016 0.004 0.0001 0.041 0.006 0.150 0.943 0.007 0.002</td>
<td>CMFT 0.001 0.001</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>tamoxifen 0.033 0.005 &lt; 0.0001 0.092 0.009 0.065 0.873 0.010 0.0001</td>
<td>CMFT 0.004 0.002</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>tamoxifen 0.048 0.007 &lt; 0.0001 0.137 0.011 0.106 0.814 0.012 0.0001</td>
<td>CMFT 0.009 0.003</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>tamoxifen 0.060 0.008 &lt; 0.0001 0.176 0.013 0.245 0.763 0.014 0.0004</td>
<td>CMFT 0.014 0.004</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>tamoxifen 0.071 0.009 &lt; 0.0001 0.209 0.015 0.544 0.720 0.016 0.004</td>
<td>CMFT 0.019 0.005</td>
<td></td>
</tr>
</tbody>
</table>

4.3 Model adequacy

Model adequacy was investigated using signed deviance residuals with censored observations (Efron, 1988) to assess goodness of fit (Table 3). These are obtained by calculating the expected number of events $e_y$ during each year $y$ as $n_y h_y$, where $n_y$ is the number of patients at risk at the beginning of each year, and $h_y$ is the hazard function integrated over the time interval. The goodness of fit between the observed number of events $S_y$ and expected events $e_y$ for each of the models can be assessed through signed deviance residuals, which follows a $\chi^2$ distribution:

$$R_y = \sqrt{2} \times \text{sign}(S_y - e_y) \left[ S_y / e_y + (N_y - S_y) \log \left( \frac{N_y - S_y}{N_y - e_y} \right) \right]^{1/2}.$$ (21)

The models tended to provide a poorer fit to the data for the endpoint of specific events, in large part due to their relative scarcity; the models provided better fits to the other events endpoint and when all events were grouped together.

4.4 Comparison of standard error of cumulative incidence estimates and disease-free survival between Weibull-based models

The standard errors of the cumulative incidence estimates at 10 months, as obtained by inverting the matrix of minus the second derivatives of the log-likelihood function, are displayed in Table 4. The cumulative incidence function estimates obtained from extensions of the Weibull distribution have a noticeably lower variance than those obtained from the two-parameter Weibull distribution. In particular, the three-parameter Weibull models had comparable variance to the four-parameter models. It can be noted that, despite the large amount of uncertainty when estimating the parameters of the exponentiated Weibull distribution, the standard errors of the cumulative incidence estimates remain small compared to the other extensions of the Weibull model. Conversely, the standard errors of the survival function estimates were similar between all
### Table 3

**Goodness of fit for breast cancer data**

<table>
<thead>
<tr>
<th>Years at risk</th>
<th>tamoxifen group</th>
<th>CMFT group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>local or regional</td>
<td>other events</td>
</tr>
<tr>
<td></td>
<td>obs (N&lt;sub&gt;y&lt;/sub&gt;)</td>
<td>exp (e&lt;sub&gt;y&lt;/sub&gt;)</td>
</tr>
<tr>
<td>[0-1]</td>
<td>770 4 6.6 4.3</td>
<td>14 15.1 9.9 11.5</td>
</tr>
<tr>
<td>(1-2)</td>
<td>752 8 6.3 10.0</td>
<td>7.1 16.8 20.5</td>
</tr>
<tr>
<td>(2-3)</td>
<td>722 12 6.0 9.4</td>
<td>8.4 6.8 19 17.0</td>
</tr>
<tr>
<td>(3-4)</td>
<td>689 8 5.7 7.4</td>
<td>6.7 6.3 17 16.8</td>
</tr>
<tr>
<td>(4-5)</td>
<td>663 1 5.5 5.9</td>
<td>5.6 5.8 14 16.5</td>
</tr>
<tr>
<td>(5-6)</td>
<td>641 7 5.3 4.8</td>
<td>4.8 5.4 16 16.3</td>
</tr>
<tr>
<td>(6-7)</td>
<td>615 2 3.9 4.2</td>
<td>5.0 12 15.9</td>
</tr>
<tr>
<td>(7-8)</td>
<td>597 1 4.9 3.3</td>
<td>3.8 4.6 16 15.7</td>
</tr>
<tr>
<td>(8-9)</td>
<td>571 4 4.6 2.8</td>
<td>3.3 4.3 15 15.2</td>
</tr>
<tr>
<td>(9-10)</td>
<td>539 6 4.4 2.4</td>
<td>3.0 3.9 11 14.5</td>
</tr>
<tr>
<td>Total</td>
<td>6549 53 9.7</td>
<td>156 18.5</td>
</tr>
</tbody>
</table>

\[ \sum R_y^2 \]

<table>
<thead>
<tr>
<th>p</th>
<th>( \chi^2 )</th>
<th>df</th>
<th>p</th>
<th>( \chi^2 )</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.008</td>
<td>0.025</td>
<td>0.016</td>
<td>0.010</td>
<td>0.831</td>
<td>0.617</td>
</tr>
</tbody>
</table>

\[ \sum R_y^2 \]

| df | 8 | 7 | 6 | 7 | 8 | 7 | 6 | 7 | 8 | 7 | 6 | 7 |
Table 4

<table>
<thead>
<tr>
<th></th>
<th>tamoxifen group</th>
<th>2-par</th>
<th>3-par</th>
<th>4-par</th>
<th>exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative incidence (specific)</td>
<td>0.0229</td>
<td>0.0093</td>
<td>0.0093</td>
<td>0.0094</td>
<td></td>
</tr>
<tr>
<td>Cumulative incidence (other)</td>
<td>0.0201</td>
<td>0.0147</td>
<td>0.0147</td>
<td>0.0148</td>
<td></td>
</tr>
<tr>
<td>Survival</td>
<td>0.0164</td>
<td>0.0163</td>
<td>0.0163</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>CMFT group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative incidence (specific)</td>
<td>0.0189</td>
<td>0.0051</td>
<td>0.0051</td>
<td>0.0051</td>
<td></td>
</tr>
<tr>
<td>Cumulative incidence (other)</td>
<td>0.0214</td>
<td>0.0147</td>
<td>0.0147</td>
<td>0.0147</td>
<td></td>
</tr>
<tr>
<td>Survival</td>
<td>0.0152</td>
<td>0.0153</td>
<td>0.0152</td>
<td>0.0152</td>
<td></td>
</tr>
</tbody>
</table>

four types of models; the additional parameters from the three and four-parameter extensions of the Weibull distribution made no contribution to reducing the standard errors relative to the two-parameter model.

5. Conclusion

This paper described approaches based on the Weibull model and several of its generalizations for fitting right-censored data with competing risks. We introduced an extension of the Weibull distribution which incorporates additional shape parameters to the two-parameter Weibull function. This extension can be easily obtained by applying a beta distribution to the two-parameter Weibull model and hence can be implemented easily using standard statistical software packages. The previously introduced exponentiated Weibull is a specific case of the beta-Weibull model.

The two-parameter, exponentiated and generalized models were fit to a previously published breast cancer data set. The basic Weibull model produced simple hazard shapes that did not reflect the true nature of the data when the behavior of the hazard was more complex. On the other hand, the extensions of the Weibull model were able to more accurately represent the behavior of the hazard function of the model. In all cases, when drawing inference on cumulative incidence or survival all the Weibull-based methods led to the same conclusion.

There were some problems with convergence that were encountered when finding the maximum likelihood estimates of the parameters. The maximum likelihood estimates were at times dependent on the initial values as the number of parameters increased. As well, the adjustments of the iterative steps used in the widely used Newton-Raphson procedure were often inflated. This appeared to be due to extremely small estimates of the negative of second derivatives of the log-likelihood, or the Fisher information, which would result in the corresponding variance estimates for the parameters being very high. However, these problems could be rectified by adjusting the scale of these iterative step lengths used in the computational algorithms.
REFERENCES


Optimum Designs for Estimation in Accelerated Life Testing Problems

Manisha Pal and Nripes Kumar Mandal
Department of Statistics, Calcutta University

ABSTRACT

In accelerated life testing experiments, the life of an item is affected by the stress on it, which may be caused by a number of factors. In this paper we use exponential regression model to represent the lifetime of an item as a function of the stress factors and attempt to find the optimum choice of factor levels for inferring about the parameters of the model.

Key words and phrases: Accelerated life testing; Type 1 censored data; Exponential regression; Optimal design.

AMS Subject Classification (2000): 90B25, 62K05.

1. INTRODUCTION

The basic problem in lifetime tests is that of specifying models to represent lifetime distributions and of making statistical inferences on the basis of these models. Several models are used to represent lifetime, like exponential, Weibull, gamma, lognormal, etc. Of these, the exponential model is most easy to handle and extensive statistical methods have been developed for it. The main drawback of the model is that its hazard rate is constant. Nonetheless, it has been found to be very useful in many real-life situations. (See Lawless (1982)).

In accelerated life testing problems, the lifetime is affected by one or more types of stresses, which should be included in the model defining the lifetime. This results in a regression model, where the stress factors are the regressor or concomitant variables. Chernoff (1962) probably is the first to study the problem of choosing the optimum design to estimate the mean of the exponential distribution for a single stress factor. Thereafter many researchers carried out studies on accelerated life test plans. Some recent works in this line are by Yang (1994), Islam and Ahmad (1994), Khamis and Higgins (1996), Khamis (1997), Park and Yum (1998), Ismail (2000), Van Dorp and Mazzuchi (2004), Ahmad et al. (2006). However, in all these studies, only one type of stress has been considered and the stress levels have been assumed to be known.

Lifetime data are generally censored. In Type I censoring, the experiment is run over a fixed period of time in such a way that an item’s lifetime will be known exactly only if it is less than some predetermined value. In this paper, we consider two and three stress factors and attempt to obtain optimum level combination of the factors for estimating the parameters in an exponential regression model when the experiment is Type I censored. In Section 2 we formulate the problem and find the Fisher information matrix for the maximum likelihood estimate of the parameters. The matrix naturally will depend on the unknown parameters. To get rid of this difficulty, we assume a suitable prior distribution for the unknown parameters. In Section 3, we attempt to find the optimum design using D-optimality criterion. (For optimality criteria see e.g. Shah and Sinha (1989) and Pukelsheim (1993)).

2. FORMULATION OF THE PROBLEM

Let \( T \) be the lifetime of an item with fixed censoring time \( L \), and let there be \( k \) stress factors acting on it. We assume that, given the levels \((x_1, x_2, \ldots, x_k) = x\) of the \( k \) factors, the
distribution of $T$ is exponential with density function
\[
f(t|x) = \theta_x^{-1} \exp\left(-\frac{t}{\theta_x}\right), \quad t > 0,
\]
where $\theta_x$ is a function of $x$, $x$ being also known as regression vector.

Various functional forms for $\theta_x$ are possible, but the most useful is
\[
\theta_x = \exp(x'\beta),
\]
where $\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$ is a vector of regression parameters.

The log lifetime $Y = \ln(T)$ then has the p.d.f.
\[
g(y|x) = \exp\{y - x'\beta\} - \exp\{y - x'\beta\}, \quad -\infty < y < \infty.
\]

Suppose $n$ ($2k$ ) test items are taken at random. Associated with the $i$th item, let $L_i$ be the censoring time and $x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijk})'$ be the regression vector, $i = 1, 2, \ldots, n$. Without loss of generality, we may take $x_o \in [-L_1]$, $i = 1, 2, \ldots, n$. Without loss of generality, we may take $x_o \in [-L_1]$, $i = 1, 2, \ldots, n$. Without loss of generality, we may take $x_o \in [-L_1]$, $i = 1, 2, \ldots, n$.

The likelihood function for the censored sample is given by
\[
L(\beta) = \prod_{i \in C} \exp\{y_i - x_{ij}', \beta\} - \exp\{y_i - x_{ij}', \beta\}\prod_{i \in D} \exp\{-\exp\{y_i - x_{ij}', \beta\}\}
\]
where $D$ = set of $i$ values for which $T_i < L_i$ and $C$ = set of $i$ values for which $T_i \geq L_i$.

In situations where censoring time is known for each individual, the expected information matrix of the maximum likelihood estimate $\hat{\beta}$ of $\beta$ is given by
\[
I = ((I_{rs}),)
\]
where
\[
I_{rs} = E\left[-\frac{\partial^2 \ln L}{\partial \beta_r \partial \beta_s}\right] = \sum_{i \in D} x_{ij} x_{ij}' \left[-\exp\{-L_i \exp\{y_i - x_{ij}', \beta\}\}\right]
\]

For the uncensored case, the expected information matrix has the entries
\[
I_{rs} = \sum_{i \in C} x_{ij} x_{jk}, \quad r, s = 1, 2, \ldots, k.
\]

The problem of designing is that of selecting $x_{ij}$’s so as to optimize some real-valued function of $I$. However, difficulty arises owing to the fact that $I_{rs}$, given by (2.5), depends on the unknown parameter vector $\beta$ for all $r$ and $s$. To resolve this, we assume a prior distribution of $\beta$. Then, the optimum design may be obtained by optimizing a suitable function of the expected information matrix, i.e. $E(I)$.

Let the prior distribution of $\beta$ be such that $U = x'\beta$ has an extreme value distribution with location parameter $\eta(x)$ and scale parameter 1. Then, $\exp(-U)$ has an exponential distribution with mean, say, $\xi(x)$, so that
\[
E[1 - \exp\{-L_i \exp(-U)\}] = \frac{\xi(x_{ij})}{1 + \xi(x_{ij})L_i} = w(x_{ij}), \quad \text{say.}
\]

Hence,
\[
E(I_{rs}) = \sum_{i=1}^{n} w(x_{ij}) x_{ij} x_{jk}.
\]

We, therefore, get
\[
E(I) = XD^\prime X,
\]
where
\[
X = \begin{pmatrix} x_{i1}, x_{i2}, \ldots, x_{in} \end{pmatrix}'
\]
\[
D^\prime = \text{Diag}\{w(x_{i1}), w(x_{i2}), \ldots, w(x_{in})\}.
\]

3. OPTIMUM DESIGN

Suppose there are $k$ stresses acting on an item.

We shall use D-optimality to find the optimal design. For this, the criterion function is
\[
\psi(E(I)) = |E(I)|
\]
and the optimum design is obtained so as to maximize $\psi(E(I))$.

For simplicity sake, we shall assume that $n = k$.

Then,
\[
\psi(E(I)) = |X|^2 \prod_{i=1}^{n} w(x_{ij}) = \left|X^2 \left(1 + \frac{1}{\xi(x_{ij})L_i}\right) \right|
\]

3.1 Suppose $k = 2$. 

97
Then,

$$X = \begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{pmatrix},$$

and 

$$\psi(E(I)) = |X^TDX| = \frac{(x_{11}x_{22} - x_{12}x_{21})^2}{(1 + \frac{1}{\xi(x_i)L_x}) (1 + \frac{1}{\xi(x_i)L_y})}\quad \text{for}\quad i = 1, 2.$$ 

(3.2)

For general form of $\xi$, it is difficult to find the optimal design. However if the function $\xi(x_i)$ possesses some properties, it is possible to find the optimum support points. Towards this, we have the following result:

**Theorem 3.1.1** If $\xi^{-1}(x_{0})$’s are concave and increasing in $x_{(i)}$’s, then (3.2) is maximized at $x_{0} = \pm 1$.

**Proof.** It is clearly seen that the numerator of (3.2) is maximized at $x_{ij} = \pm 1$ with 4 as the maximum value. Again, since $\xi^{-1}(x_{i})$ is increasing in $x_{ij}$’s, the minimum of the denominator is achieved when $x_{ij} = -1$. But it is not possible that all $x_{ij} = -1$ since then, the numerator becomes zero.

Let us assume that $L_{1} \leq L_{2}$. Then the denominator will be lower for $x_{11} = x_{12} = -1$ than for $x_{21} = x_{22} = -1$. In fact, the first term in the denominator is lowest for $x_{11} = x_{12} = -1$. Again, in the second term of the denominator at least one of $x_{21}$ and $x_{22}$ must be -1 for utmost reduction in its value. Let us fix $x_{21} = -1$ and put $x_{22} = a$, where $a \in [-1, 1]$. Then (3.2) takes the form

$$\psi(a) = \frac{(a+1)^2}{D(a)} = (a+1)^2, \quad \text{say,}$$

Hence,

$$\frac{\partial \psi(a)}{\partial a} = \frac{2(a+1)D(a) - (a+1)^2 d_{i}\xi^{-i}/L_{i}}{D'(a)} \geq 0,$$

where

$$d_{i} = \frac{\partial \xi^{-i}(-1,-1)}{\partial a}, \quad \text{and}$$

$$\xi^{-i} = \frac{\partial^2 \xi^{-i}(-1,a)}{\partial a^2} \leq 0$$

as $\xi^{-i}$ is a concave function.

This shows that $\psi$ is a convex function of $a$, and hence is maximized either at $a = -1$ or at $a = 1$. But at $a = -1$, value of $\psi$ is zero. Hence, $a = 1$ and for this value of $a$ the numerator attains its maximum. Hence the theorem.

Thus, from the theorem we have that if $\xi^{-1}(x_{i})$ is concave and increasing in $x_{(i)}$, then the optimal design points are at $x_{(1)} = (-1,-1)^{'}$ and $x_{(2)} = (-1, 1)^{'}$, provided $L_{1} \leq L_{2}$. It automatically follows that for $L_{2} \leq L_{1}$, $x_{(1)} = (-1, 1)^{'}$ and $x_{(2)} = (-1,-1)^{'}$ are the optimal design points when $\xi^{-1}(x_{(i)})$ is a concave, increasing function of $x_{(i)}$.

**Theorem 3.1.2** Maximum $\psi$ is an increasing function of $L_{2}$ for given $L_{1}$ and an increasing function of $L_{1}$ for given $L_{2}$.

**Proof.** The proof is obvious.

We now consider some specific forms of $\xi(x_{(i)})$’s, with $\xi^{-1}(x_{(i)})$ concave and increasing in $x_{(i)}$.

**Case 1:** $\xi(x_{(i)}) = \frac{1}{c + x_{1i} + x_{2i}}, c > 2, i = 1, 2$.

Then,

$$\psi(E(I)) = \frac{(x_{11}x_{22} - x_{12}x_{21})^2}{\left(1 + \frac{c + x_{11} + x_{21}}{L_{1}}\right) \left(1 + \frac{c + x_{12} + x_{22}}{L_{2}}\right)}.$$
Table 3.1.1: Showing maximum $\psi$ values for different $(L_1, L_2)$ when $L_1 \leq L_2$ and $c = 3$.

<table>
<thead>
<tr>
<th>$L_2$</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>3.00</th>
<th>5.00</th>
<th>7.00</th>
<th>9.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.062</td>
<td>0.114</td>
<td>0.160</td>
<td>0.200</td>
<td>0.500</td>
<td>0.560</td>
<td>0.600</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.190</td>
<td>0.267</td>
<td>0.333</td>
<td>0.667</td>
<td>0.833</td>
<td>0.933</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0.343</td>
<td>0.428</td>
<td>0.857</td>
<td>1.071</td>
<td>1.200</td>
<td>1.400</td>
<td>1.500</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.500</td>
<td>1.000</td>
<td>1.250</td>
<td>1.400</td>
<td>1.500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_1$</td>
<td>3.00</td>
<td>1.500</td>
<td>1.875</td>
<td>2.100</td>
<td>2.250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td></td>
<td>2.083</td>
<td>2.333</td>
<td>2.500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td></td>
<td>2.450</td>
<td>2.625</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td></td>
<td>2.700</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Case 2: $\xi(x_{(i)}) = \exp[\exp\{-(x_{i1} + x_{i2})\}]$, $i = 1, 2$.

Here

$$\psi(E(I)) = \frac{(x_{i1}x_{i2} - x_{i2}x_{i1})^2}{1 + \exp[-\exp\{-x_{i1} + x_{i2}\}]} \frac{1 + \exp[-\exp\{-x_{i1} + x_{i2}\}]}{L_1} \frac{1 + \exp[-\exp\{-x_{i1} + x_{i2}\}]}{L_2}.$$ 

Table 3.1.2: Showing maximum $\psi$ values for different combinations of $(L_1, L_2)$ when $L_1 \leq L_2$.

<table>
<thead>
<tr>
<th>$L_2$</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>3.00</th>
<th>5.00</th>
<th>7.00</th>
<th>9.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>1.614</td>
<td>2.299</td>
<td>2.677</td>
<td>2.917</td>
<td>3.554</td>
<td>3.717</td>
<td>3.791</td>
<td>3.833</td>
</tr>
<tr>
<td>0.50</td>
<td>2.301</td>
<td>2.681</td>
<td>2.921</td>
<td>3.559</td>
<td>3.721</td>
<td>3.795</td>
<td>3.838</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>2.681</td>
<td>2.922</td>
<td>3.560</td>
<td>3.723</td>
<td>3.797</td>
<td>3.840</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>2.922</td>
<td>3.561</td>
<td>3.724</td>
<td>3.798</td>
<td>3.841</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_1$</td>
<td>3.00</td>
<td>3.562</td>
<td>3.725</td>
<td>3.799</td>
<td>3.842</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td></td>
<td>3.725</td>
<td>3.800</td>
<td>3.842</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td></td>
<td>3.800</td>
<td>3.843</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td></td>
<td>3.843</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The condition of concavity of $\xi^{-1}(x_{(i)})$ is really a sufficient condition for Theorem 3.1.1 to hold. We give below two examples where $\xi^{-1}(x_{(i)})$ is a convex increasing function of $x_{(i)}$ but the theorem holds.

Case 3: Let $\xi(x_{(i)}) = \exp[-(x_{i1} + x_{i2})]$, $i = 1, 2$.

Then,

$$\psi(E(I)) = \frac{(x_{i1}x_{i2} - x_{i2}x_{i1})^2}{1 + \frac{\exp(x_{i1} + x_{i2})}{L_1} \frac{\exp(x_{i1} + x_{i2})}{L_2}}.$$ 

Arguing in the same way as in the proof of Theorem 3.1.1, we have that in the optimal situation $x_{i1} = x_{i2} = x_{21} = -1$. Then, $\psi(E(I))$ will be an increasing function of $x_{22}$, and hence we must have $x_{22} = 1$. 

99
Table 3.1.1: Showing maximum $\psi$ values for various combinations of $(L_1, L_2)$ when $L_1 \leq L_2$.

<table>
<thead>
<tr>
<th>$L_2$</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>3.00</th>
<th>5.00</th>
<th>7.00</th>
<th>9.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.519</td>
<td>0.865</td>
<td>1.112</td>
<td>1.297</td>
<td>1.946</td>
<td>2.163</td>
<td>2.271</td>
<td>2.335</td>
</tr>
<tr>
<td>0.50</td>
<td>1.049</td>
<td>1.349</td>
<td>1.574</td>
<td>2.361</td>
<td>2.623</td>
<td>2.754</td>
<td>2.833</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>1.452</td>
<td>1.694</td>
<td>2.541</td>
<td>2.824</td>
<td>2.965</td>
<td>3.050</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>1.762</td>
<td>2.642</td>
<td>2.936</td>
<td>3.083</td>
<td>3.171</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td></td>
<td></td>
<td></td>
<td>2.870</td>
<td>3.189</td>
<td>3.349</td>
<td>3.445</td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td></td>
<td></td>
<td></td>
<td>3.245</td>
<td>3.408</td>
<td>3.505</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td></td>
<td></td>
<td></td>
<td>3.434</td>
<td>3.532</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td></td>
<td></td>
<td></td>
<td>3.547</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Case 4: 
$$\xi(x_{(i)}) = \exp[-(x_{1i} + x_{2i})]/c + x_{1i} + x_{2i}, \quad c > 2, \; i = 1,2.$$ Then,
$$\psi(E(I)) = \frac{\left((x_{1i}x_{2i} - x_{1i}x_{2i})\right)^2}{L_1} \left(1 + \frac{c + x_{1i} + x_{2i}}{L_1}\right)^\frac{1}{2} \left(1 + \frac{c + x_{1i} + x_{2i}}{L_2}\right)^\frac{1}{2}$$

Here again, in the optimal situation $x_{1i} = x_{12} = x_{2i} = -1$. Then, $\psi(E(I))$ will be an increasing function of $x_{22}$, and hence we must have $x_{22} = 1$.

Table 3.1.4: Showing the optimal values of $\psi$ for different combinations of $(L_1, L_2)$ when $L_1 \leq L_2$ and $c = 3$.

<table>
<thead>
<tr>
<th>$L_2$</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
<th>3.00</th>
<th>5.00</th>
<th>7.00</th>
<th>9.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.199</td>
<td>0.371</td>
<td>0.519</td>
<td>0.649</td>
<td>1.298</td>
<td>1.622</td>
<td>1.817</td>
<td>1.946</td>
</tr>
<tr>
<td>0.50</td>
<td>0.450</td>
<td>0.630</td>
<td>0.787</td>
<td>1.574</td>
<td>1.967</td>
<td>2.204</td>
<td>2.361</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0.678</td>
<td>0.847</td>
<td>1.694</td>
<td>2.118</td>
<td>2.372</td>
<td>2.541</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.881</td>
<td>1.762</td>
<td>2.202</td>
<td>2.466</td>
<td>2.642</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td></td>
<td></td>
<td></td>
<td>1.914</td>
<td>2.392</td>
<td>2.679</td>
<td>2.870</td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td></td>
<td></td>
<td></td>
<td>2.434</td>
<td>2.726</td>
<td>2.921</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.00</td>
<td></td>
<td></td>
<td></td>
<td>2.747</td>
<td>2.943</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td></td>
<td></td>
<td></td>
<td>2.956</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2: Consider $k=3$. Let us assume that and $\psi(E(I)) = |X'DX|$ = $L_i = L$, for $i = 1,2,3$.

Here,
$$X = \begin{pmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \\ x_{13} & x_{23} & x_{33} \end{pmatrix},$$
$$\frac{|X|}{\left(1 + \frac{1}{L(x_{11})L_1}\right)^\frac{1}{2}} \left(1 + \frac{1}{L(x_{21})L_2}\right)^\frac{1}{2} \left(1 + \frac{1}{L(x_{31})L_3}\right)^\frac{1}{2}$$

(3.3)
Theorem 3.2.1 If $\xi^{-1}(x_{(i)})$'s are concave and increasing in $x_{(i)}$'s, then (3.3) is maximized at $x_{ij} = \pm 1$.

Proof. The numerator of (3.3) is maximized at $x_{ij} = \pm 1$ with $4^2$ as the maximum value (cf. Wojtas, 1964). Now, since $\xi^{-1}(x_{(i)})$ is increasing in $x_{ij}$'s, the minimum of the denominator is achieved when $x_{ij} = -1$ for all $i, j$. But this is not possible since then, the numerator becomes zero.

Since the numerator in (3.3) is invariant with respect to the multiplication of any row or column by -1, without loss of generality we may take $X = \begin{pmatrix} -1 & -1 & -1 \\ -1 & a_2 & a_3 \\ -1 & b_2 & b_3 \end{pmatrix}$ where $a_2, a_3, b_2, b_3 \in [-1,1]$.

Then (3.3) takes the form

$$\psi(a_2, a_3, b_2, b_3) = \frac{1}{L} \left( 1 + \frac{\xi^{-1}(-1-1)}{L} \right) \left( 1 + \frac{\xi^{-1}(-1,1)}{L} \right) \left( 1 + \frac{\xi^{-1}(-1,1)}{L} \right)$$

It can be easily checked that given any three of $a_2, a_3, b_2, b_3$, $\psi$ is a convex function of the remaining variable. Hence, $\psi$ must be maximized when each of $a_i, b_j, i = 1,2$ is $\pm 1$. Hence the theorem.

Now, since $X$ is non-singular and its first row is (-1, -1, -1), one of $a_2, a_3$ must be -1 and the other +1, and similarly for $b_2, b_3$. Hence, we get the following theorem:

Theorem 3.2.2 If $\xi^{-1}(x_{(i)})$ be concave and increasing in $x_{(i)}$, then the optimal design which maximizes $\psi$ has the support points at $x_{(i)} = (-1,-1,-1)^T, x_{(i)} = (-1,1,1)^T$ and $x_{(3)} = (-1,1,1)^T$.

4. CONCLUSIONS

In this investigation, we have considered the problem of determining the optimal levels of stress factors in accelerated life testing experiment with type I censoring for the purpose of estimating the parameters involved in the mean lifetime. The information matrix for the maximum likelihood estimate naturally is a function of the unknown parameters, as also of the stress levels. Assuming a suitable prior for the unknown parameters, and using the D-optimality criterion, we have found the optimum stress levels when there are two and three stress factors. It has been shown that when $\xi^{-1}(x_{(i)})$ is a concave, increasing function of the stress levels $x_{ij}$, the support points of the optimum design are at the extreme points of the factor space.

It is conjectured that for general $k$ stress factors when $\xi^{-1}(x_{(i)})$ has the property as stated earlier, the support points of the optimum design will also be at the extreme points of the factor space.

REFERENCES


Some Variations of Blackwell Martingale Inequality

Rasul A. Khan
Cleveland State University

Abstract

A certain martingale inequality due to Blackwell is revisited and certain exponential variations of this inequality are presented. The presented variations embody the exponential versions of Dubins-Savage inequality for Blackwell’s inequality.

1. Introduction

Let \( S_n = \sum^n_i X_i, \ S_0 = 0, \mathcal{F}_n, \ n \geq 0 \) be a martingale such that \( |X_n| \leq 1 \). Blackwell (1997) has shown that for any positive constants \( a \) and \( b \),

\[
P(S_n \geq a + b n \text{ for some } n \geq 1) \leq \exp(-2ab),
\]

and for any positive \( b (0 < b < 1) \),

\[
P(S_n \geq b n \text{ for some } n \geq m) \leq \rho^m \leq \exp(-mb^2/2),
\]

where \( \rho = 1/[ (1 + b)^{1/b} (1 - b)^{1-1/b} ]^{1/b} \).

Ross (1999) obtained extensions of (1) and (2) for martingales \( S_n = \sum^n_i X_i \) when \(-\alpha \leq X_n \leq \beta \) \((\alpha, \beta > 0)\). Under suitable conditions, these inequalities were recently generalized by Khan (2007) without the boundedness condition. In particular, (1) and the second part of (2) hold, for example, when \( X_1, X_2, ... \) are iid sub-normal random variables. Here we revisit (1) and (2) and present some other variations of these exponential inequalities. Some of these variations include the exponential form of the Dubins-Savage inequality, namely, how do the bounds in (1) and (2) change if we replace \( n \) by \( \sum^n_i V_i \), where \( V_i = E(X_i^2 | \mathcal{F}_{i-1}) \) \((i \geq 1)\) is the conditional variance and \( V_1 \) is constant. Similar exponential versions were explored by Khan and Tomkins (2000). This note considers such variations of (1) and (2) with exponential bounds including the case when \( V_i \) is replaced by \( X_i^2 \). To present the intended variations, we give some preliminaries in the next section, and the main results are given in Section 3.

2. Preliminaries.

The basic tool for proving the presented variations of the preceding inequalities is the following well known supermartingale inequality.

Lemma 1. Let \( \{Z_n, \mathcal{F}_n, \ n \geq 0\} \) be a positive supermartingale. Then for \( m \geq 1 \),

\[
P(\max_{n \geq m} Z_n \geq 1) \leq EZ_m \leq EZ_1 \leq EZ_0.
\]

Proof. Let \( t = \inf\{n \geq m : Z_n \geq 1\} \) and \( t(k) = \min(t, k), \ k \geq m \geq 1 \). Then \( \{Z_{t(k)}, \mathcal{F}_k\} \) is a supermartingale by optional stopping theorem and (3) follows from

\[
P(t \leq k) \leq EZ_{t(k)} \leq EZ_m \leq EZ_1 \leq EZ_0
\]

by letting \( k \to \infty \).

Our intended variations result from the next two elementary lemmas.
Lemma 2. Let $X$ be a random variable such that $|X| \leq 1$, and $EX = 0$, $V = EX^2$. Let $\psi(\theta) = e^{\theta} - \theta - 1$. Then for $\theta > 0$,

(i) $\phi(\theta) = Ee^{\theta X} \leq \cosh \theta \leq \exp(\theta^2/2)$.

(ii) $\phi(\theta) \leq \exp(V \psi(\theta))$.

(iii) $\psi(\theta) \leq h(\theta) = \frac{\theta^2}{2(1-\theta)}$, $0 < \theta < 1$.

Proof. (i) is well known and follows from the convexity of $e^{\theta x}$ ($\leq l(x)$, a line between $x = \pm 1$). (ii) is also known (cf. Freeman (1974), and Khan (2007)). Similarly, (iii) follows from the expansion of $\psi(\theta)$ and the bounding of expanded series by a geometric series for $0 < \theta < 1$.

Lemma 3. Assume the conditions and notations of Lemma 2. In addition we assume that the distribution of $X$ is symmetric about 0. Then for $\theta > 0$,

(i) $\phi(\theta) \leq \exp(V \cosh \theta - 1)$.

(ii) $E\exp(\theta X - \frac{1}{2}\theta^2 X^2) \leq 1$,

where $X$ is not necessarily bounded for the second part.

Proof. To prove (i) expand $e^{\theta X}$ as infinite series and note that $X^{2m} \leq X^2$ for each $m = 1, 2, \ldots$, and the result follows by taking expectation by invoking the symmetry of the distribution. To see (ii) we note that the symmetry of the distribution implies that

$$E\exp(\theta X - \frac{1}{2}\theta^2 X^2) = \frac{1}{2} \left[ (\exp(\theta X) + \exp(-\theta X)) \exp(-\frac{1}{2}\theta^2 X^2) \right]$$

where the last inequality uses the fact that $\cosh(u) \leq \exp(\frac{1}{2}u^2)$.

The next lemma describes couple of supermartingales needed in the sequel.

Lemma 4. Let $\{S_n = \sum_{i=1}^{n} X_i, \mathcal{F}_n\}$ be a martingale. Let $\phi_n(\theta) = E(e^{\theta X_n} | \mathcal{F}_{n-1})$ be finite for $\theta > 0$, and let $V_n$ be the conditional variance of $X_n$ given $\mathcal{F}_{n-1}$. If $\phi_n(\theta) \leq \exp(V_n f(\theta))$ for some suitable $f$, then $Z_n(\theta) = \exp(\theta S_n - a\theta - b\theta \sum_{i=1}^{n} V_i)$ is a supermartingale provided that $f(\theta) \leq b\theta$.

Proof. Since $E(Z_n | \mathcal{F}_{n-1}) = Z_{n-1} \exp(-bV_n) \phi_n(\theta)$, the lemma follows from the required conditions on $\phi_n(\theta)$ and $f(\theta)$.

3. The Main results.

Our main results show the changes resulting from Lemmas 2 and 3 in Blackwell’s inequalities. For simplicity of notation we define

$$P(m) = P(S_n \geq a + b \sum_{i=1}^{n} V_i \text{ for some } n \geq m), a \geq 0, b > 0.$$  (4)

Theorem 1. Let $\{S_n = \sum_{i=1}^{n} X_i, \mathcal{F}_n, n \geq 0\}$ be a martingale such that $|X_n| \leq 1$, and let $V_n$ be the conditional variance of $X_n$. Then assuming $V_n$’s to be constants, and letting $T_m = \sum_{i=1}^{m} V_i$, we have

$$P(m) \leq \frac{1}{(1+b)^a} \exp(-T_m((b+1) \ln(b+1) - b)) \leq \frac{1}{(1+b)^a} \exp\left(-\frac{b^2 T_m}{2(b+1)}\right).$$

Moreover,

$$P(m) \leq \exp(-T_m(a^2/(b(1 + \sqrt{1 + \frac{2a}{b}})) + a)).$$

Proof. Let $Z_n(\theta) = \exp(\theta S_n - a\theta - b\theta \sum_{i=1}^{n} V_i), \theta > 0$. Using Lemma 2 (ii) we have

$$E(Z_n(\theta) | \mathcal{F}_{n-1}) = Z_{n-1}(\theta) \exp(-bV_n \phi_n(\theta)) \leq Z_{n-1}(\theta) \exp(-V_n(b \theta - \psi(\theta))) \text{ provided } g(\theta) = \frac{1}{a} \phi_n(\theta) \leq Z_{n-1}(\theta),$$

$\theta - \psi(\theta) \geq 0$. Let $\theta_0 = \ln(b+1)$, which maximizes $g(\theta)$. Thus $Z_n(\theta_0)$ is a positive supermartingale.

104
and hence (3) gives
\[ P(S_n \geq a + b \sum_{i=1}^{n} V_i \text{ for some } n \geq m) \leq EZ_m(\theta_0). \]

Clearly, for \( \theta = \theta_0 \),
\[ EZ_m(\theta) \leq \exp(-V_m(b\theta - \psi(\theta)))EZ_{m-1}(\theta) \leq \ldots \leq \exp(-a\theta - T_m(b\theta - \psi(\theta))), \]
and the first inequality follows by setting \( \theta_0 = \ln(b+1) \), and the second part of the inequality follows by using \((b + 1)\ln(b + 1) - b > b^2/2(b + 1)\). The last inequality follows by using \( h(\theta) \) (instead of \( \psi(\theta) \)) in Lemma 2 (iii), and maximizing \( b\theta - h(\theta) \).

Our next result uses Lemma 3.

**Theorem 2.** Assume the conditions of Theorem 1, and also assume that the conditional distribution of \( X_n \) given \( \mathcal{F}_{n-1} \) is symmetric about 0. Then letting \( P(1) = P(S_n \geq a + b \sum_{i=1}^{n} V_i \text{ for some } n \geq 1) \) we have
\[ P(1) \leq \frac{1}{(b + \sqrt{1 + b^2})^a} \exp(-b\ln(b + \sqrt{1 + b^2}) - b + 1)), \]
and
\[ P(1) \leq \exp(-ab) \exp(-V_1(b^2 - \cosh b + 1)) \quad 0 < b < 1. \]
Moreover,
\[ P(S_n \geq a + b \sum_{i=1}^{n} X_i^2 \text{ for some } n \geq 1) \leq \exp(-2ab). \]

**Proof.** Let \( \psi(\theta) = \cosh \theta - b\theta - 1 \), and let \( Z_n(\theta) = \exp(\theta S_n - a\theta - b\theta \sum_{i=1}^{n} V_i) \). Then in view of Lemma 3 (i), it follows from Lemma 4 that \( Z_n(\theta) \) is a supermartingale provided \( \psi(\theta) \leq 0 \). Let \( g(\theta) = b\theta - \cosh \theta + 1 \). A \( \theta_0 \) maximizing \( g(\theta) \) is \( \ln(b + \sqrt{1 + b^2}) \), hence the supermartingale inequality (3) gives
\[ P(1) \leq \exp(-a\theta_0) \exp(\theta_0 X_1 - b V_1 \theta_0). \]
Since by Lemma 3 (i),
\[ \exp(\theta_0 X_1) \leq \exp(V_1(\cosh \theta_0 - 1)), \]
hence it follows that
\[ P(1) \leq \exp(-a\theta_0 - V_1(b\theta_0 - \cosh \theta_0 + 1)), \]
and substituting the value of \( \theta_0 \) gives the desired inequality (5). The second part of (5) follows from the fact if \( b \leq 1 \), then \( \cosh b - 1 \leq b^2 \) so that one can use \( \theta_0 = b \) to satisfy \( \psi(b) \leq 0 \). Hence the second part of the inequality follows from the supermartingale inequality with \( Z_n(\theta) \) as the supermartingale. To prove the last inequality (6) we define
\[ Z_n(\theta) = \exp(\theta S_n - a\theta - \frac{1}{2}b^2 \sum_{i=1}^{n} X_i^2). \]
Then using the conditional symmetry and Lemma 3 (ii), it follows that \( Z_n(\theta) \) is a supermartingale. In particular, \( Z_n(2b) \) is a supermartingale. Hence (5) follows from the supermartingale inequality (3).

**References**
A Note on the Modified Box-Cox Transformation

Mezbahur Rahman and Larry M. Pearson
mezbahur.rahman@mnsu.edu and larry.pearson@mnsu.edu
Minnesota State University, Mankato, MN 56001, USA

Abstract

The Box-Cox transformation is a well known family of power transformations that brings a set of
data into agreement with the normality assumption of the residuals and hence the response variable
of a postulated model. This paper compares the original Box-Cox transformation with a modified
version. The comparison is made using the Shapiro-Wilk W statistic. Estimation of the parameter
also compared between the maximum likelihood method and the maximization of the Shapiro-Wilk
W statistic method.

Key Words: Expected values and covariances for the ordered standard normal variates; Goodness-
of-fit for a normal distribution.

1. Introduction

In regression analysis, often the key assumption regarding normality of the error variable and
hence the response variable are violated. The commonly used remedy is the Box-Cox family of power
transformations (Box and Cox (1964)). The process is to select a parameter in the Box-Cox transform-
which maximizes the normal likelihood using the data at hand and then apply regression
analysis on the transformed response variable. There is no role of the estimates of the location
and the scale parameters which were derived in the process of estimating the power transformation
parameter in regression analysis. In practice, the regression model parameters are usually estimated
separately after the necessary Box-Cox power transformation parameter is selected.

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size $n$ from a population whose functional form is
unknown. Box and Cox (1964) suggested that if the transformation

$$X = \begin{cases} \frac{Y^{\lambda}-1}{\ln(Y)}, & \lambda \neq 0 \\ \lambda = 0 & \end{cases}$$

is performed on the data then $X$ will have an approximate normal distribution with mean $\mu$ and
variance $\sigma^2$. In equation (1), $\lambda$ is unknown and considered as the Box-Cox power transforma-
tion parameter and ‘ln’ represents the natural logarithm.

Yeo and Johnson (2000) suggested a slight modification of the Box-Cox transformation as

$$X = \begin{cases} \frac{(Y+1)^{\eta}-1}{\eta}, & \eta \geq 0, \eta \neq 0, \\
\ln(Y+1) & \eta \geq 0, \eta = 0, \\
\frac{-(Y+1)^{2-\eta}-1}{2-\eta} & \eta < 0, \eta \neq 2, \\
-\ln(-Y+1) & \eta < 0, \eta = 2. \end{cases}$$

which is well defined on the whole real line and is appropriate for reducing skewness and to approx-
itmate normality.
2. Parameter Estimation

In literature, the estimation procedures of the Box-Cox power transformation parameter are considered by many authors. The notable ones are the normal likelihood method of Box and Cox (1964), the robustified version of the normal likelihood method of Carroll (1980) and of Bickel and Doksum (1981), the transformation to symmetry method of Hinkley (1975), the quick estimate of Hinkley (1977) and of Taylor (1985). Lin and Vonesh (1989) constructed a nonlinear regression model which is used to estimate the transformation parameter such that the normal probability plot of the data on the transformed scale is as close to linearity as possible. Following the footsteps of Box and Cox (1982) and Lin and Vonesh (1989), Halawa (1996) considered the power transformation parameter estimation procedure using an artificial regression model which gives the estimates with very small variabilities compared to the normal likelihood procedure. Halawa (1996) conducted an exhaustive comparative study with normal likelihood procedure. In that study, he also considered estimation procedures of the location and the scale parameters in the likelihood.

Rahman (1999) introduced a method of estimating the Box-Cox power transformation parameter using maximization of the Shapiro-Wilk W (Shapiro and Wilk (1965)) statistic along with a comparison study of the normal likelihood method (Carroll (1980)), and of the artificial regression model method (Halawa (1996)). In Section 3, the estimation procedure for the power transformation parameters in (1) and (2) are considered using maximization of the normal likelihood along with the Newton-Raphson root finding method. In Section 4, an introduction to the Shapiro-Wilk W statistic is presented which will be used as indicator of closeness to normality and to compare the maximum likelihood estimates to the maximization of the Shapiro-Wilk W statistic estimates.

3.1. Maximum Likelihood Estimate for \( \lambda \)

For a fixed \( \lambda \) in (1), the log-likelihood function \( \ell_\lambda = \ell_\lambda(\mu, \sigma^2; y_1, y_2, \ldots, y_n) \), can be written as

\[
\ell_\lambda = \begin{cases} 
-\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left( \frac{y_i^\lambda - 1}{\lambda} - \mu \right)^2 + (\lambda - 1) \sum_{i=1}^{n} \ln(y_i), & \lambda \neq 0, \\
-\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\ln(y_i) - \mu)^2 - \sum_{i=1}^{n} \ln(y_i), & \lambda = 0.
\end{cases}
\]

Equation (3) is maximized when the partial derivatives of (3) with respect to \( \mu \) and \( \sigma^2 \) are equated to zero and the solution to the corresponding system is found. This leads to solving the following two equations,

\[
\frac{\partial \ell_\lambda}{\partial \mu} \bigg|_{\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2} = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i^\lambda - 1}{\lambda} \right)
\]

and

\[
\frac{\partial \ell_\lambda}{\partial \sigma^2} \bigg|_{\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i^\lambda - 1}{\lambda} \right)^2 - \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i^\lambda - 1}{\lambda} \right).
\]

Then the pseudo log-likelihood \( \ell_\lambda^* = \ell_\lambda(\lambda; y_1, y_2, \ldots, y_n) \) can be written as

\[
\ell_\lambda^* = -\frac{n}{2} \left[ \ln(2\pi) + \ln \left( \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i^\lambda - 1}{\lambda} \right)^2 - \frac{1}{n} \sum_{j=1}^{n} \left( \frac{y_j^\lambda - 1}{\lambda} \right) \right) \right] + 1
\]
\[ + (\lambda - 1) \sum_{i=1}^{n} \ln y_i. \]  

The steps for maximizing \( \ell^*_\lambda \) using the Newton-Raphson method are as follows:

\[
\frac{\partial \ell^*_\lambda}{\partial \lambda} \bigg|_{\lambda = \hat{\lambda}} = 0 \implies h(\hat{\lambda}) = -n \frac{\sum_{i=1}^{n} (P_i - \bar{P}) (Q_i - \bar{Q})}{\sum_{i=1}^{n} (P_i - \bar{P})^2} + \sum_{i=1}^{n} \ln y_i = 0
\]  

and

\[ \hat{\lambda}^{(t+1)} = \hat{\lambda}^{(t)} - \frac{h(\hat{\lambda}^{(t)})}{h'(\hat{\lambda}^{(t)})}, \]

where

\[ h'(\hat{\lambda}) = -n \left[ \frac{\sum_{i=1}^{n} (P_i - \bar{P}) (R_i - \bar{R}) + \sum_{i=1}^{n} (Q_i - \bar{Q})^2}{\sum_{i=1}^{n} (P_i - \bar{P})^2} \right]. \]

\[ P_i = \left( \frac{y_i \lambda + 1}{\lambda^2} \right), \quad \bar{P} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i \lambda + 1}{\lambda^2} \right), \quad Q_i = \left( \frac{\lambda (\ln y_i) y_i \lambda - y_i + 1}{\lambda^2} \right), \]

\[ \bar{Q} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\lambda (\ln y_i) y_i \lambda - y_i + 1}{\lambda^2} \right), \quad R_i = \left( \frac{\lambda^2 (\ln y_i) y_i \lambda - 2\lambda^2 (\ln y_i) y_i \lambda + 2\lambda y_i - 2\lambda}{\lambda^2} \right), \]

and

\[ \bar{R} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\lambda^2 (\ln y_i) y_i \lambda - 2\lambda^2 (\ln y_i) y_i \lambda + 2\lambda y_i - 2\lambda}{\lambda^2} \right). \]

The computation algorithm starts with an initial value of \( \hat{\lambda} \) (an obvious choice is 1) iteratively using (8) and then \( \hat{\mu} \) and \( \sigma^2 \) are obtained using (4) and (5) by substituting \( \hat{\lambda} \) for \( \lambda \), if desired.

### 3.2. Maximum Likelihood Estimate for \( \eta \)

For a fixed \( \eta \) and \( y \geq 0 \) in (2), the log-likelihood function \( \ell_\eta \)

\[ \ell_\eta = \ell_\eta(\mu, \sigma^2; y_1, y_2, \cdots, y_n) \], can be written as

\[
\ell_\eta = \begin{cases} 
-\frac{n}{2} \ln(2\pi\sigma^2) & - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left( \frac{(y_i + 1)^{n-1}}{\eta} - \mu \right)^2 + (\eta - 1) \sum_{i=1}^{n} \ln(y_i + 1), \\
-\frac{n}{2} \ln(2\pi\sigma^2) & - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \ln(y_i + 1) - \frac{1}{\eta} \sum_{i=1}^{n} \ln(y_i + 1), & \eta \neq 0, \\
-\frac{n}{2} \ln(2\pi\sigma^2) & - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \ln(y_i + 1) - 1, & \eta = 0.
\end{cases}
\]  

Equation (9) is maximized when the partial derivatives of (9) with respect to \( \mu \) and \( \sigma^2 \) are equated to zero and the solution to the corresponding system is found. This leads to solving the following two equations,

\[
\frac{\partial \ell_\eta}{\partial \mu} \bigg|_{\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2} = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(y_i + 1)^{n-1}}{\eta} - 1 \right),
\]

and

\[
\frac{\partial \ell_\eta}{\partial \sigma^2} \bigg|_{\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(y_i + 1)^{n-1}}{\eta} - 1 \right)^2.
\]

Then the pseudo log-likelihood \( \ell^*_\eta = \ell_\eta(\eta; y_1, y_2, \cdots, y_n) \) can be written as

\[
\ell^*_\eta = -\frac{n}{2} \left[ \ln(2\pi) + \ln \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(y_i + 1)^{n-1}}{\eta} - 1 \right) \right\}^2 + 1 \right]
\]
\[ (\eta - 1) \sum_{i=1}^{n} \ln(y_i + 1). \]  

(12)

The steps for maximizing \( \ell_{\eta}^* \) using the Newton-Raphson method are as follows:

\[ \frac{\partial \ell_{\eta}^*}{\partial \eta} |_{\eta = \hat{\eta}} = 0 \implies h(\hat{\eta}) = -n \frac{\sum_{i=1}^{n} (P_i - \hat{P}) (Q_i - \hat{Q})}{\sum_{i=1}^{n} (P_i - \hat{P})^2} + \sum_{i=1}^{n} \ln(y_i + 1) = 0 \]  

(13)

and

\[ \hat{\eta}^{(t+1)} = \hat{\eta}^{(t)} - \frac{h(\hat{\eta}^{(t)})}{h'(\hat{\eta}^{(t)})}, \]  

(14)

where

\[ h'(\hat{\eta}) = -n \left[ \frac{\sum_{i=1}^{n} (P_i - \hat{P}) (R_i - \bar{R}) + \sum_{i=1}^{n} (Q_i - \hat{Q})^2}{\sum_{i=1}^{n} (P_i - \hat{P})^2} - 2 \left\{ \sum_{i=1}^{n} (P_i - \hat{P}) (Q_i - \hat{Q}) \right\}^2 \right]. \]

\[ P_i = \left( \frac{(y_i + 1)^{n-1}}{\eta} \right), \quad \bar{P} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(y_i + 1)^{n-1}}{\eta} \right), \]

\[ Q_i = \left( \frac{\ln(y_i + 1)(y_i + 1)^{n} - (y_i + 1)^{n+1}}{\eta^2} \right), \quad \bar{Q} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\ln(y_i + 1)(y_i + 1)^{n} - (y_i + 1)^{n+1}}{\eta^2} \right), \]

\[ R_i = \left( \frac{y_i^3 \ln(y_i + 1)^2(y_i + 1)^n - 2\eta^3 \ln(y_i + 1)(y_i + 1)^n + 2\eta y_i^n - 2\eta}{\eta^4} \right), \]

and

\[ \bar{R} = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{y_i^3 \ln(y_i + 1)^2(y_i + 1)^n - 2\eta^3 \ln(y_i + 1)(y_i + 1)^n + 2\eta y_i^n - 2\eta}{\eta^4} \right). \]

The computation algorithm starts with an initial value of \( \hat{\eta} \) (an obvious choice is 1) iteratively using (14) and then \( \hat{\mu} \) and \( \hat{\sigma}^2 \) are obtained using (10) and (11) by substituting \( \hat{\eta} \) for \( \eta \), if desired.

For a fixed \( \eta \) and \( y < 0 \) in (2), the log-likelihood function

\[ \ell_{\eta} = \ell_{\eta} (\mu, \sigma^2; y_1, y_2, \cdots, y_n), \]

can be written as

\[ \ell_{\eta} = \begin{cases} 
-\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left( \frac{(-y_i + 1)^{2-\eta} - 1}{2-\eta} - \mu \right)^2 + (1 - \eta) \sum_{i=1}^{n} \ln(-y_i + 1), & \eta \neq 2, \\
-\frac{n}{2} \ln(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (-\ln(-y_i + 1) - \mu)^2 - \sum_{i=1}^{n} \ln(-y_i + 1), & \eta = 2.
\end{cases} \]  

(15)

Equation (15) is maximized when the partial derivatives of (15) with respect to \( \mu \) and \( \sigma^2 \) are equated to zero and the solution to the corresponding system is found. This leads to solving the following two equations,

\[ \frac{\partial \ell_{\eta}}{\partial \mu} |_{\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2} = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \frac{(-y_i + 1)^{2-\eta} - 1}{2-\eta} \]  

(16)

and

\[ \frac{\partial \ell_{\eta}}{\partial \sigma^2} |_{\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(-y_i + 1)^{2-\eta} - 1}{2-\eta} - \left( \frac{(-y_i + 1)^{2-\eta} - 1}{2-\eta} \right)^2 \right). \]  

(17)

Then the pseudo log-likelihood \( \ell_{\eta}^* = \ell_{\eta}(\eta; y_1, y_2, \cdots, y_n) \) can be written as

\[ \ell_{\eta}^* = -\frac{n}{2} \ln(2\pi) \]

(109)
\[ + \ln \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(-y_i + 1)^{2-n} - 1}{2-\eta} \right) - \frac{1}{n} \sum_{j=1}^{n} \left( \frac{(-y_j + 1)^{2-n} - 1}{2-\eta} \right) \right\} \right] ^2 + 1 \]

\[ + (1 - \eta) \sum_{i=1}^{n} \ln(-y_i + 1), \quad (18) \]

The steps for maximizing \( \ell^*_\eta \) using the Newton-Raphson method are as follows:

\[ \frac{\partial \ell^*_\eta}{\partial \eta} \bigg|_{\eta=\hat{\eta}} = 0 \implies h(\hat{\eta}) = -n \frac{\sum_{i=1}^{n} (P_i - \bar{P}) (Q_i - \bar{Q})}{\sum_{i=1}^{n} (P_i - \bar{P})^2} + \sum_{i=1}^{n} \ln(y_i + 1) = 0 \quad (19) \]

and

\[ \hat{\eta}^{(t+1)} = \hat{\eta}^{(t)} - \frac{h(\hat{\eta}^{(t)})}{h'(\hat{\eta}^{(t)})}, \quad (20) \]

where

\[ h'(\hat{\eta}) = -n \left[ \frac{\sum_{i=1}^{n} (P_i - \bar{P}) (R_i - \bar{R}) + \sum_{i=1}^{n} (Q_i - \bar{Q})^2}{\sum_{i=1}^{n} (P_i - \bar{P})^2} \right] \]

\[ P_i = \left( \frac{(-y_i + 1)^{2-n} - 1}{2-\eta} \right), \quad \bar{P} = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{(-y_j + 1)^{2-n} - 1}{2-\eta} \right), \]

\[ Q_i = \left( \frac{(-y_j + 1)^{2-n} - 1}{2-\eta} \right), \quad \bar{Q} = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{(-y_j + 1)^{2-n} - 1}{2-\eta} \right), \]

\[ R_i = \left( \frac{(-y_j + 1)^{2-n} + 2(2-\eta)\ln((-y_j + 1)^{2-n} - 1)}{(2-\eta)^3} \right), \]

\[ \bar{R} = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{(-y_j + 1)^{2-n} + 2(2-\eta)\ln((-y_j + 1)^{2-n} - 1)}{(2-\eta)^3} \right), \]

The computation algorithm starts with an initial value of \( \hat{\eta} \) (an obvious choice is 1) iteratively using (20) and then \( \hat{\mu} \) and \( \hat{\sigma}^2 \) are obtained using (16) and (17) by substituting \( \hat{\eta} \) for \( \eta \), if desired.

### 4. Shapiro-Wilk W Statistic

The Shapiro-Wilk W test statistic (Shapiro and Wilk (1965)) is obtained by dividing the square of an appropriate linear combination of the sample order statistics by the usual symmetric estimate of the variance.

Let \((X_1, X_2, \ldots, X_n)\) be a random sample to be tested for normality, ordered \(X_{(1)} < X_{(2)} < \cdots < X_{(n)}\). Define

\[ W = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]

where the vector \( a = \frac{m(V^{-1})}{(m - V^{-1})^2}, m \) is the vector of the expected values and \( V \) is the variance covariance matrix of the standard normal order statistics.

The value of \( W \) is closer to 1 means that the data is closer to normality and the maximum value of \( W \) is 1. The Shapiro and Wilk (1965) W statistic has been shown to yield a powerful test of
normality for a variety of nonnormal distributions (Pearson, D’Agostino, and Bowman (1977) and Shapiro, Wilk, and Chen (1968)).

The values of the $a_i$’s are tabulated in Shapiro and Wilk (1965) for $n = 2(1)50$. For other sample sizes, the $a_i$’s can be estimated using the following suggested approximations:

$$\hat{a}_i^n = 2m_i, \quad i = 2, 3, \ldots, n - 1, \quad \text{and}$$

$$\hat{a}_i^n = \hat{a}_n^n = \begin{cases} \frac{\Gamma \left( \frac{n}{2} \right)}{\sqrt{2^n \Gamma \left( \frac{n+1}{2} \right)}} & n \leq 20, \\ \frac{\Gamma \left( \frac{n}{2} \right)}{\sqrt{2^n \Gamma \left( \frac{n+1}{2} \right)}} & n > 20, \end{cases}$$

then $a_i^n$ for $i = 2, 3, \ldots, n - 1$ are normalized by dividing by $C = \sqrt{-2.722 + 4.0832n}$ as suggested by Shapiro and Wilk (1965).

The values of the $m_i$’s are tabulated in Harter (1961) for $n = 2(1)100$, 125(25)250, 300, 350, and 400. More accurate values of the $m_i$’s and $V$ are also given in Parish (1992a and 1992b) for $n = 2(1)50$.

5. Simulation Study

The choices of $\lambda$ values, in combination with certain choices of the other model parameters are used to generate data as

$$Y = (1 + \lambda(\mu + \sigma \epsilon))^\frac{1}{\beta}$$

where $\epsilon$ is a pseudo $N(0,1)$ random vector and $\lambda \neq 0$. The choices of parameters are made in such a way that the $Y$ vector is always positive, even in cases when $\lambda \neq 0$.

To see the effects of the three estimates under study, data from skewed distributions are generated using Gamma and Weibull distributions under certain choices of their scale and shape parameters.

For a Gamma distribution

$$f(y; \alpha, \beta) = \frac{1}{\beta \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y}{\beta}}, \quad y > 0$$

is considered and for a Weibull distribution

$$f(y; \beta, \eta) = \frac{\beta}{\eta} \left( \frac{y}{\eta} \right)^{\beta-1} e^{-\left(\frac{y}{\eta}\right)^\beta}$$

is considered in generating random samples.

The simulation study was performed using MATLAB software. The MATLAB codes are available from the authors. For each parameter configuration in Table 1, 1000 samples were generated. The $W$ statistic is computed for the generated samples, then estimates of $\lambda$ and $\eta$ are computed using the maximum likelihood and the maximization of the $W$ statistic methods. The maximization of the $W$ statistic method is implemented using the grid search within the three standard deviation range for the maximum likelihood estimate. $W$ statistics for the transformed data are also computed. Means and standard deviations of the measures are displayed.
<table>
<thead>
<tr>
<th>$W_0$</th>
<th>$W(\lambda_L)$</th>
<th>$W(\lambda_W)$</th>
<th>$W(\eta_L)$</th>
<th>$W(\eta_W)$</th>
<th>$\lambda_L$</th>
<th>$\lambda_W$</th>
<th>$\eta_L$</th>
<th>$\eta_W$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal: $\mu=6$, $\sigma=1$, $\lambda=-2$, $n=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9322</td>
<td>0.9660</td>
<td>0.9658</td>
<td>0.9586</td>
<td>-1.68</td>
<td>-10.68</td>
<td>-1.87</td>
<td>-7.23</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0472</td>
<td>0.0175</td>
<td>0.0176</td>
<td>0.0168</td>
<td>0.0230</td>
<td>2.77</td>
<td>12.48</td>
<td>3.18</td>
</tr>
<tr>
<td></td>
<td>Normal: $\mu=5$, $\sigma=1$, $\lambda=-1$, $n=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9130</td>
<td>0.9667</td>
<td>0.9657</td>
<td>0.9475</td>
<td>-0.85</td>
<td>-10.54</td>
<td>-0.89</td>
<td>-4.67</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0631</td>
<td>0.0160</td>
<td>0.0164</td>
<td>0.0163</td>
<td>0.0311</td>
<td>1.27</td>
<td>8.31</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>Normal: $\mu=5$, $\sigma=1$, $\lambda=1$, $n=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9522</td>
<td>0.9664</td>
<td>0.9664</td>
<td>0.9671</td>
<td>0.91</td>
<td>0.92</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0219</td>
<td>0.0168</td>
<td>0.0168</td>
<td>0.0152</td>
<td>0.0152</td>
<td>1.29</td>
<td>1.52</td>
<td>1.43</td>
</tr>
<tr>
<td></td>
<td>Normal: $\mu=5$, $\sigma=1$, $\lambda=2$, $n=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9474</td>
<td>0.9659</td>
<td>0.9660</td>
<td>0.9667</td>
<td>1.71</td>
<td>1.98</td>
<td>1.86</td>
<td>2.13</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0314</td>
<td>0.0173</td>
<td>0.0172</td>
<td>0.0166</td>
<td>0.0167</td>
<td>2.42</td>
<td>3.17</td>
<td>2.63</td>
</tr>
<tr>
<td></td>
<td>Normal: $\mu=6$, $\sigma=1$, $\lambda=-2$, $n=50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9499</td>
<td>0.9826</td>
<td>0.9825</td>
<td>0.9829</td>
<td>0.9756</td>
<td>-1.85</td>
<td>-11.66</td>
<td>-1.92</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0385</td>
<td>0.0079</td>
<td>0.0079</td>
<td>0.0075</td>
<td>0.0140</td>
<td>1.57</td>
<td>7.02</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td>Normal: $\mu=5$, $\sigma=1$, $\lambda=-1$, $n=50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9492</td>
<td>0.9821</td>
<td>0.9815</td>
<td>0.9823</td>
<td>0.9558</td>
<td>-0.99</td>
<td>-11.79</td>
<td>-0.95</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0346</td>
<td>0.0088</td>
<td>0.0088</td>
<td>0.0075</td>
<td>0.0293</td>
<td>0.70</td>
<td>4.42</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Normal: $\mu=5$, $\sigma=1$, $\lambda=1$, $n=50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9755</td>
<td>0.9818</td>
<td>0.9819</td>
<td>0.9829</td>
<td>0.9829</td>
<td>0.93</td>
<td>0.93</td>
<td>1.00</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0126</td>
<td>0.0084</td>
<td>0.0084</td>
<td>0.0076</td>
<td>0.0076</td>
<td>0.75</td>
<td>0.88</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Normal: $\mu=5$, $\sigma=1$, $\lambda=2$, $n=50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9717</td>
<td>0.9820</td>
<td>0.9820</td>
<td>0.9822</td>
<td>0.9822</td>
<td>1.89</td>
<td>2.19</td>
<td>1.90</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0161</td>
<td>0.0085</td>
<td>0.0085</td>
<td>0.0078</td>
<td>0.0079</td>
<td>1.27</td>
<td>1.67</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td>Gamma: $\alpha=4$, $\beta=10$, $n=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9177</td>
<td>0.9660</td>
<td>0.9659</td>
<td>0.9664</td>
<td>0.9664</td>
<td>0.27</td>
<td>0.25</td>
<td>0.33</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0503</td>
<td>0.0168</td>
<td>0.0168</td>
<td>0.0172</td>
<td>0.0172</td>
<td>0.40</td>
<td>0.41</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Gamma: $\alpha=9$, $\beta=5$, $n=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9369</td>
<td>0.9661</td>
<td>0.9661</td>
<td>0.9659</td>
<td>0.9659</td>
<td>0.27</td>
<td>0.26</td>
<td>0.30</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0407</td>
<td>0.0169</td>
<td>0.0169</td>
<td>0.0163</td>
<td>0.0163</td>
<td>0.63</td>
<td>0.65</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Gamma: $\alpha=9$, $\beta=10$, $n=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9378</td>
<td>0.9667</td>
<td>0.9667</td>
<td>0.9661</td>
<td>0.9661</td>
<td>0.29</td>
<td>0.28</td>
<td>0.33</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0380</td>
<td>0.0157</td>
<td>0.0157</td>
<td>0.0165</td>
<td>0.0165</td>
<td>0.64</td>
<td>0.64</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Gamma: $\alpha=16$, $\beta=5$, $n=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>0.9438</td>
<td>0.9668</td>
<td>0.9668</td>
<td>0.9668</td>
<td>0.9668</td>
<td>0.27</td>
<td>0.26</td>
<td>0.32</td>
</tr>
<tr>
<td>$s$</td>
<td>0.0342</td>
<td>0.0162</td>
<td>0.0162</td>
<td>0.0158</td>
<td>0.0159</td>
<td>0.89</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>( W_0 )</td>
<td>( W(\lambda_L) )</td>
<td>( W(\lambda_W) )</td>
<td>( W(\hat{\eta}_L) )</td>
<td>( W(\hat{\eta}_W) )</td>
<td>( \lambda_L )</td>
<td>( \hat{\lambda}_L )</td>
<td>( \hat{\eta}_L )</td>
<td>( \hat{\eta}_W )</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Gamma: ( \alpha=4, \beta=10, n=50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9336</td>
<td>0.9821</td>
<td>0.9821</td>
<td>0.9825</td>
<td>0.9825</td>
<td>0.31</td>
<td>0.29</td>
<td>0.32</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0353</td>
<td>0.0078</td>
<td>0.0078</td>
<td>0.0075</td>
<td>0.0075</td>
<td>0.23</td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td>Gamma: ( \alpha=9, \beta=5, n=50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9574</td>
<td>0.9823</td>
<td>0.9823</td>
<td>0.9823</td>
<td>0.9823</td>
<td>0.32</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0252</td>
<td>0.0079</td>
<td>0.0079</td>
<td>0.0085</td>
<td>0.0085</td>
<td>0.35</td>
<td>0.36</td>
<td>0.39</td>
</tr>
<tr>
<td>Gamma: ( \alpha=9, \beta=10, n=50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9567</td>
<td>0.9820</td>
<td>0.9820</td>
<td>0.9824</td>
<td>0.9824</td>
<td>0.33</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0255</td>
<td>0.0080</td>
<td>0.0080</td>
<td>0.0078</td>
<td>0.0078</td>
<td>0.36</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>Gamma: ( \alpha=16, \beta=5, n=50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9641</td>
<td>0.9823</td>
<td>0.9823</td>
<td>0.9823</td>
<td>0.9823</td>
<td>0.29</td>
<td>0.28</td>
<td>0.32</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0208</td>
<td>0.0080</td>
<td>0.0080</td>
<td>0.0077</td>
<td>0.0077</td>
<td>0.48</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>Weibull: ( \beta=2, \eta=10, n=20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9353</td>
<td>0.9656</td>
<td>0.9652</td>
<td>0.9654</td>
<td>0.9646</td>
<td>0.46</td>
<td>0.32</td>
<td>0.54</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0355</td>
<td>0.0168</td>
<td>0.0169</td>
<td>0.0165</td>
<td>0.0169</td>
<td>0.33</td>
<td>0.46</td>
<td>0.39</td>
</tr>
<tr>
<td>Weibull: ( \beta=3, \eta=4, n=20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9518</td>
<td>0.9644</td>
<td>0.9646</td>
<td>0.9648</td>
<td>0.9647</td>
<td>0.72</td>
<td>0.68</td>
<td>0.79</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0260</td>
<td>0.0180</td>
<td>0.0179</td>
<td>0.0178</td>
<td>0.0178</td>
<td>0.33</td>
<td>0.70</td>
<td>0.59</td>
</tr>
<tr>
<td>Weibull: ( \beta=2, \eta=6, n=20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9348</td>
<td>0.9649</td>
<td>0.9646</td>
<td>0.9652</td>
<td>0.9647</td>
<td>0.46</td>
<td>0.37</td>
<td>0.53</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0356</td>
<td>0.0172</td>
<td>0.0173</td>
<td>0.0172</td>
<td>0.0173</td>
<td>0.35</td>
<td>0.44</td>
<td>0.39</td>
</tr>
<tr>
<td>Weibull: ( \beta=3, \eta=6, n=20 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9529</td>
<td>0.9646</td>
<td>0.9647</td>
<td>0.9652</td>
<td>0.9652</td>
<td>0.73</td>
<td>0.71</td>
<td>0.81</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0250</td>
<td>0.0175</td>
<td>0.0174</td>
<td>0.0168</td>
<td>0.0168</td>
<td>0.50</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Weibull: ( \beta=2, \eta=4, n=50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9543</td>
<td>0.9810</td>
<td>0.9801</td>
<td>0.9738</td>
<td>0.9717</td>
<td>0.51</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0217</td>
<td>0.0080</td>
<td>0.0083</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.18</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>Weibull: ( \beta=3, \eta=4, n=50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9750</td>
<td>0.9811</td>
<td>0.9811</td>
<td>0.9805</td>
<td>0.9805</td>
<td>0.76</td>
<td>0.72</td>
<td>0.96</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0116</td>
<td>0.0082</td>
<td>0.0082</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.28</td>
<td>0.38</td>
<td>0.00</td>
</tr>
<tr>
<td>Weibull: ( \beta=2, \eta=6, n=50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9540</td>
<td>0.9807</td>
<td>0.9801</td>
<td>0.9718</td>
<td>0.9705</td>
<td>0.51</td>
<td>0.40</td>
<td>0.49</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0218</td>
<td>0.0087</td>
<td>0.0088</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.18</td>
<td>0.24</td>
<td>0.00</td>
</tr>
<tr>
<td>Weibull: ( \beta=3, \eta=6, n=50 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>0.9753</td>
<td>0.9810</td>
<td>0.9810</td>
<td>0.9890</td>
<td>0.9890</td>
<td>0.78</td>
<td>0.76</td>
<td>0.69</td>
</tr>
<tr>
<td>( s )</td>
<td>0.0110</td>
<td>0.0084</td>
<td>0.0084</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.27</td>
<td>0.34</td>
<td>0.00</td>
</tr>
</tbody>
</table>
10. Concluding Remarks

In both transformations $W$ statistics improved equally and their variabilities are also similar. For normal samples, maximum likelihood estimates have smaller variances. For gamma samples, all the variances are similar. For Weibull samples, estimates of $\eta$ have smaller variances for $n = 50$ compared to the estimates for $\lambda$.

11. Bibliography


Inference about a Common Mean in One Way Random Effects Model

Guido Knapp\textsuperscript{1}, Pranab K. Mitra\textsuperscript{2}, Bimal K. Sinha\textsuperscript{2}

\textsuperscript{1}Department of Statistics, University of Dortmund, Germany
\textsuperscript{2}Department of Mathematics and Statistics, University of Maryland, Baltimore County, USA

Abstract

It is well known that the inference about a common mean in a one way random effects model essentially depends on how the associated variance components are estimated. In this paper we compare different methods of estimating the variance components in terms of their effects on the common mean inference problem based on some simulation studies, and apply these methods to three classic data sets of which one deals with an environmental issue.

Keywords and phrases: Common mean, Generalized p-value, Helmert matrix, Mixed effect, Moore-Penrose inverse, Variance components, WLSE.

1 Introduction

We consider the one way random effects model of ANOVA, i.e.,
\[ y_{ij} = \mu + a_i + e_{ij} \]
for \( i = 1, 2, \ldots, r \), \( j = 1, 2, \ldots, n_i \), where \( y_{ij} \) denotes the observable variable, \( a_i \) denotes the unobservable random effect with mean 0 and variance \( \sigma^2_r \), \( e_{ij} \) is the error term with mean 0 and variance \( \sigma^2_e \), independent of \( a_i \)'s. Here \( \mu \) is a fixed unknown quantity, called the common mean. We denote \( n = \sum_{i=1}^{r} n_i \), \( n > r \), as the total number of observations.

In this context, usually \( a_i \)'s and \( e_{ij} \)'s are assumed to be independent normal, and \( \sigma^2_r \) and \( \sigma^2_e \), \( i = 1, 2, \ldots, r \) are known as variance components. When these variance components are known, one can carry out a normal theory based optimum inference by using weighted least squares estimate (WLSE) of \( \mu \) given by
\[ \hat{\mu} = \bar{y}_w = \sum_{i=1}^{r} \frac{w_i}{w} y_{i} \cdot \bar{y}_i, \quad \bar{y}_i = \frac{n_i}{r} y_{ij}, \quad w = \sum_{i=1}^{r} w_i \]  \hspace{1cm} (1.1)
where, for the case of homogeneous error variances, i.e., \( \sigma^2_i = \sigma^2_e \) \( \forall i \), \( w_i \)'s are given by
\[ w_i = \frac{n_i}{n_i \sigma^2_e + \sigma^2_e} \]
while in the case of non-homogeneous error variances, \( w_i \)'s are given by
\[ w_i = \frac{n_i}{n_i \sigma^2_e + \sigma^2_i} \]
Note that in the case of homogeneous error variances, \( w_i/w \) can be treated as a function of \( \eta = \sigma_t^2/\sigma_e^2 \) and we can write \( \hat{\mu} \) as

\[
\hat{\mu} = \bar{y}_n = \sum_{i=1}^{r} \frac{w_i(\eta)}{w(\eta)} \bar{y}_n.
\]

(1.2)

where \( w_i(\eta) = \frac{\bar{y}_n}{1+n_i\eta} \) and \( w(\eta) = \sum_{i=1}^{r} w_i(\eta) \). When the variance components are known, since

\[
\text{var}(\hat{\mu}) = 1/\sum_{i=1}^{r} w_i,
\]

one can apply the standard z method to test a hypothesis about \( \mu \) or to derive a confidence interval for \( \mu \).

When the variance components are unknown, one can estimate them based on the sufficient statistics, between sum of squares (BSS) and within sum of squares (WSS), and replace \( w_i \)'s by their corresponding estimates \( \hat{w}_i \)'s in both the WLS estimate of \( \mu \) as well as in its variance expression, thus yielding

\[
\hat{\mu} \approx \sum_{i=1}^{r} \frac{\hat{w}_i}{\hat{w}} \bar{y}_n, \quad \text{var}(\hat{\mu}) \sim 1/\sum_{i=1}^{r} \frac{\hat{w}_i}{\hat{w}}, \quad \hat{\mu}_i = \frac{n_i}{n_i \sigma_t^2 + \sigma_e^2} \text{ or } \frac{n_i}{n_i \sigma_t^2 + \sigma_e^2}.
\]

(1.3)

The same normal test statistic \( z \) can again be used, which of course is an approximate test. In this context, a slight modification of the \( z \) test has been proposed by Hartung and Knapp (2001) who suggested use of an approximate t test. This is based on a different estimate of \( \text{var}(\hat{\mu}) \) given by

\[
q = \text{var}(\hat{\mu}) \approx \frac{1}{r-1} \sum_{i=1}^{r} \frac{\hat{w}_i}{\hat{w}} (\bar{y}_n - \bar{y}_w)^2.
\]

(1.4)

\( q \) can be considered as a weighted extension of the usual empirical variance of the study-specific estimates, \( \bar{y}_i \), for \( i = 1, 2, \ldots, r \). One can then define the test statistic, \( T \), as \( T = \frac{\hat{\mu}}{\sqrt{q}} \sim t_{r-1} \), and use it in the usual way.

The crux of the problem now is to describe various ways to estimate the unknown variance components which can be used in the point estimate of \( \hat{\mu} \) of \( \mu \) as well as in the estimate of \( \text{var}(\hat{\mu}) \). We discuss some methods below to estimate the variance components in two cases of homogeneous and heterogeneous error variances separately.

In the case of homogeneous error variance, there are two variance components, \( \sigma_t^2 \) and \( \sigma_e^2 \), and we can either estimate them directly or estimate the error variance \( \sigma_e^2 \) and the ratio \( \eta = \sigma_t^2/\sigma_e^2 \). Of course, the error variance is estimated by pooled mean squared error and the main problem is really to estimate \( \sigma_e^2 \) or the ratio \( \eta \). Once these are estimated and plugged in \( \hat{\mu} \) and also in the estimate of \( \text{var}(\hat{\mu}) \), we can perform simulation studies to check on the validity of the standard \( z \) method and the approximate \( t \) method. We have discussed some well known methods of estimation of \( \sigma_e^2 \) or \( \eta \) in Section 2.

In the case of heterogeneous error variances, we estimate the error variances by respective sample variances and so the main problem is again estimation of \( \sigma_e^2 \), the treatment variance. We have discussed several methods of estimation of \( \sigma_e^2 \) in Section 3.

Simulation results comparing Monte-Carlo estimates of mean, its variances, median, coverage probabilities, expected lengths of confidence intervals and powers are reported in Section 4. The hypotheses are taken as \( H_0 : \mu = 0 \) versus \( H_1 : \mu = 1 \).

Section 5 contains three novel applications of which one deals with efficiently measuring the Reid Vapor Pressure (RVP) of gasoline.
2 Homogeneous Error Variances

There are several competing methods of estimating or constructing confidence interval of $\sigma^2_e$ or $\eta$. As mentioned earlier, we use pooled sample variance to estimate $\sigma^2_e$, the common error variance.

2.1 Thomas-Hultquist (1978) Confidence Interval

Let $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ and $\bar{y}_{..} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} y_{ij}$, and let the mean sum of squares within groups be given by $MS_2 = \frac{1}{n-r} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ whose expected value is $\sigma^2_e$. Also

$$(n-r)MS_2/\sigma^2_e \sim \chi^2_{n-r}$$

Thomas and Hultquist (1978) considered $MS_3$ which is given by

$$MS_3 = \frac{1}{r-1} \sum_{i=1}^{r} \left( \bar{y}_{..} - \frac{1}{r} \sum_{i=1}^{r} \bar{y}_i \right)^2.$$

Let $\bar{n}$ be the harmonic mean of the sample sizes $n_1, n_2, \ldots, n_r$, given by $\frac{1}{\bar{n}} = \sum_{i=1}^{r} \frac{1}{n_i}$.

They showed that

$$\frac{(r-1)MS_3}{\sigma^2_e + \sigma^2_e/\bar{n}} \approx \chi^2_{r-1}.$$  

They also proved that $MS_2$ and $MS_3$ are independent so that

$$\frac{\sigma^2_e}{\sigma^2_e + \sigma^2_e/\bar{n}} \frac{MS_3}{MS_2} \approx \frac{F_{r-1, n-r}}{F_{r-1, n-r}}.$$  

We can apply the method of moments to estimate the variance ratio, $\eta$, by solving the following Equation:

$$\frac{1}{\eta + 1/\bar{n}} MS_3 = \frac{n-r}{n-r-2}$$

which readily yields an estimate of $\eta$ as

$$\hat{\eta} = \frac{n-r-2}{n-r} MS_3 - \frac{1}{\bar{n}}.$$  

(2.1)

(2.2)

Thomas and Hultquist proposed the following confidence interval for $\sigma^2_e$:

$$\left[ \frac{(r-1)}{\chi^2_{r-1, 1-\alpha/2}} \left( MS_3 - \frac{MS_2}{n} F_{r-1, n-r-1-\alpha/2} \right), \frac{(r-1)}{\chi^2_{r-1, 1+\alpha/2}} \left( MS_3 - \frac{MS_2}{n} F_{r-1, n-r+\alpha/2} \right) \right].$$  

(2.3)

If we put $\alpha = 1$ in Equation (2.3), we get what can be called an approximate median unbiased estimate of $\sigma^2_e$, given as $\hat{\sigma}^2_{e, M} = \frac{(r-1)}{\chi^2_{r-1, 0.5}} \left( MS_3 - \frac{MS_2}{n} F_{r-1, n-r, 0.5} \right)$. Then we can estimate $\eta$ as $\hat{\eta}_M = \hat{\sigma}^2_{e, M}/\hat{\sigma}^2_e$ where $\hat{\sigma}^2_e = MSE$. Thomas and Hultquist observed that their method does not perform well for small values of $\eta$.  

118
2.2 Burdick-Eickman (1986) Confidence Interval

Burdick and Eickman (1986) suggested a confidence interval for the ratio $\eta$ which overcomes the problem with small ratios in the Thomas-Hultquist method and has a confidence coefficient of at least $1 - \alpha$.

The interval is given by

$$\left[ \frac{MS_3}{MS_2} \frac{1}{F_{r-1,n-r,1-\alpha/2}} - \frac{1}{n_{min}} \frac{MS_3}{MS_2} \frac{1}{F_{r-1,n-r,\alpha/2}} - \frac{1}{n_{max}} \right]$$

(2.4)

where $n_{min} = \min(n_1, n_2, \cdots, n_r)$ and $n_{max} = \max(n_1, n_2, \cdots, n_k)$. Again, by putting $\alpha = 1$, we can obtain an approximate median unbiased estimate

$$\hat{\eta}_{HK} = \frac{1}{2} \left( \frac{MS_3}{MS_2} \frac{1}{F_{r-1,n-r,0.5}} - \frac{1}{n_{min}} \frac{MS_3}{MS_2} \frac{1}{F_{r-1,n-r,0.5}} - \frac{1}{n_{max}} \right).$$

2.3 Hartung-Knapp (2000) Confidence Interval based on Wald’s method

Instead of approximate confidence intervals for $\eta$, Hartung and Knapp (2000, 2005) considered the exact confidence interval for $\eta$ as given in Wald (1940) to construct a confidence interval for $\sigma^2$.

Following Wald (1940), upon defining

$$(r - 1)MS_4(\eta) = \sum_{i=1}^{r} w_i (\bar{y}_i - \bar{y}_w)^2,$$

it follows that

$$(r - 1)MS_4(\eta)/\sigma^2 \sim \chi^2_{r-1}.$$

Furthermore, MS$_4$ and MS$_2$ are independent so that

$$F_w(\eta) = \frac{MS_4(\eta)}{MS_2} \sim F_{r-1,n-r}.$$

Wald (1940) showed that $(r - 1)MS_4$ is strictly monotonically decreasing in $\eta$, and so the bounds of an exact confidence interval for $\eta$ are given as the solutions of the following system of Equations:

$$\begin{cases}
\text{lower bound:} & F_w(\eta) = F_{r-1,n-r,1-\alpha/2} \\
\text{upper bound:} & F_w(\eta) = F_{r-1,n-r,\alpha/2}
\end{cases}$$

(2.5)

We take $\alpha = 1$ as usual to get an estimate $\hat{\eta}_{HK}$ of $\eta$ and plug that in $\hat{\mu}$. Note that there is a positive chance of getting a negative value of $\hat{\eta}_{HK}$. If this happens, we replace it by 0. This method is described in the paper by Hartung and Knapp (2005).

2.4 Generalized $p$-value Method

A generalized $p$-value method in the context of one way random effects model with homoskedastic error variances and unequal replications is described in Weerahandi (2004). For details of generalized-$p$-value method and computational procedure, we refer to Weerahandi (1995) and Mitra and Sinha (2006).

Recall that $\eta = \sigma^2/\sigma^2_\epsilon$, $w_i(\eta) = \frac{n_i}{\mu_i + n_i \eta}$, $i = 1(1)r$,

$$\tilde{Y}(\eta) = \tilde{Y}_w = \frac{\sum_{i=1}^{r} w_i \tilde{Y}_i}{\sum_{i=1}^{r} w_i}, \quad \tilde{Y}_i = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i}, \quad \text{and} \quad \tilde{Y}(\eta) \sim N\left( \mu_i, \frac{\sigma^2}{\sum_{i=1}^{r} w_i(\eta)} \right).$$

119
Define \( S_u B(\eta) = \sum_{i=1}^{r} w_i (\bar{Y}_i - \bar{Y}_u)^2 \sim \sigma^2 \chi^2_{r-1} \),
and \( S_E = \text{Error sum of squares} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \sim \sigma^2 \chi^2_{N-r} \). Obviously, \( S_u B \) is independent of \( S_E \). In the context of generalized \( p \)-value procedure, upper case denotes the random variables and lower case denotes the corresponding observed values.

Let \( W_1 = \frac{s^2}{s^2} \sim \chi^2_{N-r} \), \( W_2 = \frac{s_{u B}}{s^2} \sim \chi^2_{r-1} \) and \( Z \sim N(0, 1) \).
Hence \( \sigma^2 = \frac{s^2}{W_1} \) and \( \eta = \frac{1}{s_{u B}} (\frac{W_2}{W_1}) \).
Note that the distributions of \( W_1 \), \( W_2 \) and \( Z \) do not involve any parameter and they are independent.
Let \( s_E \), \( s_{u B} \) and \( \bar{Y} \) be the observed (computed from the observed data) values of the random variables \( S_E \), \( S_{u B} \) and \( \bar{Y} \).
A generalized pivot can be defined as \(\) (Weerahandi (2004), page 96),

\[
R = \frac{\eta \left( \frac{S_{u B}(\eta)}{S_E} \right) - (\bar{Y} - \mu) \left( \frac{s_E}{S_E} \sum_{i=1}^{r} \frac{w_i(\eta)}{s_{u B}} \left( \frac{S_{u B}(\eta)}{S_E} \right) \right)}{S_E \sum_{i=1}^{r} \frac{w_i}{s_{u B}} \left( \frac{S_{u B}(\eta)}{S_E} \right) \left( \frac{W_2}{W_1} \right)} \tag{2.6}
\]

\[
= \frac{\eta \left( \frac{1}{s_{u B}} \left( \frac{W_2}{W_i} \right) \right) - Z \left( \frac{s_E}{W_1} \right)}{S_E \sum_{i=1}^{r} \frac{w_i}{s_{u B}} \left( \frac{W_2}{W_1} \right)} \tag{2.7}
\]

From Equation 2.6, it can be observed that the distribution of \( R \) is free of any unknown parameters. Also \( R_{obs} = \mu \) can be obtained from Equation 2.7. So \( R \) is indeed a generalized pivot. So we can simulate \( R \) by simulating \( W_1 \), \( W_2 \) and \( Z \). A generalized confidence interval of \( \mu \) with confidence coefficient will be \([R(x; \alpha/2), R(x; 1 - \alpha/2)]\), where \( R(x; \alpha/2) \) stands for \( \alpha/2 \)th quantile and \( x \) stands for observed values.

Also \( R_T = R - \mu \) is a generalized test statistic, which is stochastically decreasing in \( \mu \).
For details of methods and computations, we refer to Weerahandi (1995, 2004).

### 2.5 Other methods

Some estimates of \( \sigma^2 \) with certain desirable properties have been discussed by Mathew, Sinha and Sutradhar (1992). We considered those estimates also in our simulation study. To describe these estimates, let \( n = \sum_{i=1}^{r} n_i \) and define \( \epsilon_1 \) and \( \epsilon_2 \) as

\[
\epsilon_1 = n^2 - \sum_{i=1}^{r} n_i^2
\]

and

\[
\epsilon_2 = n^2 \sum_{i=1}^{r} n_i^2 + \left( \sum_{i=1}^{r} n_i^2 \right)^2 - 2n \sum_{i=1}^{r} n_i^3.
\]
Note that in the one way random effects set up, \( y_{i \times n} = (y_{ij} : i = 1, 2, \ldots, r, j = 1, 2, \ldots, n_i) \),
\( E(y) = \mu I_n \) and \( \text{Cov}(y) = \sigma^2 I_n + \sigma^2 V \), where \( V = \text{diag}(J_{n_1}, \ldots, J_{n_r}) \). Also
\[
SS_{tr} = y^T \left[ \text{diag} \left( \frac{1}{n_1} J_{n_1}, \ldots, \frac{1}{n_r} J_{n_r} \right) - \frac{1}{n} J_n \right] y.
\]

Here \( J_k \) is a \( k \times k \) matrix of ones. Let \( Z \) be an \( n \times n - 1 \) matrix such that \( Z'1_n = 0, Z'Z = I_n \). Also let \( u = Z'y \) and \( V_1 = Z'VZ \) where \( V \) defined as a block diagonal matrix. Suppose \( V_1^+ \) denotes the Moore-Penrose inverse of \( V_1 \).

The four estimates, described in the paper by Mathew et al. (1992), are given below.
\[
\begin{align*}
\hat{\sigma}^2_{r,1} &= \frac{1}{r+1} u^T V_1^+ u \\
\hat{\sigma}^2_{r,2} &= \begin{cases} 0 & \text{if } r \leq 3 \\
\frac{(r-3) u^T V_1^+ u}{(r-1)(r+1)} & \text{if } r \geq 4
\end{cases} \\
\hat{\sigma}^2_{r,3} &= \frac{n c_1}{c_1 + 2 c_2} SS_{tr} \\
\hat{\sigma}^2_{r,4} &= \begin{cases} 0 & \text{if } c_1^2 \leq 2 c_2 \\
\frac{n(c_1^2 - 2 c_2)}{c_1 (c_1^2 + 2 c_2)} SS_{tr} & \text{if } c_1^2 > 2 c_2
\end{cases}
\end{align*}
\]

3 Heterogeneous Error Variances

In this section we discuss different methods of estimating the between study variability \( \sigma^2 \) under the more realistic scenario of heterogeneous error variances. We can estimate \( \sigma^2 \) from replicated observations from the \( i \)th population as \( \hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1} \). However, efficient estimation of \( \sigma^2 \) poses some difficulties.

A review of some solutions is presented below.

3.1 DerSimonian and Laird method, 1986

The standard estimate of \( \sigma^2 \) in meta analysis is a moment estimate suggested by DerSimonian and Laird (1986), viz.,
\[
\hat{\sigma}^2_{r,1} = \frac{\sum_{i=1}^{r} v_i (\bar{y}_i - \bar{y}_.)^2 - (r - 1)}{\sum_{i=1}^{r} v_i - \sum_{i=1}^{r} v_i^2 / \sum_{i=1}^{r} v_i}
\]
where \( \bar{y}_i = \frac{\sum_{j=1}^{r} v_i y_{ij}}{\sum_{j=1}^{r} v_i} \) and \( v_i = n_i / \hat{\sigma}_i^2 \). It should be noted that the above formula can yield negative estimates with a positive probability.

So the adjusted estimate is given by
\[
\hat{\sigma}^2_{r,DL} = \max(0, \hat{\sigma}^2_{r,1})
\]
See DerSimonian and Laird (1986) for details.
3.2 Rao, Kaplan and Cochran method, 1981

In Rao, Kaplan and Cochran (1981), different methods of estimation of variance components such as ANOVA-type estimate, unweighted and nonnegative estimates were discussed. To describe these estimates, let \( \bar{y}_i = \frac{\sum_{j=1}^{r} y_{ij}}{n_i} \) and \( \bar{y}_i = \frac{\sum_{j=1}^{r} y_{ij} / \sum_{i=1}^{r} l_i} {\sum_{i=1}^{r} l_i} \), where \( l_i = n_i / (n_i + 1) \).

The ANOVA-type estimate of \( \sigma^2 \) is given by

\[
\hat{\sigma}^2_{r, AN} = \left( n - \frac{\sum_{i=1}^{r} n_i}{n} \right)^{-1} \left\{ \sum_{i=1}^{r} n_i (\bar{y}_i - \bar{y})^2 - \frac{1}{r} \sum_{i=1}^{r} \frac{\sigma^2}{n_i} \right\}.
\]

(3.2)

The procedure with unweighted sums of squares (USS) yields

\[
\hat{\sigma}^2_{r, un} = (r - 1)^{-1} \left\{ \sum_{i=1}^{r} (\bar{y}_i - \bar{y})^2 - \frac{r - 1}{r} \sum_{i=1}^{r} \frac{\sigma^2}{n_i} \right\}.
\]

(3.3)

A nonnegative estimate of \( \sigma^2 \) is given by

\[
\hat{\sigma}^2_{r, nd} = \frac{1}{r} \sum_{i=1}^{r} l_i^2 (\bar{y}_i - \bar{y})^2.
\]

(4.4)

For details, we refer to Rao, Kaplan and Cochran (1981).

3.3 Mandel-Paule method, 1970

Mandel and Paule (1970) and Paule and Mandel (1982) suggested the following method to obtain \( \hat{\sigma}^2 \). Recall from (1.1) that \( \bar{y}_w = \frac{1}{w} \sum_{i=1}^{r} w_i \bar{y}_i \), where \( w_i = \frac{1}{(\sigma^2 + \sigma^2 / n_i)} \). Let

\[
Q(y) = \sum_{i=1}^{r} \frac{w_i (\bar{y}_i - \bar{y}_w)^2}{(\sigma^2 + \sigma^2 / n_i)}
\]

where \( w_i = \frac{1}{(\sigma^2 + \sigma^2 / n_i)} \).

The expectation of \( Q(y) \) is approximately equal to \( r - 1 \). This fact motivates the following Equation:

\[
Q(\sigma^2) = r - 1,
\]

(3.5)

which can be solved to yield \( \sigma^2 \) provided \( \sigma^2 \)'s are known or estimated by \( \hat{\sigma}^2 \). It can be shown that \( Q(y) \) is monotonically decreasing function in \( \sigma^2 \) and moreover it is convex, thus providing at most one positive root, \( \sigma^2_{MP} \), which is taken to be zero when the Equation yields negative solution. Thus Mandel-Paule (MP) estimate of \( \mu \) can be obtained by using \( \sigma^2_{MP} \) in the expression of \( \bar{y}_w \).

In the papers by Rukhin and Vangel (1998) and Rukhin et. al. (2000), it is shown that MP method is very close to the method of maximum likelihood (ML) as well as the method of restricted maximum likelihood (REML). So in this context, we did not consider ML or REML which are very complicated compared to MP method. For details, we refer to Rukhin and Vangel (1998) and Rukhin et. al. (2000).
3.4 Kenward-Roger method, 1997

Another method suggested by Kenward and Roger (1997) can be implemented for one way random effects model with heteroskedastic errors. Recall that the dispersion matrix of $y_{1 \times n} = (y_{ij} : i = 1, 2, \cdots, r, j = 1, 2, \cdots, n_i)$ can be written as

$$
\Sigma = \text{Diag}\{\sigma^2_r J_{n_i} + \sigma^2_r I_{n_i}\}
$$

which gives

$$
\Sigma^{-1} = \text{Diag}\left\{\frac{1}{\sigma^2_r} \left( I_{n_i} - \frac{\sigma^2_r}{\sigma^2_r + n_i \sigma^2} \right) \right\}
$$

where $n = \sum_{i=1}^{r} n_i$.

Let $\hat{\Phi}(v) = \left( (I_n^T \Sigma^{-1} I_n)^{-1} \right)$ where $v = (\sigma^2_r, \sigma^2_1, \cdots, \sigma^2_r)$ is the vector of variance components. Let $\tilde{v}$ be the MLE of $v$. The asymptotic variance-covariance matrix of $\tilde{v}$ is the inverse of Fisher information matrix $I_F(v)$, i.e. $W = W(v) = I_F^{-1}(v)$.

If variance components are unknown, we obtain a plug-in estimate of $\mu$ as

$$
\hat{\mu}^{(REML)} = \hat{\Phi} \sum_{i=1}^{r} \frac{n_i \tilde{v}_i}{\sigma^2_i + n_i \sigma^2_r}
$$

where $\hat{\Phi} = \Phi(\tilde{v})$ and the estimates of variance components are REML (Restricted Maximum Likelihood) estimates of $v$.

Following the small sample correction suggested by Kenward and Roger (1997), the adjusted estimate of the small sample variance of $\hat{\mu}^{(REML)}$ is given by

$$
\text{var}(\hat{\mu}^{(REML)}) = \hat{\Phi}_A = \hat{\Phi} + 2\hat{\Lambda},
$$

where

$$
\hat{\Lambda} = \Lambda(\tilde{v}) = \hat{\Phi}^2 \sum_{i-0}^{r} \sum_{j-0}^{r} \hat{W}_{ij} (\hat{Q}_{ij} - \hat{P}_i \hat{P}_j).
$$

Here $i = 0$ corresponds to $\sigma^2_r$ which can be denoted as $\sigma^2_0$ for sake of notational simplicity.

The 100(1 - $\alpha$)% Kenward-Roger confidence interval for the common mean $\mu$ is then given by

$$
\hat{\mu}^{(REML)} \pm \sqrt{\hat{\Phi}_A t_{m;1-\alpha/2}},
$$

where

$$
\hat{m} = \frac{2\hat{\Phi}^2}{\text{var}({\hat{\Phi}_A})}, \quad \text{and} \quad \text{var}(\hat{\Phi}) = \left( \frac{\partial \Phi}{\partial v} \right)' W \left( \frac{\partial \Phi}{\partial v} \right).
$$

Finally $\hat{m}$ is obtained as

$$
\hat{m} = \frac{2}{\Phi^2({\hat{P}' W \hat{P}})},
$$

where $\hat{P} = (\hat{P}_0, \hat{P}_1, \cdots, \hat{P}_r)'$. If $\hat{m} < 1$, we replaced it by 1.

It is suggested in Witkovsky et. al. (2003) that instead of using REML of $\mu$, the Mandel-Paule (MP) estimate $\hat{\mu}_{MP}$ can be used. The variance components are also estimated by MP method which is described in Section 3.3. In our simulation study, we employed this method along with MP estimates of $\sigma^2_r$. This method is denoted by MPKR.

For the sake of completeness we describe below the terms used in this section which can be found in Witkovsky et. al. (2003) with details of computations.
\[ P_0 = \sum_{i=0}^{r} \left( \frac{n_i}{\sigma_i^2} \right)^2, \]
\[ P_i = \sum_{i=0}^{r} \left( \frac{n_i}{\sigma_i^2} \right)^2, \ i = 1, \ldots, r, \]
\[ Q_{0,0} = \left( \sum_{i=0}^{r} \frac{n_i}{\sigma_i^2} \right)^3, \]
\[ Q_{0,i} = \sum_{i=0}^{r} \left( \frac{n_i}{\sigma_i^2} \right)^3, \ i = 1, \ldots, r, \]
\[ Q_{i,j} = \left( \frac{n_i}{\sigma_i^2} \right)^3, \ i = j \text{ and } Q_{i,j} = 0, \ i \neq j, \ i, j = 1, \ldots, r. \]

The Fisher information matrix for the REML of variance components is defined by its elements as
\[
I_{F_{i,j}} = \frac{1}{2} \left[ tr \left( \frac{\partial \Sigma^{-1}}{\partial \sigma_i^2} \Sigma \frac{\partial \Sigma^{-1}}{\partial \sigma_j^2} \right) - tr \left( 2 \Phi Q_{i,j} - \Phi P_i \Phi P_j \right) \right] = \frac{1}{2} \left[ S_{i,j} - R_{i,j} \right], \ i, j = 0, 1, \ldots, r.
\]

Here \( R_{i,j} = \Phi (2 Q_{i,j} - \Phi P_i \Phi P_j) \). The elements of \( S \), reported below, are also computed in Witkovsky et. al (2003).

\( S_{0,0} = -P_0, S_{0,i} = S_{i,0} = -P_i \) for \( i = 1, 2, \ldots, r \), and
\[ S_{i,j} = \frac{1}{\sigma_i^2} \left( n_i - \frac{2 n_i n_j}{(\sigma_i^2 + \sigma_j^2)} + \frac{n_j^2 \sigma_i^2}{(\sigma_i^2 + \sigma_j^2)} \right), \ i = j \text{ and } S_{i,j} = 0 \ otherwise, \ i, j = 0, 1, \ldots, r.
\]

### 3.5 Other methods

Two improved estimates of \( \sigma^2 \) were proposed by Mathew, Nahtman, Rosen and Sinha (2006). They considered the general estimate of the form
\[ \hat{\sigma}^2 = Y' A Y + \sum_{i=1}^{r} c_i S_i, \]
where \( Y = (\bar{y}_1, \ldots, \bar{y}_r) \), \( A = (a_{ij}) \) is assumed to be symmetric and \( A1 = 0 \) so that estimate will be translation invariant. The unbiasedness \( \hat{\sigma}^2 \) requires \( tr(A) = 1 \) and \( c_i = -a_{ii}/n_i (n_i - 1) \), \( i = 1, 2, \ldots, r \).

They proposed two estimates, namely, \( \bar{\sigma}^2_{r,1} \) and \( \bar{\sigma}^2_{r,2} \). If we choose \( a_{ii} = 1/r \) and \( a_{ij} = -1/(r - 1) \), \( i \neq j \), i.e.,
\[ A = \frac{1}{r - 1} (I - \frac{1}{r} (1, 1, \ldots, 1)^T), \]
we get
\[ \bar{\sigma}^2_{r,1} = \frac{1}{r - 1} (\sum_{i=1}^{r} (\bar{y}_i - \bar{y}_{\text{known}})^2 / (r - 1) - \sum_{i=1}^{r} S_i^2 / n_i (n_i - 1)), \]
where \( \bar{y}_{\text{known}} = \sum_{i=1}^{r} \bar{y}_i / r \).

On the other hand, the choice \( a_{ii} = \frac{n_i (n - n_i)}{n^2 - \sum n_i^2} \), and \( a_{ij} = -\frac{n_i n_j}{n^2 - \sum n_i^2} \), leads to
\[ \bar{\sigma}^2_{r,2} = \frac{1}{n - \sum n_i^2} \left[ \sum_{i=1}^{r} n_i (\bar{y}_i - \bar{y}_..)^2 - \sum_{i=1}^{r} \frac{(n - n_i) S_i^2}{n_i (n_i - 1)} \right]. \]

Observe that in the balanced case, i.e., when \( n_1 = n_2 = \cdots = n_r \), the estimates are identical.
They also considered a perturbed estimate given by
\[ \hat{\sigma}_2^{2,p} = CY'IY - \sum_{i=1}^{r} \frac{d_i a_i S_i^2}{n_i(n_i - 1)}, \]  
(3.9)

where \( c \) and \( d_i \)'s are chosen as \( d_i = (n_i - 1)/(n_i + 1) \) and \( c = (1 + 2 \sum_{i,j} a_{ij}^2)^{-1} \).

Using the above form of \( \hat{\sigma}_2^{2,p} \) with specific choices of \( c, d_i \)'s and \( A \), we can get \( \hat{\sigma}_2^{2,p1} \) and \( \hat{\sigma}_2^{2,p2} \) corresponding to \( \hat{\sigma}_2^{2,1} \) and \( \hat{\sigma}_2^{2,2} \). We used these two perturbed estimates of \( \sigma_2^2 \) and then we incorporated them in \( \hat{\mu} \).

For details, we refer to a technical report by Mathew et. al. (2006).

For heterogeneous error variances with equal or unequal replications, a solution based on generalized p-value is reported in Iyer, Wang and, Mathew (2004). But the method is quite complicated to implement. We omit this method in our simulation study.

4 Simulation

A brief description of the simulation procedure is given in this Section, followed by the Tables of simulated values of coverage probabilities and powers.

Throughout we have taken \( \mu = 0 \) to generate samples in order to obtain Monte-Carlo estimate of \( \mu \) (Mean), its variance (Var) and median (Med), expected length of confidence intervals (EL) and coverage probabilities (CP). To simulate power, we select \( H_0 : \mu = 0 \) as the null hypothesis and \( H_1 : \mu = 1 \) as the alternative hypothesis. In the case of homogeneous error variances, we took \( \sigma^2 = 1 \).

The following steps briefly sketch the procedure of simulation.

Step 1. Generate \( n_i \ (i = 1, 2, \cdots, r) \) random samples of \( y \) from \( i \)th population, following the basic model:

\[ y_{ij} = \mu + a_i + e_{ij} \]  
(4.1)

This is done by generating \( (a) \) \( n_i \) random samples from the error distribution, \( e_{ij} \), which is normal with mean 0 and variance \( \sigma^2 \) (homogeneous) or \( \sigma^2_i \) (heterogeneous); \( b \) \( a_i \) from another normal with mean 0 and variance \( \sigma^2 \) independently from the generation of errors. Of course, we take \( \mu = 0 \) or \( \mu = 1 \) depending on computation of coverage probability or power.

Step 2. Compute \( \hat{\sigma}^2 \) (homogeneous) or \( \hat{\sigma}^2_i \) (heterogeneous) based on the sample error sum of squares.

Step 3. Use different methods described in Sections 2 and 3 to estimate \( \eta = \sigma^2 / \sigma^2 \) (homogeneous) or \( \sigma^2_i \).

Step 4. Compute \( \hat{\mu} \) [Equation (1.1)] to get a point estimate of the common mean \( \mu \).

Step 5. Compute \( \hat{\var}(\mu) \) by using Equations (1.3) and (1.4).

Step 6. Compute the confidence intervals based on standard \( z \) statistic and \( t \) statistic described in Section 1.

Step 7. Repeat steps 1-6 10,000 times.

Step 8. The simulated coverage probabilities are then obtained in the usual way by counting the number of times the confidence intervals in step 6 contain the true parameter \( \mu = 0 \). As a byproduct of our simulation, we compute Monte Carlo estimate of mean, variance, median and also expected length.

To compute power, samples are generated under the alternative hypothesis, and at step 8, we count the proportion of times the test statistics, computed under different methods described in the preceding Sections, are contained in the corresponding critical regions. The coverage probabilities and powers are given in percentages( %). For example, 94.39 should be read as 94.39%.
lengths are rounded to the second decimal place and other figures are rounded to the fourth decimal place.

Several notations, which have been used in different Tables, are described below.

**Homogeneous case:**
HK: Hartung-Knapp method, Section 2.3, Equation (2.5)
TH: Thomas Hultquist method, Section 2.1, Equation (2.3)
BE: Burdick Eckman method, Section 2.2, Equation (2.4)
A: Method of moments, Section 2.1, Equation (2.2)
GP: Generalized p-value method, Section 2.4, Equations (2.6), (2.7)
P1–P4: methods proposed by Mathew, Sinha and Sutradhar (1992), Section 2.4, Equations (2.8), (2.9), (2.10), (2.11)

**Observations and Conclusions: homogeneous case**
In terms of coverage probabilities, it is found that the t-method performs quite well. On the other hand, the simulated coverage probability under standard z-method never exceeds 90%, which is well below the stipulated 95%. Powers are also reasonable in t-method although we cannot compare the powers between z and t methods have different sizes.

In our simulation, P1–P4 methods are producing very similar results although P1 yields maximum coverage probabilities. The methods HK, TH, BE are also producing similar coverage probabilities although they differ slightly from each other in expected lengths (EL). We can conclude methods P1–P4 are producing uniformly better EL than other methods (HK, TH etc.). Also Generalized p-value method (GP) is performing better than all others although it is producing CP larger than the target. This is evident from Tables 3 – 7.

Notations used in the heterogeneous case are given below.

**Heterogeneous case:**
DL: DerSimonian and Laird method, Section 3.1, Equation (3.1)
UN: Rao, Kaplan and Cochran method, Section 3.2, Equation (3.3)
AN: Rao, Kaplan and Cochran method, Section 3.2, Equation (3.2)
POS: Rao, Kaplan and Cochran method, Section 3.2, Equation (3.4)
MP: Mandel and Paul method, Section 3.3, Equation (3.5)
MPKR: Kenward-Roger method with MP estimates, Section 3.4, Equation (3.6)
NNA, NNB: Other method, Section 3.4, Equation (3.9)
The values of the parameters are reported in each table along with the values of r (the total number of populations) and n_i’s (the number of samples from each population).

Note that in MPKR procedure, we have only t method with an approximate degrees of freedom. There is no z method and thus "**" appears in the tables, representing blanks.

**Observations and Conclusions: heterogeneous case**
In terms of coverage probability, it is found that the t-method is performing better (targeted CP = 95%) than the standard z-method. It is to be noted that the methods AN, UN and MP are producing very similar outputs in the case of two populations. The method POS is performing better than other methods although in two population case its coverage probability is not always satisfactory. But when we have three populations and sample sizes are moderate, all methods are performing reasonably well in terms of coverage probabilities although POS method and sometimes MPKR method are the best among all these methods. When number of populations is more than 2, MPKR method is performing slightly better than MP method in terms of CP although it has larger EL. This is again evident from Tables 8–14.
5 Applications

Three real life datasets are considered as applications of the above methods. The three datasets are due to Yu, Sun and Sinha (2002), Meier (1953), and Eberhardt et al. (1989), respectively. The results based on several methods are reported in Tables EPA2, EPA3, 16, 17, 18, and 19. We use HK and ST to denote the estimated variances computed using the Equations 1.4 and 1.3.

Our first data set deals with the problem of estimating the mean RVP of new reformulated gasoline in U.S. Since the laboratory analysis of gasoline sample to measure RVP is quite costly, usually data are collected from both labs and fields, and the problem is to combine the two data sets for efficient inference. A summary of a sample data set is given in Table EPA1.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>( n_i )</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory</td>
<td>15</td>
<td>8.283</td>
<td>0.2454</td>
</tr>
<tr>
<td>Field</td>
<td>15</td>
<td>8.239</td>
<td>0.2848</td>
</tr>
</tbody>
</table>

Our second data set is concerned with the outcomes of four experiments used to estimate the mean percentage of albumin in the plasma protein of normal human subjects [see Table 1]. The third example deals with the problem of estimation of mean selenium in non-fat milk powder by combining results of four methods [see Table 2].

<table>
<thead>
<tr>
<th>Experiment</th>
<th>( n_i )</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>12</td>
<td>62.3</td>
<td>12.380</td>
</tr>
<tr>
<td>B</td>
<td>15</td>
<td>60.3</td>
<td>7.840</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>59.5</td>
<td>33.433</td>
</tr>
<tr>
<td>D</td>
<td>16</td>
<td>61.5</td>
<td>18.513</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>methods</th>
<th>( n_i )</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomic absorption spectrometry</td>
<td>8</td>
<td>105.0</td>
<td>85.711</td>
</tr>
<tr>
<td>Neutron activation:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrumental</td>
<td>12</td>
<td>109.75</td>
<td>20.748</td>
</tr>
<tr>
<td>Radiochemical</td>
<td>14</td>
<td>109.5</td>
<td>2.729</td>
</tr>
<tr>
<td>Isotope dilution mass spectrometry</td>
<td>8</td>
<td>113.25</td>
<td>33.640</td>
</tr>
</tbody>
</table>

We can see from the Tables EPA2, EPA3, 16, 17, 18, and 19 that Hartung-Knapp (2005) method yields the shortest interval based on t-method and also yields smallest variances in the case of homogeneous error variances. It turns out that the two variance estimates, namely, HK and ST, differ considerably in the case of homogeneous error variances while they are quite close in the case of heterogeneous error variances. Moreover, the performance of GP method is satisfactory although its implementation requires detailed simulation. We should also point out that the method NNB yields the smallest confidence intervals although its coverage level is short of the desired level.

127
Table 3: Comparison of different methods in homogeneous case when \( r = 2, n_1 = 2, n_2 = 10 \)

<table>
<thead>
<tr>
<th>( \sigma_0^2 = 0.5 )</th>
<th>Mean</th>
<th>Var</th>
<th>Mol</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>0.0125</td>
<td>0.3905</td>
<td>0.0225</td>
<td>12.39</td>
<td>3.16</td>
<td>93.81</td>
</tr>
<tr>
<td>TH</td>
<td>0.0125</td>
<td>0.3909</td>
<td>0.0229</td>
<td>12.40</td>
<td>3.25</td>
<td>93.81</td>
</tr>
<tr>
<td>BE</td>
<td>0.0125</td>
<td>0.3905</td>
<td>0.0225</td>
<td>12.39</td>
<td>3.16</td>
<td>93.81</td>
</tr>
<tr>
<td>A</td>
<td>0.0123</td>
<td>0.3882</td>
<td>0.0194</td>
<td>11.84</td>
<td>2.07</td>
<td>93.81</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>9.64</td>
<td>*</td>
<td>97.45</td>
</tr>
<tr>
<td>P1</td>
<td>0.0118</td>
<td>0.3809</td>
<td>0.0191</td>
<td>12.92</td>
<td>1.88</td>
<td>93.93</td>
</tr>
<tr>
<td>P2</td>
<td>0.0043</td>
<td>0.4394</td>
<td>0.0079</td>
<td>9.60</td>
<td>1.13</td>
<td>93.81</td>
</tr>
<tr>
<td>P3</td>
<td>0.0118</td>
<td>0.3809</td>
<td>0.0191</td>
<td>12.92</td>
<td>1.88</td>
<td>93.93</td>
</tr>
<tr>
<td>P4</td>
<td>0.0043</td>
<td>0.4394</td>
<td>0.0079</td>
<td>9.60</td>
<td>1.13</td>
<td>93.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \sigma_0^2 = 1.0 )</th>
<th>Mean</th>
<th>Var</th>
<th>Mol</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>0.0132</td>
<td>0.6456</td>
<td>0.0248</td>
<td>15.97</td>
<td>3.97</td>
<td>93.69</td>
</tr>
<tr>
<td>TH</td>
<td>0.0132</td>
<td>0.6456</td>
<td>0.0242</td>
<td>15.98</td>
<td>4.10</td>
<td>93.70</td>
</tr>
<tr>
<td>BE</td>
<td>0.0132</td>
<td>0.6454</td>
<td>0.0248</td>
<td>15.97</td>
<td>3.97</td>
<td>93.69</td>
</tr>
<tr>
<td>A</td>
<td>0.0136</td>
<td>0.6482</td>
<td>0.0213</td>
<td>15.47</td>
<td>2.56</td>
<td>93.68</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>11.57</td>
<td>96.58</td>
<td>7.44</td>
</tr>
<tr>
<td>P1</td>
<td>0.0131</td>
<td>0.6428</td>
<td>0.0210</td>
<td>15.58</td>
<td>2.16</td>
<td>93.78</td>
</tr>
<tr>
<td>P2</td>
<td>0.0052</td>
<td>0.7954</td>
<td>0.0077</td>
<td>12.20</td>
<td>1.32</td>
<td>93.68</td>
</tr>
<tr>
<td>P3</td>
<td>0.0131</td>
<td>0.6428</td>
<td>0.0210</td>
<td>15.58</td>
<td>2.16</td>
<td>93.78</td>
</tr>
<tr>
<td>P4</td>
<td>0.0052</td>
<td>0.7954</td>
<td>0.0077</td>
<td>12.20</td>
<td>1.32</td>
<td>93.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \sigma_0^2 = 1.5 )</th>
<th>Mean</th>
<th>Var</th>
<th>Mol</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>0.0138</td>
<td>0.8969</td>
<td>0.0244</td>
<td>18.89</td>
<td>4.65</td>
<td>93.44</td>
</tr>
<tr>
<td>TH</td>
<td>0.0138</td>
<td>0.8971</td>
<td>0.0238</td>
<td>18.90</td>
<td>4.80</td>
<td>93.44</td>
</tr>
<tr>
<td>BE</td>
<td>0.0138</td>
<td>0.8969</td>
<td>0.0234</td>
<td>18.89</td>
<td>4.65</td>
<td>93.44</td>
</tr>
<tr>
<td>A</td>
<td>0.0138</td>
<td>0.9019</td>
<td>0.0198</td>
<td>18.44</td>
<td>2.97</td>
<td>93.36</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>13.03</td>
<td>96.41</td>
<td>6.68</td>
</tr>
<tr>
<td>P1</td>
<td>0.0138</td>
<td>0.8986</td>
<td>0.0191</td>
<td>18.51</td>
<td>2.41</td>
<td>93.66</td>
</tr>
<tr>
<td>P2</td>
<td>0.0059</td>
<td>1.151</td>
<td>0.0017</td>
<td>14.34</td>
<td>1.33</td>
<td>93.35</td>
</tr>
<tr>
<td>P3</td>
<td>0.0138</td>
<td>0.8986</td>
<td>0.0191</td>
<td>18.51</td>
<td>2.41</td>
<td>93.66</td>
</tr>
<tr>
<td>P4</td>
<td>0.0061</td>
<td>1.149</td>
<td>0.0018</td>
<td>14.37</td>
<td>1.33</td>
<td>93.35</td>
</tr>
</tbody>
</table>

Table EPA2: Comparison of different methods (homogeneous case) using an EPA dataset

<table>
<thead>
<tr>
<th>Mean</th>
<th>Variance</th>
<th>t-method</th>
<th>z-method</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>0.0005</td>
<td>0.0333</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>TH</td>
<td>0.0005</td>
<td>0.0333</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>BE</td>
<td>0.0005</td>
<td>0.0333</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>A</td>
<td>0.0005</td>
<td>0.0333</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>P1</td>
<td>0.0005</td>
<td>0.0333</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>P2</td>
<td>0.0005</td>
<td>0.0333</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>P3</td>
<td>0.0005</td>
<td>0.0333</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>P4</td>
<td>0.0005</td>
<td>0.0333</td>
<td>(7.98, 8.54)</td>
</tr>
</tbody>
</table>

Table EPA3: Comparison of different methods (heterogeneous case) using an EPA dataset

128
Table 4: Comparison of different methods in homogeneous case when $r = 2, n_1 = 4, n_2 = 6$

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>0.0077</td>
<td>0.3561</td>
<td>0.0083</td>
<td>12.05</td>
<td>3.09</td>
<td>95.04</td>
</tr>
<tr>
<td>TH</td>
<td>0.0077</td>
<td>0.3561</td>
<td>0.0079</td>
<td>12.05</td>
<td>3.19</td>
<td>95.04</td>
</tr>
<tr>
<td>BE</td>
<td>0.0077</td>
<td>0.3561</td>
<td>0.0083</td>
<td>12.05</td>
<td>3.09</td>
<td>95.04</td>
</tr>
<tr>
<td>A</td>
<td>0.0079</td>
<td>0.3563</td>
<td>0.0097</td>
<td>12.01</td>
<td>2.06</td>
<td>95.04</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>9.14</td>
<td></td>
<td>98.07</td>
</tr>
<tr>
<td>P1</td>
<td>0.0080</td>
<td>0.3559</td>
<td>0.0099</td>
<td>12.02</td>
<td>1.84</td>
<td>95.04</td>
</tr>
<tr>
<td>P2</td>
<td>0.0088</td>
<td>0.3630</td>
<td>0.0092</td>
<td>11.83</td>
<td>1.24</td>
<td>95.04</td>
</tr>
<tr>
<td>P3</td>
<td>0.0080</td>
<td>0.3559</td>
<td>0.0099</td>
<td>12.02</td>
<td>1.84</td>
<td>95.04</td>
</tr>
<tr>
<td>P4</td>
<td>0.0088</td>
<td>0.3630</td>
<td>0.0092</td>
<td>11.83</td>
<td>1.24</td>
<td>95.04</td>
</tr>
<tr>
<td>$\sigma^2 = 1.0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>0.0106</td>
<td>0.6094</td>
<td>0.0145</td>
<td>15.76</td>
<td>3.94</td>
<td>94.99</td>
</tr>
<tr>
<td>TH</td>
<td>0.0106</td>
<td>0.6094</td>
<td>0.0144</td>
<td>15.76</td>
<td>4.09</td>
<td>94.99</td>
</tr>
<tr>
<td>BE</td>
<td>0.0106</td>
<td>0.6094</td>
<td>0.0145</td>
<td>15.76</td>
<td>3.94</td>
<td>94.99</td>
</tr>
<tr>
<td>A</td>
<td>0.0107</td>
<td>0.6096</td>
<td>0.0157</td>
<td>15.73</td>
<td>2.54</td>
<td>94.99</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>11.19</td>
<td></td>
<td>96.94</td>
</tr>
<tr>
<td>P1</td>
<td>0.0109</td>
<td>0.6096</td>
<td>0.0151</td>
<td>15.73</td>
<td>2.13</td>
<td>95.00</td>
</tr>
<tr>
<td>P2</td>
<td>0.0122</td>
<td>0.6261</td>
<td>0.1934</td>
<td>15.47</td>
<td>1.24</td>
<td>94.99</td>
</tr>
<tr>
<td>P3</td>
<td>0.0109</td>
<td>0.6096</td>
<td>0.0151</td>
<td>15.73</td>
<td>2.13</td>
<td>95.00</td>
</tr>
<tr>
<td>P4</td>
<td>0.0122</td>
<td>0.6261</td>
<td>0.1934</td>
<td>15.47</td>
<td>1.24</td>
<td>94.99</td>
</tr>
<tr>
<td>$\sigma^2 = 1.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>0.0129</td>
<td>0.8622</td>
<td>0.0198</td>
<td>18.75</td>
<td>4.64</td>
<td>94.93</td>
</tr>
<tr>
<td>TH</td>
<td>0.0129</td>
<td>0.8624</td>
<td>0.0196</td>
<td>18.75</td>
<td>4.83</td>
<td>94.93</td>
</tr>
<tr>
<td>BE</td>
<td>0.0129</td>
<td>0.8624</td>
<td>0.0198</td>
<td>18.75</td>
<td>4.64</td>
<td>94.93</td>
</tr>
<tr>
<td>A</td>
<td>0.0129</td>
<td>0.8629</td>
<td>0.0233</td>
<td>18.72</td>
<td>2.95</td>
<td>94.93</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>12.69</td>
<td></td>
<td>96.78</td>
</tr>
<tr>
<td>P1</td>
<td>0.0131</td>
<td>0.8629</td>
<td>0.0228</td>
<td>18.72</td>
<td>2.38</td>
<td>94.95</td>
</tr>
<tr>
<td>P2</td>
<td>0.0148</td>
<td>0.8894</td>
<td>0.0287</td>
<td>18.39</td>
<td>1.24</td>
<td>94.93</td>
</tr>
<tr>
<td>P3</td>
<td>0.0131</td>
<td>0.8629</td>
<td>0.0228</td>
<td>18.72</td>
<td>2.38</td>
<td>94.95</td>
</tr>
<tr>
<td>P4</td>
<td>0.0148</td>
<td>0.8894</td>
<td>0.0287</td>
<td>18.39</td>
<td>1.24</td>
<td>94.93</td>
</tr>
</tbody>
</table>
Table 5: Comparison of different methods in homogeneous case when $r = 3$, $n_1 = 2$, $n_2 = 4$, $n_3 = 6$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma^2 = 0.5$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Var</td>
<td>Mod</td>
<td>EL</td>
<td>CP</td>
<td>POW</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
</tr>
<tr>
<td>HK</td>
<td>0.0010</td>
<td>0.2638</td>
<td>-0.0003</td>
<td>3.88</td>
<td>2.32</td>
<td>94.52</td>
<td>89.39</td>
<td>22.27</td>
</tr>
<tr>
<td>TH</td>
<td>0.0011</td>
<td>0.2640</td>
<td>-0.0005</td>
<td>3.88</td>
<td>2.38</td>
<td>94.56</td>
<td>89.51</td>
<td>22.15</td>
</tr>
<tr>
<td>BE</td>
<td>0.0011</td>
<td>0.2639</td>
<td>-0.0003</td>
<td>3.89</td>
<td>2.29</td>
<td>94.54</td>
<td>88.95</td>
<td>22.19</td>
</tr>
<tr>
<td>A</td>
<td>0.0010</td>
<td>0.2638</td>
<td>-0.0018</td>
<td>3.86</td>
<td>1.83</td>
<td>94.47</td>
<td>85.02</td>
<td>22.61</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>4.02</td>
<td>97.36</td>
<td></td>
<td>18.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>0.0013</td>
<td>0.2616</td>
<td>-0.0018</td>
<td>3.86</td>
<td>1.89</td>
<td>94.68</td>
<td>89.43</td>
<td>21.94</td>
</tr>
<tr>
<td>P2</td>
<td>0.0024</td>
<td>0.2744</td>
<td>0.0006</td>
<td>3.69</td>
<td>1.13</td>
<td>94.40</td>
<td>71.90</td>
<td>23.41</td>
</tr>
<tr>
<td>P3</td>
<td>0.0013</td>
<td>0.2614</td>
<td>-0.0018</td>
<td>3.86</td>
<td>1.84</td>
<td>94.64</td>
<td>88.78</td>
<td>22.27</td>
</tr>
<tr>
<td>P4</td>
<td>0.0024</td>
<td>0.2744</td>
<td>0.0006</td>
<td>3.69</td>
<td>1.13</td>
<td>94.40</td>
<td>71.90</td>
<td>23.41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma^2 = 1.0$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Var</td>
<td>Mod</td>
<td>EL</td>
<td>CP</td>
<td>POW</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
</tr>
<tr>
<td>HK</td>
<td>0.0015</td>
<td>0.4327</td>
<td>0.0002</td>
<td>4.98</td>
<td>2.92</td>
<td>94.60</td>
<td>87.64</td>
<td>16.14</td>
</tr>
<tr>
<td>TH</td>
<td>0.0015</td>
<td>0.4328</td>
<td>0.0006</td>
<td>4.99</td>
<td>3.01</td>
<td>94.65</td>
<td>87.85</td>
<td>16.03</td>
</tr>
<tr>
<td>BE</td>
<td>0.0015</td>
<td>0.4329</td>
<td>0.0001</td>
<td>4.98</td>
<td>2.89</td>
<td>94.62</td>
<td>87.06</td>
<td>16.94</td>
</tr>
<tr>
<td>A</td>
<td>0.0012</td>
<td>0.4334</td>
<td>-0.0002</td>
<td>4.96</td>
<td>2.26</td>
<td>94.49</td>
<td>82.06</td>
<td>16.54</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>5.05</td>
<td>96.40</td>
<td></td>
<td>13.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>0.0017</td>
<td>0.4309</td>
<td>0.0004</td>
<td>4.98</td>
<td>2.20</td>
<td>94.77</td>
<td>85.23</td>
<td>15.97</td>
</tr>
<tr>
<td>P2</td>
<td>0.0029</td>
<td>0.4651</td>
<td>0.0023</td>
<td>4.74</td>
<td>1.13</td>
<td>94.27</td>
<td>59.70</td>
<td>17.20</td>
</tr>
<tr>
<td>P3</td>
<td>0.0029</td>
<td>0.4311</td>
<td>0.0004</td>
<td>4.95</td>
<td>2.15</td>
<td>94.72</td>
<td>84.52</td>
<td>16.24</td>
</tr>
<tr>
<td>P4</td>
<td>0.0029</td>
<td>0.4651</td>
<td>0.0023</td>
<td>4.74</td>
<td>1.13</td>
<td>94.27</td>
<td>59.70</td>
<td>17.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma^2 = 1.5$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Var</td>
<td>Mod</td>
<td>EL</td>
<td>CP</td>
<td>POW</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
</tr>
<tr>
<td>HK</td>
<td>0.0021</td>
<td>0.6003</td>
<td>0.0027</td>
<td>5.87</td>
<td>3.42</td>
<td>94.89</td>
<td>86.71</td>
<td>13.28</td>
</tr>
<tr>
<td>TH</td>
<td>0.0021</td>
<td>0.6003</td>
<td>0.0019</td>
<td>5.88</td>
<td>3.53</td>
<td>94.97</td>
<td>87.01</td>
<td>13.15</td>
</tr>
<tr>
<td>BE</td>
<td>0.0021</td>
<td>0.6004</td>
<td>0.0036</td>
<td>5.88</td>
<td>3.39</td>
<td>94.92</td>
<td>86.24</td>
<td>13.18</td>
</tr>
<tr>
<td>A</td>
<td>0.0013</td>
<td>0.6016</td>
<td>0.0029</td>
<td>5.86</td>
<td>2.63</td>
<td>94.75</td>
<td>80.46</td>
<td>13.44</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>5.89</td>
<td>95.96</td>
<td></td>
<td>11.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>0.0021</td>
<td>0.5991</td>
<td>0.0048</td>
<td>5.87</td>
<td>2.47</td>
<td>95.04</td>
<td>82.57</td>
<td>13.17</td>
</tr>
<tr>
<td>P2</td>
<td>0.0034</td>
<td>0.6560</td>
<td>0.0032</td>
<td>5.59</td>
<td>1.13</td>
<td>94.45</td>
<td>51.75</td>
<td>14.39</td>
</tr>
<tr>
<td>P3</td>
<td>0.0021</td>
<td>0.5994</td>
<td>0.0040</td>
<td>5.84</td>
<td>2.41</td>
<td>94.98</td>
<td>81.70</td>
<td>13.26</td>
</tr>
<tr>
<td>P4</td>
<td>0.0034</td>
<td>0.6560</td>
<td>0.0032</td>
<td>5.59</td>
<td>1.13</td>
<td>94.45</td>
<td>51.75</td>
<td>14.39</td>
</tr>
</tbody>
</table>
Table 6: Comparison of different methods in homogeneous case when $r = 4, n_1 = 2, n_2 = 3, n_3 = 4, n_4 = 5$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma^2 = 0.5$</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$t$</td>
<td>$z$</td>
<td>$t$</td>
<td>$z$</td>
<td>$t$</td>
<td>$z$</td>
</tr>
<tr>
<td>HK</td>
<td>0.0065</td>
<td>0.2063</td>
<td>0.0029</td>
<td>2.63</td>
<td>1.96</td>
<td>94.39</td>
<td>90.71</td>
</tr>
<tr>
<td>TH</td>
<td>0.0064</td>
<td>0.2063</td>
<td>0.0031</td>
<td>2.63</td>
<td>2.00</td>
<td>94.43</td>
<td>90.83</td>
</tr>
<tr>
<td>BE</td>
<td>0.0065</td>
<td>0.2062</td>
<td>0.0027</td>
<td>2.63</td>
<td>1.93</td>
<td>94.41</td>
<td>90.17</td>
</tr>
<tr>
<td>A</td>
<td>0.0065</td>
<td>0.2061</td>
<td>0.0004</td>
<td>2.62</td>
<td>1.65</td>
<td>94.30</td>
<td>87.10</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>2.72</td>
<td></td>
<td>97.23</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>0.0064</td>
<td>0.2054</td>
<td>0.0018</td>
<td>2.63</td>
<td>1.79</td>
<td>94.54</td>
<td>92.02</td>
</tr>
<tr>
<td>P2</td>
<td>0.0066</td>
<td>0.2052</td>
<td>0.0005</td>
<td>2.61</td>
<td>1.36</td>
<td>94.37</td>
<td>84.91</td>
</tr>
<tr>
<td>P3</td>
<td>0.0064</td>
<td>0.2053</td>
<td>0.0011</td>
<td>2.62</td>
<td>1.75</td>
<td>94.51</td>
<td>91.72</td>
</tr>
<tr>
<td>P4</td>
<td>0.0066</td>
<td>0.2055</td>
<td>0.0014</td>
<td>2.60</td>
<td>1.31</td>
<td>94.33</td>
<td>83.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma^2 = 1.0$</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$t$</td>
<td>$z$</td>
<td>$t$</td>
<td>$z$</td>
<td>$t$</td>
<td>$z$</td>
</tr>
<tr>
<td>HK</td>
<td>0.0079</td>
<td>0.3334</td>
<td>0.0026</td>
<td>3.94</td>
<td>2.45</td>
<td>94.43</td>
<td>89.35</td>
</tr>
<tr>
<td>TH</td>
<td>0.0080</td>
<td>0.3334</td>
<td>0.0024</td>
<td>3.35</td>
<td>2.52</td>
<td>94.47</td>
<td>89.59</td>
</tr>
<tr>
<td>BE</td>
<td>0.0080</td>
<td>0.3334</td>
<td>0.0027</td>
<td>3.55</td>
<td>2.43</td>
<td>94.46</td>
<td>88.82</td>
</tr>
<tr>
<td>A</td>
<td>0.0080</td>
<td>0.3335</td>
<td>0.0044</td>
<td>3.35</td>
<td>2.03</td>
<td>94.33</td>
<td>84.52</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>3.40</td>
<td></td>
<td>96.56</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>0.0080</td>
<td>0.3327</td>
<td>0.0028</td>
<td>3.35</td>
<td>2.09</td>
<td>94.58</td>
<td>88.59</td>
</tr>
<tr>
<td>P2</td>
<td>0.0084</td>
<td>0.3335</td>
<td>0.0049</td>
<td>3.33</td>
<td>1.51</td>
<td>94.32</td>
<td>78.74</td>
</tr>
<tr>
<td>P3</td>
<td>0.0081</td>
<td>0.3327</td>
<td>0.0035</td>
<td>3.34</td>
<td>2.05</td>
<td>94.52</td>
<td>88.10</td>
</tr>
<tr>
<td>P4</td>
<td>0.0081</td>
<td>0.3341</td>
<td>0.0053</td>
<td>3.32</td>
<td>1.43</td>
<td>94.26</td>
<td>76.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma^2 = 1.5$</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$t$</td>
<td>$z$</td>
<td>$t$</td>
<td>$z$</td>
<td>$t$</td>
<td>$z$</td>
</tr>
<tr>
<td>HK</td>
<td>0.0091</td>
<td>0.4599</td>
<td>0.0089</td>
<td>3.94</td>
<td>2.87</td>
<td>94.48</td>
<td>88.66</td>
</tr>
<tr>
<td>TH</td>
<td>0.0091</td>
<td>0.4599</td>
<td>0.0088</td>
<td>3.94</td>
<td>2.95</td>
<td>94.51</td>
<td>89.14</td>
</tr>
<tr>
<td>BE</td>
<td>0.0091</td>
<td>0.4599</td>
<td>0.0087</td>
<td>3.94</td>
<td>2.85</td>
<td>94.53</td>
<td>88.41</td>
</tr>
<tr>
<td>A</td>
<td>0.0092</td>
<td>0.4600</td>
<td>0.0098</td>
<td>3.94</td>
<td>2.37</td>
<td>94.45</td>
<td>83.80</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>3.97</td>
<td></td>
<td>96.21</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>0.0096</td>
<td>0.4594</td>
<td>0.0112</td>
<td>3.94</td>
<td>2.35</td>
<td>94.63</td>
<td>86.37</td>
</tr>
<tr>
<td>P2</td>
<td>0.0096</td>
<td>0.4626</td>
<td>0.0080</td>
<td>3.92</td>
<td>1.64</td>
<td>94.51</td>
<td>74.79</td>
</tr>
<tr>
<td>P3</td>
<td>0.0093</td>
<td>0.4593</td>
<td>0.0109</td>
<td>3.93</td>
<td>2.31</td>
<td>94.56</td>
<td>86.08</td>
</tr>
<tr>
<td>P4</td>
<td>0.0097</td>
<td>0.4615</td>
<td>0.0077</td>
<td>3.90</td>
<td>1.55</td>
<td>94.43</td>
<td>72.62</td>
</tr>
</tbody>
</table>

131
Table 7: Comparison of different methods in homogeneous case when $r = 5, n_1 = 2, n_2 = 2, n_3 = 4, n_4 = 4, n_5 = 6$

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td>t</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>-0.0038</td>
<td>0.1654</td>
<td>-0.0044</td>
<td>2.10</td>
<td>1.70</td>
<td>94.15</td>
</tr>
<tr>
<td>TH</td>
<td>-0.0037</td>
<td>0.1655</td>
<td>-0.0041</td>
<td>2.11</td>
<td>1.73</td>
<td>94.20</td>
</tr>
<tr>
<td>BE</td>
<td>-0.0035</td>
<td>0.1655</td>
<td>-0.0043</td>
<td>2.10</td>
<td>1.70</td>
<td>94.19</td>
</tr>
<tr>
<td>A</td>
<td>-0.0035</td>
<td>0.1657</td>
<td>-0.0051</td>
<td>2.10</td>
<td>1.49</td>
<td>94.10</td>
</tr>
<tr>
<td>GP</td>
<td></td>
<td></td>
<td></td>
<td>2.16</td>
<td>96.72</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>-0.0037</td>
<td>0.1645</td>
<td>-0.0042</td>
<td>2.11</td>
<td>1.66</td>
<td>94.35</td>
</tr>
<tr>
<td>P2</td>
<td>-0.0035</td>
<td>0.1644</td>
<td>-0.0053</td>
<td>2.10</td>
<td>1.36</td>
<td>94.33</td>
</tr>
<tr>
<td>P3</td>
<td>-0.0037</td>
<td>0.1645</td>
<td>-0.0050</td>
<td>2.10</td>
<td>1.60</td>
<td>94.24</td>
</tr>
<tr>
<td>P4</td>
<td>-0.0034</td>
<td>0.1646</td>
<td>-0.0049</td>
<td>2.08</td>
<td>1.28</td>
<td>94.30</td>
</tr>
<tr>
<td>$\sigma^2 = 1.0$</td>
<td></td>
<td></td>
<td></td>
<td>t</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>-0.0058</td>
<td>0.2681</td>
<td>-0.0026</td>
<td>2.57</td>
<td>2.14</td>
<td>94.24</td>
</tr>
<tr>
<td>TH</td>
<td>-0.0058</td>
<td>0.2682</td>
<td>-0.0032</td>
<td>2.68</td>
<td>2.18</td>
<td>94.30</td>
</tr>
<tr>
<td>BE</td>
<td>-0.0058</td>
<td>0.2681</td>
<td>-0.0031</td>
<td>2.68</td>
<td>2.14</td>
<td>94.29</td>
</tr>
<tr>
<td>A</td>
<td>-0.0055</td>
<td>0.2687</td>
<td>-0.0033</td>
<td>2.67</td>
<td>1.85</td>
<td>94.21</td>
</tr>
<tr>
<td>GP</td>
<td></td>
<td></td>
<td></td>
<td>2.71</td>
<td>95.89</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>-0.0056</td>
<td>0.2674</td>
<td>-0.0032</td>
<td>2.68</td>
<td>1.95</td>
<td>94.45</td>
</tr>
<tr>
<td>P2</td>
<td>-0.0053</td>
<td>0.2680</td>
<td>-0.0018</td>
<td>2.67</td>
<td>1.54</td>
<td>94.34</td>
</tr>
<tr>
<td>P3</td>
<td>-0.0056</td>
<td>0.2675</td>
<td>-0.0031</td>
<td>2.67</td>
<td>1.89</td>
<td>94.32</td>
</tr>
<tr>
<td>P4</td>
<td>-0.0052</td>
<td>0.2687</td>
<td>-0.0030</td>
<td>2.65</td>
<td>1.45</td>
<td>94.27</td>
</tr>
<tr>
<td>$\sigma^2 = 1.5$</td>
<td></td>
<td></td>
<td></td>
<td>t</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>-0.0074</td>
<td>0.3698</td>
<td>-0.0057</td>
<td>3.14</td>
<td>2.51</td>
<td>94.56</td>
</tr>
<tr>
<td>TH</td>
<td>-0.0074</td>
<td>0.3698</td>
<td>-0.0055</td>
<td>3.15</td>
<td>2.56</td>
<td>94.61</td>
</tr>
<tr>
<td>BE</td>
<td>-0.0074</td>
<td>0.3698</td>
<td>-0.0052</td>
<td>3.15</td>
<td>2.50</td>
<td>94.61</td>
</tr>
<tr>
<td>A</td>
<td>-0.0072</td>
<td>0.3703</td>
<td>-0.0075</td>
<td>3.14</td>
<td>2.17</td>
<td>94.48</td>
</tr>
<tr>
<td>GP</td>
<td></td>
<td></td>
<td></td>
<td>3.17</td>
<td>95.55</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>-0.0071</td>
<td>0.3693</td>
<td>-0.0064</td>
<td>3.15</td>
<td>2.20</td>
<td>94.64</td>
</tr>
<tr>
<td>P2</td>
<td>-0.0070</td>
<td>0.3703</td>
<td>-0.0057</td>
<td>3.14</td>
<td>1.71</td>
<td>94.53</td>
</tr>
<tr>
<td>P3</td>
<td>-0.0071</td>
<td>0.3695</td>
<td>-0.0060</td>
<td>3.14</td>
<td>2.14</td>
<td>94.60</td>
</tr>
<tr>
<td>P4</td>
<td>-0.0067</td>
<td>0.3713</td>
<td>-0.0060</td>
<td>3.12</td>
<td>1.59</td>
<td>94.45</td>
</tr>
</tbody>
</table>
Table 8: Comparison of different methods in heterogeneous case when \( r = 2 \)

\[
\sigma_2^2 = 0.5, \ (\sigma_1^2, \sigma_2^2) = (1,0.5) \\
\begin{array}{|c|c|c|c|c|c|c|c|c|}
\hline
(n_1, n_2) & \text{method} & \text{Mean} & \text{Var} & \text{Med} & \text{EL} & \text{CP} & \text{POW} \\
\hline
\hline
(2, 10) & DL & -0.0006 & 0.5116 & -0.0022 & 11.36 & 1.89 & 91.28 & 70.84 & 17.86 & 57.67 \\
& AN & -0.0121 & 0.3893 & -0.0153 & 11.36 & 1.89 & 91.28 & 70.84 & 17.86 & 57.67 \\
& UN & -0.0121 & 0.3893 & -0.0153 & 11.36 & 1.89 & 91.28 & 70.84 & 17.86 & 57.67 \\
& POS & 0.0105 & 0.3857 & 0.0138 & 11.50 & 1.44 & 92.08 & 68.17 & 15.67 & 67.95 \\
& MP & -0.0121 & 0.3893 & 0.0153 & 11.36 & 1.89 & 91.28 & 70.84 & 8.72 & 29.16 \\
& MPKR & -0.0121 & 0.3893 & 0.0153 & 12.99 & * & 85.93 & * & 28.35 & * \\
& NNA & 0.0104 & 0.3923 & 0.0134 & 11.13 & 1.38 & 91.50 & 65.67 & 17.29 & 69.33 \\
& NNB & -0.0061 & 0.4501 & -0.0056 & 9.05 & 0.89 & 90.30 & 45.38 & 19.38 & 80.81 \\
\hline
\hline
(10, 2) & DL & -0.0031 & 0.4829 & -0.0086 & 11.11 & 1.87 & 92.56 & 74.31 & 14.85 & 57.10 \\
& UN & -0.0020 & 0.3520 & -0.0073 & 11.11 & 1.87 & 92.56 & 74.31 & 14.85 & 57.10 \\
& AN & -0.0020 & 0.3520 & -0.0073 & 11.11 & 1.87 & 92.56 & 74.31 & 14.85 & 57.10 \\
& POS & -0.0019 & 0.3574 & -0.0084 & 11.10 & 1.41 & 93.26 & 70.30 & 13.89 & 68.36 \\
& MP & -0.0020 & 0.3520 & -0.0073 & 11.11 & 1.87 & 92.56 & 74.31 & 14.85 & 57.10 \\
& MPKR & -0.0020 & 0.3520 & -0.0073 & 11.87 & * & 92.70 & * & 20.40 & * \\
& NNA & -0.0019 & 0.3588 & 0.0076 & 11.03 & 1.41 & 93.12 & 69.77 & 14.28 & 67.90 \\
& NNB & -0.0004 & 0.4428 & -0.0068 & 9.19 & 0.88 & 90.63 & 47.99 & 18.35 & 81.16 \\
\hline
\end{array}
\]

Table 9: Comparison of different methods in heterogeneous case when \( r = 2 \)

\[
\sigma_2^2 = 0.5, \ (\sigma_1^2, \sigma_2^2) = (1,1.5) \\
\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
(n_1, n_2) & \text{method} & \text{Mean} & \text{Var} & \text{Med} & \text{EL} & \text{CP} & \text{POW} \\
\hline
\hline
(2, 10) & DL & 0.0043 & 0.5791 & 0.0138 & 12.11 & 2.11 & 92.47 & 76.11 & 14.60 & 52.39 \\
& AN & 0.0117 & 0.4267 & 0.0215 & 12.11 & 2.11 & 92.47 & 76.11 & 14.60 & 52.39 \\
& UN & 0.0117 & 0.4267 & 0.0215 & 12.11 & 2.11 & 92.47 & 76.11 & 14.60 & 52.39 \\
& POS & 0.0103 & 0.4311 & 0.0189 & 12.17 & 1.67 & 93.26 & 73.64 & 13.60 & 60.70 \\
& MP & 0.0117 & 0.4267 & 0.0215 & 12.11 & 2.11 & 92.47 & 76.11 & 14.60 & 52.39 \\
& MPKR & 0.0117 & 0.4267 & 0.0215 & 13.11 & * & 93.77 & * & 20.50 & * \\
& NNA & 0.0120 & 0.4362 & 0.0177 & 11.93 & 1.63 & 93.02 & 72.04 & 14.05 & 61.99 \\
& NNB & 0.0039 & 0.5517 & 0.0128 & 9.75 & 1.04 & 90.39 & 51.47 & 17.59 & 76.59 \\
\hline
\hline
(10, 2) & DL & -0.0001 & 0.6052 & -0.0050 & 12.13 & 2.10 & 92.24 & 73.45 & 15.07 & 53.03 \\
& AN & -0.0041 & 0.6465 & -0.0068 & 12.13 & 2.10 & 92.24 & 73.45 & 15.07 & 53.03 \\
& UN & -0.0041 & 0.6465 & -0.0068 & 12.13 & 2.10 & 92.24 & 73.45 & 15.07 & 53.03 \\
& POS & -0.0035 & 0.4606 & -0.0053 & 12.32 & 1.69 & 92.95 & 72.19 & 13.56 & 59.99 \\
& MP & -0.0041 & 0.4645 & -0.0068 & 12.13 & 2.10 & 92.24 & 73.45 & 15.07 & 53.03 \\
& MPKR & -0.0041 & 0.4645 & -0.0068 & 13.43 & * & 88.30 & * & 25.50 & * \\
& NNA & -0.0033 & 0.4717 & -0.0055 & 11.76 & 1.57 & 92.54 & 68.58 & 14.67 & 63.00 \\
& NNB & -0.0082 & 0.5332 & 0.0062 & 9.64 & 1.03 & 91.07 & 49.86 & 16.91 & 76.67 \\
\hline
\end{array}
\]
Table 10: Comparison of different methods in heterogeneous case when \( r = 2 \)

<table>
<thead>
<tr>
<th>(( n_1, n_2 ))</th>
<th>method</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 10)</td>
<td>DL</td>
<td>0.0020</td>
<td>0.8973</td>
<td>0.0023</td>
<td>15.04</td>
<td>2.44</td>
<td>91.15</td>
</tr>
<tr>
<td></td>
<td>AN</td>
<td>0.0135</td>
<td>0.6498</td>
<td>0.0235</td>
<td>15.04</td>
<td>2.44</td>
<td>91.15</td>
</tr>
<tr>
<td></td>
<td>UN</td>
<td>0.0135</td>
<td>0.6498</td>
<td>0.0235</td>
<td>15.04</td>
<td>2.44</td>
<td>91.15</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>0.0122</td>
<td>0.6509</td>
<td>0.0142</td>
<td>15.09</td>
<td>1.71</td>
<td>92.14</td>
</tr>
<tr>
<td></td>
<td>MP</td>
<td>0.0135</td>
<td>0.6498</td>
<td>0.0235</td>
<td>15.04</td>
<td>2.44</td>
<td>91.15</td>
</tr>
<tr>
<td></td>
<td>MPKR</td>
<td>0.0135</td>
<td>0.6498</td>
<td>0.0235</td>
<td>16.85</td>
<td>*</td>
<td>85.13</td>
</tr>
<tr>
<td></td>
<td>NNA</td>
<td>0.0123</td>
<td>0.6611</td>
<td>0.0053</td>
<td>14.71</td>
<td>1.66</td>
<td>91.40</td>
</tr>
<tr>
<td></td>
<td>NNB</td>
<td>0.0063</td>
<td>0.7827</td>
<td>0.0074</td>
<td>11.95</td>
<td>1.06</td>
<td>90.17</td>
</tr>
<tr>
<td>(10, 2)</td>
<td>DL</td>
<td>-0.0039</td>
<td>0.8421</td>
<td>-0.0084</td>
<td>14.95</td>
<td>2.43</td>
<td>92.77</td>
</tr>
<tr>
<td></td>
<td>AN</td>
<td>-0.0022</td>
<td>0.6049</td>
<td>-0.0055</td>
<td>14.95</td>
<td>2.43</td>
<td>92.77</td>
</tr>
<tr>
<td></td>
<td>UN</td>
<td>-0.0022</td>
<td>0.6049</td>
<td>-0.0055</td>
<td>14.95</td>
<td>2.43</td>
<td>92.77</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>-0.0022</td>
<td>0.6129</td>
<td>-0.0039</td>
<td>14.90</td>
<td>1.68</td>
<td>93.46</td>
</tr>
<tr>
<td></td>
<td>MP</td>
<td>-0.0022</td>
<td>0.6049</td>
<td>-0.0055</td>
<td>14.95</td>
<td>2.43</td>
<td>92.77</td>
</tr>
<tr>
<td></td>
<td>MPKR</td>
<td>-0.0022</td>
<td>0.6049</td>
<td>-0.0055</td>
<td>15.25</td>
<td>*</td>
<td>90.40</td>
</tr>
<tr>
<td></td>
<td>NNA</td>
<td>-0.0022</td>
<td>0.6148</td>
<td>-0.0030</td>
<td>14.83</td>
<td>1.71</td>
<td>93.25</td>
</tr>
<tr>
<td></td>
<td>NNB</td>
<td>-0.0024</td>
<td>0.7391</td>
<td>-0.0103</td>
<td>12.59</td>
<td>1.02</td>
<td>90.77</td>
</tr>
</tbody>
</table>

Table 11: Comparison of different methods in heterogeneous case when \( r = 2 \)

<table>
<thead>
<tr>
<th>(( n_1, n_2 ))</th>
<th>method</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 10)</td>
<td>DL</td>
<td>0.0032</td>
<td>0.7408</td>
<td>0.0142</td>
<td>15.65</td>
<td>2.61</td>
<td>92.23</td>
</tr>
<tr>
<td></td>
<td>AN</td>
<td>0.0136</td>
<td>0.6817</td>
<td>0.0244</td>
<td>15.65</td>
<td>2.61</td>
<td>92.23</td>
</tr>
<tr>
<td></td>
<td>UN</td>
<td>0.0136</td>
<td>0.6817</td>
<td>0.0244</td>
<td>15.65</td>
<td>2.61</td>
<td>92.23</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>0.0121</td>
<td>0.6894</td>
<td>0.0202</td>
<td>15.66</td>
<td>1.92</td>
<td>92.89</td>
</tr>
<tr>
<td></td>
<td>MP</td>
<td>0.0136</td>
<td>0.6817</td>
<td>0.0244</td>
<td>15.65</td>
<td>2.61</td>
<td>92.23</td>
</tr>
<tr>
<td></td>
<td>MPKR</td>
<td>0.0136</td>
<td>0.6817</td>
<td>0.0244</td>
<td>17.03</td>
<td>*</td>
<td>91.77</td>
</tr>
<tr>
<td></td>
<td>NNA</td>
<td>0.0123</td>
<td>0.6975</td>
<td>0.0201</td>
<td>15.40</td>
<td>1.89</td>
<td>92.71</td>
</tr>
<tr>
<td></td>
<td>NNB</td>
<td>0.0048</td>
<td>0.8722</td>
<td>0.0176</td>
<td>12.61</td>
<td>1.11</td>
<td>90.08</td>
</tr>
<tr>
<td>(10, 2)</td>
<td>DL</td>
<td>-0.0013</td>
<td>0.9933</td>
<td>-0.0088</td>
<td>15.54</td>
<td>2.58</td>
<td>91.99</td>
</tr>
<tr>
<td></td>
<td>AN</td>
<td>-0.0042</td>
<td>0.7344</td>
<td>-0.0158</td>
<td>15.54</td>
<td>2.58</td>
<td>91.99</td>
</tr>
<tr>
<td></td>
<td>UN</td>
<td>-0.0042</td>
<td>0.7344</td>
<td>-0.0158</td>
<td>15.54</td>
<td>2.58</td>
<td>91.99</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>-0.0040</td>
<td>0.7354</td>
<td>-0.0138</td>
<td>15.66</td>
<td>1.93</td>
<td>92.78</td>
</tr>
<tr>
<td></td>
<td>MP</td>
<td>-0.0042</td>
<td>0.7344</td>
<td>-0.0158</td>
<td>15.54</td>
<td>2.58</td>
<td>91.99</td>
</tr>
<tr>
<td></td>
<td>MPKR</td>
<td>-0.0042</td>
<td>0.7344</td>
<td>-0.0158</td>
<td>17.23</td>
<td>*</td>
<td>86.70</td>
</tr>
<tr>
<td></td>
<td>NNA</td>
<td>-0.0039</td>
<td>0.7535</td>
<td>-0.0124</td>
<td>15.04</td>
<td>1.83</td>
<td>92.17</td>
</tr>
<tr>
<td></td>
<td>NNB</td>
<td>-0.0002</td>
<td>0.8954</td>
<td>-0.0187</td>
<td>12.22</td>
<td>1.12</td>
<td>90.80</td>
</tr>
</tbody>
</table>
Table 12: Comparison of different methods in heterogeneous case when \( r = 3 \)

<table>
<thead>
<tr>
<th>( (n_1, n_2, n_3) )</th>
<th>method</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (5, 10, 15) )</td>
<td></td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
</tr>
<tr>
<td>DL</td>
<td>-0.0056</td>
<td>0.2458</td>
<td>-0.0018</td>
<td>3.25</td>
<td>1.51</td>
<td>93.32</td>
<td>79.58</td>
</tr>
<tr>
<td>AN</td>
<td>-0.0018</td>
<td>0.2008</td>
<td>-0.0012</td>
<td>3.29</td>
<td>1.51</td>
<td>93.54</td>
<td>80.11</td>
</tr>
<tr>
<td>UN</td>
<td>-0.0018</td>
<td>0.2010</td>
<td>-0.0018</td>
<td>3.37</td>
<td>1.51</td>
<td>93.66</td>
<td>79.70</td>
</tr>
<tr>
<td>POS</td>
<td>-0.0019</td>
<td>0.2002</td>
<td>0.0010</td>
<td>3.31</td>
<td>1.29</td>
<td>94.04</td>
<td>79.43</td>
</tr>
<tr>
<td>MP</td>
<td>-0.0017</td>
<td>0.2008</td>
<td>-0.0014</td>
<td>3.29</td>
<td>1.52</td>
<td>93.48</td>
<td>80.49</td>
</tr>
<tr>
<td>MPKR</td>
<td>-0.0017</td>
<td>0.2008</td>
<td>-0.0014</td>
<td>3.76</td>
<td>*</td>
<td>93.93</td>
<td>*</td>
</tr>
<tr>
<td>NNA</td>
<td>-0.0014</td>
<td>0.1995</td>
<td>-0.0020</td>
<td>3.38</td>
<td>1.71</td>
<td>94.32</td>
<td>85.53</td>
</tr>
<tr>
<td>NNB</td>
<td>-0.0056</td>
<td>0.2486</td>
<td>-0.0022</td>
<td>2.93</td>
<td>0.50</td>
<td>92.11</td>
<td>57.56</td>
</tr>
</tbody>
</table>

Table 13: Comparison of different methods in heterogeneous case when \( r = 3 \)

<table>
<thead>
<tr>
<th>( (n_1, n_2, n_3) )</th>
<th>method</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (5, 10, 15) )</td>
<td></td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
</tr>
<tr>
<td>DL</td>
<td>-0.0063</td>
<td>0.2620</td>
<td>-0.0066</td>
<td>3.34</td>
<td>1.57</td>
<td>93.42</td>
<td>81.39</td>
</tr>
<tr>
<td>AN</td>
<td>-0.0016</td>
<td>0.2098</td>
<td>-0.0016</td>
<td>3.36</td>
<td>1.55</td>
<td>93.50</td>
<td>81.05</td>
</tr>
<tr>
<td>UN</td>
<td>-0.0015</td>
<td>0.2096</td>
<td>0.0001</td>
<td>3.37</td>
<td>1.56</td>
<td>93.49</td>
<td>81.66</td>
</tr>
<tr>
<td>POS</td>
<td>-0.0018</td>
<td>0.2087</td>
<td>-0.0003</td>
<td>3.38</td>
<td>1.37</td>
<td>94.20</td>
<td>81.46</td>
</tr>
<tr>
<td>MP</td>
<td>-0.0017</td>
<td>0.2093</td>
<td>-0.0009</td>
<td>3.37</td>
<td>1.57</td>
<td>93.45</td>
<td>81.80</td>
</tr>
<tr>
<td>MPKR</td>
<td>-0.0017</td>
<td>0.2093</td>
<td>-0.0009</td>
<td>3.84</td>
<td>*</td>
<td>93.59</td>
<td>*</td>
</tr>
<tr>
<td>NNA</td>
<td>-0.0014</td>
<td>0.2076</td>
<td>-0.0003</td>
<td>3.40</td>
<td>1.73</td>
<td>94.37</td>
<td>86.84</td>
</tr>
<tr>
<td>NNB</td>
<td>-0.0063</td>
<td>0.2621</td>
<td>-0.0067</td>
<td>3.03</td>
<td>0.60</td>
<td>92.50</td>
<td>45.05</td>
</tr>
</tbody>
</table>

Table 14: Comparison of different methods in heterogeneous case when \( r = 3 \)

<table>
<thead>
<tr>
<th>( (n_1, n_2, n_3) )</th>
<th>method</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (5, 10, 15) )</td>
<td></td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
</tr>
<tr>
<td>DL</td>
<td>-0.0083</td>
<td>0.4749</td>
<td>-0.0052</td>
<td>4.49</td>
<td>2.06</td>
<td>93.77</td>
<td>79.20</td>
</tr>
<tr>
<td>AN</td>
<td>-0.0030</td>
<td>0.3711</td>
<td>-0.0005</td>
<td>4.52</td>
<td>2.06</td>
<td>94.00</td>
<td>79.91</td>
</tr>
<tr>
<td>UN</td>
<td>-0.0032</td>
<td>0.3710</td>
<td>-0.0008</td>
<td>4.54</td>
<td>2.06</td>
<td>94.11</td>
<td>80.22</td>
</tr>
<tr>
<td>POS</td>
<td>-0.0032</td>
<td>0.3710</td>
<td>-0.0005</td>
<td>4.54</td>
<td>1.67</td>
<td>94.41</td>
<td>76.55</td>
</tr>
<tr>
<td>MP</td>
<td>-0.0030</td>
<td>0.3710</td>
<td>-0.0009</td>
<td>4.53</td>
<td>2.07</td>
<td>94.04</td>
<td>80.39</td>
</tr>
<tr>
<td>MPKR</td>
<td>-0.0030</td>
<td>0.3710</td>
<td>-0.0009</td>
<td>4.87</td>
<td>*</td>
<td>95.28</td>
<td>*</td>
</tr>
<tr>
<td>NNA</td>
<td>-0.0028</td>
<td>0.3699</td>
<td>-0.0004</td>
<td>4.56</td>
<td>2.19</td>
<td>94.60</td>
<td>84.06</td>
</tr>
<tr>
<td>NNB</td>
<td>-0.0082</td>
<td>0.4876</td>
<td>-0.0052</td>
<td>4.01</td>
<td>0.50</td>
<td>92.18</td>
<td>45.58</td>
</tr>
</tbody>
</table>
Table 15: Comparison of different methods in heterogeneous case when $r = 3$

<table>
<thead>
<tr>
<th>$(n_1, n_2, n_3)$</th>
<th>method</th>
<th>Mean</th>
<th>Var</th>
<th>Med</th>
<th>EL</th>
<th>CP</th>
<th>POW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
<td>t</td>
<td>z</td>
</tr>
<tr>
<td>(5, 10, 15)</td>
<td>DL</td>
<td>-0.0097</td>
<td>0.4890</td>
<td>-0.0106</td>
<td>4.55</td>
<td>2.10</td>
<td>93.81</td>
</tr>
<tr>
<td></td>
<td>AN</td>
<td>-0.0029</td>
<td>0.3792</td>
<td>-0.0039</td>
<td>4.58</td>
<td>2.09</td>
<td>93.82</td>
</tr>
<tr>
<td></td>
<td>UN</td>
<td>-0.0028</td>
<td>0.3790</td>
<td>-0.0025</td>
<td>4.59</td>
<td>2.09</td>
<td>93.87</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>-0.0031</td>
<td>0.3789</td>
<td>-0.0037</td>
<td>4.59</td>
<td>1.73</td>
<td>94.25</td>
</tr>
<tr>
<td></td>
<td>MP</td>
<td>-0.0029</td>
<td>0.3787</td>
<td>-0.0023</td>
<td>4.58</td>
<td>2.10</td>
<td>93.89</td>
</tr>
<tr>
<td></td>
<td>MPKR</td>
<td>-0.0029</td>
<td>0.3787</td>
<td>-0.0023</td>
<td>4.93</td>
<td>*</td>
<td>94.86</td>
</tr>
<tr>
<td></td>
<td>NNA</td>
<td>-0.0027</td>
<td>0.3775</td>
<td>-0.0012</td>
<td>4.60</td>
<td>2.24</td>
<td>94.48</td>
</tr>
<tr>
<td></td>
<td>NNB</td>
<td>-0.0097</td>
<td>0.4890</td>
<td>-0.0010</td>
<td>4.09</td>
<td>0.61</td>
<td>92.45</td>
</tr>
</tbody>
</table>

Table 16: Comparison of different methods (homogeneous case) using data set due to Meier (1953)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>t-method</th>
<th>z-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HK</td>
<td>ST</td>
<td>CI</td>
<td>L</td>
</tr>
<tr>
<td>HK</td>
<td>61.03</td>
<td>0.328</td>
<td>0.026</td>
<td>(59.21, 62.86)</td>
</tr>
<tr>
<td>TH</td>
<td>61.02</td>
<td>0.385</td>
<td>0.028</td>
<td>(59.18, 62.86)</td>
</tr>
<tr>
<td>BE</td>
<td>61.03</td>
<td>0.327</td>
<td>0.025</td>
<td>(59.21, 62.85)</td>
</tr>
<tr>
<td>A</td>
<td>61.04</td>
<td>0.318</td>
<td>0.021</td>
<td>(59.25, 62.84)</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>(59.04, 62.83)</td>
</tr>
<tr>
<td>P1</td>
<td>60.90</td>
<td>0.386</td>
<td>1.55</td>
<td>(58.92, 62.88)</td>
</tr>
<tr>
<td>P2</td>
<td>60.91</td>
<td>0.383</td>
<td>0.533</td>
<td>(58.94, 62.88)</td>
</tr>
<tr>
<td>P3</td>
<td>61.01</td>
<td>0.339</td>
<td>0.032</td>
<td>(59.16, 62.86)</td>
</tr>
<tr>
<td>P4</td>
<td>61.04</td>
<td>0.324</td>
<td>0.024</td>
<td>(59.23, 62.85)</td>
</tr>
</tbody>
</table>

Table 17: Comparison of different methods (heterogeneous case) using data set due to Meier (1953)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>t-method</th>
<th>z-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HK</td>
<td>ST</td>
<td>CI</td>
<td>L</td>
</tr>
<tr>
<td>DL</td>
<td>60.99</td>
<td>0.276</td>
<td>0.280</td>
<td>(59.34, 62.68)</td>
</tr>
<tr>
<td>AN</td>
<td>60.99</td>
<td>0.271</td>
<td>0.256</td>
<td>(59.34, 62.65)</td>
</tr>
<tr>
<td>UN</td>
<td>60.99</td>
<td>0.271</td>
<td>0.256</td>
<td>(59.34, 62.65)</td>
</tr>
<tr>
<td>POS</td>
<td>61.08</td>
<td>0.305</td>
<td>0.549</td>
<td>(59.32, 62.83)</td>
</tr>
<tr>
<td>MP</td>
<td>61.01</td>
<td>0.275</td>
<td>0.275</td>
<td>(59.34, 62.68)</td>
</tr>
<tr>
<td>MPKR</td>
<td>61.01</td>
<td>0.442</td>
<td>*</td>
<td>(58.15, 63.87)</td>
</tr>
<tr>
<td>NNA</td>
<td>60.99</td>
<td>0.271</td>
<td>0.256</td>
<td>(59.34, 62.65)</td>
</tr>
<tr>
<td>NNB</td>
<td>60.99</td>
<td>0.271</td>
<td>0.256</td>
<td>(59.34, 62.65)</td>
</tr>
</tbody>
</table>
Table 18: Comparison of different methods (homogeneous case) using data set due to Eberhardt et al. (1989)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>t-method</th>
<th>z-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HK</td>
<td>ST</td>
<td>CI</td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>109.39</td>
<td>2.715</td>
<td>0.117</td>
<td>(104.14, 114.63)</td>
</tr>
<tr>
<td>TH</td>
<td>109.39</td>
<td>2.723</td>
<td>0.121</td>
<td>(104.13, 114.64)</td>
</tr>
<tr>
<td>BE</td>
<td>109.39</td>
<td>2.723</td>
<td>0.123</td>
<td>(104.13, 114.64)</td>
</tr>
<tr>
<td>A</td>
<td>109.39</td>
<td>2.673</td>
<td>0.093</td>
<td>(104.18, 114.59)</td>
</tr>
<tr>
<td>GP</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>(104.11, 114.52)</td>
</tr>
<tr>
<td>P1</td>
<td>109.38</td>
<td>2.847</td>
<td>1.340</td>
<td>(104.01, 114.74)</td>
</tr>
<tr>
<td>P2</td>
<td>109.38</td>
<td>2.823</td>
<td>0.463</td>
<td>(104.03, 114.72)</td>
</tr>
<tr>
<td>P3</td>
<td>109.39</td>
<td>2.619</td>
<td>0.070</td>
<td>(104.24, 114.55)</td>
</tr>
<tr>
<td>P4</td>
<td>109.40</td>
<td>2.425</td>
<td>0.038</td>
<td>(104.45, 114.37)</td>
</tr>
</tbody>
</table>

Table 19: Comparison of different methods (heterogeneous case) using data set due to Eberhardt et al. (1989)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>t-method</th>
<th>z-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HK</td>
<td>ST</td>
<td>CI</td>
<td></td>
</tr>
<tr>
<td>DL</td>
<td>109.60</td>
<td>1.115</td>
<td>0.816</td>
<td>(106.45, 113.17)</td>
</tr>
<tr>
<td>AN</td>
<td>109.80</td>
<td>1.829</td>
<td>2.016</td>
<td>(105.50, 114.11)</td>
</tr>
<tr>
<td>UN</td>
<td>109.76</td>
<td>1.999</td>
<td>2.567</td>
<td>(105.27, 114.27)</td>
</tr>
<tr>
<td>POS</td>
<td>109.77</td>
<td>1.967</td>
<td>2.446</td>
<td>(105.31, 114.24)</td>
</tr>
<tr>
<td>MP</td>
<td>109.82</td>
<td>1.698</td>
<td>1.698</td>
<td>(105.67, 113.97)</td>
</tr>
<tr>
<td>MPKR</td>
<td>109.82</td>
<td>2.153</td>
<td>*</td>
<td>(104.03, 115.61)</td>
</tr>
<tr>
<td>NNA</td>
<td>109.83</td>
<td>1.424</td>
<td>1.203</td>
<td>(106.04, 113.63)</td>
</tr>
<tr>
<td>NNB</td>
<td>109.60</td>
<td>0.287</td>
<td>0.166</td>
<td>(107.89, 111.31)</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>Variance</td>
<td>t-method</td>
<td>z-method</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td>HK</td>
<td>ST</td>
<td>CI</td>
<td>L</td>
</tr>
<tr>
<td>DL</td>
<td>8.262</td>
<td>0.0005</td>
<td>0.0088</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>AN</td>
<td>8.262</td>
<td>0.0005</td>
<td>0.0088</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>UN</td>
<td>8.262</td>
<td>0.0005</td>
<td>0.0088</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>POS</td>
<td>8.262</td>
<td>0.0005</td>
<td>0.0089</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>MP</td>
<td>8.262</td>
<td>0.0005</td>
<td>0.0088</td>
<td>(7.98, 8.54)</td>
</tr>
<tr>
<td>MPKR</td>
<td>8.262</td>
<td>0.0100</td>
<td>*</td>
<td>(8.06, 8.47)</td>
</tr>
<tr>
<td>NNA</td>
<td>8.262</td>
<td>0.0005</td>
<td>0.0089</td>
<td>(7.99, 8.54)</td>
</tr>
<tr>
<td>NNB</td>
<td>8.262</td>
<td>0.0005</td>
<td>0.0088</td>
<td>(7.98, 8.54)</td>
</tr>
</tbody>
</table>

6 Acknowledgement.

The research of the third author (B.K. Sinha) was supported by grant R01-OH03628 from the National Institute of Occupational Safety and Health.
References


LM-stationary Processes to analyse time series data with linearly compressing periodic behavior

Md. Jobayer Hossain*
Biostatistics Division, HHI Clinical Research
Hunt Valley, MD-21031.
E-mail: mhossain@smu.edu

Wayne A. Woodward and Henry L. Gray
Department of Statistical Sciences, Southern Methodist University
Dallas, TX-75275.

Abstract
One type of nonstationarity of a time series occurs when the frequency behavior of the process changes over time. Window based methods are commonly used for the spectral analysis of this type of nonstationary time series. Gray, Woodward and their co-researchers have recently introduced techniques based on time deformation for analyzing nonstationary processes with monotonic time varying frequencies. Their methods outperform window based methods in this case. We introduce the LM-stationary process to analyse time series data with linearly compressing periodic behavior using the time deformation technique. The forecasts and the spectral analysis of this model show improved results compared to commonly used methods in the case of this particular type of data.

1 Introduction
An observed time series is a realization or sample function of an ensemble, infinite set of realizations, of a certain time series process. In time series analysis, it is of interest to make inferences on statistical properties of the ensemble itself, not a single realization. Without some kind of statistical equilibrium, it is hard to make inferences on the statistical properties of an ensemble on the basis of a single observation of each variable at time $t$. A particular assumption of statistical equilibrium is stationarity. The common type of stationarity assumption used in practice, weak stationarity, implies constant mean, finite variance, constant correlation structure of the successive observations. One common type of nonstationarity occurs when frequencies and periodicities change over time with out violating the assumptions of constant mean and finite variance. Common methods for analyzing this type of nonstationary process are based on windowing the entire data in to many small pieces in order to achieve approximate stationarity in these short time intervals. Instantaneous spectrum is the main tool to characterize a nonstationary process in window based methods. However, most of these methods lead to distortion of the spectral estimate due to sharp truncation of the data. Another, relatively new, method of analyzing this type of time series data is time deformation technique. This technique converts a nonstationary process with monotonic time varying properties to a stationary one using some time transformations and then use standard analytical techniques for a stationary process. Gray, Woodward et al developed a number of models to analyse time series data with monotonic time varying frequencies. They defined a class of these types of processes as

---

*Professor Ali was my graduate advisor and instructor of five graduate courses during my student life at the Department of Mathematical Sciences, Ball State University, Muncie, Indiana, USA. His wonderful teaching skill, endless patience, and capability of proper guidance made him unique as a professor and graduate advisor. My supreme regards to this extraordinary individual.
G-stationary processes. Models of G-stationary processes are capable of analyzing time series data with monotonic time varying frequencies with a high level of precision. This article introduces the left multiplicative stationary (LM-stationary) process, a type of G-stationary processes, to analyse processes with linearly compacting cyclical behavior.

## 2 Left Multiplicative Stationary (LM-stationary) Processes

Nonstationary processes with linearly compressing periodic behavior (increasing frequency) over time can be transformed to classical stationary processes by taking a logarithmic transformation of time. For analyzing this type of processes we consider negative time domain $t$ and take a logarithmic transformation of $-t$ in order to obtain a stationary process. The original process is called an LM-stationary process. This process satisfies a multiplicative composition law.

**Definition 2.1.** Let $\{X(t), t < 0\}$ be a stochastic process such that

(i) $E(X(t)) = \mu,$

(ii) $\text{Var}(X(t)) = \sigma_X^2 < \infty,$ and

(iii) $E[(X(t) - \mu)(X(t\tau) - \mu)] = R_X^\tau(\tau).$

Then $X(t)$ is called a weakly LM-stationary process with LM-autocovariance $R_X^\tau(\tau)$ at lag $\tau$, where $\tau > 0$ and $t\tau \in (-\infty, 0)$. If $\tau < 0$, and $0 < \tau \leq 1$, then $t\tau \geq t$, while if $\tau \geq 1$, then $t\tau \leq t$. Moreover $R_X^\tau(\tau) = R_X^{-\tau}(\frac{1}{\tau}).$

**Definition 2.2.** Let $Y(u)$ be a stationary stochastic process over $(-\infty, \infty)$. Let $u = g(t) = \ln(-t)$ and $X(t) = Y(u)$ on $t \in (-\infty, 0)$. Then $\{Y(u), u \in (-\infty, \infty)\}$ is called the stationary dual process of $\{X(t), t < 0\}$.

It should be noted that $X(t)$ and $Y(u)$ in Definition 2.2 move in opposite directions. That is, $u$ increases as $t$ decreases. Thus the starting point of a realization from the process $X(t)$ is the end point of the associated realization from the dual $Y(u)$ and vice versa. However, reversal of the transformed time axis results in a dual with the same direction of time. Figure 1(a) - Figure 1(c) represent a generated LM-stationary process and its dual along with reverse dual, where the reverse dual is the time reversed dual process.

It is clear that LM-stationary data of Figure 1(a) and its dual of Figure 1(b) are in the opposite directions of time. However the LM-stationary data and its time reverse dual, Figure 1(a), are in the same time direction. According to the theory of the time reversibility of a stationary process, data in Figure 1(b) and Figure 1(c) possess the same statistical properties and the “reversed dual” in Figure 1(c) will be referred to as the dual of the data in Figure 1(a) as well.

**Theorem 2.1.** Let $Y(u)$ be a time series defined over $(-\infty, \infty)$ and let $Y(u) = X(t)$, where $t = -e^u$. Then the process $\{X(t), t < 0\}$ is weakly LM-stationary iff $Y(u)$ is stationary on $(-\infty, \infty)$.

**Definition 2.3.** Let $\epsilon(u)$ be a white noise process for $-\infty < u < \infty$, i.e. $E[\epsilon(u) \epsilon(u + h)] = C \delta(h)$, where $C$ is a positive constant and $\delta$ is the Dirac delta function. Let $a(t)$ be its LM-stationary dual i.e. $a(t) = \epsilon(\ln(-t))$ for $t < 0$. Then $a(t)$ is referred to as LM-white noise.

**Example 2.1.** For $t \in (-\infty, 0)$, the process $X(t) = A \cos(2\pi\beta\ln(-t) + \phi)$ is LM-stationary, where $\phi \sim \text{Uniform}(0, 2\pi)$, $A$ and $\beta$ be constants. The dual $Y(u) = A \cos((2\pi\beta u) + \phi)$ is well known to be stationary.

**Definition 2.4.** For the LM-stationary process $\{X(t), t < 0\}$, the LM-linear process is defined as $X(t) = \int_0^\infty h(u) z(u) d(\ln |u|)$ where $\int_{-\infty}^0 \frac{|h(u)|^2}{1 - \frac{u}{t}} dt < \infty$.

**Definition 2.5.** Suppose $X(t), t \in (-\infty, 0)$ is defined as the LM-stationary solution of the stochas-
tic differential equation
\[ t^k X^{(k)}(t) + \psi_1 t^{k-1} X^{(k-1)}(t) + \ldots + \psi_k X(t) = a(t) \]  \hspace{1cm} (1)

where \( a(t) \) is LM-white noise, \( E(a(t)) = 0 \), and the \( \psi_i \) are constants. Then \( X(t) \) is called a \( k \)th order continuous left Euler (LEuler) process. The dual of this LEuler process is obtained by taking the transformation \( u = \ln(-t) \), that is \( Y(u) = Y(\ln(-t)) \). The dual is given by
\[ Y^{(k)}(u) + \phi_1 Y^{(k-1)}(u) + \ldots + \phi_k Y(u) = z(u) \] \hspace{1cm} (2)

which is an autoregressive model of order \( k \). It is notable that LEuler process is an example of LM-stationary process.

**Definition 2.6.** The LM-spectrum of the LM-stationary process \( \{X(t), t < 0\} \) is the Mellin transformation of the LM-autocovariance function \( R_X(\tau) \) and can be defined as,
\[ G_X(f) = \int_0^\infty \tau^{-2\pi j f - 1} R_X(\tau) \, d\tau. \] \hspace{1cm} (3)

The LM-spectrum has the same role as the traditional spectrum in that it indicates the most important LM-periodic components in the data.

**Theorem 2.2.** If \( X(t) \) is the \( k \)th order left Euler process, then for \( \tau > 0 \), \( R_X(\tau) \) satisfies the ordinary differential equation
\[ \tau^k R_X^{(k)}(\tau) + \psi_1 \tau^{k-1} R_X^{(k-1)}(\tau) + \ldots + \psi_k R_X(\tau) = 0. \] \hspace{1cm} (4)

Moreover, the real part of the roots of the equation \( r_k + \psi_1 r_{k-1} + \ldots + \psi_{k-1} r_1 + \psi_k = 0 \), where \( r_k = r(r-1) \ldots \ldots (r-k+1) \), are less than zero iff the process is LM-stationary.

**Property 2.1.** Let \( \{X(t), t < 0\} \) be an LM-stationary process and let \( Y(u) \) be its dual, Then
\[ R_X(\tau) = C_Y(\ln \tau) \]
(ii) \( R_X^{-}(\tau) = R_X^{-}(\tau^{-1}) \)
(iii) \( R_X^{-}(e^{-\tau}) = C_Y(-\tau) \)
(iv) \( G_X^{-}(f) = S_Y(f \ln \tau) \)
(v) \( G_X^{-}(f) = G_X^{-}(-f) \).

**Definition 2.7.** A function \( f(t) \), \( t < 0 \) is LM-periodic over the interval \( I \), measured in the negative direction, with LM-period \( \delta \), if \( 0 < \delta < 1 \) is the maximum value of \( \delta \) such that \( f(t) = f(t\delta) \) for all \( t\delta \in I \). The associated LM-frequency is defined as \( f^* = -\ln(\delta)^{-1} \).

**Example 2.2.** Let \( f(t) = \cos(\beta \ln(-t)) \) be an LM-stationary process with \( t < 0 \). Then

\[
\begin{align*}
\frac{\partial}{\partial t} f(t) &= \cos(\beta \ln(-t)) \\
&= \cos \left( \beta \left[ \ln(-t) - \frac{2\pi}{\beta} \right] \right) \\
&= \cos \beta \left[ \ln(-te^{\frac{2\pi}{\beta}}) \right] \\
&= f(t\delta).
\end{align*}
\]

This indicates that the LM-period of this signal is \( e^{-\frac{2\pi}{\beta}} \) and the LM-frequency is \( f^* = -\ln \left( e^{-\frac{2\pi}{\beta}} \right)^{-1} = \frac{\beta}{2\pi} \).

**Definition 2.8.** The general instantaneous period (GIP) of an LM-stationary process at time \( t \) is given by,

\[
I_P(t) = te^{-\tau} - t, \text{ where } t < 0 \text{ and } \tau > 0
\]

\[
= t(e^{-\tau} - 1), \quad 0 < e^{-\tau} < 1,
\]

which is a linear decreasing function of time \( t \).

**Definition 2.9.** The general instantaneous frequency (GIF) of an LM-stationary process at time \( t \) is the reciprocal of the GIP at that time. It is defined by

\[
\bar{f}_t = \frac{1}{I_P(t)} = \frac{1}{te^{-\tau} - 1}.
\]

Let \( \delta \) be the LM-period and \( f^* \) be the LM-frequency of the LM-stationary process \( \{X(t), t < 0\} \). Then,

\[
\bar{f}_t = \frac{1}{t(\delta - 1)}
\]

\[
f^* = -\ln(\delta)^{-1} \Rightarrow \delta = e^{-\frac{1}{f^*}}
\]

and

\[
I_P(t) = t \left( e^{\frac{1}{f^*}} - 1 \right)
\]

\[
\bar{f}_t = \frac{1}{t \left( e^{\frac{1}{f^*}} - 1 \right)}.
\]

**Definition 2.10.** Let \( \{X(t), t < 0\} \) be an LM-stationary process, and without loss of generality let \( \mu = 0 \). The instantaneous spectrum of \( X(t) \) is defined as,

\[
G_X^{-}(f, t) = \left\{ \begin{array}{ll}
G_X^{-}(0), & \text{if } f = 0 \\
G_X^{-}(f^*), & \text{where } f^* = -\frac{1}{\ln(1 + \frac{1}{f_t})}
\end{array} \right.
\]

i.e. \( G_X^{-}(f, t) \) is the LM-spectrum with frequency \( f^* \), where \( f^* \) is the LM-frequency of \( X(t) \) and \( f_t \) is the instantaneous frequency. By Definition 2.9, \( f_t \) is given by,

\[
f_t = \frac{1}{t \left( e^{\frac{1}{f^*}} - 1 \right)}.
\]
It follows that

\[ f^* = \frac{1}{\ln \left( 1 + \frac{1}{f_1} \right)} \, . \]

Figure 2 shows the instantaneous spectrum of the LM-stationary data in Figure 1(a). Frequency in this figure increases in a convex pattern over time. Figure 3 is the Wigner-Ville time frequency distribution of the same data. Spectral analysis of LM-stationary process using the Definition 2.10 captures frequency change pattern of this data more precisely than that of Wigner-Ville distribution.

**Theorem 2.3.** Let \( X(t_k) \) be the sample data from an LM-stationary process at the time points

\[ t_k = -e^{\Delta k} = -h_k \quad \text{with} \quad k = 0, \pm 1, \pm 2, \ldots \quad \text{and} \quad h > 0. \]

Then \( V_h = X(t_k) = X(-h_k) \) is a discrete stationary process. The process \( V_h \) is referred to as the discrete dual of the LM-stationary process \( X(t) \) at the LM-sample rate \( h = e^{\Delta} \).
Corollary 2.1. If \( Y(u) \) and \( V_k \) are the continuous and discrete stationary dual of the LM-stationary process \( X(t) \) and \( Z_k \) is the reverse process of \( V_k \), then

\[
C_Z(k) = C_V(k) = C_Y(k\Delta),
\]

where \( C_Z(k) \), \( C_V(k) \) and \( C_Y(k\Delta) \) are the corresponding covariances of the processes \( Z_k \), \( V_k \) and \( Y(u) \). That is, \( Z_k \) can be considered to be the equally spaced sample obtained from the reverse process of \( Y(u) \) at the equal space interval \( \Delta \).

**Definition 2.11.** Let \( X(t), t < 0 \) be an LM-stationary process that is sampled at \( t_k = -h\Delta_k \). If the Nyquist frequency of the discrete stationary dual \( V_k \) is \( \frac{1}{2\Delta} \), then \( \frac{1}{2\Delta} = \frac{1}{2T \pi} \) is the LM-Nyquist frequency, where \( h \) is the sample rate for the L\( \text{Euler} \) process, a particular case of LM-stationary process.

**Theorem 2.4.** Let \( X(t), t < 0 \) be a continuous L\( \text{Euler}(p,q) \) process, where \( 0 \leq q < p \). If the LM-Nyquist frequency is greater than the highest frequencies corresponding to the characteristic equation of the L\( \text{Euler}(p,q) \) process, then the discrete dual of \( X(t) \) is a discrete \( \text{ARMA}(p,r) \) process, where \( r \leq p - 1 \). The coefficients of the discrete \( \text{ARMA}(p,r) \) model depend both on the L\( \text{Euler}(p,q) \) and the LM-sampling interval.

**Definition 4.12.** Let \( X(t_k) \) be a discrete LM-stationary process, and let the associated dual \( Y(k) \) have the sample rate \( h \). Then the instantaneous Nyquist frequency \( f_{Ny\Delta} \) associated with this LM-stationary process is given by \( \frac{h^2}{(1-k\Delta)^2} \).

**Forecast of LM-stationary process** One important problem in the study of time series is that of predicting a future value of a process, given a record of its past values. Given LM-stationary data, following are the steps to forecast future values of the series.

(i) Obtain an stationary dual of the original data.
(ii) Select an \( \text{ARMA}(p,q) \) model and estimate the parameters.
(iii) Find L-step ahead forecasts of the dual data by finding the best linear predictor for \( X_{t+L} \) given \( X_1, X_{t-1}, \ldots \) based on the \( \text{ARMA}(p,q) \) model fit to the dual in step (ii).
(iv) Convert the forecast in step (iii) to the LM-time scale by interpolation.

Figure 4 exhibits the comparison of 20-step ahead forecasts of LM-stationary data in Figure 1(a) using an LM-stationary model and an AR model. The last 20 observations of the data have been used for validation. Forecast performance will be measured by mean square forecast error. Let
MSE1 denote the mean square LM-stationary forecast error and MSE2 be mean square forecast error based on AR model. Then the relative improvement in LM-stationary forecasts over AR forecasts can be given (in percent) by,
\[ Q = \frac{MSE2 - MSE1}{MSE2} \times 100\% . \]
Based on this criterion, a positive \( Q \) implies \( Q\% \) improvement of LM-stationary forecasts over AR forecasts while a negative \( Q \) implies worse performance for LM-stationary forecasts. The \( Q \) for forecasts in Figure 4 is 75.85 which implies a 75.85% improvement of LM-stationary forecasts over AR forecasts. A comparative study of LM-stationary forecasts and AR forecasts for 20 realizations from an LM-stationary process with a discrete dual \((1 - 1.732B + .98B^2)X_t = \epsilon_t\), LM-sample rate \( h=1.01 \), and \( \text{var} (\epsilon_t) = 1 \) shows an average of 47.41% relative improvement of LM-stationary forecasts over AR forecasts. For each realization, the first 90% of the data is used for model selection and estimation and the remaining 10% of the data is used to compare forecast performance.

3 Application to Seismo-Acoustic Signal

As an example of real data with LM-stationary behavior, we consider a data set collected from the seismo-acoustic station BRDAR located on an island off the northwest coast of the Republic of Korea. The data consist of the part of one signal out of eight strongly dispersed infrasound signals recorded at one seismo-acoustic array on May 30, 2005 (Herrin, Kim, and Stump, 2006). The raw data were sampled at frequency 40 Hz after a lowpass filter at a frequency 10 Hz.

Figure 5(a) is the plot of these seismic-acoustic data. The linearly compacting periodic behavior over time of these data is clearly visible from this plot. Figure 5(b) is the dual with an origin offset 27 and LM-sampling rate \( h = 1.0114 \). Figure 6(a) and Figure 6(b) show the spectrum and LM-spectrum of the original data. It is clear that the ordinary spectrum of these data fails totally to tell us anything about the periodicities or frequency content of this data. However, the LM-spectrum clearly shows that data have a multicomponent signal with at least two strong dominant time varying frequencies. The instantaneous spectrum of these data is shown in Figure 7, where it can be seen how frequencies change over time. The Wigner-Ville distribution in Figure 8 also shows the change of frequency over time for these data. It fails to clearly show more than a single frequency component in the data.

4 Conclusions

The instantaneous spectrum is an important tool for the spectral analysis of a nonstationary process with time varying frequencies. Spectral analysis using this tool for LM-stationary processes shows superior results compared to the commonly used window-based methods, especially in the case of multicomponent time series data. The Wigner-Ville time frequency distribution is a standard window based method and does a good job of estimating the instantaneous spectrum of a monocomponent time series. But the Wigner-Ville distribution often fails to explain the time varying frequency content in the case of more than one time varying components.

Another goal of time series analysis is to predict future values based on past values. Developed time deformation method applies to LM-stationary data outperforms commonly used AR methods for this purpose.
Figure 5: (a) Seismic Signal; (b) Dual of Signal in (a).

Figure 6: (a) Spectrum of data in Figure 5(a); (b) LM-spectrum of data in Figure 5(a)
Figure 7: Instantaneous Spectrum of Seismic data in 5(a)

Figure 8: Wigner-Ville time frequency distribution of Seismic data in 5(a)
References


A Permutation Test for the Stimuli Effect  
In The Spatio-Spectral Profile of Brain Signals

Zhewen Fan * & Hernando Ombao †

Abstract

Numerous studies in neuroscience have demonstrated links between the oscillatory behavior of neuronal populations (as quantified by spectral analysis) and cognition, sleep, disease diagnosis and treatment outcomes. Neuroscientists are interested in investigating stimulus-induced differences in the variation of neuronal oscillatory activity across the brain cortical surface. An estimator for the spatially-varying spectrum is developed and a permutation test is used to identify brain locations where the oscillatory activity is different for two conditions in a spatial- verbal Stroop task experiment.

Keywords: Event-related optical signals, Permutation test, Spatio-temporal random process, Spectrum.

1 Introduction

Neuroscientists are able to observe different facets of the brain activity underlying cognitive function through non-invasive modalities. In this paper, we study the oscillatory activity of signals recorded from optical imaging. Our goal is to investigate the variation of oscillatory activity across the brain cortical surface and compare the spatio-spectral profile across experimental conditions, namely the position (spatial) and meaning (verbal) conditions.

Optical signals measure changes in the light absorption and scattering properties of brain tissue following local event-related alterations in neuronal activity. As discussed in Cohen (1972), Stepnoski, Porta, Raccuia-Behling, Blonder, Slusher, and Kleinfeld (1991) and Gratton and Fabiani (1998), the phenomenon can plausibly be due to movement of ions across the neuronal membrane and the reorientation of the membrane proteins during neuronal activity. Event related optical signals (EROS) possess both good temporal and spatial resolution and thus are excellent for studying the time course activity in localized cortical areas. Gratton, Goodman-Wood and Fabiani (2001) demonstrated that EROS has good temporal correspondence with event related brain potentials (ERP). Moreover, Gratton, Fabiani, Corballis, Hood, Goodman-Wood, Hirsh, Kim, Friedman and Gratton (1997) demonstrated that EROS has good spatial correspondence with fMRI. EROS has been used in several neuroscience investigations including cognitive aging (Fabiani, Low, Wee, Sable and Gratton, 2006) and active and passive oddball

*University of Illinois at Urbana-Champaign, Department of Statistics, 117 Illini Hall, 725 South Wright Street, Champaign, IL 61820, USA. Email: zfan3@uiuc.edu
†University of Illinois at Urbana-Champaign, Department of Statistics, 117 Illini Hall, 725 South Wright Street, Champaign, IL 61820, USA. Email: ombao@uiuc.edu. Telephone: (217) 244-7200. Fax: (217) 244-7190.
tasks (Low, Leaver, Kramer, Fabiani and Gratton, 2006). Recently, Rhyklevskaia, Fabiani and Gratton (2006) developed a lagged covariance model that uses EROS to study functional connectivity.

EROS, due to its good spatial (sub-centimeter) and temporal (millisecond scale) localization properties, may provide ideal data to investigate changes in the oscillatory behavior of the signal across extended cortical areas. The standard approach to quantifying brain oscillations is through spectral analysis [see Muthuswamy and Thakor (1998) and Dummermuth and Molinari (1987) among others]. Most of these conventional approaches do not consider the spatial variation of the temporal spectrum. However, methods that incorporate spatial and topographic information in the spectral analysis of brain signals are being developed - see, for example, Koenig, Marti-Lopez and Valdes-Sosa (2001) and Hoogenboom, Schoffelen, Oostenveld, Parkes and Fries (2006).

A substantial amount of research has focused on oscillatory activity, such as rhythms observed in the electroencephalogram (EEG). Some of them have investigated how these oscillatory phenomena vary over space (e.g., Gratton, Villa, Fabiani, Colombis, Palin, Bocioni and Fiori, 1992). By and large, these studies have treated the time series observed at individual locations as separate measures, and then either compared the spectral properties of individual locations, thus generating amplitude or power maps observed at different locations. However, in many cases measures from different locations are not really independent, which may lead to problems in interpretation of the results.

A rigorous framework for spectral analysis of spatio-temporal data is recently developed in Ombao, Shao, Rykhlevskaia, Fabiani and Gratton (2007) [ORRF, hereafter]. They develop a general representation of spatio-temporal processes and rigorously define the location-dependent (spatially-varying) temporal spectrum which is the primary quantity of interest. A nonparametric method for estimating the spatially-varying temporal spectrum was developed and they put forward an asymptotic framework under which the estimator is asymptotically mean-squared consistent. In this paper, our goal is to identify localized cortical regions and specific frequency bands where group oscillatory behavior significantly differ.

The remainder of this paper is organized as follows. In Section 2, we review the Cramér representation of a spatio-temporal process and the estimation procedure in OSRF. In Section 3, we describe the permutation test for spatio-spectral differences and analyze the EROS data set. We conclude in Section 4 with discussion of the results and future work.

2 Spatio-Temporal Process

To study the oscillatory properties of brain signals, we first consider the Cramér representation of a stationary univariate temporal process

\[ X(t) = \int_{-\pi}^{\pi} A(\lambda) \exp(i\lambda t) d\lambda, \]
where \( A(\lambda) \) is the complex valued Hermitian transfer function and \( Z(\lambda) \) is a stochastic process with zero mean uncorrelated increments, that is,

\[
E[dZ(\lambda)] = 0 \quad \text{and} \quad \text{cov}[dZ(\lambda), dZ(\omega)] = \delta(\omega - \lambda)d\omega.
\]

Thus, a stationary temporal process may be viewed as a linear combination of infinitely many sinusoids [i.e., Fourier waveforms \( \exp(i\lambda t) \)] having random coefficients \( A(\lambda)dZ(\lambda) \). The variance of the random coefficient \( A(\lambda)dZ(\lambda) \) is the spectrum of the process at frequency \( \lambda \). That is, \( f(\lambda) = E[A(\lambda)dZ(\lambda)]^2 = |A(\lambda)|^2 \). The oscillatory content of signals is characterized by the spectrum which is related to the variance decomposition of \( X(t) \). Due to the orthonormality of the increments \( dZ(\lambda) \), the variance of the stochastic process has the decomposition \( \text{var}[X(t)] = f_{-\infty}^{\infty} f(\lambda)d\lambda \). Thus, one may interpret the temporal spectrum at some frequency \( \lambda_0 \) to be approximately the variation in \( X(t) \) that is “explained” by the Fourier waveform \( \exp(i\lambda_0 t) \) which oscillates at frequency \( \lambda_0 \).

### 2.1 Spectral representation of spatio-temporal processes

We now consider the Cramér representation of random processes that are observed across both space and time. We summarize the essential ideas presented in OSRFG. In this paper, we will consider the EROS data set (optical time series observed from several locations in the cortical surface). The locations are given in Figure (1) and the data recorded from one subject is plotted in Figure (2).

Let \( X(t) \) be the spatio-temporal process defined on location \( s = (s_1, s_2) \in \mathbb{R}^2 \) and time point \( t \in \mathbb{Z} \). Suppose that the spatio-temporal data at hand is observed at locations \( s \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\} \) and time points \( t \in \{1, \ldots, T\} \). There are a total of \( n = n_1n_2 \) locations in space and a total of \( T \) time points. For this particular study, it will be assumed that, for each
Figure 2: Event-related optical signals at different cortical locations recorded from one subject. Each time series has length $T = 125$. These signals are obtained by the following procedure. First, we computed the average across trials that were time locked to the position (P) cue that were preceded by meaning (M) cue. Second, we computed the average across trials that were time locked to the position (P) cue that were preceded by the position (P) cue. The plotted time series is the difference between the first (switched cue) and the second (no-switch cue). Sampling rate: one time point per 16 milliseconds. The specific locations on the cortex where the time series where extracted are the following: Top left: left medio-frontal gyrus ($x = -40$ mm, $y = 20$ mm). Top right: right medio-frontal gyrus ($x = 40$ $y = 20$). Center: superior-frontal gyurs ($x = 0, y = -12$). Bottom Left: left posterior-central gyrus ($x = -50, y = -28$). Bottom Right: right posterior-central gyrus ($x = 50, y = -28$).
location, the process is stationary over time. Over space, on the other hand, the process is not restricted to be stationary. The Cramér representation of the spatio-temporal process \( X_t(s) \) is

\[
X_t(s) = \int_{-\pi}^{\pi} A(u, \lambda) \exp(i \lambda t) dZ_t(\lambda).
\]

where \( s = (s_1, s_2) \) is the spatial index in the observation space and \( u_s = u_s, n = (s_1/n_1, s_2/n_2) \) is the corresponding location in the re-scaled space. The elements of the above model are as follows.

1. \( A(u, \lambda) \) is the location-dependent complex-valued transfer function. It is defined on a fixed re-scaled spatial domain \( u = (u_1, u_2) \in [0, 1]^2 \) and frequency domain \( \lambda \in [-\pi, \pi] \).

   Moreover, it is Hermitian, that is, \( A(\cdot ; -\lambda) = A^*(\cdot; \lambda) \).

2. \( Z_t(\lambda) \) is a complex-valued stochastic process that satisfies

   \[
   \mathbb{E}[dZ_t(\lambda)] = 0 \quad \text{and} \quad \text{cov}[dZ_t(\lambda), dZ_{t'}(\lambda')] = \rho(s, s') \delta(\lambda - \lambda') d\lambda
   \]

   with \( \rho(s, s) = 1 \).

The primary quantity of interest in the above model is the location-dependent temporal spectrum

\[
f(u, \lambda) = |A(u, \lambda)|^2
\]

which is defined on the re-scaled space \( u \in [0, 1]^2 \) and frequency \( \lambda \in [-\pi, \pi] \). One may interpret \( f(u, \lambda) \) as the variation of the time series at location \( s \) that is explained by the oscillation at frequency \( \lambda \).

Remark 2.1. An example of the above process is the first order temporal auto-regressive (AR(1)) process with location-dependent coefficient: \( X_t(s) = \phi(u) X_{t-1}(s) + \xi_t(s) \) where \( \phi(u) \) is the AR(1) coefficient at location \( u \), \( \{\xi_t(s)\} \) is a spatial process that are independent over time with \( \mathbb{E}\xi_t(s) = 0 \) and \( \text{var}(\xi_t(s)) = 1 \). The location-dependent transfer function is

\[
A(u, \lambda) = \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{\infty} [\phi(u) \exp(-i\lambda j)]^2
\]

which gives rise to the equivalent Cramér representation \( X_t(s) = \int_{-\pi}^{\pi} \frac{1}{\sqrt{2\pi}} [1 - \phi(u) e^{-i\lambda}]^{-1} \exp(i\lambda t) dZ_t(\lambda) \). The corresponding location-dependent temporal spectrum is \( f(u, \lambda) = |1 - \phi(u) e^{-i\lambda}|^{-2}/(2\pi) \). In the simulation study, we used \( \phi(u_1, u_2) = 0.2 + 0.3 u_1 + 0.2 u_2 + 0.2 u_1 u_2 \).

2.2 Estimating the spatially-varying temporal spectrum

For univariate temporal stationary processes, the spectrum can be consistently estimated by noting its relationship with the auto-covariance. Let \( X_t \) be a zero mean stationary process with an auto-covariance sequence \( \gamma(k) = \mathbb{E}[X_{t+k}X_t] \) that satisfies \( \sum_k |\gamma(k)| < \infty \). The spectrum of \( X_t \) is

\[
f(\lambda) = \frac{1}{2\pi} \sum_k \gamma(k) \exp(-i\lambda k).
\]
Denote the sample auto-covariance to be \( \bar{\gamma}(k) = \frac{1}{T} \sum_{t=1}^{T-|k|} X_t X_{t+|k|} \). The lag window estimator [Brockwell and Davis (1991)] for \( \hat{f}(\lambda) \) is defined to be

\[
\hat{f}(\lambda) = (2\pi)^{-1} \sum_{k=-B_T}^{B_T} a(kB_T) \bar{\gamma}(k) \exp(-i\lambda k),
\]

where \( a(. \) is a window function, \( B_T = b^{-1}_T \) is the bandwidth, \( B_T \to \infty \) and \( B_T = o(T) \) as \( T \to \infty \).

For a fixed spatial location \( u \) and frequency \( \lambda \), the location-dependent temporal spectrum \( f(u, \lambda) \) can be estimated as follows. We first compute the lag window estimator at each location \( s_j, j = 1, \ldots, n = n_1 n_2 \), which is denoted by

\[
\hat{f}(u_{s_j}, \lambda) = \frac{1}{2\pi} \sum_{k=-B_T}^{B_T} \bar{\gamma}(s_j; k) a(kB_T) e^{-i\lambda k},
\]

where \( \bar{\gamma}(s_j; k) = \frac{1}{T} \sum_{t=1}^{T-|k|} X_t(s_j) X_{t+|k|}(s_j) \). We then form an estimator of \( f(u, \lambda) \) by smoothing the lag window estimates within a spatial neighborhood of \( u \), that is,

\[
\hat{f}(u, \lambda) = \sum_{j=1}^{n} w_{jn}(u) \hat{f}(u_{s_j}, \lambda)
\]

where the weights \( w_{jn}(u) \) are nonnegative, and \( \sum_{j=1}^{n} w_{jn}(u) = 1 \). Under regularity conditions in OSRFG, the above spatio-spectral estimator is shown to be asymptotically consistent and, moreover, follows a Gaussian distribution.

In this paper, we consider \( f^1(u, \lambda) \) and \( f^2(u, \lambda) \) to be the spatially-varying temporal spectrum from the position (spatial) and meaning (verbal) conditions, respectively. Our goal is to identify the locations \( u \) in the brain cortical surface and the frequency bands \( \lambda \) where the two conditions have different oscillatory behavior. For this particular EROS data set, we focus on the delta band (0.5 – 4.0 Hertz). To do so, one may construct formal statistical tests using the derived asymptotic distributions. However, this would require specifying correctly the temporal and spatial covariance structure of the noise in the EROS can be a difficult task. Thus, statistical inference can be invalid or inefficient when these assumptions are violated. In the next section, we propose a permutation test based on the conditions.

3 Permutation Test for Condition Differences

Permutation tests do not require strong assumptions on the distribution of spatio-spectral-temporal noise. Here, we propose permutation tests of spatio-spectral estimates based on permutation of within-subject experimental conditions. Various authors have proposed permutation tests for analysis of brain images and genomic data. An introduction to the this methodology and its applications to neuroimaging data is given in Nichols and Holmes (2001).

Consider this set up. Let \( X_{1i}(u, \lambda), \ldots, X_{Mi}(u, \lambda) \) be the spatio-spectral data (estimate) of \( M \) subjects under the position condition and let \( X_{1i}(u, \lambda), \ldots, X_{Mi}(u, \lambda) \) be the spatio-spectral data (estimate) of the same subjects under the meaning condition. In general, for all \( i = 1, \ldots, M \),

\[
\mathbf{E} X_{1i}(u, \lambda) = f^1(u, \lambda)
\]
\[ E X^2_i(u, \lambda) = f^2(u, \lambda). \]

Thus, under the null hypothesis \( H_0 : f^1(u, \lambda) = f^2(u, \omega) \) we may consider these two conditions to be exchangeable.

For this data set, will only consider a specific frequency band of 0.5 – 4 Hertz. We compute over all locations \( s = 1, \ldots, S \) (\( S = 240 \)) the following quantities

\[
\begin{align*}
\bar{X}^1(u_s, \lambda) &= \frac{1}{M} \sum_{i=1}^{M} X_i^1(u_s, \lambda) \text{ condition 1 average} \\
\bar{X}^2(u_s, \lambda) &= \frac{1}{M} \sum_{i=1}^{M} X_i^2(u_s, \lambda) \text{ condition 2 average} \\
D_i(u_s, \lambda) &= X_i^1(u_s, \lambda) - X_i^2(u_s, \lambda) \text{ subject-specific difference} \\
\overline{D}(u_s, \lambda) &= \frac{1}{n} \sum_{i=1}^{n} D_i(u_s, \lambda) \text{ averaged difference} \\
V(u_s, \lambda) &= \frac{1}{n-1} \sum_{i=1}^{n} [D_i(u_s, \lambda) - \overline{D}(u_s, \lambda)]^2 \text{ sample variance of difference} \\
t(u_s, \lambda) &= \frac{|D(u_s, \lambda)|}{\sqrt{V(u_s, \lambda)}}.
\end{align*}
\]

For data set with \( n \) subjects whose recordings come from a paired design (position vs meaning conditions), there are \( B = 2^n \) possible possible permutations. For each permutation \( b = 1, \ldots, B \), we compute all of the above quantities and finally obtain

\[ t^b(u_s, \lambda) = \frac{|D^b(u_s, \lambda)|}{\sqrt{V^b(u_s, \lambda)}}. \]

However, an added complication here is that there are a total of \( V = 240 \) locations and thus we would need to control for multiple hypothesis testing. To accomplish this, our test statistic is

\[ T^b_{\max} = \max_s t^b(u_s, \lambda). \]

To determine the threshold, we arrange the values \( T^b_{\max} \) from all permutations in increasing order. Thus, for a probability of Type I error controlled at \( \alpha \), the threshold is the \( (\lfloor \alpha B \rfloor + 1) \)-th largest value of all \( T^b_{\max} \) values.

4 Analysis of the EROS data set

The event related optical signal (EROS) data set was recorded from 16 participants in a spatial-verbal Stroop task experiment conducted at the Cognitive Neuroimaging Laboratory at the University of Illinois at Urbana-Champaign. Each trial included the visual presentation of the words “Above” or “Below” that were located, respectively, above or below a central fixation cross. The subject was instructed to press one button for above and another for below. Across the different trials, the subjects were instructed to pay attention to either the word position or the word meaning (denoted by “P” or “M” and presented 2 seconds prior to the actual stimulus). Therefore, during the two seconds following the cue, the participant had to prepare to respond to the spatial (position) or verbal (meaning) features of the stimulus. In the current
paper, we will consider data obtained during this 2s period. The EROS data, specifically the phase delay data, was recorded with a multi-channel frequency-domain optical instrument, with a modulation frequency of 220 Hertz and an effective sampling rate of 16 milliseconds, over 160 source-detector pairs located on both the left and right frontal cortex. A temporal low pass filter was applied with cutoff of 8 Hertz. Separate averages, time-locked to the cues, were computed for each type of cue (P or M), switch (from P to M or from M to P) or no-switch trial types and recording channel. Difference time series between the switch and no-switch trial types were then computed to isolate brain activity related to task switching. The source and detector locations were digitized and then mapped onto a surface image of the brain, thus allowing for surface image constructions of the optical activity for each data point in Talairach space. This yielded a three dimensional data matrix (x and y surface locations and time). Each time series was associated with a particular location, with 289 locations sampled on a grid with a 5 millimeter step size. For this particular analysis, we only used 20 × 12 time series located on the (x, y) grid [−50, 50] × [−30, 24] square millimeter surface that was common to all 16 subjects (age is 55 years or older) in the experiment.

The condition-specific spatio-spectral estimates are given in Figure 3 along with the “bright spots” that exceed the threshold (and hence declared statistically significant). It is clear that the two experimental conditions require or induce different intensities in the oscillatory activity at the delta band at the superior-frontal and left posterior-central areas.

5 Conclusion

This paper is motivated by the need investigate the variation of the oscillatory activity across cortical regions and the difference across experimental conditions. Previous work in this area has not taken specific advantage of the spatial nature of brain imaging data. OSRFG develop non-parametric estimation method is proposed to estimate the location-dependent temporal spectrum, a quantity that describes the spatial variation of temporal oscillatory behavior. In this work, we took the further step of comparing spatial oscillatory pattern across experimental conditions. We used permutation tests which do not require strong assumptions on the distribution of spatio-spectral-temporal noise. We identified local “bright spots” under the delta band which were highly spatially localized to the superior-frontal gyrus and the left posterior-central gyrus. Our next step is to compare this profile with younger subjects to determine how age modifies brain oscillatory spatial pattern.

References

Figure 3: Top: Spatio-spectral estimate for the position (spatial) condition. Middle: Spatio-spectral estimate for the meaning (verbal) condition. Bottom: Bright spots indicate significant difference between the position and meaning conditions.


Inference on limiting availability of a one-unit repairable system

Fang Li and Jyotirmoy Sarkar

Department of Mathematical Sciences, Indiana University Purdue University Indianapolis, 402 N
Blackford Street, Indianapolis, IN 46202-3216, USA

Abstract
We consider a one-unit repairable system under continuous monitoring and a perfect repair policy, with instantaneous commencement of repair and installation to operation. For arbitrary continuous failure time and repair time distributions, we study the problems of point estimation, one-sided confidence interval estimation and hypothesis testing for the limiting availability, based on a random sample of failure and repair times.

MSC: primary 90B25; secondary 62N05

Keywords: Continuous monitoring; Perfect repair; Mean time to failure; Mean time to repair; Key renewal theorem; Slutsky’s theorem; Bias correction; Jackknife; Bootstrap estimation; Acceleration; One-sided confidence interval; One-sided test

1. Introduction

We consider a one-unit repairable system under continuous monitoring. In the beginning, the system is put on operation, and as soon as it fails it goes to a repair facility for repair (this is called instantaneous commencement of repair). We assume that repair takes a random amount of time, after which the unit is restored back to a level equivalent to a new unit (this is called the perfect repair policy). Then it is immediately put on operation (this is called instantaneous installation). Thus, the system alternates between on and off states according as it remains operational or undergoes repair.

As a benchmark to evaluate the long-term performance of a repairable system, one often considers the limiting average availability

$$A_{av} := \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}},$$

(1.1)

where MTTF stands for mean time to failure and MTTR stands for mean time to repair. Thus, $A_{av}$ equals the limiting expected proportion of time the system is up. For an account on the subject see Barlow and Proschan (1975, p. 206).

We assume all lifetimes are independent and identically distributed (IID) with an absolutely continuous cumulative distribution function (CDF) $F$ (with density function $f$), while all repair times are IID with an absolutely continuous CDF $G$ (with density function $g$). Furthermore, the lifetimes and the repair times are stochastically independent. Under these assumptions, the limiting availability $A$, which is the probability that the system will be found in a functioning state much later in future, exists and, by the Key Renewal Theorem, we have

$$A = A_{av} = \frac{\mu_F}{\mu_F + \mu_G},$$

(1.2)
where $\mu_F$ is the mean time to failure under CDF $F$, and $\mu_G$ is the mean time to repair under CDF $G$. See, for example, Hoyland and Rausand (1994, p. 224) and Ross (1996, p. 115). In most practical applications, however, $F$ and $G$ are unknown. Hence, $A$ is unknown. We consider the problem of statistical inference on $A$ based on two independent random samples of equal size $n$ from $F$ and $G$ respectively.

In the literature, there has been many papers over the last two decades studying the statistical inference on $A$ for specific lifetime and repair time distributions. See Ananda (2003), Chandrasekhar et. al. (2004) and the references therein. In this paper, we present a comprehensive study of point estimation, confidence interval estimation and hypothesis testing on $A$ in a non-parametric setting with fixed samples.

In the non-parametric setting with arbitrary lifetime and repair time distributions, Sen and Bhattacharjee (1986) correctly point out that the natural estimator of $A$, given by its sample version, may not be strictly unbiased, and therefore they use a jackknife estimator to reduce the effective bias. In this paper, we additionally present several bootstrap methods to correct the bias. Sen and Bhattacharjee (1986) also point out that a fixed sample level $\alpha$ test may not achieve a specified power against a specified alternative, or the width of a specified $(1 - \alpha)$ level confidence interval may not have pre-fixed width. Hence, they developed sequential methods of confidence interval estimation and testing procedures. In the fixed sample set up, Lim, et. al. (2004) compares the bootstrap methods proposed in DiCiccio and Efron (1996) to produce confidence intervals for $A$. However, none of the proposed bootstraps come out as a clear winner. In this paper, we consider another bootstrap method, the ratio bootstrap-\(t\), and exhibit its superiority over those considered in Lim, et. al. (2004).

In Section 2, we show that the natural estimator $\hat{A}$, given by the ratio of sample mean lifetime to the total of sample mean lifetime and sample mean repair time, is consistent and asymptotically normal. However, for small to moderate sample sizes the natural estimator is biased and skewed to the left, and therefore is not useful for statistical inference in most ordinary situations. In Section 3, we exhibit that the bias in $\hat{A}$ is corrected by each of the two jackknife estimators: ordinary and infinitesimal, and three bootstrap estimators: natural, centered and balanced. In Section 4, we use several bootstrap based techniques to construct improved confidence intervals, and in Section 5 we carry out the corresponding hypothesis tests for $A$. We compare the performance of each inference procedure in a simulation set up using tables and graphs. The ratio bootstrap-\(t\) emerges as the best method for constructing confidence intervals and carrying out hypotheses tests for small to moderate sample sizes. Section 6 provides a summary and some concluding remarks.

2. The natural estimator of $A$

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of lifetimes with common CDF $F$ and $Y_1, Y_2, \ldots, Y_n$ is a random sample of repair times with common CDF $G$. Assume that the two random samples are stochastically independent. Let $\bar{X}$ and $\bar{Y}$ denote the sample means, and $S_X^2$ and $S_Y^2$ denote the sample variances. By Khintchine’s strong law of large numbers $\bar{X} (\bar{Y})$ is a strongly consistent estimator of $\mu_F (\mu_G)$. Hence, the natural estimator

$$\hat{A} = \frac{\bar{X}}{\bar{X} + \bar{Y}}$$

is a strongly consistent estimator of $A$. Moreover, since $\hat{A}$ is bounded (between 0 and 1), $E[\hat{A}]$ converges to $A$ as $n \rightarrow \infty$. In other words, $\hat{A}$ is asymptotically unbiased.

Only in the special case of exponential lifetimes and exponential repair times, one can derive the exact distribution of the natural estimator $\hat{A}$. However, $F$ and $G$ are usually unknown. Hence, the exact distribution of $\hat{A}$ can not be derived. Nonetheless, it is straight-forward to establish the asymptotic normality of $\hat{A}$ using the central limit theorem and Slutsky’s theorem. Note that

$$\sqrt{n}(\hat{A} - A) = \sqrt{n} \left\{ \frac{\bar{X}}{\bar{X} + \bar{Y}} - \frac{\mu_F}{\mu_F + \mu_G} \right\}$$

164
\[
\hat{A} = \frac{\sqrt{n} \left[ \mu_G (\bar{X} - \mu_F) - \mu_F (\bar{Y} - \mu_G) \right]}{(\mu_F + \mu_G)(\bar{X} + \bar{Y})}
\]

converges in distribution to \( N(0, \sigma^2) \) where

\[
\sigma^2 = \frac{\mu_F^2 \sigma_G^2 + \mu_G^2 \sigma_F^2}{(\mu_F + \mu_G)^4}.
\]

Furthermore, since a strongly consistent estimator of \( \sigma^2 \) is given by

\[
\hat{\sigma}^2 = \frac{\bar{X}^2 \sigma_G^2 + \bar{Y}^2 \sigma_F^2}{(\bar{X} + \bar{Y})^4},
\]

\( \hat{A} \) is asymptotically normal with approximate variance \( \hat{\sigma}^2 / n \).

However, as pointed out by Sen and Bhattacharjee (1986, p. 284), in view of the fact that \( \hat{A} \) is not a linear function of \( \mu_F \) and \( \mu_G \) (rather a ratio involving them), the natural estimator may not be strictly unbiased (although it is asymptotically unbiased). In our simulation study, we exhibit the bias of \( \hat{A} \) even for moderately large sample sizes. See the \( \hat{A} \)-column in Table 1. Furthermore, while \( \hat{A} \) is asymptotically normal, the convergence to normality is rather slow, as exhibited in Figure 1. Therefore, for small to moderate sample sizes, confidence interval estimates and hypothesis testing based on asymptotic normality of \( \hat{A} \) perform poorly in terms of lack of proper coverage and inaccurate level of test respectively, as exhibited by the \( \hat{A}_{\text{std}} \)-column of Table 2 and the solid lines in Figure 3.

3. Jackknife and bootstrap estimators of \( \hat{A} \)

In this Section, we apply different resampling schemes to construct the standard jackknife, the infinitesimal jackknife, the natural bootstrap, the centered bootstrap and the balanced bootstrap estimators of \( \hat{A} \) in order to effectively reduce (but we cannot completely eliminate) the bias of the natural estimator \( \hat{A} \).

3.1 Jackknife estimators of \( \hat{A} \)

For \( i = 1, \ldots, n \), let \( \bar{X}^{(i)} \) and \( \bar{Y}^{(i)} \) denote the sample means that are obtained when \( (X_i, Y_i) \) is omitted from the original sample. As in Sen and Bhattacharjee (1986, p. 286), we define the jackknife estimator of \( \hat{A} \) and its variance estimator as

\[
\hat{A}_J = \hat{A} + \frac{n-1}{n} \sum_{i=1}^{n} (\hat{A} - \hat{A}^{(i)}) \quad \text{with} \quad \hat{A}^{(i)} = \frac{\bar{X}^{(i)}}{\bar{X}^{(i)} + \bar{Y}^{(i)}},
\]

\[
s_J^2 = (n-1) \left[ \sum_{i=1}^{n} \left( \hat{A}^{(i)} - \frac{1}{n} \sum_{j=1}^{n} \hat{A}^{(j)} \right)^2 \right].
\]

The infinitesimal jackknife estimator (see Efron (1979, pp. 12-14) and Jaeckel, 1972) of \( \hat{A} \), which corrects the bias of \( \hat{A} \) through a ratio (see some details in Subsection 4.4), and its variance estimator are given by

\[
\hat{A}_{IJ} = \hat{A} \left[ 1 - \frac{1}{n^2} \sum_{i=1}^{n} \left( \frac{X_i}{\bar{X}} - 1 \right) \left( \frac{X_i + Y_i}{\bar{X} + \bar{Y}} - 1 \right) - \sum_{i=1}^{n} \left( \frac{X_i + Y_i}{\bar{X} + \bar{Y}} - 1 \right)^2 \right]^{-1}
\]

\[
s_{IJ}^2 = \frac{\hat{A}_{IJ}^2}{n^2} \sum_{i=1}^{n} \left[ \frac{X_i}{\bar{X}} - \frac{X_i + Y_i}{\bar{X} + \bar{Y}} \right]^2.
\]
3.2 Bootstrap estimators of $\hat{A}$

Let $(X_1^1, \ldots, X_n^1), \ldots, (X_1^B, \ldots, X_n^B)$ and $(Y_1^1, \ldots, Y_n^1), \ldots, (Y_1^B, \ldots, Y_n^B)$ denote $B$ independent bootstrap samples chosen at random with replacement from the original samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ respectively. For $k = 1, \ldots, B$ let $\bar{X}^{*k}$ and $\bar{Y}^{*k}$ denote the corresponding means of those independent resamples and let

$$\hat{A}^{*k} = \frac{\bar{X}^{*k}}{\bar{X}^{*k} + \bar{Y}^{*k}} \quad \text{and} \quad \hat{A}^{**} = \frac{\sum_{k=1}^{B} \bar{X}^{*k}}{\sum_{k=1}^{B} (\bar{X}^{*k} + \bar{Y}^{*k})}.$$  

(3.5)

Note that $\hat{A}^{**}$ is equivalent to a centered $\hat{A}^{*k}$ obtained by pooling all $B$ independent resamples together as one big resample of size $Bn$. Following Efron (1990), we define the natural bootstrap and centered bootstrap estimators of $\hat{A}$ as

$$\hat{A}_{nb} = \hat{A} - \left( \frac{1}{B} \sum_{k=1}^{B} \hat{A}^{*k} - \hat{A} \right) = 2 \hat{A} - \frac{1}{B} \sum_{k=1}^{B} \hat{A}^{*k},$$  

(3.6)

$$\hat{A}_{cb} = \hat{A} - \left( \frac{1}{B} \sum_{k=1}^{B} \hat{A}^{*k} - \hat{A}^{**} \right) = \hat{A} + \hat{A}^{**} - \frac{1}{B} \sum_{k=1}^{B} \hat{A}^{*k}.$$  

(3.7)

The balanced bootstrap estimator $\hat{A}_{nb}$ of $\hat{A}$ is based on another resampling scheme which ensures that each of $X_1, \ldots, X_n$ and each of $Y_1, \ldots, Y_n$ appears exactly $B$ times when the $B$ resamples are pooled together. The balanced bootstrap sample is easily obtained as follows: First pool $B$ copies of the original $X$-sample together, then take a random permutation of elements of the pooled sample, and finally separate them into $B$ blocks of $n$ elements. Do the same with the $Y$-sample. Clearly, with this sampling scheme, $\hat{A}^{**} = \hat{A}$, which implies that $\hat{A}_{nb} = \hat{A}_{cb}$, and the common estimator in this case is denoted by $\hat{A}_{nb}$.

3.3 Comparing various estimators of $\hat{A}$

We conduct a Monte Carlo simulation study to compare the natural estimator $\hat{A}$ and the above-mentioned five bias-corrected estimators of $\hat{A}$ in terms of their bias and standard deviation. For the simulation, we choose a Weibull lifetime distribution with different values of the shape parameter $b$ and corresponding value of the scale parameter so that the mean lifetime becomes 10, and we choose an exponential repair time with mean 1. Hence, the limiting availability is $\hat{A} = 10/(10 + 1) = .9091$ for each lifetime distribution. We consider three values of the shape parameter of the lifetime distribution: $b = 0.5, 1$ and 2; representing decreasing, constant and increasing failure rates respectively. We also consider three sample sizes: $n = 20, 60$ and 100, to represent small, moderate and large samples. We use $B = 1000$ bootstrap samples. Also, each computation is repeated 1000 times.

Figure 1 depicts the densities of the natural estimator $\hat{A}$ based on samples of size $n = 60$, under the simulation set up. Note the bias and left skweness of the densities, which render the use of $\hat{A}$ inaccurate for construction of confidence interval and for hypothesis testing.

Table 1 shows the mean departure from the true value $\hat{A} = .9091$ and the standard deviation (in parenthesis) of 1000 replicates of various estimators. As anticipated, both bias and standard deviation of each estimator decrease with increase in sample size. The bias of $\hat{A}$ is effectively reduced by each of the five proposed estimators. The standard jackknife estimator $A_J$ performed the best in bias reduction for small sample size $n = 20$. For moderate to large sample sizes, all five proposed bias-corrected estimators performed equally well. The simulated standard deviations of all five bias-corrected estimators are almost the same as that of the natural estimator $\hat{A}$, except for the natural bootstrap estimator $\hat{A}_{nb}$, which has a slightly larger standard deviation. This is reflective of the fact that the main purpose of the proposed bias-corrected estimators is to reduce the bias of $\hat{A}$ without changing the corresponding variances appreciably.

166
4. Confidence intervals

In this section, we construct \((1 - \alpha)\) level left-sided confidence intervals for the limiting availability of a system with unknown lifetime and repair time distributions. The reason for considering only left-sided confidence intervals is that in case the true limiting availability of the system exceeds the upper bound there is no concern by the system user. Therefore, our objective is to obtain the lower bound of the confidence interval (the upper bound being 1 always).

Since the natural estimator \(\hat{A}\) is biased and skewed to the left, the standard lower confidence bound (LCB)

\[
\hat{A}_{\text{std}}[\alpha] = \hat{A} - z_{1-\alpha} \frac{\hat{\sigma}}{\sqrt{n}},
\]

where \(z_{1-\alpha}\) is the \((1 - \alpha)\)-th quantile of a standard normal distribution and \(\hat{\sigma}\) is given in (2.4), is not accurate in the sense that \((\hat{A}_{\text{std}}[\alpha], 1]\) does not achieve the desired coverage probability of \(1 - \alpha\).

To improve the accuracy of the standard LCB, we use several bootstrap methods: percentile bootstrap (PB), bias-corrected and accelerated (BCa) bootstrap, bootstrap-t (Bt) and ratio bootstrap-t (RBt). Throughout the paper, by the \(\alpha\)-th quantile of a set of real numbers \(\{a_1, a_2, \ldots, a_B\}\) we mean the average of the \([aB]\)-th and the \((\lfloor aB \rfloor + 1)\)-st order statistics, where \(\lfloor x \rfloor\) denotes the largest integer not exceeding \(x\).

4.1 Percentile bootstrap interval

The PB interval is the simplest to motivate and is very easy to implement. See Efron and Tibshirani (1993). Let \(G_{\hat{A}}\) represent the empirical CDF of the bootstrap estimates \(\{\hat{A}^{*k}\} : 1 \leq k \leq B\) of \(\hat{A}\) given in (3.5). Then the lower bound of the \((1 - \alpha)\) left-sided PB interval is given by

\[
\hat{A}_P[\alpha] = G_{\hat{A}*}^{-1}(\alpha)
\]

which is the \(\alpha\)-th quantile of \(G_{\hat{A}*}\). Hall (1988) showed that the PB confidence intervals are only first-order accurate. That is,

\[
P(\hat{A} \leq \hat{A}_P[\alpha]) = \alpha + O(1/\sqrt{n}),
\]
Table 1: Bias and standard deviation of various estimators of $A$.

<table>
<thead>
<tr>
<th>$b$</th>
<th>$n$</th>
<th>$A$</th>
<th>$A_I$</th>
<th>$A_{I,0}$</th>
<th>$A_{I,1}$</th>
<th>$A_{I,2}$</th>
<th>$A_{II}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>20</td>
<td>-0.0185</td>
<td>-0.0006</td>
<td>-0.0068</td>
<td>-0.0022</td>
<td>-0.0049</td>
<td>-0.0046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0528)</td>
<td>(0.0555)</td>
<td>(0.0517)</td>
<td>(0.0610)</td>
<td>(0.0532)</td>
<td>(0.0530)</td>
</tr>
<tr>
<td>60</td>
<td>-0.0056</td>
<td>0.0004</td>
<td>-0.0006</td>
<td>0.0006</td>
<td>0.00005</td>
<td>-0.0004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0274)</td>
<td>(0.0275)</td>
<td>(0.0272)</td>
<td>(0.0313)</td>
<td>(0.0273)</td>
<td>(0.0273)</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>-0.0041</td>
<td>-0.0004</td>
<td>-0.0008</td>
<td>0.0068</td>
<td>0.0008</td>
<td>-0.0007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0206)</td>
<td>(0.0205)</td>
<td>(0.0244)</td>
<td>(0.0205)</td>
<td>(0.0200)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.2 BCa bootstrap interval

DiCiccio and Efron (1996, pp. 191-195) defines the LCB of a $(1 - a)$ one-sided BCa interval for $A$, as

$$
\hat{A}_{BCa}[a; z_0, a] = G^{-1}_{A} \left[ \Phi \left( z_0 + \frac{z_0 + z_a}{1 - a(z_0 + z_a)} \right) \right],
$$

where $z_0$ is the bias correction, $a$ is the acceleration, $G_A$ is the empirical CDF of the $B$ bootstrap estimates $\{\hat{A}^k : 1 \leq k \leq B\}$ of $\hat{A}$ and $\Phi$ is the CDF of a standard normal variate. Note that the BCa bootstrap is a generalization of the PB method in the sense that when $z_0 = 0$ and $a = 0$, the BCa LCB reduces to the PB LCB.

However, $z_0$ and $a$ are unknown. Therefore, the BCa method additionally requires estimation of $z_0$ and $a$. Indeed, $\Phi(z_0) = P\{\hat{A} < A\}$, and so

$$
\hat{z}_0 = \Phi^{-1} \left( \frac{1}{B} \sum_{k=1}^{B} I[\hat{A}^k < \hat{A}] \right).
$$

Thus, $\hat{z}_0$ can be thought of as the number of normal units by which $\hat{A}$ differs from the median of $\{\hat{A}^k : 1 \leq k \leq B\}$. Interpretation and estimation of $a$ are relatively more difficult. See DiCiccio and Efron (1996) for a detailed explanation. In short, $a$ is the rate of change of standard deviation of $\hat{A}$ on the normalized scale. Because of its relation to Fisher’s score function, one way to estimate $a$ is to use

$$
\hat{a} = \frac{1}{6} \left( \frac{\sum_{i=1}^{n} (\hat{X}^i - \hat{A}^i)^3}{\sum_{i=1}^{n} (\hat{A}^i - \hat{X}^i)^2} \right)^{3/2},
$$

with $\hat{A}^i = \frac{\hat{X}^i}{X^i + Y^i}$ and $\hat{X}^i = \frac{1}{n} \sum_{j=1}^{n} \hat{A}^j$.

Thus, $\hat{a}$ is $\sqrt{n}/6$ times the skewness of $\{\hat{A}^i : 1 \leq i \leq n\}$.

The BCa LCB is given by $A_{BCa}[a] = \hat{A}_{BCa}[a; z_0, a]$. DiCiccio and Efron (1996) showed that the BCa one-sided confidence interval $(\hat{A}_{BCa}[a], 1]$ is second-order accurate. That is,

$$
P(A \leq \hat{A}_{BCa}[a]) = a + O(1/n).
$$
4.3 Bt interval

The Bt confidence interval is conceptually simpler to describe than the BCa confidence interval and it also has the second-order accuracy property. See Hall (1988) and DiCiccio and Efron (1996, pp. 199-200). By analogy to Student’s t-statistics, we define

\[ T = \frac{\hat{A} - A}{\hat{\sigma}_A}, \]  

(4.4)

where \( \hat{\sigma}_A \) is an estimated standard deviation of \( \hat{A} \). The LCB of an \( (1 - \alpha) \) level left-sided confidence interval for \( A \) then is

\[ \hat{A} - T_{1-\alpha} \hat{\sigma}_A, \]  

(4.5)

where \( T_{1-\alpha} \) is the \( (1 - \alpha) \)-th quantile of \( T \).

However, the \( T \)-quantiles are unknown in most situations. The idea of the Bt method is to compute \( \hat{\sigma}_A \) and the quantiles of \( T \) by bootstrapping. Recall that \( \hat{A}^1, \ldots, \hat{A}^B \) are \( B \) bootstrap estimates of \( \hat{A} \). Therefore,

\[ \hat{\sigma}_A = \left\{ \frac{1}{B-1} \sum_{k=1}^{B} \left( \frac{1}{B} \sum_{l=1}^{B} \hat{A}^l - \hat{A}^k \right)^2 \right\}^{1/2}. \]  

(4.6)

Likewise, the standard deviation \( \sigma_{A^k} \) of each \( \hat{A}^k \) may be estimated by \( \hat{\sigma}_{A^k} \) using a second level nested bootstrap resample. To be more specific, from \( (X_1^{sk}, \ldots, X_n^{sk}) \) and \( (Y_1^{sk}, \ldots, Y_n^{sk}) \) we generate \( M \) second level nested bootstrap resamples \( (X_1^{sk,1}, \ldots, X_n^{sk,1}), \ldots, (X_1^{sk,M}, \ldots, X_n^{sk,M}) \) and \( (Y_1^{sk,1}, \ldots, Y_n^{sk,1}), \ldots, (Y_1^{sk,M}, \ldots, Y_n^{sk,M}) \). We take \( M = 30 \) in our simulation study (as recommended by Efron and Tibshirani, 1986). Then we compute

\[ \hat{\sigma}_{A^k} = \left\{ \frac{1}{M-1} \sum_{m=1}^{M} \left( \hat{A}^{sk,m} - \frac{1}{M} \sum_{m=1}^{M} \hat{A}^{sk,m} \right)^2 \right\}^{1/2}, \quad \hat{A}^{sk,m} = \frac{\hat{X}^{sk,m}}{\hat{X}^{sk,m} + \hat{Y}^{sk,m}}. \]  

(4.7)

Then we obtain \( B \) bootstrap replicates of \( T \) in (4.4) as

\[ T^{sk} = \frac{\hat{A}^{sk} - \hat{A}}{\hat{\sigma}_{A^k}}, \quad k = 1, \ldots, B, \]  

(4.8)

\( T_{1-\alpha} \) is estimated by the \( (1 - \alpha) \)-th quantile of \( \{T^1, \ldots, T^B\} \), and the \( (1 - \alpha) \) level Bt LCB for \( A \) is given by

\[ \hat{A}_{Bt}[\alpha] = \hat{A} - \hat{T}_{1-\alpha} \hat{\sigma}_A. \]  

(4.9)

4.4 Ratio Bt interval

For the Bt method we used \( (1 - \alpha) \)-th quantile of \( T \) given in (4.4). But this \( T \) is biased. The RBt method corrects this bias via a ratio as described below. The RBt interval also is second-order accurate just as the BCa and the Bt intervals are, though it is conceptually as well as computationally simpler than the other two.

Let \( \rho = E \left[ \hat{A}/A \right] \) and let \( \hat{\tau} \) be a consistent estimator of the standard deviation of \( \hat{A}/A \). We modify the \( T \) in (4.4) as

\[ T_{BC} = \frac{\hat{A}/A - \rho}{\hat{\tau}}. \]  

(4.10)
The LCB of \((1 - \alpha)\) level left-sided RBT interval for \(A\) is then given by

\[
\hat{A}_{RB}[\alpha; \rho, \tau] = \frac{\hat{A}}{\rho + T_{BC,1-\alpha} \tau},
\]  

(4.11)

where \(T_{BC,1-\alpha}\) is the \((1 - \alpha)\)-th quantile of \(T_{BC}\).

However, \(\rho\) and the \(T_{BC}\)-quantiles are unknown in most situations, and they have to be estimated. Also we must choose a suitable \(\tau\). Let \(\rho^*\) be the expectation and \(\tau^*\) be the standard deviation of \(\hat{A}^{*k}/\hat{A}\) under the empirical CDFs \(F\) of \(\{X_i : i = 1, \ldots, n\}\) and \(G\) of \(\{Y_i : i = 1, \ldots, n\}\). Efron (1979, pp. 12-14) showed that, by expanding \(\hat{A}^{*k}/\hat{A}\) in a Taylor series, we can approximate \(\rho^*\) and \(\tau^*\) as

\[
\rho^* \doteq \rho^*_1 := 1 - \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\bar{X}} \right) \left( \frac{X_i + Y_i}{\bar{X} + \bar{Y}} - 1 \right) - \sum_{i=1}^{n} \left( \frac{X_i + Y_i}{\bar{X} + \bar{Y}} - 1 \right)^2 \right\},
\]

\[
\tau^* \doteq \tau^*_1 := \left\{ \frac{1}{n^2} \sum_{i=1}^{n} \left[ \frac{X_i - \bar{X}}{\bar{X}} \right] \left( \frac{X_i + Y_i}{\bar{X} + \bar{Y}} - 1 \right)^2 \right\}^{1/2}.
\]

Therefore, we can use \(\rho^*_1\) as the bootstrap estimates of \(\rho\) and choose \(\tau = \tau^*_1\).

To estimate the \(T_{BC}\)-quantile, we construct bootstrap replicates of \(T_{BC}\) in (4.10). Let \(\rho^{*k}\) be the expectation and \(\tau^{*k}\) be the standard deviation of \(\hat{A}^{*k,m}/\hat{A}^{*k}\) under the empirical CDF \(\hat{F}^{*k}\) of \(\{X_i^k : i = 1, 2, \ldots, n\}\) and \(\hat{G}^{*k}\) of \(\{Y_i^k : i = 1, 2, \ldots, n\}\). As above, \(\rho^{*k}\) and \(\tau^{*k}\) have the following bootstrap estimators

\[
\rho^{*k}_1 := 1 - \frac{1}{n^2} \left\{ \sum_{i=1}^{n} \left( \frac{X_i^k - \bar{X}^k}{\bar{X}^k} \right) \left( \frac{X_i^k + Y_i^k}{\bar{X}^k + \bar{Y}^k} - 1 \right) - \sum_{i=1}^{n} \left( \frac{X_i^k + Y_i^k}{\bar{X}^k + \bar{Y}^k} - 1 \right)^2 \right\},
\]

\[
\tau^{*k}_1 := \left\{ \frac{1}{n^2} \sum_{i=1}^{n} \left[ \frac{X_i^k - \bar{X}^k}{\bar{X}^k} \right] \left( \frac{X_i^k + Y_i^k}{\bar{X}^k + \bar{Y}^k} - 1 \right)^2 \right\}^{1/2}.
\]

We compute the \(B\) bootstrap replicates of \(T_{BC}\) in (4.10) as

\[
T^{*k}_{BC} = \frac{\hat{A}^{*k}}{\hat{A}} \rho^{*k}_1, \quad k = 1, 2, \ldots, B
\]

(4.14)

and estimate the \(T_{BC,1-\alpha}\) by the \((1 - \alpha)\)-th quantile of \(\{T^{*1}_{BC}, \ldots, T^{*B}_{BC}\}\). Finally, the LCB of the \((1 - \alpha)\) level left-sided RBT interval for \(A\) is given by

\[
\hat{A}_{RB}[\alpha] = \hat{A}_{RB}[\alpha; \rho^*_1, \tau^*_1] = \frac{\hat{A}}{\rho^*_1 + T_{BC,1-\alpha} \tau^*_1},
\]

(4.15)

Indeed, one may wish to construct a bias corrected bootstrap estimator of \(A\) as follows

\[
\hat{A}_{BC} = \frac{\hat{A}}{\rho^*_1}.
\]

(4.16)

This bias corrected estimator of \(A\) in (4.16) turns out to be the same as the infinitesimal jackknife estimator of \(A\) given in (3.3). Likewise, \(\tau^*_1\) of (4.12) is both the bootstrap estimator and the infinitesimal jackknife estimator of the standard deviation of \(\hat{A}/\hat{A}\). Not only is the RBT interval a bias corrected one, but also it is easier to compute than the Bt interval since it avoids the second level nested bootstrap resamples.
4.5 Comparing various confidence interval estimators of $\hat{A}$

Under the same simulation set up as in Subsection 3.3, we now compare the various confidence interval estimators in terms of the distributions of the LCBs and the coverage probabilities. Figure 2 gives the density curves of the 1000 simulated estimates of these various 95% LCBs when $b = .5$ and $n = 60$. The figure shows that all bootstrap LCBs are stochastically to the left of $\hat{A}_{std}[.05]$, which has a coverage probability smaller than the nominated value of .95. The distributions of $\hat{A}_{BCa}[.05]$ and $\hat{A}_{PB}[.05]$ are very similar with coverage probabilities close to 95%. However, the distributions of $\hat{A}_{BT}[.05]$ and $\hat{A}_{PB}[.05]$ shift further left resulting in coverage probabilities in excess of 95%.

Table 2 gives the simulated coverage probabilities of the various left-sided 95% confidence intervals, with standard error $\sqrt{.95(.05)/n}$. An asterisk indicates a statistically significant difference based on a two-sided .05-level binomial test, since each confidence interval includes $A = 10/11$ with a 95% probability. The table shows that all bootstrap intervals provide more accurate coverage probabilities than the standard interval. However, the PB intervals do not perform as well as the other three bootstrap intervals, attesting to the fact that the PB interval is only first-order accurate, whereas the BCa, Bt and Rbt intervals are second-order accurate. Also the Rbt intervals perform better than the other bootstrap intervals in the sense that it’s simulated coverage probabilities are never significantly different from the nominated value of .95.

Table 2: Coverage probabilities of various left-sided 95% confidence intervals (and the corresponding attained significance levels of $\alpha = .05$ tests)

<table>
<thead>
<tr>
<th>$b$</th>
<th>$n$</th>
<th>$\hat{A}_{std}[.05]$</th>
<th>$\hat{A}_{PB}[.05]$</th>
<th>$\hat{A}_{BCa}[.05]$</th>
<th>$\hat{A}_{BT}[.05]$</th>
<th>$\hat{A}_{RBT}[.05]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>20</td>
<td>0.916*</td>
<td>0.972*</td>
<td>0.953</td>
<td>0.966*</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.920*</td>
<td>0.969*</td>
<td>0.947</td>
<td>0.961</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.940</td>
<td>0.969*</td>
<td>0.953</td>
<td>0.964*</td>
<td>0.957</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>0.911*</td>
<td>0.945</td>
<td>0.940</td>
<td>0.969</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.920*</td>
<td>0.937*</td>
<td>0.939</td>
<td>0.954</td>
<td>0.946</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.936*</td>
<td>0.943</td>
<td>0.947</td>
<td>0.951</td>
<td>0.950</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>0.894*</td>
<td>0.912*</td>
<td>0.933*</td>
<td>0.947</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.920*</td>
<td>0.935*</td>
<td>0.946</td>
<td>0.952</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.926*</td>
<td>0.925*</td>
<td>0.943</td>
<td>0.950</td>
<td>0.945</td>
</tr>
</tbody>
</table>
5. Hypothesis testing

In addition to the point and interval estimation, we also are interested in testing hypotheses on $A$. Since the purpose is to demonstrate statistically that the system limiting availability exceeds the specification imposed by the user, we consider a right-sided alternative hypothesis against a left-sided null hypothesis.

$$H_0: A \leq A_0 \quad \text{versus} \quad H_a: A > A_0,$$  

(5.1)

where $A_0$ is the specified limiting availability.

Because of its asymptotic normality, the natural estimator $\hat{A}$ leads to a standard test statistics which reject $H_0$ at significance level $\alpha$, if

$$\sqrt{n}(\hat{A} - A_0)/\hat{\sigma} > z_{1-\alpha},$$

or equivalently, if the left-sided confidence interval does not include $A_0$; that is,

$$\hat{A}_{\text{std}[\alpha]} > A_0.$$  

(5.2)

However, since $\hat{A}$ is biased and left skewed, the standard testing rule is not very accurate (does not achieve the specified significance level). The testing rule can be improved by using bootstrap intervals instead.

Likewise, the PB, BCa, Bt, RBt intervals respectively yield the rejection regions:

$$\hat{A}_P[\alpha] > A_0, \quad \hat{A}_{BCa}[\alpha] > A_0, \quad \hat{A}_{Bt}[\alpha] > A_0, \quad \hat{A}_{RBt}[\alpha] > A_0.$$  

(5.3)

Continuing in the same simulation set up as in Subsections 3.3 and 4.5, we specify $H_0 : \mu_F = 10$, and consider several alternative hypotheses which assign $\mu_F = 12, 14, 16, 18$. Throughout we take $\mu_G = 1$. We carry out the test in (5.1) at significance level $\alpha = .05$ based on $n = 60$ observations. Under $H_0 : \mu_F = 10$, the simulated significance levels of the various .05-level tests can be obtained by subtracting from 1 the simulated coverage probabilities given in Table 2. Figure 3 depicts the simulated powers, at the chosen values $A = 10/11, 12/13, 14/15, 16/17, 18/19$, of these various .05-level tests. To save space, the case for $b = 1$ is omitted from the figure, as its behavior is intermediate between the cases of $b = 0.5$ and $b = 2$. The figure shows that the power of each test increases with the true availability $A$. Hence, all tests are consistent. For moderate sample size, the powers of the tests based on BCa and RBt are close and they both have approximate size $\alpha$. Their powers are slightly lower than that of the test based on the standard LCB, but this is only because the later has a size larger than $\alpha$. The power of the test based on PB is the smallest when $b = 0.5$ as it attains a much smaller size than $\alpha$, but its power is a little larger than those of the other bootstrap tests when $b = 2$ as it does not achieve the correct $\alpha$ level. The power of the test based on Bt is smaller than those of the tests based on BCa and RBt, both for $b = 0.5$ when it undershoots the level correction and for $b = 2$ when it corrects the $\alpha$ level equally well as the other two. Hence, for moderate sample size it is better to use tests based on BCa and RBt, and between these two methods, the RBt is easier to motivate and compute.

6. Summary and conclusions

In this paper, we studied a one-unit repairable system under continuous monitoring and a perfect repair policy. We considered a non-parametric set up in which the lifetime and the repair time distributions have arbitrary continuous densities. We gave a comprehensive account of statistical inference on the limiting availability of the system based on fixed sized independent samples from lifetime and repair time distributions. We exhibited the drawbacks of inference using the natural estimator, and instead advocate the use of computationally more intensive jackknife and bootstrap estimators. Regarding confidence intervals and hypotheses testing, we demonstrated the superiority of the ratio bootstrap-t
method, which is easier to comprehend and implement than its closest competitor, the BCa bootstrap method.

The ideas in this paper can be extended to incorporate censored lifetime data, or to a more general one-unit repairable system that is supported by one or more identical spare units. The details will be reported in a forthcoming paper.

A more detailed measure of system performance is given by the instantaneous availability function $A(t)$, which is the probability that the system is up at a specified time $t > 0$. For an excellent account on the subject see, for example, Hoyland and Rausand (1994). Explicit evaluation of $A(t)$ is often very difficult. For some examples see Sarkar and Chaudhuri (1999), Biswas and Sarkar (2000) and Sarkar and Sarkar (2000, 2001) and Biswas et al. (2003). Inference on $A(t)$ for arbitrary lifetime and repair time distributions is delegated to a future research project.

REFERENCES


174
On the UMVUE of Reliability in Some Discrete Distributions

Jungsoo Woo*
Department of Statistics
Yeungnam University
Gyongsan, South Korea

Abstract

In this paper, we consider uniformly minimum variance unbiased estimator (UMVUE) of reliability in such important discrete distributions as Poisson, Binomial, and Negative Binomial distributions.

2000 AMS Subject Classification: 62N05; 62F10.

Key Words and Phrases: Poisson; Binomial; Negative Binomial; Tail Probability.

1 Reliability

Many authors have considered the tail probabilities in continuous distributions for reliability studies. However, for example, since the curtate-future-lifetime random variable X has non-negative values, it is natural in actuarial studies to consider the right tail probability of a discrete random variable to apply the reliability to evaluation of life insurance premiums [see Bowers et al (1997), for example]. In this paper, we consider the reliability of such important discrete random variables as Poisson, Binomial and Negative Binomial random variables and derive the UMVUE of the reliability in each case.

1.1 Poisson Model

Let $X_1, X_2, \ldots, X_n$ be independent Poisson random variables with parameter $\lambda$. Then it is well-known that $S = \sum_{i=1}^{n} X_i$ is a complete sufficient statistic for $\lambda$. Define a statistic by

$$u(X_1) = \begin{cases} 0, & \text{if } X_1 > t_0 \\ 1, & \text{if } X_1 \leq t_0. \end{cases}$$

Then $u(X_1)$ is an unbiased estimator of $P(X \leq t_0)$ in the Poisson model and hence from Lehmann-Scheffé Theorem in Rohatgi (1976), $E(u(X_1)|S)$ is an UMVUE of $P(X \leq t_0)$ in the Poisson model. Since the conditional pdf of $X_1$ given $S = s$ can be obtained as

$$P(X_1 = t|S = s) = \binom{s}{t} \left(1 - \frac{1}{n}\right)^{s-t} \left(\frac{1}{n}\right)^t, \quad t = 0, 1, 2, \ldots, s.$$
Therefore,

$$E(u(X_1)|S) = \sum_{t=0}^{[\theta]} \binom{S}{t} \left(1 - \frac{1}{n}\right)^{S-t} \left(\frac{1}{n}\right)^t$$

(1.1)

is the UMVUE of $P(X \leq t_0)$. Also, we can derive the UMVUE of the tail probability $P(X > t_0) = 1 - P(X \leq t_0)$ from result (1.1).

Especially, if $0 < t_0 < 1$, then $P(X \leq t_0) = e^{-\lambda}$. From (1.1), it is well-known [Rohatgi (1976)] that the UMVUE of $P(X \leq t_0) = e^{-\lambda}$ is $\hat{P}(X \leq t_0) = e^{-\lambda} = e^{-\sum_{i=1}^n X_i/n}$ and the first and second moments of the MLE are

$$E[\hat{P}(X \leq t_0)] = \exp(n\lambda\exp(-1/n) - 1)$$
and

$$E[(\hat{P}(X \leq t_0))^2] = \exp(n\lambda(\exp(-2/n) - 1)),$$ respectively.

When $\lambda = 1$ and $n = 10, 20, 30$, Table 1 provides the MSE of the UMVUE and the MSE of $P(X \leq t_0) = e^{-\lambda}$.

Table 1: MSE of the UMVUE and the MLE of $\hat{P}(X \leq t_0)$ when $0 < t_0 < 1$ and $\lambda = 1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMVUE</td>
<td>.01423</td>
<td>.00694</td>
<td>.00459</td>
</tr>
<tr>
<td>MLE</td>
<td>.04737</td>
<td>.00701</td>
<td>.00462</td>
</tr>
</tbody>
</table>

For this special case in Table 1, if $0 < t_0 < 1$, we observe that UMVUE of $\hat{P}(X \leq t_0) = e^{-\lambda}$ has smaller MSE than the MLE in the Poisson model.

1.2 Binomial Model

Let $X_1, X_2, \ldots, X_n$ be $n$ independent random variables each being distributed as binomial with parameter $(k, p)$, where $k$ is assumed known. Then it is well-known that $S = \sum_{i=1}^n X_i$ is a complete sufficient for $p$.

As in the case of Poisson model in Section 1.1, we can obtain the UMVUE of $P(X \leq t_0)$ in the binomial model as:

$$\sum_{t=0}^{[\theta]} \binom{k}{t} \binom{(n-1)k}{S-t} \binom{nk}{S},$$

(1.2)

where $S = \sum_{i=1}^n X_i$.

We can also derive the UMVUE of the reliability $P(X > t_0)$ from (1.2).

The MLE of $P(X \leq t_0)$ is $\hat{P}(X \leq t_0) = \sum_{i=0}^{[\theta]} \binom{k}{t} \hat{p}^i (1-\hat{p})^{k-i}$, where $\hat{p} = \frac{1}{nk} \sum_{i=1}^n X_i$.

Especially, if $0 < t_0 < 1$, then $P(X \leq t_0) = (1-p)^k$ and hence, from the result in (1.2)

$$\binom{(n-1)k}{S} \binom{nk}{S}$$

is an UMVUE of $P(X \leq t_0) = (1-p)^k,$

(1.3)
where \( S \leq (n - 1)k \) and

\[
\hat{P}(X \leq t_0) = (1 - \hat{p})^k
\]
is the MLE of \( P(X \leq t_0) = (1 - p)^k \). \hspace{1cm} (1.4)

Especially, if \( k = 1 \), i.e., each \( X_i \) is a Bernoulli random variable, it is a well-known result that the UMVUE of \( P(X \leq t_0) \) equals the MLE of \( P(X \leq t_0) \) when \( 0 < t_0 < 1 \).

From the results (1.3) and (1.4) in Section 1.2, Table 2 provides the numerical values of MSE of the UMVUE and the MLE of \( P(X \leq t_0) \) in the binomial case when \( 0 < t_0 < 1 \) and \( k = 5 \).

**Table 2:** MSE of UMVUE and MLE of \( P(X \leq t_0) \) in the binomial case when \( 0 < t_0 < 1 \) and \( k = 5 \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p )</th>
<th>( 0.25 )</th>
<th>( 0.50 )</th>
<th>( 0.75 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>MLE</td>
<td>0.01034</td>
<td>0.000743</td>
<td>0.0000048</td>
</tr>
<tr>
<td></td>
<td>UMVUE</td>
<td>0.01327</td>
<td>0.001396</td>
<td>0.0000181</td>
</tr>
<tr>
<td>20</td>
<td>MLE</td>
<td>0.00491</td>
<td>0.000305</td>
<td>0.0000014</td>
</tr>
<tr>
<td></td>
<td>UMVUE</td>
<td>0.00563</td>
<td>0.000438</td>
<td>0.0000033</td>
</tr>
<tr>
<td>30</td>
<td>MLE</td>
<td>0.00324</td>
<td>0.000189</td>
<td>0.0000008</td>
</tr>
<tr>
<td></td>
<td>UMVUE</td>
<td>0.00413</td>
<td>0.000273</td>
<td>0.0000016</td>
</tr>
</tbody>
</table>

From Table 2, we observe that the MLE of the reliability has less MSE than the UMVUE of the reliability in the binomial case when \( 0 < t_0 < 1 \), \( k = 5 \) and \( p = 0.25, 0.50, 0.75 \).

### 1.3 Negative Binomial Model

The pmf of negative binomial random variable with parameters \( r \) and \( p \) is given by

\[
f(x) = \binom{x + r - 1}{x} p^r (1 - p)^x, \hspace{1cm} x = 0, 1, 2, 3, \ldots
\]

We now consider the UMVUE of the tail probability in the negative binomial distribution. Let \( X_1, X_2, \ldots, X_n \) be \( n \) independent negative binomial random variables each with parameters \( (r_i, p) \), \( i = 1, 2, \ldots, n \) when the parameters \( r_i \) are known. Again, it is well-known that \( S = \sum_{i=1}^{n} X_i \) is a complete sufficient statistic for the parameter \( p \), \( 0 < p < 1 \).

In the similar fashion as in Section 1.1, we obtain the UMVUE of \( P(X \leq t_0) \) in the negative binomial case as follows.

\[
\frac{\sum_{i=0}^{t_0} \binom{r_i + t - 1}{t} \left( \frac{S - t + \sum_{i=2}^{n} r_i - 1}{S - t} \right)}{\binom{S + \sum_{i=1}^{n} r_i - 1}{S}}
\hspace{1cm} (1.5)
\]

where \( S = \sum_{i=1}^{n} X_i \). We can also get the UMVUE of the reliability \( P(X > t_0) \) from (1.5).

The MLE of \( P(X \leq t_0) \) is \( \hat{P}(X \leq t_0) = \sum_{i=0}^{t_0} \binom{t + r - 1}{t} \hat{p}^r (1 - \hat{p})^x \), where \( \hat{p} = n/[n + \sum_{i=1}^{n} X_i ] \).

Especially, if \( 0 < t_0 < 1 \) and each \( r_i = 1 \), then \( P(X \leq t_0) = p \) and hence from (1.5),

\[
\frac{n - 1}{n - 1 + \sum_{i=1}^{n} X_i}
\]
is an UMVUE of \( P(X \leq t_0) = p \), \hspace{1cm} (1.6)

and the MLE of \( P(X \leq t_0) = p \) is

\[
\hat{P}(X \leq t_0) = \hat{p} = \frac{n}{n + \sum_{i=1}^{n} X_i}
\hspace{1cm} (1.7)
\]

177
Table 3: MSE of UMVUE and MLE of $P(X \leq t_0)$ in the negative binomial case when $0 < t_0 < 1$ and $r_i = 1, \forall i$.

From the results (1.6) and (1.7) in Section 1.3, Table 3 provides the numerical values of MSE of the UMVUE and the MLE of $P(X \leq t_0)$ in the negative binomial case when $0 < t_0 < 1$ and $r_i = 1, \forall i$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>MLE</td>
<td>0.00635</td>
<td>0.014149</td>
<td>0.013521</td>
</tr>
<tr>
<td></td>
<td>UMVUE</td>
<td>0.00548</td>
<td>0.013805</td>
<td>0.014745</td>
</tr>
<tr>
<td>20</td>
<td>MLE</td>
<td>0.00265</td>
<td>0.006693</td>
<td>0.006917</td>
</tr>
<tr>
<td></td>
<td>UMVUE</td>
<td>0.00242</td>
<td>0.006569</td>
<td>0.007205</td>
</tr>
<tr>
<td>30</td>
<td>MLE</td>
<td>0.00123</td>
<td>0.004387</td>
<td>0.004649</td>
</tr>
<tr>
<td></td>
<td>UMVUE</td>
<td>0.00101</td>
<td>0.004307</td>
<td>0.004765</td>
</tr>
</tbody>
</table>

From Table 3, for $r_i = 1, \forall i$, we observe that the UMVUE of the reliability has less MSE than the MLE of the reliability in the negative binomial case when $0 < t_0 < 1$ and $p = 0.25$, and 0.50 while the MLE of the reliability has less MSE than the UMVUE of the reliability in the negative binomial case when $0 < t_0 < 1$ and $p = 0.75$.

REFERENCES


A Stochastic Representation of Matrix Variate Skew Normal Models

John T. Chen
Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, Ohio 43403
E-mail: jchen@bgsu.edu
Phone: 419-372-7461
Fax: 419-372-6092

Summary

The skew normal distribution and its extensions have been thoroughly investigated in the literature. However, there is no result regarding the matrix form of stochastic representation, which reveals insightful probabilistic interpretation for the matrix variate skew normal model. In this paper, after reviewing stochastic representations for the univariate and the vector variate skew normal models, we present a result of a matrix variate stochastic formulation using a truncated linear combination of two independent normal random matrices. After setting the connection with other skew normal models, we formulate a matrix variate skew model that can be used to amalgamate other versions of skew normal families.

AMS 1999 Subject Classification: Primary 62H10; Secondary 62H11

Keywords: Skew normal model, stochastic representation, matrix variate normal distribution, truncated random matrix.

1 Introduction

Let $X$ be a skew normal random variable, according to Azzalini (1985), $X$ has density

$$f(x) = 2\phi(x)\Phi(\lambda x),$$

where $\phi(x)$ and $\Phi(x)$ are, respectively, the density and cumulative distribution function of the standard normal model. One of the main interesting parts of the skew normal model is the use of the parameter $\lambda$ that models the asymmetry of the data. Henes(1986) showed that the skew normal model is endowed with the following probabilistic interpretation,

$$X = \lambda |Z_1| + Z_2,$$

where $Z_1$ and $Z_2$ are two independent standard normal random variables. Equation (1) opens a path to simulate the skew normal model using the standard normal random variables.
For the multivariate extension of the skew normal model, Azzalini (1985) defined a skew normal vector as follows. A random vector $\mathbf{y}$ follows a vector variate skew normal distribution if the joint density of $\mathbf{y}$ can be expressed as:

$$f(\mathbf{y}) = c\phi_{k,\Omega}(\mathbf{y}) \prod_{j=1}^{k} \Phi(\beta_j y_j)$$

(2)

where $\phi_{k,\Omega}(\cdot)$ is the density of a multivariate normal random vector with correlation matrix $\Omega = (\rho_{ij})$, $\mathbf{y} = (y_1, \ldots, y_k)' \in \mathbb{R}^k$, $\beta_1, \ldots, \beta_k$ are $k$ real numbers, $c^{-1}$ is the orthant probability of a standardized normal random variable. Obviously the family of random vectors following such joint distribution utilizes parameters $\Omega$ and $\beta$ to model the skew distribution of the data, where vector $\beta = (\beta_1, \ldots, \beta_k)'$.

As pointed out by Azzalini and Dalla Valle (1996), the definition in (2) is just a direct extension of the univariate skew normal distribution. To model multivariate data, they put forward another version of vector variate skew normal model as follows. A random vector $\mathbf{y}$ follows a vector skew normal distribution if the joint density of $\mathbf{y}$ can be formulated as

$$f(\mathbf{y}) = 2\phi_{k}(\mathbf{y}, \Omega)\Phi(\mathbf{\alpha}'\mathbf{y}) \text{ for } \mathbf{y} \in \mathbb{R}^k,$$

(3)

where $\mathbf{\alpha}$ is a vector of $k$ real numbers. In this definition, the parameters used are systematically designed in $\Omega$ and $\mathbf{\alpha}$ to model the skew property of the data.

Azzalini and Capitanio (1999) applied model (3) and showed that it possesses statistical properties such as the quadratic form of such skew normal random vector is $\chi^2_k$, along with a probabilistic interpretation on the basis of the stochastic representation. However, in the definition (3), $\Omega$ is not the correlation matrix of $\mathbf{y}$, and $\Omega = I$ does not imply that all components of $\mathbf{y}$ are independent due to the skew vector involved in the model.

Although the modified definition for the skew normal family (3) has many interesting properties, Gupta and Chen (2001, 2003) observed that if $Y_1, \ldots, Y_k$ constitute a random sample from skew normal population $SN(\lambda)$, the distribution of $\mathbf{y} = (Y_1, \ldots, Y_k)'$ is not included in the skew family (3). Consequently, the distribution of the mean of a random sample does not belong to the same family. For the multivariate extension of the skew normal model, Gupta and Chen (2004) showed that another possible extension of the univariate skew normal model into vector variate skew normal models, is to consider the vector $\mathbf{x}$ follows the density

$$f(\mathbf{x}) = 2^k \phi_{k,0,\Omega}(\mathbf{x}) \prod_{j=1}^{k} \Phi(\varDelta_j \mathbf{x})$$

where $\phi_{k,0,\Omega}$ is the density of a multivariate normal random vector with correlation matrix $\Omega$. $\varDelta_1, \ldots, \varDelta_k$ are real vectors satisfying

$$(\varDelta_1, \ldots, \varDelta_k) = \Omega^{-1/2}diag(\delta_1, \ldots, \delta_k)$$

for a skew vector $\varDelta = (\delta_1, \ldots, \delta_k)'$. For this model, Gupta and Chen (2004) pointed out that it has the following stochastic representation,

$$\mathbf{z} = \text{diag}(\frac{\delta_1}{\sqrt{1+\delta_1^2}}, \ldots, \frac{\delta_k}{\sqrt{1+\delta_k^2}})\mathbf{x} + \text{diag}(\frac{1}{\sqrt{1+\delta_1^2}}, \ldots, \frac{1}{\sqrt{1+\delta_k^2}})\mathbf{y},$$

(4)

where $|\mathbf{x}| = (|X_1|, \ldots, |X_k|)'$. The stochastic representation (4) naturally develops from (1) in the sense that letting $k = 1$ results in the univariate model (1). It also possesses statistical properties as the one defined by Azzalini and Dalla Valle (1996).

In spite of many vector versions of the skew normal family that may be applied to model multivariate data with skew vectors, in practice, one is frequently confronted with matrix variate models. For instance, in the case where the experimenter is interested in comparing the treatment effects of $n$ dosages of a drug (with certain number of patients in each treatment group), in which each patient has $m$ spontaneous measurements for the symptoms or responses corresponding to the drug (for example, blood pressure, cholesterol level, weight, height, abdomen measurement, body fat percentage, to list a few), we have to consider the matrix variate model. Especially when the treatments
are correlated and the distribution of the data is skew, the study of the matrix variate skew model is called.

Let \( X_{ij} \) represent the result of the \( i \)th measurement \((i = 1, \ldots, m)\) from the \( j \)th treatment \((j = 1, \ldots, n)\). In the literature, the normal random matrix refers to the distribution family in which \( X = (X_{ij}) \sim N_{m,n}(M, \Sigma \otimes \Psi) \) where \( M \in R^{m \times n} \) is an \( m \times n \) mean matrix, \( \Sigma \) is an \( m \times m \) positive definite matrix (describing the dependence structure among the measurements) and \( \Psi \) is an \( n \times n \) positive definite matrix (describing the dependence structure among treatments). The density of a normal matrix variate, \( X \sim N_{m,n}(M, \Sigma \otimes \Psi) \), takes the following form: (See, for example, Gupta and Nagar (2000) and Nguyen (1997))

\[
f(X) = (2\pi)^{-\frac{mn}{2}} \det(\Sigma)^{-n/2} \det(\Psi)^{-m/2} \text{etr} \left[ -\frac{1}{2} \Sigma^{-1} (X - M) \Psi^{-1} (X - M)' \right]
\]

(5)

where \( M \) is an \( m \times n \) mean matrix.

Chen and Gupta (2005) extended the vector skew normal model to matrix variates. However, the corresponding stochastic representation for that model is not available. In this paper, to search for a stochastic representation, we investigate the distribution of a linear combination of two normal matrices with one truncated at the original point. The result of the investigation leads to a stochastic representation, which naturally brings in a new matrix variate model to unify the univariate and vector variate skew normal models.

2 Truncated Linear Combination of Normal Matrices

As shown in (4) in the preceding section, the stochastic representation of the univariate (or vector variate) skew normal models takes the form of being a linear combination of two standard normal variables (or vectors) with one variable (or vector) truncated at the original. In this section we derive the distribution of a matrix form of this type of stochastic representations, namely a linear combination of two normal random matrices with one truncated at the original. To this end, we obtain the following result.

For any positive definite matrix \( \Psi > 0 \), let \( X, Y \) be two independent random matrices following \( N_{m,n}(0, I_m \otimes \Psi) \). For any \( d = (\delta_1, \ldots, \delta_m)' \in R^m \), denote

\[
Z = \text{diag}(\frac{\delta_1}{\sqrt{1 + \delta_1^2}}, \ldots, \frac{\delta_m}{\sqrt{1 + \delta_m^2}}) X + \text{diag}(\frac{1}{\sqrt{1 + \delta_1^2}}, \ldots, \frac{1}{\sqrt{1 + \delta_m^2}}) Y,
\]

(6)

where \( \|X\| = (|x_{ij}|) \) is the matrix formed by the absolute value of every element in \( X \). With this setting, we have

**Theorem 1.** The random matrix \( Z \) defined in (6) has the following density

\[
f(Z) = 2^{-mn} \phi_{mn,0,1 \otimes \Psi}(Z) \Phi_{mn,0,1 \otimes \Psi}(\Lambda Z),
\]

(7)

where \( \Lambda = \text{diag}(\delta_1, \ldots, \delta_m) \) is the parameter matrix consisting of skew factors \( \delta_1, \ldots, \delta_m \).

**Proof:** For convenience, denote the matrices

\[
U = \text{diag}(\frac{\delta_1}{\sqrt{1 + \delta_1^2}}, \ldots, \frac{\delta_k}{\sqrt{1 + \delta_k^2}})
\]

and

\[
V = \text{diag}(\frac{1}{\sqrt{1 + \delta_1^2}}, \ldots, \frac{1}{\sqrt{1 + \delta_k^2}}),
\]

then the truncated linear combination can be expressed as

\[
Z = U X + V Y
\]

(8)

and

\[
V Y \sim N_{mn}(0, (V' V) \otimes \Psi).
\]

(9)
For any real matrix $T \in R^{m \times n}$, the probability associated with the matrix variate $Z$ can be evaluated as follows.

$$P(Z \leq T) = E_{X \mid \{P(Z \leq T \mid X)\}}$$

$$= \int_{R^{m \times n}} P(Z \leq T)2^{-mn}\phi_{mn, I \otimes \Psi}(X)dX$$

(Notice that $R_{+}^{mn} = (0, +\infty)^{mn}$).

Now consider the matrix equation (7) in conjunction with (9) and (10), we have

$$F_Z(T) = \int_{R^{m \times n}} P(\mathbf{VY} \leq T - UX)2^{-mn}\phi_{mn, I \otimes \Psi}(X)dX$$

$$= \int_{R^{m \times n}} \Phi_{mn, \mathbf{V} \otimes \Psi}(T - UX)2^{-mn}\phi_{mn, I \otimes \Psi}(X)dX.$$

Recall that the CDF of the normal random matrix

$$\Phi_{mn, \Sigma \otimes \Psi}(X) = \int_{R^{m \times n}} \phi_{mn, \Sigma \otimes \Psi}(Y)dY.$$

We obtain the joint density of $Z$ as follows.

$$f(T) = d(F_Z(T))/dT$$

$$= \int_{R^{m \times n}} \{d(\Phi_{mn, \mathbf{V} \otimes \Psi}(T - UX))/dT\}2^{-mn}\phi_{mn, I \otimes \Psi}(X)dX$$

$$= \int_{R^{m \times n}} (2\pi)^{-mn/2}|\mathbf{V}|^{-n/2}2^{-mn}\exp\{-\frac{1}{2}tr((T - UX)'(V^{-1})(T - UX))\Psi^{-1}\}dX.$$  

Now, since $\mathbf{V}$ is a diagonal matrix satisfying (9), the density of the random matrix $Z$ can be written as follows.

$$f(T) = \int_{R^{m \times n}} (2\pi)^{-mn/2}|\mathbf{V}|^{-n/2}2^{-mn}\exp\{-\frac{1}{2}tr(A\Psi^{-1})\}dX,$$

where the matrix $A$ takes the following form

$$A = (T - UX)'(\mathbf{V} \otimes \Psi)^{-1}(T - UX) + X'X$$

$$= T'V^{-2}T - T'V^{-2}UX + X'U'V^{-2}T + X'U'V^{-2}UX + X'X.$$  

In the expression above, the matrices $U$ and $V$ are diagonal Matrices with the relationship

$$U'V^{-2}U + I = V^{-2}.$$  

Therefore the matrix $A$ can be expressed as

$$A = (X - UT)'V^{-2}(X - UT) + T'T.$$  

(11) in conjunction with (12) yields the expression of the density of the random matrix $Z$ at point $T$ as follows.

$$f(t_{ij}, i = 1, \ldots, m, j = 1, \ldots, n)$$

$$= \int_{R^{m \times n}} (2\pi)^{-mn/2}|\mathbf{V}|^{-n/2}2^{-mn}\exp\{-\frac{1}{2}tr((X - UT)'V^{-2}(X - UT) + T'T)\Psi^{-1}\}dX$$

$$= 2^{-mn}\phi_{mn, \mathbf{0} \otimes \Psi}(T)\int_{R^{m \times n}} (2\pi)^{-mn/2}|\mathbf{V}|^{-n/2}\Psi^{-m/2}\exp\{-\frac{1}{2}tr((Q - UV^{-1}T)'(Q - UV^{-1}T)\Psi^{-1})\}dQ.$$  

Letting $Q = V^{-1}X$ yields

$$f(t_{ij}, i = 1, \ldots, m, j = 1, \ldots, n)$$

$$= 2^{-mn}\phi_{mn, \mathbf{0} \otimes \Psi}(T)\int_{R^{m \times n}} (2\pi)^{-mn/2}|\Psi|^{-m/2}\exp\{-\frac{1}{2}tr((Q - UV^{-1}T)'(Q - UV^{-1}T)\Psi^{-1})\}dQ.$$  

182
Applying the special structure of matrices $U$ and $V$ gets the density of $Z$ as

$$2^{-mn}\phi_{mn,0}\Pi\Psi(T)\Phi_{mn,0}\Pi\Psi(\Lambda T),$$

where $\Lambda = UV^{-1} = diag(\delta_1, ..., \delta_m)$, is the matrix consisting of all the skew factors. □

**Remark 1:** Theorem 1 reveals the distribution family of the standard formulation of stochastic representation for random matrices. When $m = 1$, it returns to the model of Gupta and Chen (2004), when $n = m = 1$ it returns to the univariate skew normal model. Notice that one may define a new matrix variate skew normal family on the basis of Theorem 1.

### 3 A Family of Skew Normal Matrices

Theorem 1 in Section 2 may be intuitively obtained as follows, the vectorizations of independent random matrices $vec(X)$ and $vec(Y)$ follow vector normal model with mean $0$ and covariance matrix $I \otimes \Psi$, thus the distribution of the linear combination of two independent random normal vectors with one truncated at the original follows a skew normal distribution. Thus one may conjecture that the vectorization of the random matrix $Z$ follows a multivariate skew normal model. Theorem 1 proves this conjecture. In what follows, we further verify the eligibility of the function (7) as a density function.

**Theorem 2** For any $m \geq 1$ and real number $\delta_1, ..., \delta_m$, let $\Lambda = diag(\delta_1, ..., \delta_m)$. Let $\Psi$ be a positive definitive matrix, we have

$$\int_{R^{mn}} 2^{-mn}\phi_{mn,0}\Pi\Psi(T)\Phi_{mn,0}\Pi\Psi(\Lambda T)dT = 1.$$  

**Proof:** To evaluate the integration of the function (7) over the whole domain of $R^{mn}$, notice that

$$\int_{R^{mn}} 2^{-mn}\phi_{mn,0}\Pi\Psi(T\Phi_{mn,0}\Pi\Psi(\Lambda T)dT = 2^{-mn}E_W(\Phi_{mn,0}\Pi\Psi(\Lambda W)),$$  

where $W$ is a random matrix following the normal model with mean matrix $0$ and covariance matrix $I \otimes \Psi$. Thus the right-hand side of (13) can be written as

$$2^{-mn}E_W(\Phi_{mn,0}\Pi\Psi(\Lambda W)) = 2^{-mn}E_W(P(R \leq \lambda W|W)),$$  

where $R$ is a random matrix following the normal model with mean matrix $0$, covariance matrix $I \otimes \Psi$, and the random matrices $R$ and $W$ are independent. We can thus simplify the right-hand side of (14) as

$$2^{-mn}E_W(P(R \leq \lambda W|W)) = 2^{-mn}P(R - \Lambda W \leq 0) = 2^{-mn}P(S \leq 0) = 1,$$

where $S$ is a normal random matrix with mean matrix $0$ and covariance matrix $(I + \Lambda^2) \otimes \Psi$. □

**Remark** Theorem 2 covers many versions of the skew normal models. When $n = m = 1$ it returns to the univariate skew normal model, when $m=1$, it returns to the multivariate skew normal model of Gupta and Chen (2004).

Theorem 2 actually specifies a family of skew normal random matrices. In univariate case the stochastic representation is a linear combination with one normal variable truncated at zero to form the asymmetric structure of the distribution. In the vector version, it comes from the same format except the skew factors formed in two diagonal matrices. And for the matrix variate cases, the stochastic representation is constructed in the same way with special diagonal matrices almost identical to the vector variate cases.
References


Statistics courses are a required part of many graduate programs in the social sciences and education. At the same time, for many students statistics is one of the most anxiety-inducing courses in their programs of study (Onwuegbuzie & Wilson, 2003). Indeed, this anxiety may be so overwhelming that many students put off taking statistics until the end of their college career, if at all. Onwuegbuzie and Wilson (2003) have posited that perhaps the failure to complete required statistics courses due to their anxiety-inducing nature may be partly responsible for many students failure to complete their graduate programs. Onwuegbuzie, DaRos, and Ryan (1997) have defined statistics anxiety as a very generalized fear, occurring any time a student encounters statistics in any form and at any level. On the other hand, Zeidner (1990) provided a more specific definition in which statistics anxiety is characterized by a variety of psychological disturbances (e.g., worry, stress) when students must deal with the subject in an academic context. Cruise, Cash, and Bolton (1985) defined statistics anxiety as consisting of six factors: worth of statistics, interpretation anxiety, test and class anxiety, computational self-concept, fear of asking for help, and fear of statistics teachers. Another definition of statistics anxiety was presented by Schau, Stevens, Dauphinee, and Del Vecchio (1995) who identified four factors of statistics anxiety: affect, cognitive competence, value, and difficulty. In short, while statistics anxiety has been described in a variety of ways over time, there are certain commonalities that appear to be present across definitions: psychological stress, fear of academic failure and feelings of incompetence.

The causes of statistics anxiety have been classified into three broad categories: situational, dispositional, and environmental (Baloglu & Zelhart, 2003; Onwuegbuzie et al., 1997). Situational causes are associated with the actual participation in a statistics course, such as the statistics instructor and his/her teaching style, overall satisfaction with the statistics course, grade received in the statistics course, and status of the course (Onwuegbuzie & Wilson, 2003). The dispositional causes include characteristics that the student brings to the class, most particularly self-esteem, mathematics self-concept, perfectionism (Onwuegbuzie & Daley, 1999) and procrastination (Onwuegbuzie, 2004). Finally, environmental causes include demographic factors such as gender, age, race, and academic major (Onwuegbuzie et al., 1997, p. 19), among others. Regardless of the cause of the statistics anxiety, it has been shown to have a negative relationship with course performance, so that greater anxiety was associated with lower course performance (Onwuegbuzie & Daley, 1996; Zeidner, 1991). Indeed, it has been shown to be the best predictor of achievement in research methodology (Onwuegbuzie, Slate, Paterson, Watson, & Schwartz, 2000) and statistics courses (Fitzgerald, Jurs, & Hudson, 1996).

Given this clear link between statistics anxiety and performance in statistics courses, and the increasing prominence of statistics in many graduate programs, it is important for instructors to have a sense of students’ levels of anxiety as they begin studying the subject. The goal of the current study was to ascertain the level of anxiety. The research questions of interest in this study are:

- What is the general level of anxiety upon entering a graduate statistics course?
- Is this anxiety related to prior statistics coursework, or lack thereof?
- Can students be classified into a few discrete groups based upon their expectations regarding graduate statistics?
Statistics anxiety was assessed using 6 questions, which appear in Appendix A. These items were not part of a formal scale, and were designed to measure different aspects of generalized statistics anxiety. In addition, the questionnaire contained items asking students to provide their program of study and the number of statistics courses they had taken in their undergraduate program.

**Methods**

Data were collected from 178 students enrolled in one of three graduate level statistics courses offered by the Department of Educational Psychology at a Midwestern university over a period of 4 semesters. None of the students had taken a graduate level statistics course previously. Each subject was asked to complete a survey on the first day of class in order to obtain information regarding previous statistics coursework they have completed, as well as their responses to 6 items designed to gain an understanding of their anxiety regarding statistics. The items were on a 5-point Likert scale from strongly agree to strongly disagree. As can be seen from an examination of the items, greater agreement corresponds to greater anxiety about some aspect of the statistics course.

The first research question was answered using simple descriptive statistics for each of the 6 anxiety items. In order to determine whether previous exposure to statistics during their undergraduate training was related to responses on any of these anxiety items, Spearman’s rho correlation coefficient was used. Finally, cluster analysis was used to determine whether or not definite groups could be identified given the response patterns to the 6 anxiety items. Specifically, K Means clustering was used in an experimental fashion, with a determination as to the number of clusters present in the data based upon a maximization of the distance among cluster centers, and a contextual examination of the cluster response patterns to the items.

**Results**

Of the 178 subjects included in this study, 54 (30.5%) were doctoral candidates, 118 (66.1%) were masters students and 6 (3.4%) were undergraduates. The mean number of statistics courses taken prior to this one was 0.8107 (SD=0.9998), with 85 (47.9%) having never taken one before. The distribution of subjects by home department appears in Table 1. As can be seen, the majority of students came from one of three social science departments, Education, Educational Psychology and Counseling and Clinical Psychology. The Other category included individuals from a wide range of departments and programs including linguistics, dietetics, environmental science, physical education, sociology and gerontology, among others.

<table>
<thead>
<tr>
<th>Department</th>
<th>Number</th>
<th>(Percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>44</td>
<td>(24.7%)</td>
</tr>
<tr>
<td>Educational Psychology</td>
<td>37</td>
<td>(20.8%)</td>
</tr>
<tr>
<td>Counseling/Clinical Psychology</td>
<td>39</td>
<td>(21.9%)</td>
</tr>
<tr>
<td>Exercise Science/Physiology</td>
<td>23</td>
<td>(13.0%)</td>
</tr>
<tr>
<td>Other</td>
<td>35</td>
<td>(19.6%)</td>
</tr>
</tbody>
</table>

The first research question to be addressed by this study focused on the overall level of anxiety students had regarding the statistics course at its outset. The means and standard deviations for each of the items appear in Table 2. When interpreting these results, it is important to remember that the range of possible responses was from 1 to 5, with larger values indicating greater anxiety regarding that aspect of the statistics course. It appears that the students were least anxious about failing the course, and most anxious regarding the amount of work they would be required to do relative to other courses in their graduate program. They were also somewhat more concerned about the relative difficulty of the course compared to others that they would take, but given that the mean was less than 3.0, they could not be considered generally very anxious.

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Concepts)</td>
<td>2.60</td>
<td>1.09</td>
</tr>
<tr>
<td>2 (Symbols)</td>
<td>2.47</td>
<td>0.95</td>
</tr>
<tr>
<td>3 (Fail)</td>
<td>1.73</td>
<td>0.76</td>
</tr>
<tr>
<td>4 (Anxiety)</td>
<td>2.32</td>
<td>1.01</td>
</tr>
<tr>
<td>5 (Difficult)</td>
<td>2.96</td>
<td>0.89</td>
</tr>
<tr>
<td>6 (Study)</td>
<td>3.11</td>
<td>0.84</td>
</tr>
</tbody>
</table>

The second research question of interest in this case was whether the number of statistics courses a student had taken prior to the current one was related to any of these anxiety items. Table 3 contains the values of the Spearman’s rho correlation coefficients between the number of courses students reported having taken prior to the current one, and their responses to the 6 anxiety items on the survey. From these correlations, it is possible to conclude that the number of courses taken by individual students was only related to their anxiety of failing the class. Specifically, those who had taken more statistics classes previously were less likely to be concerned about failing the current course. The number of statistics courses

Table 1: Distribution of home departments
previously taken was not related to any other of the anxiety items.

**Table 3:** Spearman’s rho values for the correlation between number of courses previously taken and responses to anxiety items

<table>
<thead>
<tr>
<th>Item</th>
<th>Number of courses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Concepts)</td>
<td>-0.09</td>
</tr>
<tr>
<td>2 (Symbols)</td>
<td>-0.10</td>
</tr>
<tr>
<td>3 (Fail)</td>
<td>-0.16*</td>
</tr>
<tr>
<td>4 (Anxiety)</td>
<td>-0.09</td>
</tr>
<tr>
<td>5 (Difficult)</td>
<td>-0.06</td>
</tr>
<tr>
<td>6 (Study)</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

* p-value < 0.05

The results of the K-Means cluster analysis, using the 6 item responses to develop the clusters, revealed the presence of three distinct clusters of students. This determination as to the number of factors was based on both the maximization of the between cluster distances and descriptive analyses of the clusters. Table 4 displays the means for each of the items by cluster.

**Table 4:** Item means by cluster membership

<table>
<thead>
<tr>
<th>Item</th>
<th>Cluster 1 (103)</th>
<th>Cluster 2 (45)</th>
<th>Cluster 3 (30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Concepts)</td>
<td>1.87</td>
<td>3.40</td>
<td>3.90</td>
</tr>
<tr>
<td>2 (Symbols)</td>
<td>1.90</td>
<td>3.31</td>
<td>3.13</td>
</tr>
<tr>
<td>3 (Fail)</td>
<td>1.41</td>
<td>1.73</td>
<td>2.83</td>
</tr>
<tr>
<td>4 (Anxiety)</td>
<td>1.77</td>
<td>2.53</td>
<td>3.90</td>
</tr>
<tr>
<td>5 (Difficult)</td>
<td>2.55</td>
<td>3.24</td>
<td>3.93</td>
</tr>
<tr>
<td>6 (Study)</td>
<td>2.81</td>
<td>3.24</td>
<td>3.97</td>
</tr>
</tbody>
</table>

Linear Discriminant Analysis (LDA) was used in order to gain insights into which of the items most differentiated the clusters. Specifically, the cluster membership served as the classification variable while the 6 items were the predictors. Two discriminant functions were found to be statistically significant at $p=0.05$. Variables were said to be important in differentiating the groups if their structure coefficients were greater than 0.4. The structure values appear in Table 5.

**Table 5:** Structure coefficients for anxiety items based on LDA

<table>
<thead>
<tr>
<th>Item</th>
<th>Structure coefficient 1</th>
<th>Structure coefficient 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Concepts)</td>
<td>0.747</td>
<td>-0.362</td>
</tr>
<tr>
<td>2 (Symbols)</td>
<td>0.501</td>
<td>-0.598</td>
</tr>
<tr>
<td>3 (Fail)</td>
<td>0.501</td>
<td>0.430</td>
</tr>
<tr>
<td>4 (Anxiety)</td>
<td>0.695</td>
<td>0.324</td>
</tr>
<tr>
<td>5 (Difficult)</td>
<td>0.422</td>
<td>0.038</td>
</tr>
<tr>
<td>6 (Study)</td>
<td>0.331</td>
<td>0.137</td>
</tr>
</tbody>
</table>

Given these results, it appears that all of the items contributed to cluster separation except for the amount of time that students expected to study for this course as compared to others in their graduate program. Indeed, based on the means in Table 4 members of all three clusters had relatively high mean values for this item. On the other hand, cluster 1 clearly had lower means on the other items than did the other groups, indicating less anxiety regarding the statistical concepts, use of symbols, fear of failure, anxiety about the course and the relative difficulty of the course. In addition, cluster 2 had lower means that cluster 3 on all of the items except for symbols. Indeed, for both fear of failure and overall anxiety, the mean for cluster 2 was more than 1 point lower than that for cluster 3.

Finally, in order to gain more insights into the nature of these clusters, they were compared with respect to the mean number of statistics courses previously taken, using Analysis of Variance (ANOVA). The results were not statistically significant, indicating that there were no differences among the clusters in terms of the mean number of statistics courses previously taken. The chi-square test of association was used to ascertain whether there was a significant relationship between students’ home departments and cluster membership. These results were statistically significant ($p=0.039$), so that such a relationship was present. Table 6 contains the distribution of home department by cluster membership. Please note that due to smaller sample sizes for the other programs, only those in Educational/School Psychology, Education and Counseling Psychology were included in this analysis.

**Table 6:** Cluster membership by home department

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Educational or School Psychology</th>
<th>Education</th>
<th>Counseling Psychology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28 (41.2%)</td>
<td>18 (26.5%)</td>
<td>22 (32.4%)</td>
</tr>
<tr>
<td>2</td>
<td>5 (15.2%)</td>
<td>17 (51.5%)</td>
<td>11 (33.3%)</td>
</tr>
<tr>
<td>3</td>
<td>4 (21.1%)</td>
<td>9 (47.3%)</td>
<td>6 (31.6%)</td>
</tr>
</tbody>
</table>

Cluster 1 appears to be dominated by Psychology students, particularly those in Educational/School Psychology. On the other hand, approximately half of the members of both clusters 2 and 3 were from a program in Education, with relatively low representation from Educational/School Psychology. For all three clusters, individuals from Counseling Psychology make up approximately one third of the membership.
Discussion

The goal of this study was to gain greater understanding regarding the level of anxiety that students bring with them into an introductory graduate level statistics class for students in non-statistics educational programs. In addition, this study focused on the relationship between prior exposure to statistics in undergraduate training and statistics anxiety. The results of the study appear to show that in fact, prior exposure to statistics courses does not relate strongly to current levels of anxiety in graduate students. The only statistically significant correlation between the individual anxiety items and number of courses taken was for fear of failure, which was negative indicating that the more prior courses taken, the less fearful students were of failing the current course. However, this relationship was not very strong (r= -0.16). An examination of the means of the items revealed that generally speaking, students were not terribly anxious about the course with no item mean larger than 3.0 except for study time.

A cluster analysis using the anxiety items identified 3 distinct groups of students. The largest of these groups (cluster 1) displayed the lowest levels of anxiety on all of the items, and appears to have been made up largely of individuals in Educational/School Psychology and Counseling Psychology. The other two groups, while much smaller, were characterized by being in Education programs. These two groups could be distinguished primarily by differences on the items asking about fear of failure and general anxiety. These results were particularly interesting in light of the fact that none of the groups had significantly different mean values for the number of statistics courses previously taken. Despite this lack of a significant difference, students coming from a Psychology background will have generally taken several technical classes, such as research design, tests and measurement and cognitive assessment. While these are not purely statistics classes, they all carry a research/quantitative component that may help ease anxiety that these students have regarding more technical subject matter. On the other hand, many of the students from purely educational programs do not typically need to have a background in testing and measurement and thus may not have had many such technical courses, thus making the prospect of a statistics class seem somewhat more daunting.

The results of this study should be helpful to instructors of statistics because they highlight potential differences in graduate students who are taking an introductory course. Specifically, even when all of the students in a class come from non-statistical programs, there are definite differences in terms of their anxiety regarding statistics. It would appear that those students studying more technical sub-specialties in relatively non-technical areas may in fact have relatively low levels of statistics anxiety. On the other hand, it would also seem that there are students whose anxiety levels are fairly high as compared to their classmates, and who therefore may need greater reassurances from the instructor. Indeed, one of the underlying findings of this research is that introductory graduate level statistics classes for social scientists, and others, are populated by individuals with a wide variety of opinions and fears about statistics. This fact stands as perhaps the greatest challenge to instructors in these courses.

References


**Appendix A**

Each item had the following 5 point likert style response:
1=Strongly Disagree
2=Disagree
3=Neutral
4=Agree
5=Strongly Disagree
1. I am anxious about not being able to understand statistical concepts in this course.
2. Statistical symbols and formulas will confuse me.
3. I am concerned that I may fail this course.
4. My anxiety level for this course is extremely high.
5. Compared to other graduate courses in my program, this course will be difficult.
6. I will spend more time studying for this course than for other courses in my program.
1. Introduction

In the last decades, across the world we observe decline of trade barriers. In addition, transportation and communication costs of trade decline, too. Both of these effects increase the degree of market integration. However, even if all explicit trade barriers disappear markets might not get fully integrated due to implicit barriers, as the border between nations.¹

Measuring explicit trade barriers is straightforward. To measure implicit trade barriers is more demanding as these are measured indirectly through their effect on market integration. Observing tendency towards price equalization might lead us to believe that both explicit and implicit trade barriers were eliminated.²

Market integration is connected to the law of one price³. Engel and Rogers (1995) write that “One of the most direct implications of rational behavior is that two identical goods selling in the same market should have the same price.” One expects the law of one price to hold in financial and currency markets because arbitrage in these markets is more explicit. However, there is a plethora of reasons which might prevent the law of one price to hold in final consumer markets. Wolf (2003) lists these reasons into four groups⁴:

- different tax rates facing wholesalers and retailers in various locations
- different transportation costs from factory to wholesalers and from wholesalers to retailers
- different local costs across wholesalers and retailers
- different mark-ups charged by producers to wholesalers, by wholesalers to retailers, and by retailers to purchasers

In this paper we survey some of these problems to the extent that they connect the price dispersion with border effect. We discuss how price dispersion of final consumer good prices changes when border is crossed. We proceed as follows. In Section Two we survey the theory concerning the border effect, and in Section Three we review some empirical literature. Finally, we conclude.

¹ Some authors are skeptical to claims about tendency of markets for integration. For example, Knetter (1994) shows that German firms charge significantly higher prices to Japanese importers than to other markets; i.e. they price-to-market. Under perfect separation of locations price discrimination is feasible, and no possibility of consumers arbitraging differences in final goods prices appears; Engel and Rogers (1995).

² Knetter and Slaughter (1999) note a disadvantage of this measurement of market integration as one cannot link the outcomes back to barriers and thus cannot differentiate if other reasons, not the barriers, prevent market integration.

³ Lamont and Thaler (2003, p. 191) write that “The first law of economics is clearly the law of supply and demand, and a fine law it is. We would nominate as the second law ‘the law of one price.’”

⁴ In addition, the degree of failure of the law of one price also depends on the distance between locations. There seems to be a stronger arbitrage pressure over shorter distances. The law of one price does seem to be weakened if across markets price discrimination is prevalent, which arises typically due to differences in demand characteristics across groups of consumers.
2. Border Effect

Engle and Rogers (1996) analyze the border effect in a context of consumer price dispersion. Following Sanyal and Jones (1982) they consider all final consumer goods as non-traded; as tradable goods contain non-traded service component (marketing, distribution, and others).\(^5\)
In addition, all goods contain a tradable intermediate component. Thus, the price of the final product \(i\), sold in location \(j\), equals:
\[
p_j^i = \beta_j^i \alpha_j^{i'} (w_j)^{\gamma_j^i} (q_j^i)^{\delta_j^i}
\]
(1)
where \(\beta_j^i\) is the mark-up over costs, \(\alpha_j^{i'}\) is the total productivity of the final-goods sector, \(w_j\) is the price of the non-traded service; \(\gamma_j^i\) is the share of non-traded service in final output; and \(q_j^i\) is the price of traded intermediate input. Engel and Rogers (1996) provide arguments why borders might matter in the final consumer pricing:
- Locations farther apart might have less similar cost structures
- Total productivity of final-goods sector might differ at various locations; these effects might be exacerbated between distant regions, and even more across borders
- Price variation of goods might differ if cities lie across borders
- Mark-ups might differ across borders
- Non-traded market services might be more integrated within a country than across countries

---

\(^5\) We meet the expression ‘border effect’ also in the spatial trade theory. Lösch writes already in 1944 that “the economic landscape, i.e. a system of different spatial market areas, is affected by introducing national borders. Borders, if they are incompletely or completely impermeable, are distortions in the market networks, divide the market area and thus negatively affect a firm’s market potential;” Niebuhr and Stiller (2001, p.5). Modern literature investigating the importance of borders begins with McCallum (1995) who finds that after accounting for distance and size trade between Canadian provinces is much greater than trade between Canadian provinces and U.S. states. The difference seems so large that it throws into doubt the view that international markets have become well integrated. Helliwell (1998) remarks that these results were not expected, since one would expect that decreasing explicit trade barriers between Canada and the U.S. would lead to as large trade among countries as among regions. Similar results were obtained by Helliwell and Verdier (2001) for Canadian provinces and Wolf (2000) for the U.S. states.

\(^6\) Consumer final prices contain costs associated with distribution and marketing. These costs have either a strong local component (store rent, local salesmen wages, etc.) or strong regional component (when for example within a region one location unit shares similar packaging, distribution and other service system for final consumer goods). For 1992 estimates suggest that producers of final goods on average receive sixty per cent of the final retail price; Wolf (2003).

---

2.1. How to Isolate the Border Effect\(^7\)

Define the real exchange rate as:
\[
Q_{j,k,t}^i = \frac{P_{j,t}^i}{S_{j,t}^i}
\]
(2)
where \(P\) is the nominal price of good \(i\) in location \(j\), or location \(k\), at time \(t\), and \(S\) is a nominal exchange rate. Furthermore, define \(T_{j,k,t}^i\) as the cost of trade between location \(j\) and \(k\), at time \(t\), per unit of good \(i\). Define the law of one price as:
\[
\frac{1}{T_{j,k,t}^i} \leq Q_{j,k,t}^i \leq T_{j,k,t}^i
\]
(3)
Transfer (3) into logs to obtain:
\[-t_{j,k,t}^i \leq q_{j,k,t}^i \leq T_{j,k,t}^i\]
(4)
i.e. the price differential between two locations adjusted by nominal exchange rate cannot be higher than the cost of trade between these two locations. Assume that the cost of trade has following form:
\[
T_{j,k,t}^i = e^{c + \phi_t \ln d_{j,k} + \beta_{j,k,t} \ln \text{Border} + \alpha_{j} + \omega_{j,k} + \epsilon_{j,k,t}}
\]
(5)
where \(c\) is constant; \(d_{j,k}\) is the distance between locations \(j\) and \(k\); \(\text{Border}\) equals to one if locations are separated by a border, and to zero otherwise. Furthermore, \(\phi_t\) represents costs of trade specific to good \(i\); \(\alpha_j\) costs of trade specific to location \(j\), and \(\omega_{j,k}\) cost of trade specific to location \(k\); the last term is an error term. Time series standard deviation of the real exchange rate, \(q_{j,k,t}^i\) is assumed to be proportional to the log of the trade costs to obtain:

\(^7\) In this exposition we follow Engel and Rogers (1996), and Gorodnichenko and Tesar (2005).
\[
\sigma(q_{ij,t}) = \gamma + \frac{1}{N-k} \sum_{k=1}^{K} \gamma_k + \frac{1}{k} \sum_{k=1}^{K} \gamma_k
\]

(6)

Panel data are needed for estimation of (6) but estimation itself is cross-sectional.\(^8\)

Crucini, Telmer and Zachariadis (2000), and Parsley and Wei (2001) use different specification to obtain the border effect. They assume that (7) holds,

\[ P_{ij} = P_{ij}^e + \sigma_{ij} \]

(7)

then

\[
\sigma_{ij} = \frac{P_{ij}^e - P_{ij}}{P_{ij}}
\]

(8)

which in logs is

\[
\ln \sigma_{ij} = \ln P_{ij}^e - \ln P_{ij}
\]

(9)

To estimate the border effect they consider the following specification:

\[
\ln \sigma_{ij} = \gamma U + \gamma C + \gamma D + \sum_{k=1}^{K} \gamma_k
\]

(10)

Gorodnichenko and Tesar (2005) show that under certain conditions - estimating (6) may confuse within-country price dispersion with the border effect. If we denote city-pairs as UU (intra-US), UC (between US and Canada) and CC (intra-Canada), and omitting the error terms and additional variables from equation (6) one obtains

\[
\sigma(q_{ij,t}) = \gamma U C_{ij} + \gamma U U_{ij} + \gamma C C_{ij} + \sum_{k=1}^{K} \gamma_k D_{ij}
\]

(11)

where UC, UU, CC are dummy variables for UC, UU, and CC city pairs, D, is a city dummy variable, and N is the number of cities. Assume for example, that there are k cities in Canada. The average volatility for a city pair is then given for the US-Canada city pair by:

\[
\sigma_{U} = \gamma U + \frac{1}{N-K} \sum_{k=1}^{K} \gamma_k + \frac{1}{k} \sum_{k=1}^{K} \gamma_k
\]

(12)

For US-US city pair by:

\[
\sigma_{uu} = \gamma U + \gamma U + \frac{1}{N-K} \sum_{k=1}^{K} \gamma_k
\]

(13)

For Canada-Canada city pair by:

\[
\sigma_{cc} = \gamma C + \gamma C + \frac{1}{k} \sum_{k=1}^{K} \gamma_k
\]

(14)

The coefficients \( \gamma \), \( \gamma_U \), and \( \gamma_C \) can take negative and positive values and they measure the contribution to volatility of being a US-Canada, a US-US pair and a Canada-Canada respectively.

Equation (11) represents an exact decomposition of variance in theory, which cannot be estimated in practice because a collinear relationship between set of dummies. One way out is to follow Engel and Rogers (1996) who substitute (13) and (14) into (11) to eliminate \( CC_{ij} \) and \( UU_{ij} \) to obtain:

\[
\sigma_{ij} = \gamma U + \gamma U C_{ij} + \gamma U U_{ij} + \gamma C C_{ij} + \sum_{k=1}^{K} \gamma_k D_{ij}
\]

(15)

Gorodnichenko and Tesar (2005) show that in case \( \gamma_U \) and \( \gamma_C \) significantly differ it might be the case that \( \beta \) in specification (6) does not properly measure the border effect. The border dummy in (6) measures the inter-country volatility as compared to the average intra-country volatility. In case that \( \gamma_U \) and \( \gamma_C \) are equal (similar) this measure seems reasonable, however in case that \( \gamma_U \) and \( \gamma_C \) are different \( \beta \) coefficient might be biased in any direction as shown in details in Gorodnichenko and Tesar (2005).

Yin and Lai (2006) re-evaluate the way to isolate the border effect. They show that if foreign and domestic prices have similar volatility then the inter-country versus intra-country analysis would yield unbiased border effect estimates. However, such cross-country homogeneity is typically violated in actual data. In a case with heterogeneity in foreign and domestic price volatility Yin and Lai (2006) show that a symmetric sampling strategy, under which the same number of cities is sampled from each country, can be used to secure the unbiased border effect estimates.

Yin and Lai (2006) also devise a method for the border effect decomposition. Their method avoids the co-linearity problem, which arises in regression estimates as discussed by Gorodnichenko and Tesar (2005). The decomposition of Yin and Lai (2006) enables to quantify the exact contribution of exchange rate fluctuations to the border effect. The border effect is thus shown to capture both the exchange rate effect and the effect of dissimilar shocks.

3. **Empirical Evidence**

Engel and Rogers (1996) use consumer price index data disaggregated into 14 categories of goods for 9 Canadian and 14 U.S. cities for the period 1978:6-1994:12. They test whether "volatility of the price of similar goods between cities should be positively related to the distance between those cities; but holding distance constant, volatility should be higher between two cities separated by the national border."(p.1115).
Engle and Rogers isolated the border effect using procedures described in previous section, measuring prices as two-month differences, and measuring volatility as the standard deviation of the relative price series.\(^9\) The authors find substantial difference between the intra-national price volatilities, i.e. the price volatility among U.S. cities was considerable larger than the price volatility among Canadian cities. However, the price volatility was highest among American-Canadian city pairs.\(^10\) Also they find strong evidence that distance explains large portion of price dispersion across cities for most of goods. Most importantly, the coefficients on the dummy variable for the border are of the hypothesized sign and highly significant for all goods.\(^11\)

The border effect might be significant because it picks up the effect of nominal exchange rate volatility. To tackle the problem Engle and Rogers (1996) construct real prices by dividing the nominal price ratios by the aggregate city price indexes. The Border effect remains important, i.e. the coefficients on the border dummy are all positive and significant. They also consider calculating the individual goods prices in each city relative to the national-level producer price index. Again, the results remain qualitatively similar to previous regressions. These results for the U.S. and Canada are puzzling since free trade prevails between them and in addition, they share common language, similar cultural and political traditions.

Beck and Weber (2001) test the importance of borders in European markets. Their data contain monthly price indexes for 86 locations in 7 European countries.\(^12\) The sample period for aggregated data is 1991:1-2002:12, and for the disaggregated data is 1995:1-2002:12. They measure real price volatility as the two-month change in log of relative prices. In their results price volatility is lower for intra-country city pairs than for inter-country city pairs.\(^13\)

Beck and Weber (2001) estimate the following regression:

\[
v(\Delta q_{ij}) = \sum \alpha(c) \Delta(c) + \beta \ln(d_{ij}) + \delta B_{ij} + \eta \Delta(V_s) + u_{ij}
\]

(16)

where \(v(\Delta q_{ij})\) is the measure for volatility of prices, \(D(c)\) is the dummy explaining the city-specific effects, \(d_{ij}\) depicts distance, \(B_{ij}\) is the border effect dummy, and \(v(\Delta V_{ij})\) is the volatility of nominal exchange rate changes between location \(i\) and \(j\). \(u_{ij}\) is the error term. First, Beck and Weber (2001) estimate (16) without nominal exchange rate and find that border dummy is positive and significant for aggregated as well as disaggregated price data.\(^14\) As the second step, they estimate (16) including the exchange rate variable: the effect of the border dummy weakens but remains significant. After including dummies for all country pairs, Beck and Weber find that the smallest border effect is at the German-Austrian border and the biggest at the Italian-Swiss border.

Beck and Weber are also interested whether the border effect weakens after the introduction of euro. Their findings support weakening of the border effect after the common currency is introduced, however the border dummy still remains significant even if smaller in size.\(^15\) The EMU is found to have greatly reduced but not completely eliminated the importance of intra-EMU borders.\(^16\)

Engel, Rogers and Yi (2003) re-examine the border effect using actual data for different U.S. and Canadian cities instead of price index data as in Engel and Rogers (1996). Their dataset contains annual prices – including taxes - of one-

---

\(^9\) They have also calculated volatility as the spread between the 10\(^{th}\) and 90\(^{th}\) percentile, however their results have not changed significantly.

\(^10\) For the pooled sample of goods crossing the border adds to the average standard deviation of prices between pairs of cities. In order to generate that much volatility by distance, the cities would have to be 75 000 miles apart. This indicates that crossing the border adds substantially to volatility.

\(^11\) To check robustness of their results Engle and Rogers (1996) also consider a filtered measure of the real exchange rate including relative wage volatility; after the common currency is introduced, however the border dummy still remains significant even if smaller in size.\(^15\) The EMU is found to have greatly reduced but not completely eliminated the importance of intra-EMU borders.\(^16\)

\(^12\) These countries are Germany (West and East), Austria, Finland, Italy, Spain, Portugal and Switzerland.

\(^13\) One interesting result of Beck and Weber (2001) is that relative price volatility between Germany and Austria is lower than the within-country price volatility in Portugal.

\(^14\) They interpret these results as evidence in favor of European markets segmentation.

\(^15\) Beck and Weber (2001) also estimate the importance of the border effect between East and West Germany. They find that after the German unification the border effect has disappeared.

Engel, Rogers and Yi (2003) estimate two specifications. First, they estimate
\[
[p_{i,j,t} - p_{i,k,t}] = \beta_1 \text{dist}_{jk} + \beta_2 \text{pop}_{jk} + \beta_3 \text{tax}_{jk} + \sum_{h} \lambda_h \text{citdum}_h + u_{i,j,k},
\]
(17)
where \(p_{i,j,t} - p_{i,k,t}\) is the absolute price difference in price of good \(i\), in cities \(j\) and \(k\), at time \(t\); \(\text{dist}_{jk}\) is the log of distance between cities; \(\text{pop}_{jk}\) is the absolute value difference in the log of the population between cities \(j\) and \(k\); \(\text{tax}_{jk}\) is the absolute value of the tax rate difference between cities \(j\) and \(k\); \(\text{bord}_{jk}\) is the border variable; \(\text{citdum}_h\) represent the city dummies intended to capture idiosyncratic city-effects. Second, they estimate (18):
\[
[p_{i,j,t} - p_{i,j,t-1}] = \eta_1 \text{dist}_{jk} + \eta_2 \text{pop}_{jk} + \eta_3 \text{tax}_{jk} + \eta_4 \text{bord}_{jk} + \sum_{h} \lambda_h \text{citdum}_h + u_{i,j,k},
\]
(18)
where \(p_{i,j,t} - p_{i,j,t-1}\). In (17) the coefficient of the dummy variable \(\text{bord}_{jk}\) is significant and large in size. Similarly, coefficients of \(\text{dist}_{jk}\) and \(\text{pop}_{jk}\) are also significantly different from zero, while the coefficient of \(\text{tax}_{jk}\) is not significant. The variables \(\text{pop}_{jk}\) and \(\text{bord}_{jk}\) are also significant in (18), but the effect is smaller \(\text{dist}_{jk}\) is not significant any more, which is in contrast to Engel and Rogers (1996) where distance remained significant across different specifications. Engel, Rogers and Yi (2003) estimate the same specifications for different sub-periods and for different groups of goods. The border dummy is significant and has the correct sign, but varies in magnitude.

Rogers and Smith (2001) compare the border effect between the U.S. and Mexico with the border effect between the U.S. and Canada. They use monthly and semi-monthly price data for the period 1980:1-1997:12 for fourteen U.S., ten Canadian, and 14 Mexican cities (regions). They divide the total sample period into two sub-periods to differentiate between the stable and volatile periods of the peso-dollar exchange rates.

Rogers and Smith (2001) estimate the following regression:
\[
V(\Delta P(j,k)) = \sum \alpha(m) D(m) + \beta r(j,k) + X \theta + u(j,k)
\]
(19)
where \(V(\Delta P(j,k))\) is the volatility of \(\Delta P(j,k)\), \(D(m)\) is the dummy variable for each city, \(r(j,k)\) log of distance between cities \(j\) and \(k\); \(X\) is a vector of containing border dummies, \(V(\Delta s(j,k))\) volatility of nominal exchange rate changes between location \(j\) and \(k\). \(u(j,k)\) is the error term.

In the full sample the border dummy is highly significant; if separate dummies are taken, then the US-MX and CA-MX border dummies are greater in size. But as the volatility of the exchange rate is included the border dummies lose their explanatory power, and are not anymore significant. Results of Rogers and Smith (2001) suggest that large part of the border effect is caused by nominal exchange volatility and sticky prices.

However, in sub-sample results - which differentiate between the stable and un-stable peso periods - the coefficients of the border dummy for the US-MX and CA-MX are notably smaller, and the inclusion of exchange rate volatility doesn’t make the border dummies insignificant, even as they get smaller. Rogers and Smith (2001) also compare border effect before and after the NAFTA agreement. The US-MX border dummy for the full period is larger than for the post NAFTA period, so the removal of explicit trade barriers decreased the border effect. The inclusion of relative wage into the regression doesn’t change the main results.

Cegłowski (2003) finds that Canadian provincial borders account for a significant fraction of the discrepancy of prices across provinces.

Gorodnichenko and Tesar (2005) argue that results as obtained in Engle and Rogers (1996) are not the true border effects due to reasons shown in equation (15). They re-estimate the results using the same data and obtain two ‘quasi-border’ coefficients. One estimates a small width of the border, if one considers it from the US perspective (where the intra-country volatility was high) and the other estimates a large width of the border, if one considers it from the Canadian perspective (where the intra-country volatility was low).

Horvath, Ratfai and Dome (2007) evaluate the importance of the border between Hungary and Slovakia. Their sample uses monthly prices of twenty goods and services,

\[17\] In their sample 42 items represent food products, 9 clothing items, 6 durables, 21 non-tradable services and 22 miscellaneous products.
observed in a total of fifty-six locations over a period of 1997:5-2001:12. The volatility of prices
is quite similar at district pairs in Hungary and Slovakia; cross-border district pairs show much higher volatility. High volatility in cross-border city pairs holds in pooled data as well as for most individual products.

As in Engel and Rogers (1996), their baseline regression equation specified separately
for each product is

$$\sigma(q_{j,k}) = e + \beta_1 \ln d_{j,k} + \beta_2 HS + \sum_{i=1}^{N-1} \alpha_i D_i + \epsilon_{j,k}$$

(20)

where the $HS$ dummy equals one if the two locations $j$ and $k$ are in different countries, and zero otherwise. When the data are pooled for the four product categories, or for the whole sample, product-specific dummies are also added. Time-invariant district-specific factors are controlled through the inclusion of district dummies.

The estimated parameters provide a strong evidence for the border effect, i.e. after controlling for distance and district-specific fixed effects, coefficients on the border dummy are significantly positive in all individual cases. The results in the pooled specifications also show significant coefficients on the border dummy, all with the expected sign. The results for distance are less pronounced. Their results are robust to inclusion of language variable\(^{10}\). They also assess the importance of the nominal exchange rate in driving the wedge in volatility in international vs. intra-national relative prices. Nominal exchange rates do not appear to be responsible for driving the border effect between these two transition countries.

Gorodnichenko and Tesar (2006) argue that the baseline specification in Engel and Rogers (1996) results in a biased estimate of the border effect, as it confounds the true effect with the impact of within-country heterogeneity. To correct for the bias, they suggest augmenting the regression equation with a dummy variable capturing country specific effects in relative price variability. Horvath, Ratfai and Dome (2007) estimate such an equation, where the only new variable relative to (20) is the dummy representing price pairs taken only from Slovakia. The results lead to a smaller but still positive and significant border effect.

Shieue (2005) uses grain prices in France, Switzerland, the Habsburg Empire of Austria, and fifteen Bavarian cities for the years 1815 to 1855 to assess the size of border effect in the context of the German Zollverein. The main result is that the estimated border effect for Europe in the mid-19th century is small as compared to estimates using contemporary data.

4. Conclusion

This paper surveys a body of previously published research on the border effect as appearing in price dispersion behavior within and across countries. First, we review the currently accepted methodology as how to isolate the border effect. Second, we review some of the papers which estimate the size and importance of the border effect. In majority of research the border effect is found to be positive and significant, in other words it is empirically supported that inter-country price dispersion is generally higher than intra-country price dispersion even after accounting for distance, city-and-product specification, and nominal exchange rate variability.

5. References


Gorodnichenko, Yuriy and Linda Tesar, A Re-examination of the Border Effect, NBER, WP 11706, October 2005


\(^{10}\) They split districts in Slovakia into two groups, ones with and without Hungarian serving as a second language.
Border Effect in Canada.” Canadian Journal of Economics, 34, 1024-1041
Knetter, Michael, N., “Why are Retail Prices in Japan so High? Evidence from German Export Prices,” NBER WP 4894, 1994
Shiue, Carol H. (2005), “From Political Fragmentation towards a Customs Union: Border Effects of the German Zollverein, 1815 to 1855” University of Texas, February
Identities Based on Probability Mass Functions

M. Forhad Hossain  
King Khalid University  
Department of Mathematics  
Abha, Saudi Arabia  
Email: forhad_55@yahoo.com

and

Anwar H. Joarder  
Department of Mathematical Sciences  
King Fahd University of Petroleum & Minerals  
Dhahran, Saudi Arabia  
Email: anwarj@kfupm.edu.sa

Abstract  Moments of discrete distributions are usually calculated by averaging the probability mass function or calculating characteristic function and then differentiating it. In this paper we try to popularize a mass identity technique for calculating moments. Since the probability mass function adds to one, an identity called density identity is obtained. Differentiating repeatedly by a continuous parameter in the density identity, we obtain higher order mass identities that may be useful beyond the calculation of moments.

Mathematics Subject Classification: 60-01, 60E10

Key Words and Phrases: Discrete distributions; identities; factorial moments; raw moments; centered moments

1. Introduction

Moments of discrete distributions are calculated variously by averaging the probability mass function, differentiating the characteristic function etc. In this paper we try to popularize a mass identity technique for calculating moments. Since the probability mass function adds to one, an identity called density identity is obvious. In case, the probability mass function has at least one continuous parameter, one can differentiate the density identity repeatedly to obtain higher order mass identities. The general mass identity of general order may be useful beyond the calculation of moments.

Consider a probability mass function $f(x; \theta) > 0$, $x \in D$ where $\theta$ is continuous. Then

$$\sum_{x \in D} f(x; \theta) = 1$$

will be called the density identity. By differentiating the density identity, some new identities can be generated which is helpful in calculating factorial moments

$$\mu_{(i)} = E(X^{(i)}),$$

$$= E[(X-1) \cdots (X-r+1)], \quad r = 1, 2, \cdots$$

is the $r$-th factorial moment. These can then be used to find raw moments by the relation

$$\mu_r = E(X^r)$$

$$= \sum_{i=1}^{r} S(r,i) \mu_{(i)}^r, \quad (r = 1, 2, \cdots)$$

where $S(r,i)$ is the Stirling number of the second kind given by

$$S(r,i) = \frac{1}{i! \sum_{j=0}^{i-1} (-1)^{i-j} \binom{i}{j} j^r}, \quad 0 \leq i \leq r.$$

The raw moments can be further used to find central moments by

$$\mu_{(i)} = \sum_{j=0}^{r} (-1)^j \binom{r}{j} \mu_{(i-j)} (\mu_{(1)})^j$$

(3)
(Johnson, Kotz and Kemp, 1993, 42-44). For a combinatorial proof of Stirling number of the second kind, the reader is referred to Robersts (1984, pp. 182-183), and for an inductive proof, the reader is referred to Joarder and Mahmood (1997).

2. Identities Based on Mass Identities

we will use the notation:

\[ x^{(a)} = x (x - 1) \cdots (x - a + 1), \quad x^{(0)} = 1, \]
\[ x_{(a)} = x (x + 1) \cdots (x + a - 1), \quad x_{(0)} = 1. \]

In this section, we derive higher order mass identities for a number of univariate discrete distributions.

(i) Consider the geometric distribution with probability mass function

\[ f(x) = pq^{x-1}, \quad 0 < p < 1, \quad p + q = 1; \quad x = 1, 2, \ldots. \]

Since \( f(x) \) is a probability mass function, we have \( \sum_{x=1}^{\infty} pq^{x-1} = 1 \) or equivalently

\[ \sum_{x=1}^{\infty} q^x = q (1-q)^{-1} \]  \hspace{1cm} (2.1)

which is an identity in \( q \).

\[ \sum_{x=1}^{\infty} x^{(a)} q^x = a! \cdot p^{-(a+1)} q^a. \]

**Theorem 2.1** For \( 0 < p < 1 \) and \( p + q = 1 \),

\[ (a) \sum_{x=1}^{\infty} q^x = p^{-3} q, \]
\[ (b) \sum_{x=1}^{\infty} xq^x = p^{-2} q, \]
\[ (c) \sum_{x=1}^{\infty} x^{(2)} q^x = 2p^{-3} q^2, \]
\[ (d) \sum_{x=1}^{\infty} x^{(3)} q^x = 6p^{-4} q^3, \]
\[ (e) \sum_{x=1}^{\infty} x^{(4)} q^x = 24p^{-5} q^4. \]

The generalization is obvious.

**Corollary 2.1** For \( 0 < p < 1 \) and \( p + q = 1 \), the following identities are true:

\[ (a) \sum_{x=1}^{\infty} x^2 q^x = p^{-3} (q^2 + q) = p^{-3} q (q + 1), \]
\[ (b) \sum_{x=1}^{\infty} x^3 q^x = p^{-4} (q^3 + 4q^2 + q), \]
\[ (c) \sum_{x=1}^{\infty} x^4 q^x = p^{-5} (q^4 + 11q^3 + 11q^2 + q). \]

**Proof.** The first identity in (a) below is equivalent to (2.1). Differentiating (2.1) with respect to \( q \) repeatedly, we have the following identities (b) to (e) follows from Theorem 2.1.

\[ (a) \sum_{x=1}^{\infty} x^2 q^x = \sum_{x=1}^{\infty} (x^{(2)} + x) q^x = 2p^{-3} q^2 + p^{-2} q, \]
\[ (b) \sum_{x=1}^{\infty} x^3 q^x = \sum_{x=1}^{\infty} (x^{(3)} + 3x^{(2)} + x) q^x = 6p^{-4} q^3 + 3(2p^{-3} q^2) + p^{-2} q, \]
\[ (c) \sum_{x=1}^{\infty} x^4 q^x = \sum_{x=1}^{\infty} (x^{(4)} + 6x^{(3)} + 7x^{(2)} + x) q^x = 24p^{-5} q^4 + 36p^{-4} q^3 + 7(2p^{-3} q^2) + p^{-2} q. \]

Raw moments are then be calculated in the following way:

\[ E(X) = \sum_{x=1}^{\infty} xf(x) = pq^{-1} \sum_{x=1}^{\infty} xq^x = pq^{-1}(p^{-2}q) = p^{-1}, \]
\[ E(X^2) = \sum_{x=1}^{\infty} x^2 f(x) = pq^{-1} \sum_{x=1}^{\infty} x^2 q^x = pq^{-1}(p^{-3}(q^2 + q)) = p^{-2}(q + 1), \]
\[ E(X^3) = \sum_{x=1}^{\infty} x^3 f(x) = pq^{-1} \sum_{x=1}^{\infty} x^3 q^x = pq^{-1}(p^{-4}(q^3 + 4q^2 + q)) = p^{-3}(q^3 + 4q + 1), \]

etc. The centered moments can then be calculated by the above raw moments by (3).
Consider the Binomial Distribution with the probability mass function

\[ f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 < p < 1, \quad p + q = 1, \quad x = 0, 1, \ldots, n. \]

so that the mass identity is given by

\[ \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (1-p)^{-x} x! . \]

**Theorem 2.2** For \( 0 < p < 1 \) and \( p + q = 1, \)

\[ \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = n^{(n)} p^a q^{-x} . \]

(iii) Consider the Poisson Distribution with the probability mass function

\[ f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad \lambda > 0, \quad x = 0, 1, \ldots. \]

so that the mass identity is

\[ \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = 1. \]

**Theorem 2.3** For \( \lambda > 0, \)

\[ \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^\lambda. \]

(iv) Consider the Negative Binomial Distribution with probability mass function

\[ f(x) = \binom{k+x-1}{k-1} p^x q^k, \quad k = 1, 2, 3, \ldots; 0 < p < 1, \quad p + q = 1, \quad x = 0, 1, 2, \ldots. \]

so that the mass identity

\[ \sum_{x=0}^{\infty} \binom{k+x-1}{k-1} p^x q^k = q^{-k} . \]

**Theorem 2.4** For \( k = 1, 2, 3, \ldots; 0 < p < 1 \) and \( p + q = 1, \)

\[ \sum_{x=0}^{\infty} \binom{k+x-1}{k-1} p^x q^k = k^{(k-1)} p^{-k} q^a . \]

(iv) Consider the Truncated Poisson Distribution with the probability mass function

\[ f(x) = \left( e^\lambda - 1 \right)^{-1} \frac{\lambda^x}{x!}, \quad \lambda > 0, \quad x = 1, 2, \ldots. \]

so that the mass identity is

\[ \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^\lambda - 1. \]

**Theorem 2.5** For \( \lambda > 0, \)

\[ \sum_{x=0}^{\infty} x^a \frac{\lambda^x}{x!} = \lambda^a e^\lambda. \]

(vi) Consider the Truncated Binomial Distribution with the probability mass function

\[ f(x) = (1-q)^{a_x} q^{n-x}, \quad 0 < p < 1, \quad p + q = 1, \quad x = 1, 2, \ldots, n \]

so that the mass identity is

\[ \sum_{x=0}^{n} x^a (n \choose x) \frac{p^x q^{n-x}}{x!} = q^{-a} - 1. \]

**Theorem 2.6** For \( 0 < p < 1 \) and \( p + q = 1, \)

\[ \sum_{x=0}^{n} x^a (n \choose x) \frac{p^x q^{n-x}}{x!} = n^{(a)} p^a q^{-x} . \]

(vii) Consider the Logarithmic Distribution with probability mass function

\[ f(x) = -\left( \ln(1-\theta) \right)^{-1} \frac{\theta^x}{x}, \quad 0 < \theta < 1, \quad x = 1, 2, 3, \ldots \]

so that the mass identity is

\[ \sum_{x=0}^{\infty} x^a \frac{\theta^x}{x} = -\ln(1-\theta) . \]

**Theorem 2.7** For \( 0 < \theta < 1, \)

\[ \sum_{x=0}^{\infty} x^a \frac{\theta^x}{x} = (k-1)! \left( \frac{\theta}{1-\theta} \right)^x . \]

3. Some General Identities

The identities developed in this paper are applicable in more general situations than in the context of deriving moments. The following identities are obtained from Section 2 by putting particular values of the continuous parameter involved.

\[ (a) \sum_{x=0}^{\infty} x^a \left( \frac{1}{2} \right)^x = a! 2^a; \]

\[ (b) \sum_{x=0}^{\infty} x^a \left( \frac{n}{x} \right) = n^{(n)} 2^{x-n}; \]

\[ (c) \sum_{x=0}^{\infty} x^a \left( \frac{1}{x!} \right) = e; \]

\[ (d) \sum_{x=0}^{\infty} x^a \left( \frac{k+x-1}{(k-1)!} \right)^x = k^{(k-1)} \left( \frac{1}{2} \right)^x. \]
Acknowledgements

The second author A.H. Joarder is grateful to the excellent research facilities available at King Fahd University of Petroleum & Minerals, Saudi Arabia.

References


File: PedaMath\IndentitiesProbMassFunctiona.doc
Abstract The standardized moments or Mahalanobis moments are easily calculated for bivariate elliptical distribution which includes bivariate normal as a special case. But for many bivariate distributions these are challenging. A set of alternative formulae is developed to derive Mahalanobis moments for any bivariate distribution. The second order Mahalanobis moment accounts for the coefficient of kurtosis. The proposed method works well if the product moments have closed forms. Ideas are illustrated with examples.

AMS Mathematics Subject Classification: 60E10, 60E05

Key Words and Phrases: bivariate distribution, standardized moments, Mahalanobis distance, product moments, kurtosis

1. Introduction

In a series of papers, Mardia (1970; 1974; 1975) defined and discussed the properties of measures of kurtosis and skewness based on Mahalanobis distance. The coefficient of kurtosis is the second order moment of standardized distance better known as Mahalanobis distance. The moments can be referred to as standardized moments, Mahalanobis moments or Mardia moments. Interested readers may go through Kotz, Balakrishnan and Johnson (2000) for an excellent discussion on multivariate skewness and kurtosis.

For some distributions, it is easy to derive the distribution of the Mahalanobis distance and calculate moments but for others these are difficult. Kotz, Balakrishnan and Johnson (2000) have provided the kurtosis of the Marshall-Olkin bivariate exponential distribution. In this paper we provide an alternative method to calculate Mahalanobis moments for bivariate distributions in terms of product moments of the components of a bivariate vector. Product moments (also called raw product moments or product moments around zero) of order \(a\) and \(b\) for two random variables \(X_1\) and \(X_2\) are defined by \(\mu(a,b) = E(X_1^a X_2^b)\) while the centered product moments (sometimes called central product moments, corrected product moments or central mixed moments) are defined by \(\mu(a,b) = E[(X_1 - E(X_1))^a (X_2 - E(X_2))^b]\).}

Interested readers may go through Johnson, Kotz and Kemp (1993, 46) or Johnson, Kotz and Balakrishnan (1997, 3). Evidently \(\mu(a,0) = E(X_1^a)\) is the \(a\)-th moment of \(X_1\), and \(\mu(0,b) = E(X_2^b)\) is the \(b\)-th moment of \(X_2\). In case \(X_1\) and \(X_2\) are independent, then \(\mu(a,b) = E(X_1^a E(X_2^b)) = \mu(a,0)\mu(0,b)\) and \(\mu(a,b) = \mu(a,0)\mu(0,b)\). The correlation coefficient \(\rho(-1 < \rho < 1)\) between \(X_1\) and \(X_2\) is denoted by

\[
\rho_{X_1,X_2} = \frac{\mu(1,1)}{\sqrt{\mu(2,0)\mu(0,2)}}. \tag{1.2}
\]

Note that \(\mu(2,0) = E((X_1 - E(X_1))^2) = \sigma_{X_1}^2\) which is popularly denoted by \(\sigma_{X_1}^2\) while the central product moment, \(\mu(1,1) = E[(X_1 - E(X_1))(X_2 - E(X_2))]\) denoted popularly by \(\sigma_{X_1X_2}^2\), is, in fact, the covariance between \(X_1\) and \(X_2\).

In this paper, we derive Mahalanobis moments in terms of centered product moments. Mahalanobis moments for a bivariate normal distribution and bivariate \(t\)-distribution are calculated. It is observed that general formulae for Mahalanobis moments for bivariate elliptical distribution, which includes bivariate normal and \(t\)-distributions as special cases, are easily obtained.
An example of bivariate chi-square distribution is considered for which the proposed method developed in Section 2 seems to be appropriate. In what follows we will rather use \( X_1 = X \) and \( X_2 = Y \) to avoid all confusion of a trivial nature, and define \( \mu(a,b) = E\left[ (X - \xi) \rho Y' - \theta \rho' \right] \) where \( \xi = E(X), \theta = E(Y) \).

## 2. Mahalanobis Moments in Terms of Product Moments

For a bivariate random vector \( W = (X, Y)' \), with mean vector \( \mu = (\xi, \theta)' \) and covariance matrix
\[
\text{Cov}(W) = E(W - \mu)(W - \mu)' = \begin{pmatrix} \mu(2,0) & \mu(1,1) \\ \mu(1,1) & \mu(0,2) \end{pmatrix} = \Omega \text{ (say)},
\]
the standardized distance is defined by
\[
Q = (W - \mu)' \Omega^{-1} (W - \mu) = \|(W - \mu)\|^2.
\]

The quantity \( Q \) is also known to be generalized distance or Mahalanobis distance. For a bivariate random vector \( W \) with \( E(W) = \mu \) and \( \text{Cov}(W) = \Omega \), we define standardized moments or Mahalanobis moments by
\[
\beta_i = E(Q^i), \quad i = 1, 2, \ldots
\]
where \( Q = (W - \mu)' \Omega^{-1} (W - \mu) = \Omega^{-1/2} (W - \mu) \|^2 \). Of special interest is the the second order Mahalanobis moments by \( \beta_2 = E(Q^2) \) which is the coefficient of kurtosis in the sense of Mardia (1979c).

In fact regardless of the distribution of the variable in question the first order standardized moments or Mahalanobis moments by \( \beta_1 = E(Q) \), may be considered for which the proposed method developed in Section 2 seems to be appropriate. Kotz, Nadarajah and Mitov (2003) presented an elegant technique for product moments of the components of any multivariate random vectors in terms of cumulative distribution function or survival function. It appears that if the cumulative distribution function or the survival function has a closed form, the Nadarajah and Mitov (2003) technique works well. For Marshall-Olkin bivariate exponential distribution with survival function
\[
P(X \geq x, Y \geq y) = \begin{cases} 1 - e^{-\lambda x} & 0 \leq x \leq y \\ 1 - e^{-\lambda y} & 0 \leq y \leq x \end{cases}
\]
where \( \lambda > 0 \), Nadarajah and Mitov (2003) calculated raw product moment of general order from which it is possible to calculate Mahalanobis moments of the distribution. Kotz, Balakrishnan and Johnson (2000, 82) mentioned that the coefficient of kurtosis of the distribution is given by
\[
\beta_2 = 2(1 + \rho)^{-3}(3\rho^4 + 9\rho^2 + 15\rho^2 + 12\rho + 4)
\]
where the correlation coefficient \( \rho \) is given by
\[
(\lambda + 2)^2 = \lambda.
\]
They also mentioned that in case \( \rho = 0 \), the components \( X \) and \( Y \) become independent, in which case \( \beta_2 = 8 \) (which is the same as that of the bivariate normal distribution). Interested readers may go through Kotz, Nadarajah and Mitov (2003) for a useful formula for product moments for any univariate distribution.

We derive standardized moments in terms of centered product moments just to demonstrate the potential of an alternative way.

### Theorem 2.1

Let \( \mu(a,b) \) be centered product moments between \( X \) and \( Y \). Then
\[
(i) \quad \left[ \mu(2,0) \mu(0,2) - \mu^2(1,1) \right]^2 E(Q^2) = \mu(4,0) \mu^2(0,2) + \mu(0,4) \mu^2(2,0) + 4 \mu^2(1,1) \mu(2,2) + 2 \mu(2,0) \mu(0,2) \mu(2,2) - 4 \mu(0,2) \mu(1,1) \mu(3,1) - 4 \mu(2,0) \mu(1,1) \mu(1,3),
\]
\[
(ii) \quad \left[ \mu(2,0) \mu(0,2) - \mu^2(1,1) \right]^2 E(Q^2) = \mu(6,0) \mu^2(0,2) + \mu^2(2,0) \mu(0,6) - 6 \mu^2(0,2) \mu(1,1) \mu(5,1) - 6 \mu^2(2,0) \mu(1,1) \mu(1,5) + 12 \mu(0,2) \mu^2(1,1) \mu(4,2) + 12 \mu(2,0) \mu^2(1,1) \mu(2,4) + 3 \mu(2,0) \mu^2(0,2) \mu(4,2) + 3 \mu^2(2,0) \mu(0,2) \mu(2,4) - 8 \mu^2(1,1) \mu(3,3) - 12 \mu(0,2) \mu(2,0) \mu(1,1) \mu(3,3).
\]

### Proof.

From (2.1) we have
\[
Q = (X - \xi Y - \theta) \left[ \begin{array}{c} \mu(2,0) \\ \mu(0,2) \end{array} \right] E(Q^2) = (X - \xi Y - \theta) \left[ \begin{array}{c} \mu(2,0) \\ \mu(0,2) \end{array} \right] E(Q^2)
\]
which can be simplified as
\[
Q = \mu(0,2)(X - \xi Y - \theta) + \mu(2,0)(X - \xi Y - \theta)' = \mu(0,2)(X - \xi Y - \theta) + \mu(2,0)(X - \xi Y - \theta)'.
\]

By taking expected values in both sides of the above identity, we have
\[
\mu(2,0) \mu(0,2) - \mu^2(1,1) E(Q) = \mu(2,0) \mu(0,2) - 2 \mu^2(1,1) + \mu(0,2) \mu(2,0) = 2 \mu(2,0) \mu(0,2) (1 - \rho^2)
\]
i.e. \( E(Q) = 2 \) which is generally true (see 2.2). By squaring both sides of (2.4) we have...
\[ \begin{align*}
&\left[ \mu(2,0)\mu(0,2) - \mu^2(1,1) \right] Q^* \\
n &= \mu^2(2,0)(X - \xi)^4 + 4\mu^2(1,1)(X - \xi)(Y - \theta)^2 \\
&\quad + \mu^2(2,0)(Y - \theta)^4 - 4\mu(0,2)\mu(1,1)(X - \xi)(Y - \theta) \\
&\quad + 2\mu(0,2)\mu(0,2)(X - \xi)^2(Y - \theta)^2 \\
&\quad - 4\mu(2,0)\mu(1,1)(X - \xi)(Y - \theta)^2.
\end{align*} \]

Then the result in (i) follows by taking expected values in both sides of the above identity. By cubing both sides of (2.4) we have:

\[ \begin{align*}
&\left[ \mu(2,0)\mu(0,2) - \mu^2(1,1) \right] Q^* \\
n &= \mu^3(0,2)(X - \xi)^6 - 6\mu(0,2)\mu(1,1)(X - \xi)^5(Y - \theta) \\
&\quad + \mu^3(2,0)(Y - \theta)^6 - 6\mu(0,2)\mu(1,1)(X - \xi)^4(Y - \theta)^2 \\
&\quad + 3\mu(0,2)\mu(0,2)(X - \xi)^3(Y - \theta)^3 \\
&\quad + 3\mu^2(2,0)\mu(0,2)(X - \xi)^3(Y - \theta)^2 \\
&\quad + 6\mu^2(2,0)\mu(1,1)(X - \xi)^3(Y - \theta)^2 \\
&\quad - 12\mu(2,0)\mu(2,0)(X - \xi)(Y - \theta)^4.
\end{align*} \]

Part (ii) follows by taking expected values of the above identity.

**Corollary 2.1** Let $\mu(a,b)$ be the centered product moment and $\rho = (\mu(2,0)\mu(0,2))^3/\mu^2(1,1)$ be the correlation coefficient between $X$ and $Y$. Then

(i) $\left[ \mu(2,0)\mu(1,1) \right] E(Q^*)$

\[ \begin{align*}
&= \mu(4,0)\mu(0,0) + \mu(0,4)\mu(2,0) \\
&\quad + (4\rho + 2)\mu(2,0)\mu(0,2) \\
&\quad - 4(\mu(2,0)\mu(0,2))^2 \{ \mu(0,2)(\mu(3,1) + \mu(2,0)\mu(1,3)) \},
\end{align*} \]

(ii) $\left[ \mu(2,0)\mu(0,2)(1 - \rho^2) \right] E(Q^*)$

\[ \begin{align*}
&= \mu(6,0)\mu(0,0) + \mu(0,6)\mu(2,0) \\
&\quad - (\mu(2,0)\mu(2,0))^2 \mu(3,3) + 4\rho(2\rho^2 + 3) \\
&\quad - 6(\mu(2,0)\mu(0,2))^2 \rho^2 \mu(0,2)\mu(5,1) \\
&\quad + \mu^2(2,0)\mu(0,2) \{ 3 \mu(2,0)\mu(0,2)(4\rho^2 + 1) \} \{ \mu(0,2)(\mu(4,2) + \mu(2,0)\mu(2,4)) \}.
\end{align*} \]

**Corollary 2.2** Let $X$ and $Y$ have a bivariate distribution with $E(XY) = E(X^Y) = E(X^{Y^*})$ and correlation coefficient $\rho$. Then

(i) $\mu^2(2,0)(1 - \rho^2)^2 E(Q^*) = 2\mu(4,0)$

\[ \begin{align*}
&\quad + (4\rho + 2)\mu(2,0) - 8\rho\mu(3,1),
\end{align*} \]

(ii) $\mu^2(0,2)(1 - \rho^2)^2 E(Q^*) = \mu(0,4) - (6\rho^2 + 12\rho)\mu(3,3) \\
&\quad - 12\rho\mu(5,1) + (24\rho^2 + 6)\mu(4,2)$.

**Corollary 2.3** Let $X$ and $Y$ have a bivariate distribution. If $X$ and $Y$ are independent, then

(i) $E(Q^*) = 2\mu(4,0)\mu(2,0)$

\[ \begin{align*}
&\quad + \frac{\mu(0,4)}{\mu^2(0,2)}.
\end{align*} \]

(ii) $E(Q^*) = \mu(6,0)\mu(0,2) + \frac{\mu(0,6)}{\mu^2(0,2)} + 3 \left( \frac{\mu(4,0)}{\mu(2,0)} + \frac{\mu(0,4)}{\mu(0,2)} \right)$

**Corollary 2.4** Let $X$ and $Y$ have a bivariate distribution. If $X$ and $Y$ are independent and identically distributed, then

(i) $E(Q^*) = \frac{2(1 + \mu(4,0))}{\mu^2(2,0)}$

(ii) $E(Q^*) = \frac{2\mu(6,0)}{\mu(2,0)} + 3 \frac{\mu(4,0)}{\mu(2,0)}$

3. Centered Product Moments of the Bivariate Normal Distribution

The pdf (probability density function) of the bivariate normal distribution is given by

\[ f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} \exp \left( \frac{-q(x, y)}{2} \right), \]

where

\[ (1 - \rho^2)q(x, y) = \frac{(x - \xi)^2}{\sigma_1^2} + \frac{(y - \theta)^2}{\sigma_2^2} - 2\rho(x - \xi)(y - \theta). \]

The following theorem is due to Kendall and Stuart (1969, 91).

**Theorem 3.1** The centered product moments $\mu(a,b) = E\left[(X - \xi^*)^a(Y - \theta^*)^b\right]$ of the bivariate normal distribution with pdf in (3.1) are given by

\[ \mu(a,b) = \sigma_1^a\sigma_2^b\lambda(a,b), \]

where

\[ \lambda(a,b) = (a + b + 1)\rho\lambda(a - 1, b - 1) + (a - 1)(b - 1)(1 - \rho^2)\lambda(a - 2, b - 2), \]

\[ \lambda(2a,2b) = \lambda(2a+1,2b+1) + \sum_{j=0}^{2b+1} (2a)!(2b+1)!(2j)! (2\rho)^j \]

\[ \lambda(2a+1,2b+1) = \lambda(2a+1,2b), \]

or

\[ \lambda(a,b) = (a + b - 1)\rho\sigma_1\sigma_2(\mu(a - 1, b - 1)) \]

\[ + (a - 1)(b - 1)(1 - \rho^2)\sigma_1^2\sigma_2^2(\mu(a - 2, b - 2)), \]

\[ \mu(2a,2b) = \sigma_1^a\sigma_2^b(2a+1)(2b+1)!(2j)! (2\rho)^j \]

\[ \mu(2a+1,2b+1) = \sigma_1^{a+1}\sigma_2^{b+1}(2a+1)(2b+1)!(2j)! (2\rho)^j \]

\[ \mu(2a+1,2b) = \mu(2a+1,2b+1) = 0. \]

The above can be rewritten as

\[ \mu(a,b) = (a + b - 1)\rho\sigma_1\sigma_2(\mu(a - 1, b - 1)) \]

\[ + (a - 1)(b - 1)(1 - \rho^2)\sigma_1^3\sigma_2^3(\mu(a - 2, b - 2)), \]

\[ \mu(2a,2b) = \sigma_1^a\sigma_2^b(2a)(2b)!(2j)! (2\rho)^j \]

\[ \mu(2a+1,2b+1) = \sigma_1^{a+1}\sigma_2^{b+1}(2a+1)(2b+1)!(2j)! (2\rho)^j \]

\[ \mu(2a+1,2b) = \mu(2a+1,2b+1) = 0. \]

Product moments that are needed for deriving standardized moments up to order 3 are provided below:
\[ \mu(2,0) = \sigma_2^2, \]
\[ \mu(0,2) = \sigma_1^2, \]
\[ \mu(4,0) = 3\sigma_1^4, \]
\[ \mu(0,4) = 3\sigma_2^4, \]
\[ \mu(6,0) = 15\sigma_1^6, \]
\[ \mu(0,6) = 3\sigma_2^6, \]
\[ \mu(1,1) = \rho \sigma_1 \sigma_2, \]
\[ \mu(2,2) = (1 + 2\rho^2)\sigma_1^2\sigma_2^2, \]
\[ \mu(1,3) = 3\rho \sigma_1 \sigma_2^3, \]
\[ \mu(3,1) = 3\rho \sigma_2 \sigma_1^3, \]
\[ \mu(2,4) = 3(1 + 4\rho^2)\sigma_1^2\sigma_2^4, \]
\[ \mu(3,3) = 3\rho(3 + 2\rho^2)\sigma_1^3\sigma_2^3, \]
\[ \mu(4,2) = 3(1 + 4\rho^2)\sigma_1^4\sigma_2^2, \]
\[ \mu(1,5) = 15\rho \sigma_1 \sigma_2 \sigma_1^3, \]
\[ \mu(5,1) = 15\rho \sigma_2 \sigma_1 \sigma_2^3. \]

4. Some Examples

(i) Bivariate Normal Distribution

Let us represent the bivariate normal distribution with pdf in (3.1) by,

\[ W = \begin{bmatrix} X \\ Y \end{bmatrix} \sim N_p(\mu, \Sigma), \quad \mu = \begin{bmatrix} \mu(2,0) \\ \mu(0,2) \end{bmatrix}. \]

It is known that for a \( p \)-variate normal distribution, \( W \sim N_p(\mu, \Sigma) \), the standardized distance \( Q = (W - \mu)^\top \Sigma^{-1}(W - \mu) \sim \chi_p^2 \), so that \( \beta_1 = E(Q) = p \), \( \beta_2 = E(Q^2) = p(p + 2) \) and \( \beta_3 = E(Q^3) = p(p + 2)(p + 4) \). That is for the univariate normal distribution, \( \beta_1 = 1, \beta_2 = 3, \beta_3 = 15 \) and for the bivariate normal distribution, \( \beta_1 = 2, \beta_2 = 8, \beta_3 = 48. \)

We derive Mahalanobis moments for bivariate normal and bivariate t-distribution by the method developed in Section 2. It may be mentioned that the standardized moments for bivariate elliptical distributions are easily obtained, but these are not easy for other distributions as it is difficult to derive the distribution of the standardized distance.

Theorem 4.1 The second and the third order standardized moments of bivariate normal distribution are given by \( \beta_2 = 8, \beta_3 = 48. \)

Proof. By the use of moments from Section 3, it follows from Theorem 2.1 (i) or preferably Corollary 2.1(i) that

\[ \left[ \sigma_1^2 \sigma_2^2 (1 - \rho^2) \right] E(Q^2) = (3\sigma_1^2)(\sigma_2^2(1 + 2\rho^2)) + 4(\rho^2\sigma_1^2\sigma_2^2) \]
\[ = 4(\rho^2\sigma_1^2\sigma_2^2) \]
\[ + 2\sigma_1^2 \sigma_2^2 (1 + 2\rho^2) \]

so that

\[ (1 - \rho^2) E(Q^2) = 3 + 4\rho^2(1 + 2\rho^2) + 2(1 + 2\rho^2) - 4\rho(3\rho) - 4\rho(3\rho) \]
\[ = 8(1 - 2\rho^2 + \rho^4). \]

Similarly by plugging in the moments from Section 3, it follows from Theorem 2.1 (ii) or preferably Corollary 2.1(ii) that

\[ \left[ \sigma_1^2 \sigma_2^2 (1 - \rho^2) \right] E(Q^3) = (15\sigma_1^2)(\sigma_2^2) + 4(15\sigma_1^4) \]
\[ - 8(\rho^2\sigma_1^2\sigma_2^2) + 12(\rho^2\sigma_1^2\sigma_2^2) + 12(\rho^2\sigma_1^2\sigma_2^2) \]
\[ + 3\sigma_1^2 \sigma_2^2(1 + 4\rho^2) + 3(\sigma_1^2 \sigma_2^2) \]

so that

\[ (1 - \rho^2) E(Q^3) = 15 + 15 - 8\rho^2(3\rho(3 + 2\rho^2)) - 12\rho(3\rho(3 + 2\rho^2)) \]
\[ - 6\rho(\rho(3\rho(3 + 2\rho^2)) - 6\rho(\rho(3\rho(3 + 2\rho^2)) \]
\[ + 12\rho(\rho(3\rho(3 + 2\rho^2)) + 12\rho(\rho(3\rho(3 + 2\rho^2)) \]
\[ + 3\rho(3\rho(3 + 2\rho^2)) + 3(\rho(3\rho(3 + 2\rho^2)) \]
\[ = 48(1 - 3\rho^2 + 3\rho^4 - \rho^6). \]

(ii) Bivariate T-Distribution

Let \( X' = (X_1, X_2) \) be the bivariate \( t \)-random vector with pdf

\[ f_t(x) = (2\pi)^{-1}|\Sigma|^{-1/2}|1 + (x - \theta)|^{-1}|\Sigma|^{-1}(x - \theta)|^{-1/2}, \]

where \( \theta = (\theta_1, \theta_2) \) is an unknown vector of location parameters and \( \Sigma \) is the 2x2 unknown positive definite matrix of scale parameters while the scalar \( V \) is assumed to be a known positive constant (Anderson, 2003, 289). For recent update on t-distributions see Kotz and Nadarajah (2005) and Kibria (2006) and the references therein.
The following theorem, due to Joarder (2006a), is needed to calculate Mahalanobis moments of the above bivariate t-distribution given by (4.1),

**Theorem 4.2** The centered product moments of the bivariate t-distribution with pdf in (4.1) are given by

\[
\mu(a,b;\nu) = (a + b - 1) \rho \sigma_1 \sigma_2 \mu (a - 1, b - 1) \gamma_2
\]

+ \((a - 1)(b - 1)(1 - \rho^2)\sigma_1^2 \sigma_2^2 \mu (a - 2, b - 2) \gamma_4,
\]

\[
\mu(2a,2b;\nu) = \sigma_1^2 \sigma_2^2 \left( \frac{2(2a)(2b)}{\nu^2} \sum (a-j)(b-j)(2j)! \gamma_{a+j} \right),
\]

\[
\mu(2a+1,2b+1;\nu) = \sigma_1^2 \sigma_2^2 \left( \frac{2(2a+1)(2b+1)}{\nu^2} \sum (a-j)(b-j)(2j)! \gamma_{a+j} \right),
\]

\[
\mu(2a,2b+1;\nu) = \mu(2a+1,2b;\nu) = 0,
\]

where \( \gamma_a = \frac{(\nu/2)^{-a/2} \Gamma(\nu/2 - a/2)}{\Gamma(\nu/2)}, \)

\( \nu > a. \)

By the use of the above moments in Theorem 2.1 or preferably in Corollary 2.1, we have the second and third order Mahalanobis moments of the bivariate t-distribution having pdf in (4.1):

\[
\beta_2 = 8 \frac{\nu - 2}{\nu - 4}, \quad \nu > 4,
\]

\[
\beta_3 = 48 \frac{(\nu - 2)^2}{(\nu - 4)(\nu - 6)}, \quad \nu > 6.
\]

(iii) **Multivariate Elliptical Distribution**

The second and third order Mahalanobis moments are calculated for p-variate elliptical distribution. With \( p = 2 \), the results boil down to bivariate elliptical distribution. Consider the multivariate elliptical distribution with pdf

\[
f_\Sigma(x) = g((x-\mu)\Sigma^{-1}(x-\mu)), \quad (4.2)
\]

where \( x \) is a \( p \)-dimensional column vector with mean \( E(X) = \mu \) and the covariance matrix \( Cov(X) = p^{-1}E(R^2)\Sigma \) where \( R^2 = ZZ \) and \( Z = \Sigma^{-1/2}(X - \mu) \). Then we have the following theorem (cf. Anderson, 2003, 103):

**Theorem 4.3** Let \( X \) have the multivariate elliptical distribution with pdf in (4.2). Then the second and the third order Mahalanobis moments of the distribution are given by

\[
\beta_2 = E(Q^2) = p^2 \frac{E(R^4)}{E(R^2)}, \quad \text{and}
\]

\[
\beta_3 = E(Q^3) = p^3 \frac{E(R^6)}{E(R^4)},
\]

respectively, where \( R^2 = ZZ \) and \( Z = \Sigma^{-1/2}(X - \mu) \).

**Proof.** The covariance matrix of the elliptical distribution is given by \( Cov(X) = p^{-1}E(R^2)\Sigma \) so that the standardized distance is given by \( Q = (X - \mu)'(p^{-1}E(R^2)\Sigma)^{-1}(X - \mu) \). Then the theorem is obvious by virtue of

\[
Q = \frac{pR^2}{E(R^2)} = \frac{pR^4}{E(R^4)} = \frac{p^2R^6}{E(R^4)},
\]

Note that if the form of \( g(.) \) is known, the second and the third order Mahalanobis moments of the distribution can be calculated by the pdf of \( R \) given by

\[
h(r) = \frac{2\pi^{-1/2}}{\Gamma(p/2)} r^{p-1} g(r^2), \quad 0 < r.
\]

It is well known that for the multivariate normal distribution \( R^2 \sim \chi^2_p \), and for the multivariate t-distribution with pdf

\[
f_\nu(x) = \frac{\Gamma((\nu+p)/2)}{\Gamma((\nu/2)(\nu\pi)^{p/2})} \left| \Sigma \right|^{1/2} (1+(x-\theta)'\Sigma^{-1}(x-\theta))^{-(\nu+p)/2},
\]

\( \nu > 2, \)

we have \( p^{-1}R^2 \sim F(p,\nu) \).

(iv) **Bivariate Chi-square Distribution**

The following bivariate chi-square distribution was introduced by Joarder (2006b).

**Theorem 4.4** The random variables \( U \) and \( V \) are said to have a correlated bivariate chi-square distribution each with \( m \) degrees of freedom, if its pdf is given by

\[
f_{\nu}(u,v) = \frac{(\nu)^{\nu/2-1} e^{-\nu/2}}{2^{\nu/2} \Gamma(m/2) \Gamma(1-\rho^2)^{1/2} \sum \Gamma((k+m)/2) \gamma_k(1-\rho^2)^{k/2}},
\]

\( m > 2, \quad -1 < \rho < 1. \)

Since it is difficult to derive the distribution of Mahalanobis distance of the above bivariate chi-square distribution, Joarder (2006b) derived the second and third order Mahalanobis moments of the above bivariate chi-square distribution by using the results developed in Section 2. In case
$\rho = 0$, the pdf of the joint probability distribution in Theorem 4.4, would be that of the product of two independent chi-square random variables $U \sim \chi^2_m$ and $V \sim \chi^2_n$.

Acknowledgement

The author gratefully acknowledges the excellent research support provided by King Fahd University of Petroleum & Minerals, Saudi Arabia through the Project FT 2004-23

References


Minimax Estimation of the Parameter of the Rayleigh Distribution

M. Kamruj Jaman Bhuiyan, Manindra Kumar Roy, M. Farouq Imam

1Department of Agricultural Statistics, Bangladesh Agricultural University, Mymensingh,
2Department of Statistics, University of Chittagong, Chittagong, Bangladesh

Correspondence e-mail: manindraroy@yahoo.com

Abstract

Minimax estimators of the parameter $\theta$ of the Rayleigh distribution have been obtained for the well known Quadratic and Modified linear Exponential (MLINEX) by applying the theorem of Lehmann [1950]. Then the obtained results have been interpreted in the light of two-person zero sum game and compared with the classical Maximum Likelihood Estimator. Efficiency of the estimators were also been studied.

Key words and Phrases: Minimax, Game theory, Rayleigh distribution.

1 Introduction

The minimax estimation is an upgraded non-classical approach in the estimation area of statistical inference, which was introduced by Abraham Wald (1945) from the concept of game theory. It opens a new dimension in statistical estimation and enriched the method of point estimations. Von Neumann [1944] introduced the word minimax in game theory which is the optimum strategy of the second player in the two person zero game. According to Abraham Wald, “minimax approach tries to guard against the worst by requiring that the chosen decision rule should provide maximum protection against the highest possible risk”. An estimator having this property is called a minimax estimator. The most important elements in the minimax approach are the specification of the prior distribution and the loss function used. In this paper, quadratic and modified linear exponential (MLINEX) loss functions have been used to obtain the minimax estimators of the parameter of the Rayleigh distribution. Rayleigh distribution is a special case of the Weibull distribution. It has been used to study the scattering of radiation, wind speeds or to make certain transformation.

The probability density function of the Rayleigh distribution is

$$f(x; \theta) = \frac{x}{\theta^2} e^{-\frac{x^2}{2\theta^2}} ; x, \theta > 0$$ (1)

where $\theta$ is the parameter of the distribution.

The properties of the Rayleigh distribution and the estimation of its parameters were extensively studied by Hirai [1966, 1972]. It is easily seen that the maximum likelihood estimator (MLE) of $\theta$ is

$$\sqrt{\frac{\sum x^2}{n}} \frac{\theta^2}{n}$$ with large sample variance $\theta^2/4n$.

In the present paper we have derived the minimax estimators of the parameter $\theta$ by using quadratic and MLINEX [1998] type loss functions. The derivation depends primarily on a theorem due to Lehmann which can be stated as follows:

Lehmann’s Theorem: If $\tau = \{F_\theta, \theta \in \Theta\}$ be a family of distribution functions and $D$ a class of estimators of $\theta$. Suppose that $d^* \in D$ is a Bayes estimator against a prior distribution $\xi^* (\theta)$ on $\Theta$, and $R(d^*, \theta) = \text{constant on } \Theta$; then $d^*$ is a minimax estimator of $\theta$.

2 Main Results

Theorem 1.1: Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from the density (1), then

$$\hat{\theta}_{\text{MQL}} = \frac{\Gamma \left( \frac{n + 1}{2} \right)}{\alpha^2(n)} \left( \frac{\sum x_i^2}{\frac{n}{\alpha}} \right)^\frac{1}{2}$$

is the minimax estimator of the parameter $\theta$ for the quadratic loss function (MQL) of the type,

$$L(\theta, d) = \left( \frac{\theta - d}{\theta} \right)^2$$

where $\theta$ is the parameter to be estimated and $d$ is the estimate of $\theta$.

Theorem 1.2: Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from the density (1), then

$$\hat{\theta}_{\text{MML}} = \frac{\left( \Gamma(n) \right)^\frac{1}{2}}{\left( \Gamma \left( \frac{n + 1}{2} \right) \right)^\frac{n}{2}} \left( \frac{\sum x_i^2}{\frac{n}{\alpha}} \right)^\frac{1}{2}$$

is the minimax estimator of $\theta$. 
the parameter $\theta$ for the MLINEX loss function of the type,
\[ L(\theta, d_1) = \frac{1}{\theta} \left( \frac{d_1}{\theta} \right)^2 - c \left( \ln \frac{d_1}{\theta} - 1 \right) ; \omega > 0, c \neq 0 \]

Where $d_1$ is the estimate of $\theta$, $\omega$ and $c$ are two known parameters the loss function.
First we have to prove the theorem (1.1). To prove the theorem we shall use Lehmann’s theorem, which has been stated before.

Here, we consider the quadratic loss function (QLF) of the form
\[ L(\theta, d) = \left( \frac{\theta - d}{\theta} \right)^2 \]
which is a non-negative symmetric and continuous loss function of $\theta$ and $d$.
In order to prove the theorem it will be sufficient to show that \( d = \frac{\Gamma \left( \frac{n + 1}{2} \right)}{n \Gamma \left( \frac{n}{2} \right)} \left( \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \) is a minimax estimator of $\theta$ for the loss function (2).

For this, first we have to find the Bayes estimator $d$ of $\theta$. Then if we can show that the risk of $d$ is constant, then the theorem (1.1) will be proved. Let us assume that $\theta$ has Jeffreys’ non-informative prior density defined as, \( g(\theta) = \frac{1}{\theta} ; \theta > 0 \)

Then the posterior distribution of $\theta$ for given the random sample $X_1, X_2, \ldots, X_n$ is
\[ f(\theta|x) = \left( \frac{\sum_{i=1}^{n} x_i^2}{\Gamma(n)} \right)^{n} e^{-\frac{\sum_{i=1}^{n} x_i^2}{\theta}} x, \theta > 0 \]

Now, for the quadratic loss function (2), the Bayes estimator of $\theta$ is given by
\[ d = \frac{\Gamma \left( \frac{n + 1}{2} \right)}{n \Gamma \left( \frac{n}{2} \right)} \left( \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} = KT^{\frac{1}{2}} \]
where \( K = \frac{\Gamma \left( \frac{n + 1}{2} \right)}{n \Gamma \left( \frac{n}{2} \right)} \) and \( T = \sum_{i=1}^{n} x_i^2 \) is a complete sufficient statistics for $\theta$.

The risk function of the estimator $d$ is
\[ R(\theta) = E[L(\theta, d)] = E \left[ \left( \frac{\theta - d}{\theta} \right)^2 \right] \]
\[ = \frac{1}{\theta^2} K^2 \Gamma(\frac{n}{2}) - 2E \left( \frac{1}{\theta^2} \right) + \theta^2 \]
Since each of $X_i$ is distributed as Rayleigh variate with parameter $\theta$, so the statistic $T = \frac{1}{\theta^2}$ is distributed as gamma distribution with the parameters $\frac{1}{\theta^2}$ and $n$. i.e., $T = G \left( \frac{1}{\theta^2}, \theta \right)$

Therefore the p.d.f. of $T$ is given by
\[ h(t) = \frac{1}{\theta^2 \Gamma(n)} e^{-\frac{t}{\theta^2}} ; t > 0, \theta > 0 \]

From where we get,
\[ E(T) = n \theta^2 \]
and
\[ \int_0^{\infty} \frac{1}{\Gamma(n)} \frac{\Gamma \left( \frac{n + 1}{2} \right)}{n \Gamma \left( \frac{n}{2} \right)} \left( \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \] \( \text{E}(T) = n \theta^2 \)

Now using these results we get, the risk function,
\[ R(\theta) = \frac{1}{\Gamma(n)} \Gamma \left( \frac{n + 1}{2} \right) \]
which is a constant.
So, according to the Lehmann’s theorem it follows that, \( d = \frac{\Gamma \left( \frac{n + 1}{2} \right)}{n \Gamma \left( \frac{n}{2} \right)} \left( \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \) is the minimax estimator for the parameter $\theta$ of the Rayleigh distribution under the quadratic loss function of the form (2).

Now we are going to prove the theorem (1.2.). To prove the theorem (1.2.), we have to show that ,
\[ \delta_{\text{MLE}} = \left( \frac{\Gamma(n)}{\Gamma \left( \frac{n + 1}{2} \right)} \right) \left( \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \]
the parameter $\theta$ for the MLINEX loss function of the type,
\[ L(\theta, d_1) = \frac{1}{\theta} \sum_{i=1}^{n} x_i^2 - c \ln \frac{d_1}{\theta} - 1 \]
which is an asymmetric loss function of $\theta$ and $d_1$.
In order to prove the theorem it will be sufficient to show that \( d_1 = \frac{\Gamma(n)}{\Gamma \left( \frac{n + 1}{2} \right)} \left( \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \)
the minimax estimator of $\theta$ for the loss function (3).

For this, first we have to find the Bayes estimator $d_1$ of $\theta$. Then if we can show that the risk of $d_1$ is constant, then the theorem (1.2) will be proved. As before, we get the posterior distribution of $\theta$ for given the random sample $X_1, X_2, \ldots, X_n$ as
\[ f(\theta|x) = \left( \frac{\sum_{i=1}^{n} x_i^2}{\Gamma(n)} \right)^{n} e^{-\frac{\sum_{i=1}^{n} x_i^2}{\theta}} x, \theta > 0 \]

Now, the Bayes estimator of $\theta$ under the MLINEX loss function (3) is given by
\[ \hat{\theta} = \left[ E_\theta \left( \theta^{-c} \right) \right]^{\frac{1}{c}} \]
where,
\[ E_\theta \left( \theta^{-c} \right) = \int_0^{\infty} \theta^{-c} f(\theta/X) d\theta \]
\[ = \frac{\Gamma(n)}{\Gamma(n)} \int_0^{\infty} e^{-\frac{\sum_{i=1}^{n} x_i^2}{\theta^2}} \theta^{-\frac{c+1}{2}} d\theta \]
\[
\frac{\left(\frac{\xi_i^2}{\xi_0^2}\right)^n}{\Gamma(n)} \frac{\Gamma\left(n + \frac{\xi}{2}\right)}{\Gamma(n)} = \frac{1}{n + \frac{\xi}{2}}
\]

By using this results we get,
\[
\hat{c} = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{\xi_i^2}{\xi_0^2} \cdot \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds
\]

where, \( p = \frac{\Gamma(n)}{\Gamma(n + \frac{\xi}{2})} \) and \( S = \sum_{i=1}^{\xi} X_i^2 \) is the complete sufficient statistics for \( \theta \).

Now the risk function under the MLINEX loss function (3) is
\[
R_M(\theta) = E[\ln \hat{\theta} - \ln \theta]
\]

\[
= \omega E\left[\frac{1}{\theta} - c \ln \frac{\theta}{\hat{\theta}} - 1\right]
\]

Where,
\[
\hat{\theta} = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{\xi_i^2}{\xi_0^2} \cdot \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds
\]

and
\[
E[\hat{\theta}] = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds
\]

Now by using these results we get,
\[
E[\hat{\theta}] = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds
\]

Using the relation,
\[
\frac{1}{\theta^2} y \Rightarrow s = \theta^2 y = ds = \theta^2 dy
\]

We get,
\[
E[\hat{\theta}] = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{-\infty}^{\infty} \frac{n + \frac{\xi}{2}}{\Gamma(n + \frac{\xi}{2})} ds
\]

\[
= \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{0}^{\infty} n y e^{-y} dy = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{0}^{\infty} n y e^{-y} dy = \frac{1}{\Gamma(n + \frac{\xi}{2})} \int_{0}^{\infty} n y e^{-y} dy
\]

where, \( \Gamma(n)' = \frac{1}{\Gamma(n)} n y e^{-y} dy \) is the first derivative of \( \Gamma(n) \) w.r.t. \( y \).

Therefore,
\[
E[\hat{\theta}] = \ln p + c \ln \frac{\theta}{\hat{\theta}} + \frac{\Gamma(n)'}{2\Gamma(n)} + \frac{1}{2\Gamma(n)}
\]

Interpretation of minimax estimators with Two-Person Zero-Sum Game

According to Wald [1950] the following statistical problem is equivalent to some two-person zero-sum-game between the Statistician (Player-II) and Nature (Player-I). Here the pure strategies of Nature are the different values of \( \theta \) in the interval \((0, \infty)\). The mixed strategies of Nature are the prior densities of \( \theta \) in the interval \((0, \infty)\). The pure strategies of Statistician are all possible decision functions in the interval \((0, \infty)\).

Expectation of the loss function \( L(\theta, d) \) is the risk function, \( R(\theta, d) = E[L(\theta, d)] \) which is the winning of player-I. \( R(\theta, d) \) is the value of \( \int R(\theta, d) d\xi(\theta) \), where \( \xi(\theta) \) is the prior density \( \theta \). If the loss function is continuous in both \( d \) and \( \theta \) convex in \( d \) for each \( \theta \) then there exist measures \( \xi^* \) and \( d^* \) for all \( \theta \) and \( d \) so that the following relation holds good.

\[
R(\xi^*, d^*) \leq R(\xi^*, d^*) \leq R(\xi^*, d^*)
\]

The number \( R(\xi^*, d^*) \) is known to be the value of the game and \( \xi^* \) and \( d^* \) are the corresponding optimum strategies of Player-I and Player-II. In
statistical terms $\xi^*$ is the least favorable prior density of $\theta$ and $d^*$ is a minimax estimator of $\theta$.

It has been shown that, here according to Wald [1950]

Player-I : Nature (Parameter)
Player-II : Statistician
Pure strategies of Nature : The different values of the parameter $\theta$ in the interval $(0, \infty)$.
Pure strategies of Statistician : All possible decision functions in the interval $(0, \infty)$.
Winning of player-I : The risk function, $R(\theta, d) = E[L(\theta, d)]$
Value of the game : The loss of the Statistician.

$$\text{i) } g \left( \frac{\theta}{\theta} \right) = \frac{\Gamma \left( n + \frac{1}{2} \right)}{\Gamma(n)} \text{ for quadratic loss function.}$$

$$\text{ii) } g \left( \frac{\ln \theta}{\ln \theta} \right) = \frac{\Gamma \left( n + \frac{c}{2} \right)}{2\Gamma(n)} \text{ for MLINEX loss function.}$$

In both the cases, $g \left( \theta \right) > 0$ is the optimum strategy for Player-I (Nature).

**Efficiency of $\hat{\theta}_{MQL}$ w.r.t $\hat{\theta}_{MLE}$ of $\theta$:**

We have

$$\hat{\theta}_{MQL} = \frac{\Gamma \left( n + \frac{1}{2} \right)}{n\Gamma(n)} \left( \sum_{i=1}^{n} x_i \right)^{\frac{1}{2}}$$

And

$$\text{Var} (\hat{\theta}_{MQL}) = \theta^2 \left[ \frac{\Gamma \left( n + \frac{1}{2} \right)}{n\Gamma(n)} \right] \left( \sum_{i=1}^{n} x_i \right)^{\frac{1}{2}}$$

And the MLE of $\theta$ is

$$\hat{\theta}_{MLE} = \left( \frac{S^2}{n} \right)^{\frac{1}{2}}$$

And

$$\text{Var} (\hat{\theta}_{MLE}) = \frac{\text{Var} (\hat{\theta}_{MLE})}{n} = \theta^2 \left[ \frac{\Gamma \left( n + \frac{1}{2} \right)}{n\Gamma(n)} \right]$$

Therefore the efficiency of $\hat{\theta}_{MQL}$ w.r.t. $\hat{\theta}_{MLE}$ of $\theta$ is

$$E = \frac{\text{Var} (\hat{\theta}_{MLE})}{\text{Var} (\hat{\theta}_{MLE})} = \frac{n(\Gamma(n))^2}{\left( \frac{n}{\Gamma(n)} \right)^{\frac{1}{2}}}$$

$$> 1 \text{ for } c > 0.$$
Efficiency of $\hat{\theta}_{MML}$ w.r.t $\hat{\theta}_{MLE}$

We have $\text{Var} ( \hat{\theta}_{MLE} ) = \frac{\theta^2}{n} \left( \frac{\Gamma\left(n + \frac{1}{2}\right)}{n\Gamma(n)} \right)^2$

We have $\text{Var} ( \hat{\theta}_{MML} ) = \theta^2 \left( \frac{\Gamma(n)}{\Gamma(n + \frac{1}{2})} \right)^2 \left( n - \frac{\Gamma\left(n + \frac{1}{2}\right)}{\Gamma(n)} \right)^2$

Therefore the efficiency of $\hat{\theta}_{MML}$ w.r.t $\hat{\theta}_{MLE}$ of $\theta$ is

$$E = \frac{\text{Var}( \hat{\theta}_{MLE} )}{\text{Var}( \hat{\theta}_{MML} )} = \left( \frac{\Gamma\left(n + \frac{1}{2}\right)}{n\Gamma(n)} \right)^2 \leq 1 \quad \text{if } c \neq 0.$$  

This means the non-classical ME under the QLF is more efficient than the classical MLE Procedure.

4 Empirical Study:

Mean-Squared errors (MSE) are considered to compare the different estimators of the parameter $\theta$ of the Rayleigh distribution that are obtained by maximum likelihood and minimax methods for Quadratic and MLINEX loss functions. The MSE of an estimator is defined by

$$\text{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$$

The estimated values of the parameter and MSE of the estimators are computed by the Monte-Carlo Simulation method from the Rayleigh distribution. It is seen that for small sample sizes ($n < 25$), minimax estimators for Quadratic loss function appear to be better than the minimax estimator under MLINEX loss function and the classical maximum likelihood estimator. But for large sample sizes ($n > 25$), all the three estimators have approximately the same MSE. The obtained results are demonstrated in Table 1, Table 2, Table 3 and Table 4 and presented in Figure 1, Figure 2, Figure 3 and Figure 4 respectively.

<table>
<thead>
<tr>
<th>n</th>
<th>Criteria</th>
<th>$\hat{\theta}_{MLE}$</th>
<th>$\hat{\theta}_{MQL}$</th>
<th>$\hat{\theta}_{MML}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Estimated value</td>
<td>1.0788</td>
<td>1.0352</td>
<td>1.2422</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0854</td>
<td>0.0742</td>
<td>0.1637</td>
</tr>
<tr>
<td>5</td>
<td>Estimated value</td>
<td>1.0872</td>
<td>1.0606</td>
<td>1.1785</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0559</td>
<td>0.0496</td>
<td>0.0886</td>
</tr>
<tr>
<td>10</td>
<td>Estimated value</td>
<td>0.9962</td>
<td>0.9840</td>
<td>1.0358</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0243</td>
<td>0.0239</td>
<td>0.0275</td>
</tr>
<tr>
<td>15</td>
<td>Estimated value</td>
<td>1.0559</td>
<td>1.0513</td>
<td>1.0876</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0197</td>
<td>0.0185</td>
<td>0.0246</td>
</tr>
<tr>
<td>20</td>
<td>Estimated value</td>
<td>1.0040</td>
<td>0.9979</td>
<td>1.0235</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0120</td>
<td>0.0118</td>
<td>0.0130</td>
</tr>
<tr>
<td>25</td>
<td>Estimated value</td>
<td>0.9905</td>
<td>0.9858</td>
<td>0.9953</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0096</td>
<td>0.0096</td>
<td>0.0096</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>Criteria</th>
<th>$\hat{\theta}_{MLE}$</th>
<th>$\hat{\theta}_{MQL}$</th>
<th>$\hat{\theta}_{MML}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Estimated value</td>
<td>2.1577</td>
<td>2.0704</td>
<td>1.7265</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.3418</td>
<td>0.2968</td>
<td>0.2777</td>
</tr>
<tr>
<td>5</td>
<td>Estimated value</td>
<td>2.1745</td>
<td>2.1213</td>
<td>1.8459</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.2237</td>
<td>0.1986</td>
<td>0.1629</td>
</tr>
<tr>
<td>10</td>
<td>Estimated value</td>
<td>1.9924</td>
<td>1.9680</td>
<td>1.7821</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0972</td>
<td>0.0958</td>
<td>0.1251</td>
</tr>
<tr>
<td>15</td>
<td>Estimated value</td>
<td>2.1198</td>
<td>2.1027</td>
<td>1.9334</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0788</td>
<td>0.0740</td>
<td>0.0580</td>
</tr>
<tr>
<td>20</td>
<td>Estimated value</td>
<td>2.0080</td>
<td>1.9959</td>
<td>1.8501</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0481</td>
<td>0.0475</td>
<td>0.0632</td>
</tr>
<tr>
<td>25</td>
<td>Estimated value</td>
<td>1.9811</td>
<td>1.9716</td>
<td>1.8368</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0385</td>
<td>0.0386</td>
<td>0.0594</td>
</tr>
</tbody>
</table>

Table 1: Estimated value and MSE of different estimates of the parameter $\theta$ of the Rayleigh distribution when $\theta = 1$ and $c = 1$

Table 2: Estimated value and MSE of different estimates of the parameter $\theta$ of the Rayleigh distribution when $\theta = 1$ and $c = -1$

Table 3: Estimated value and MSE of different estimates of the parameter $\theta$ of the Rayleigh distribution when $\theta = 2$ and $c = 7.5$
Table-4: Estimated value and MSE of different estimates of the parameter $\theta$ of the Rayleigh distribution when $n = 20$ and $c = -5$

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Criteria</th>
<th>$\hat{\theta}_{MLE}$</th>
<th>$\hat{\theta}_{MQL}$</th>
<th>$\hat{\theta}_{MML}$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Estimated value</td>
<td>1.0040</td>
<td>0.9979</td>
<td>1.0514</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0120</td>
<td>0.0118</td>
<td>0.0158</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Estimated value</td>
<td>2.0080</td>
<td>1.9959</td>
<td>2.1028</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0481</td>
<td>0.0475</td>
<td>0.0633</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Estimated value</td>
<td>3.0120</td>
<td>2.9938</td>
<td>3.1543</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.1083</td>
<td>0.1069</td>
<td>0.1424</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Estimated value</td>
<td>4.0160</td>
<td>3.9918</td>
<td>4.2057</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.1926</td>
<td>0.1901</td>
<td>0.2533</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Estimated value</td>
<td>5.0200</td>
<td>4.9898</td>
<td>5.2572</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.3009</td>
<td>0.2970</td>
<td>0.3958</td>
<td></td>
</tr>
</tbody>
</table>

References

Detection of Outliers in Non-linear Time Series: A Review

Murugason Kallianna Gounder  
Laboratory of Applied and Computational Statistics, Institute for Mathematical Research  
University Putra Malaysia, Malaysia

Mahendran Shitan  
Department of Mathematics, Faculty of Science, Laboratory of Applied and Computational Statistics Institute for Mathematical Research, University Putra Malaysia, Malaysia

A.H.M. Rahmatullah Imon  
Department of Statistics, University of Rajshahi, Rajshahi – 6205, Bangladesh

Abstract

Non-linear time series analysis has gained much attention in recent years, primarily due to the fact that linear time series models have encountered various limitations in real applications. The development in nonparametric regression has established a solid foundation for non-linear time series analysis. It is now evident that outliers have several adverse effects on the entire inferential procedure of time series analysis including tests for nonlinearity in the time series data. Detection of outliers has recently been studied in an extensive way in the statistical time series literature though it is more complicated than the detection of outliers in linear models. The detection of outliers is even more complicated in nonlinear time series and it is still a very unexplored area of research. In this paper we have presented a brief review of some recently developed outlier detection techniques in nonlinear time series data.

1. Introduction

Time series observations are generally influenced by interruptive events, such as strikes, outbreaks of war, sudden political or economic crisis, unexpected heat or cold waves, natural disasters such as Tsunami or even unnoticed errors of typing or recording. Such observations are usually referred to as outliers. Outliers are defined as observations which appear to be inconsistent with the remainder of the data set. It is well known that time series data are very often contaminated with outliers or affected by structural changes as level shifts. These can affect all the stages of time series analysis: model identification, estimation and forecasting. Thus the detection of genuine outliers and correction are vital in time series analysis (Tolvi, 2000; Battagila and Orfei, 2002; Charles and Darne, 2005). The main purpose of outlier correction is to modify the data in such a way that the normality hypothesis can be accepted, so that one can proceed with proper estimation, testing and inference. Outliers may have significant impact on the results of standard methodology for time series analysis, therefore it is important to detect them, estimate their effects and undertake the appropriate corrective actions.

In spite of clear importance of this topic, published works on the detection of outliers in time series was slow to appear. Barnett and Lewis (1994) pointed out that the major difficulties in detecting outliers are as in general linear model data, outliers in time series data are not necessarily extreme values and they can be cloaked to some extent by the general structure of the process. In particular, we may experience a smoothing-out effect when we attempt to examine outliers in terms of derived quantities such as the values of the estimated residuals about a fitted model. In the linear model any outlier does not tend to influence adjacent observations, the same need not be true for time series data in view of the correlational pattern of the basic process.

The detection of outliers is even more complicated in nonlinear time series and not much attempt is made on the topic until quite recently. We present a review of nonlinear time series models in section 2 while in section 3 we present a review on outliers in time series data. In section 4, we present a brief review of some recently developed outlier detection techniques in nonlinear time series data.

2. Non-linear Time Series Models

Many non-linear time series models have been introduced in the literature (see Kantz and Schreiber, 2005; Zivot and Wang, 2005). With the development of chaos theory, the role of nonlinearity has also become widely recognized among statisticians. The importance
of nonlinearity in statistical modeling is emphasized notably by Tong (1990). A simple but important tool for spotting non-linearities is a careful inspection of time plot. However it can be difficult to distinguish between data from:

(i) a non-linear model
(ii) a linear model with normally distributed disturbances to which some outliers have been added.
(iii) a linear process with disturbances that are not normally distributed.

If a linear model can be fitted to the transformed data, then a non-linear model is appropriate to the original data. In particular, a series which shows multiplicative seasonality, can be changed to additive seasonality by taking logs which can be handled using linear methods. However, the multiplicative model for the original data will be non-linear.

Towards the end of the seventies of last century there was an increasing demand to model more complex time series models in the presence of outliers than those given by linear autoregressive moving average (ARMA) model. One drawback of the stationary ARMA model with the Gaussian (normal) noise \( a_t \) is that it is unable to capture irreversibility. Time irreversibility is one of the major features exhibited by non-linear or non-Gaussian time series model. A stationary time series \( X_t \) is time reversible if for any integer \( n > 0 \) and any \( t_1, t_2, \ldots, t_n \) (integers), the vectors \((X_{t_1}, \ldots, X_{t_n})\) and \((X_{t_1}, X_{t_2}, \ldots, X_{t_n})\) have the same multivariate distribution. Weiss (1986) showed that stationary ARMA process with a nontrivial AR component are time reversible if and only if they are Gaussian.

Two major classes of nonlinear models were developed by the end of the 1970’s. These were the threshold model of Tong (1978) and the bilinear models by Granger and Anderson (1978). A full generalization of the threshold model occurred in Tong and Lim (1980) and full generalization of the bilinear model appeared in Subba Rao (1979). Tong (1978) proposed the threshold autoregressive (TAR) models. Haggan and Ozaki (1981) considered the exponential autoregressive (EAR) model and showed that this model is useful in modeling sound vibration. Priestly (1980) considered the state-dependent model. In econometric literature, Engle (1982) proposed the conditional heteroscedastic autoregressive (ARCH) model to capture the serial dependence in conditional variance of the time series. The model has attracted much attention since 1985, and several generalizations of the model are available in the literature such as generalized ARCH (GARCH) model of Bollerslev (1986), the exponential GARCH (EGARCH) of Nelson (1991), the conditional heteroscedastic ARMA (ARCHMA) model of Tsay (1988) etc. Hamilton (1994) proposed Markov switching models for modeling business cycles of macroeconomic time series.

### 2.1 Threshold models

The simplest form of threshold AR(1) model is given as:

\[ X_t = \phi X_{t-1} + a_t, \quad \text{if } X_{t-1} > c \]

\[ X_t = \phi' X_{t-1} + a_t, \quad \text{if } X_{t-1} \leq c \]  

(1)

where \( c \) is the threshold value, \( \phi \neq \phi' \) and \( \{a_t\} \sim WN(0, \sigma_a^2) \); \( \{a_t\} \sim WN(0, \sigma_a^2) \). The time series (1) satisfies a different AR or regime whenever the threshold \( c \) is crossed. Many hydrological series appear to satisfy this model. Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to pick an estimate for \( c \). Chan (1995) showed that the estimate of \( c \) converges at a rate \( 1/n \) to a true value, which is faster than the usual rate \( 1/\sqrt{n} \).

A good example of threshold model is the TAR model defined as:

\[ X_t = \phi_1 X_{t-1} + \ldots + \phi_{t-1} X_{t-(t-1)} + a_t \quad \text{for } X_{t-d} > c \]

\[ X_t = \phi_0 + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + a_t \quad \text{otherwise} \]

(2)

where \( \{a_t\} \sim WN(0, \sigma_a^2) \) and \( \{a_t\} \sim WN(0, \sigma_a^2) \), \( 1 \leq d \leq \max(p_1, p_2) \). TAR models can easily model features like limit cycles and time irreversibility (Tong and Lim, 1980). Because of these two features TAR model is a successful nonlinear model.

### 2.2 Bilinear models

A class of non-linear models, called bilinear class, may be regarded as plausible non-linear extension of ARMA model, rather than the AR model. Bilinear models incorporate cross product terms involving lagged values of the time series and of the innovation process. Denoting the time series \( \{X_t\} \) and the innovation process by \( \{Z_t\} \), a simple example could be

\[ X_t = \alpha X_{t-1} + \beta Z_{t-1} X_{t-1} + Z_t \]

(3)

where \( \alpha \) and \( \beta \) are constants. As well as the innovation term \( Z_t \), this model includes one AR term plus one cross-product term involving \( Z_{t-1} \) and \( X_{t-1} \).
ARCH models can account for some excess kurtosis. It has been noted, however that even after fitting an ARCH model to some time series, the residuals are still not compatible with the assumption of normally distributed errors, but in fact still have excess kurtosis. This observation has lead to many proposed solutions, one of which is to assume a heavier-tailed distribution, such as Student’s $t$, for errors. Another possibility is to assume that outliers are the reason for this excess kurtosis.

2.4 Generalized ARCH (GARCH) Model

GARCH model of order $(p,q)$ assumes conditional variance depends on the squares of the last $p$ values of the series and on the last values of $\sigma_t^2$. The conditional GARCH (1,1) may be written as:

$$\sigma_t^2 = \gamma + \alpha Z_{t-1}^2 + \beta \sigma_{t-1}^2$$

(6)

where $\gamma$, $\alpha$ and $\beta$ must satisfy $(\alpha + \beta) < 1$ for stationarity. GARCH model can be thought of as an approximation to a high order ARCH model. GARCH (1,1) model has become the ‘standard’ model for describing changing variance because of its relative simplicity. It is often found that $(\alpha + \beta) \approx 1$ so that the stationarity condition may not be satisfied. If $(\alpha + \beta) = 1$, the process does not have a finite variance. Other extensions of GARCH models are integrated GARCH (IGARCH), quadratic GARCH (QGARCH) and exponential GARCH (EGARCH) (Nelson, 1991).

2.5 Smooth Threshold Autoregressive (STAR) Model

A key feature of the TAR models is the discontinuous nature of the AR relationship as the threshold is passed. This has led the researches to consider alternative ways of allowing the AR parameter to change. STAR models were proposed by Tong (1990) so as to give a smooth continuous transition from one linear to another rather than a sudden jump. STAR models are further studied by Terasvirta (1994) and many others. Prediction error technique is used to estimate the parameters from a STAR model. To ease the estimation task the delay and the order of the AR process is fixed to their correct values, i.e. $d = 1; k = 1$.

2.6 Self-exciting Threshold (SETAR) Model

Tong (1990) defines a self-exciting TAR model (SETAR), when the choice from a various sets of parameter values is determined by just one of the past
values, say $X_{t-d}$ where $d$ is the delay. The choice is determined solely by the value of $X_{t-1}$ and so the model is indeed self-exciting when $d = 1$. The SETAR model operates with a regime of $l$ models, where each model $j$ of the regime is an AR process of some order $p^j$. Selection of the regime at time $t$ happens based on $d$ times lagged value of the univariate output of the process itself (see Holst et al., 1994).

### 2.7 Markov Modulation AR (MMAR) Model

An MMAR model is a special kind of AR regime model, where the regime variable $J_t = \{1, \ldots, l\}$ is given by a stationary Markov chain, i.e. with time invariant transition matrix $P$. The regime variable is assumed unobserved. Similar to SETAR models we use the abbreviation MMAR $(l; k_1, \ldots, k_l; P)$ to describe an MMAR process and its parameters. In case all models have the same order $p$ we simply use the notation MMAR $(l; p)$ and explicitly specify the remaining parameters when necessary.

MMAR models are a subclass of switching dynamic linear models, (see Holst, 1994; Murphy, 1998), which are used when the signals can be thought of as generated by a phenomena in an unknown discrete state. Sleep analysis (see Hhle and Kristiansen, 1999) is an example of an MMAR model.

### 2.8 Regime-switching Model

This model is widely applied in econometrics. The key feature is that the generating mechanism is different at different points in time and non-linear when the model changes, it is said to switch between regimes. The time points at which the regime changes may be known in advance or the regimes may change according to Markov process.

A simple illustration is provided by the following model for sales at time $t$ in terms of demand and supply, namely:

\[
\text{Sales}_t = \text{Minimum (demand}_t, \text{ supply}_t) \]

Here the two regimes are (i) demand exceeds supply; and (ii) supply exceeds demand. While the above equation is deterministic, as well as non-linear, separate equations are needed to model demand and supply and these are likely to be stochastic. In fact, although it is often convenient to treat regime-switching models as separate class models, they can be written as special cases of the general threshold model (Tong, 1990) using indicator variables.

### 2.9 Models for Changing Variance

These are non-linear models which do not lead to better point forecasts of the measured variable, but may lead to better estimates of the (local) variance. This, in turn, allows more reliable prediction intervals to be computed and hence better assessment of risk. This can be especially important when modeling financial time series, where there is often clear evidence of changing variance in the time plot of the data. It is customary to talk about the volatility of the series, rather than the variance, and to look in particular, for clusters of observations with high volatility.

### 2.10 Neural Networks

A completely different type of non-linear model is provided by neural networks, where the structure is thought to mimic the design of the human brain in some sense. Neural networks have been applied successfully to a wide variety of scientific problems, and increasingly to statistical applications, notably pattern recognition (Ripley, 1996). The topic is expanding research area.

A neural network can be thought of as a system connecting a set of inputs to a set of outputs in a possibly non-linear way. In a time series context, the ‘output’ could be value of a time series to be forecasted and the ‘inputs’ could be lagged values of the series and of other explanatory variables. The connections between inputs and outputs are typically made via one or more hidden layers of neuron or nodes. The structure of neural network is usually called architecture.

### 3. Outliers in Time Series Analysis

The concept of outliers in a data set is considered to be as old as the subject of statistics. To quote Peirce (1852) ‘In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions, and the introduction of which into the investigations can only serve … to perplex and mislead the inquirer.’ Barnett and Lewis (1994) pointed out that even before the development of formal statistical method, argument raged over whether, and on what basis, we should discard observations from a set of data on the grounds that they are ‘unrepresentative’, ‘spurious’, or ‘mavericks’ or ‘rogues’. But it is now evident that outliers do not inevitably ‘perplex’ or ‘mislead’; they are not necessarily ‘bad’ or ‘erroneous’. According to Barnett and Lewis (1994) ‘We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.’ Outliers may or may not be contaminants; contaminants may not be outliers. Extreme values may or may not be outliers. They may arise for different reasons, namely due to gross errors and true outliers. Gross errors are faulty observations, such as measurement, recording and typing errors.
Identifying these is the least controversial aspect of outlier detection, since gross errors should naturally be identified and corrected whenever possible. The true outliers are the values where a lot of careful measures to be taken when handling them.

3.1 Types of Outliers

Extensive study on detecting outliers in time series has been carried out by many researchers (see Burman, 1989; Chen and Liu, 1993; Choy, 2001; Struzik and Siebes, 2002). In automatic outlier detection and correction procedures, four main types are traditionally considered (Fox, 1972; Tsay, 1986).

3.1.1 Additive Outlier (AO)

The simplest and the most usually studied outlier in time series analysis is the additive outlier which is also known as Type I outlier (Fox, 1972). An AO only affects a single observation, which is either smaller or larger in value compared to the expected values in the data. After this disturbance, the series returns to its normal path as if nothing has happened. The effect of an additive outlier is independent of the ARIMA model and is bounded. Additive outliers are the most troublesome outliers because it inflates two consecutive residuals, one before the AO and the other after the AO.

An additive outlier corresponds to an external error or exogenous change of the observed value of the time series at a particular time point. In standard outlier literature (Tsay, 1986; Chen and Liu, 1993) the effect caused by AO at a time \( t = T \), with the magnitude of the effect is denoted by \( \omega \) is given by

\[
Z_t = X_t + \omega I_{t,(r)} = \frac{\theta(B)}{\phi(B)} a_t + \omega I_{t,(r)}
\]

where \( I_{t,(r)} = 1, t = T \) and \( I_{t,(r)} = 0, t \neq T \). The true rule suggests that the shock caused by an AO affects the original observation at \( t = T \) only with the magnitude of \( \omega \) and the rest remained unaffected. An additive outlier can have serious effects on the properties of the observed. It will affect the estimated residuals and also the estimates of the parameters. It can be proved in general that a large additive outlier will push all the autocorrelation coefficients towards zero. The effect of the outlier decreases for a large sample size.

3.1.2 Innovational Outliers (IO)

In contrast to the AO, innovational outlier also known as Type II outlier (Fox, 1972) affects several observations. An AO only affects only one residual, at the date of the outlier. The effect of the IO on an observed series consists of an initial shock that propagates in the subsequent observations with the weights of the moving average (MA) representation of the ARIMA model. So far as these weights are often explosive, the influence of the IO may, in some cases, increase as the date of its occurrence becomes more and more distant in the past, a rather undesirable future.

The effect of the IO depends on the particular model of the series, and for series with stationary transformation given by, for example \( \delta(B) = \nabla \nabla_{12} \) or \( \delta(B) = \nabla^2 \), the effect will be unbounded. For series that contain trend and seasonality, an IO will affect both. IO’s present some serious drawbacks and in our view, should be avoided. The problem then, is that the remaining three types of outliers cannot explain any changes in the seasonal component. This is an important limitation because in our experience, seasonal breaks are often present in actual economic time series.

Let \( Z_t \) be the observed series and \( X_t \) be an outlier free series. Assume \( \{X_t\} \) follows ARMA \((p, q)\) model

\[
\phi(B) X_t = \theta(B) a_t
\]

where \( \phi(B) = 1 - \theta_1 B - \ldots - \theta_p B^p \) and \( \theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q \) are stationary and invertible operators. The innovational outlier is defined as

\[
Z_t = X_t + \frac{\theta(B)}{\phi(B)} \omega I_{t,(r)} = \frac{\theta(B)}{\phi(B)}(a_t + \omega I_{t,(r)})
\]

An additive outlier affects only the \( T^{th} \) observation, that is, \( Z_T \), whereas an innovational outlier affects all observations \( Z_{T+1}, \ldots \), beyond time \( T \) through the memory of the system described by \( \frac{\theta(B)}{\phi(B)} \). Generally a time series containing \( k \) outliers of different types, general model is defined as

\[
Z_t = \sum_{j=1}^{k} \omega_j V_j(B) I_{t,(r)} + X_t
\]

where \( X_t = \frac{\theta(B)}{\phi(B)} a_t, V_j(B) = 1 \), for AO and \( V_j(B) = \frac{\theta(B)}{\phi(B)} \) for IO at \( t = T_j \).

3.1.3 Level Shifts Outliers (LSO)

Level shift is a step function. Level shifts are permanent. If one level shifts occurs at time \( t_0 \), the
observed series is $Z_t^* = Z_t + c$ for all $t \geq t_0$ with $c$ being a constant. Note that if the series $y^*$ has a level shift at $t_0$, the difference $Z_t^* - Z_{t-1}^*$ has an additive outlier at $t_0$. Least square estimates of the ARMA parameters will be inconsistent when a series contains level shift (Chen and Tiao, 1990). For a stationary process, a level shift implies a change in the mean of the process after a point and consequently the process is transformed into a non stationary one. The effect of an LSO on the residuals and on the parameter estimates can be strong.

3.1.4 Transitory Change Outliers (TCO)

Transitory change outlier (TCO) is a spike that takes a few periods to disappear exponentially. The effect of TCO is independent of the ARIMA model. The effect of these outliers is bounded. TCO is a generalization of AO and LSO in the sense that it causes an initial impact like an AO but the effect is also passed on to the following observations. The impact of a TCO is not a permanent however it decays exponentially.

Kaiser and Maravall (2001) argued that this type of classification has serious shortcomings. They mentioned the existence of seasonal outliers which may cause serious interpretative problems in the classification of outliers in time series data.

3.2 The Effects of Outliers

A large body of literature is now available to study the impact of outliers in the estimation of parameters of ARIMA (Denby and Martin, 1979; Martin, 1980; Peña, 1986, 1987, 1990; Peña and Maravall, 1991; Peña et al., 2001; Bianco et al., 2001, McQuarrie and Tsai, 2003). Outliers can seriously distort the autocorrelation structure of the time series (see Chernick et al.;1982). If the outliers are ignored and left in the time series they may seriously bias the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the series (Mills and Prasad, 1992). The main aims of time series modeling are to capture the essential features of the observed irregularity and to increase our understanding of the generating processes, or dynamics of the observed time series. The biases depend upon the number of outliers, magnitude and relative position of the outliers. These can lead to either under estimation or over estimation of the time series parameters. Series of short to moderate length, often in the presence of a single outlier, will result in a true AR model being falsely identified as an MA or an ARMA model, and the identified lag lengths ($p$ and $q$) may also be wrong (Deutsch et al., 1990). Least squares estimates of the ARMA will be inconsistent when a series contains a level shift (Chen and Tiao, 1990). Guttman and Tiao (1978), Miller (1980), Chang et al. (1988), Chan (1995) studied the effects of outliers in the autocorrelation of the series.

Outliers may influence forecasting in two different ways. Firstly, the optimal predictor for an ARIMA model depends on its parameters. Therefore the bias in the parameter estimates produced by outliers will decrease its efficiency. Secondly, the optimal predictor is a linear combination of the observed data, and generally, the largest coefficients are those corresponding to observations near to the forecast origin. Therefore the presence of outliers among these observations may have a large impact on forecasts. However, the prediction error will depend on the type of outliers: an AO may increase the predictor error considerably, but an IO will not have any effect. Level shifts and transitory changes have more serious effects on point forecasts even when outliers are not close to the forecast origin, and they increase the width of the prediction intervals as well (see Trivez, 1993). Balke (1993) present an interesting view and a simulation based solution to the construction of forecasting intervals in the presence of outliers. They have found that the presence of level changes may lead to misidentification and loss of test power. Outliers can help in fitting imperfect theories to complex real phenomena. Outliers can be also used to reveal where theory does not work, or to check what aspects of it need refining in order to describe the real world in a better way. Doula and Imon (2007) mentioned the existence of a set of observations, called ARCH-influential, whose presence or absence may cause a huge interpretative problem in understanding the ARCH effect of the data. They presented several examples to demonstrate the fact that outliers are the prime source of ARCH-influential observations in time series data.

In addition to all these difficulties, outlier analysis has two special problems to check, namely smearing and masking. These concepts are related to the detection of outliers in statistical data and can even be intertwined to complicate the situation even further. Smearing (popularly known as swamping in the literature of outlier identification in statistical data (Barnett and Lewis, 1994)) is talked of when one outlier affects the series in a manner that makes the other observations appear to be outliers as well, when they are actually not. Conversely, masking means that one outlier hides the others from being identified. It is generally believed that these notions are closely connected to specific outlier detection methods and not properties of data itself and smearing and masking are only deficiencies of certain methods, not types of outliers as such.
3.3 Estimation of the Outlier Effect

To motivate the procedure for detecting AO and IO, we consider simpler case when $T$ and all parameters are known.

Let

$$
\pi(B) = \frac{\phi(B)}{\theta(B)} = (1 - \pi_1 B - \pi_2 B^2 - \ldots) ,
$$

$$
e_t = \pi(B)Z_t 
$$

(11)

For AO: $e_t = \omega \pi(B)T_t^{(T)} + \alpha_t$  

(12)

IO : $e_t = \omega T_t^{(T)} + \alpha_t$  

(13)

For $n$ available observations, the AO model can be written as

$$
\begin{pmatrix}
e_1 \\
e_2 \\
\vdots \\
e_{T-1} \\
e_T \\
\vdots \\
e_{n+1}
\end{pmatrix} = \begin{pmatrix}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 1 \\
1 & \cdots & \cdots & \cdots \\
\vdots & \cdots & \cdots & \cdots \\
-\pi_1 & -\pi_2 & \cdots & -\pi_{n-T}
\end{pmatrix}
\begin{pmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_{T-1} \\
\alpha_T \\
\vdots \\
\alpha_n
\end{pmatrix}
$$

Let $\sigma_{aT}$ be the least square estimation of the $\omega$ for AO model. As it is assumed that $\{\alpha_t\}$ is white noise, from the least squares theory, for AO we obtain

$$
\sigma^2 = \frac{e_T^2 - \sum_{j=1}^{n-T} \pi_j e_{T+j}}{\sum_{j=0}^{n-T} \pi_j^2} = \frac{\pi'(F)e_e}{\tau^2}
$$

(14)

where

$$
\pi'(F) = 1 - \pi_1 F - \ldots - \pi_{n-T} F^{n-T}, F
$$

is the forward shift operator such that $F e_T = e_{T+1}$ and

$$
\tau^2 = \sum_{j=0}^{n-T} \pi_j^2
$$

and the variance estimator is

$$
\text{Var}(\sigma_{aT}) = \text{Var}\left(\frac{\pi'(F)e_e}{\pi'^2}\right) = \frac{1}{\pi'^2} \text{Var}(\pi'F\alpha_T) = \sigma_a^2
$$

(15)

However, for IO we obtain

$$
\sigma_{IT}^2 = e_T^2
$$

and

$$
\text{Var}(\sigma_{IT}) = \text{Var}(e_T) = \text{Var}(\omega T_t^{(T)} + \alpha_t) = \sigma_a^2
$$

(17)

Thus the best estimate of the effect of an AO at time $T$ and else the best estimate of the IO is a linear combination of $e_T, e_{T+1}, \ldots$ and $e_n$ with weight depending on the structure of the time series process. It is easy to show that

$$
\text{Var}(\sigma_{aT}) \leq \text{Var}(\sigma_{IT}) = \sigma_a^2.
$$

The above results help us to formulate various tests for hypothesis:

$$
H_0 : Z_T \text{ is neither AO or IO.}
$$

$$
H_1 : Z_T \text{ is an AO.}
$$

$$
H_2 : Z_T \text{ is an IO.}
$$

The likelihood ratio test statistics for AO and IO:

$$
H_1 \text{ Vs } H_0 : \lambda_{1,T} = \frac{\pi' \sigma_{aT}}{\sigma_a}
$$

$$
H_2 \text{ Vs } H_0 : \lambda_{2,T} = \frac{\sigma_{IT}}{\sigma_a}
$$

Under the null hypothesis $H_0$, both $\lambda_{1,T}$ and $\lambda_{2,T}$ are distributed as $N(0, 1)$.

4. Detection of Outliers in Non-linear Time Series

An early and most detailed examination of detection of outliers in stationary time series was done by Fox (1972). He proposed a method using the likelihood ratio test to detect additive and innovational outliers. Chernick et al. (1982) suggested detecting outliers by examining the influence matrix of the estimated autocorrelations. They also proposed a visual procedure for detecting an outlier, Tiao (1985) describes methods for outlier detection in ARMA models. An interesting graphical approach is described by Embrechts et al. (1986) to detect periodic components and outliers in time series. Muirhead (1986) presents a test of discordancy and develops a rule to distinguish whether the outlier is of AO or IO type that was not possible with the test suggested by Fox (1972). A multiple outlier detection method in time series is discussed by Schmid (1986). Chang and Tiao (1983) and Chang, Tiao and Chen (1988) extended the results of Fox (1972) to the ARIMA models and proposed likelihood ratio test and an iterative procedure for detecting outliers and estimating the model parameters. Abraham and Yatawara (1988) describe score test for detecting outliers in ARMA models. Tsay (1988) generalized the Chang-Tiao-Chen (1988) procedure to include the detection of level shifts and temporary changes. Random level shifts were studied.
by Chen and Tiao (1990). Chen and Liu (1993) presented an outlier detection and parameter estimation procedure for ARIMA models that seems to be widely used. Abraham and Chuang (1989) considered deletion statistics based on the influence measures in regression for outlier identification. Peña (1986, 1987, 1990) proposed a missing value approach to study the influence of outliers in time series. Bruce and Martin (1989) studied the identification of the outlier patches in the ARIMA models using ideas of influential observations in the time series. Abraham and Chuang (1993) applied the EM algorithm to the estimation procedure in the presence of outliers. Ljung (1982, 1989, 1993) studied the likelihood function of the ARMA models with the missing data and its relation to outlier analysis. Most outlier detection procedures have been developed under the assumption that the underlying outlier-free model is known or can be identified, which is not possible in practice. To overcome this problem, Lee and Wei (1995) proposed a model-independent outlier detection procedure. Abraham and Box (1975) take a Bayesian look at the problem of outliers in time series which is studied further by many authors.

As we have already mentioned that not much attention has been focused on the issues of the identification of outliers in non-linear time series data until quite recently. A very popular approach is, however, to use the robustified estimation techniques of linear models or linear time series models to identify outliers in non-linear time series data (see Rousseuw and Leroy, 1987; Barnett and Lewis, 1994; Gabr, 1998; van Dijk, D et al., 1999; Franses and van Dijk, 2000; Peña et al., 2001; Battaglia and Orfei, 2002; Tsay, 2002; Charles and Darne, 2005; Doornik and Ooms, 2005).

4.1 A General Procedure

It is believed that the iterative framework of Chen and Liu (1993) is quite effective in the nonlinear case. However, the choice of suitable time series model is important when searching for outliers, because a large residual variance caused by overall lack of fit would result in under identification of outliers, while a model unable to explain the local behaviour of the series would yield single large residuals resulting in over-identification of outliers. If the observed phenomenon is described by a stationary zero-mean process \( \{X_t\} \) following the model:

\[
X_t = f(x^{t-1}; \varepsilon^{t-1}) + \varepsilon_t
\]

where \( f \) is a nonlinear function also containing unknown parameters, \( x^{t-1} = (x_{t-1}, x_{t-2}, \ldots, x_{t-p})' \); \( \varepsilon^{t-1} = (\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-p})' \) and \( \{\varepsilon_t\} \) is a zero-mean Gaussian white-noise series with \( E(\varepsilon_t^2) = \sigma^2 \). The observed data \( (Z_1, Z_2, \ldots, Z_n) \) are realizations of a perturbed process with an outlier at a time \( q \), \( (1 < q < n) \) defined as follows:

**AO**: \( Z_t = X_t + \omega_q \delta_{t,q} \) where \( \delta_{t,q} = 1 \) if \( t = q \) and \( \delta_{t,q} = 0 \) if \( t \neq q \).

**IO**: \( Z_t = f(Z^{(t-1)}; \eta^{(t-1)}) + \eta_t \) where \( \eta^{(t-1)} = (\eta_{t-1}, \eta_{t-2}, \ldots, \eta_{t-p})' \) and \( Z^{(t-1)} = (Z_{t-1}, Z_{t-2}, \ldots, Z_{t-p})' \). The quantity \( \omega_q \) is assumed constant and unknown, and we shall refer to it as the amplitude of the outlier.

A strategy analogous to Chen and Liu (1993) for the detection of outliers in the nonlinear time series is as follows:

Step 1: Derive the estimates of the model parameters.

Step 2: Given the parameter values, for any \( q \) and for each type of outliers, assume that an outlier has occurred at time \( q \) and estimate its amplitude. If the largest absolute estimated is significant (i.e. larger than a priori fixed sensitivity level, usually 3.5 or 4 times its estimated standard error), identify an outlier of that type at that time; otherwise stop.

Step 3: Remove the effect of the identified outlier by subtracting its estimated amplitude from \( Z_q \) (and also correcting all subsequent observations according to the estimated model in the case of innovational outlier).

Step 4: Estimate again the model parameters on the corrected series, and iterate step 2.

In order to estimate \( \omega_q \), we adopt a conditional maximum likelihood approach, and assume that, given the model and parameter values, the likelihood function of the data is proportional to the likelihood of the residuals. The quantity \( \omega_q \) is estimated by minimizing the sum of squares

\[
\sum \varepsilon_t^2 \quad \text{(see Battaglia and Orfei; 2002).}
\]

However there is a strong reason to believe that masking effects can be even more serious in a method like this.

4.2 Detection of Outliers in Bilinear Time Series
Gabr (1998) investigated a modification of some robustified versions of methods used in linear time series models to detect outliers in bilinear models. The problem of detecting additive outliers in bilinear time series is considered by Chen (1997) using Gibbs sampler. Undoubtedly, the bilinear time series models are more complex than autoregressive models and they lead to several problems which may be difficult to resolve. Gibbs sampling approach is used to estimate the probability of the occurrence of the outliers at each time point. The Gibbs sampler is a Markovian updating scheme enabling one to obtain samples from a joint distribution via iterated sampling from full conditional distributions. Detailed investigation of the Gibbs sampler applied to general Bayesian calculation is given by Gelfand and Smith (1990).

To review the method, let us consider the case of three parameters \( (\theta_1, \theta_2, \theta_3) \) and suppose the conditional distribution of each, given the remainder, has a simple form while the joint distribution is more complicated. The conditional distribution is denoted by \( f_1(\theta_1|\theta_2, \theta_3, Y), f_2(\theta_2|\theta_3, \theta_1, Y), f_3(\theta_3|\theta_1, \theta_2, Y) \), where \( Y = (y_1, y_2, ..., y_n) \). The Gibbs sampler employed in this procedure works as follows: Given arbitrary values of \( \theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)} \), draw \( \theta_1^{(1)} \) from \( f_1(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, Y) \), then draw \( \theta_2^{(1)} \) from \( f_2(\theta_2|\theta_3^{(0)}, \theta_1^{(1)}, Y) \), and finally, the first iteration is completed by drawing \( \theta_3^{(1)} \) from \( f_3(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, Y) \). After large number, M, of iterations \( \theta_1^{(m)}, \theta_2^{(m)}, \theta_3^{(m)} \) is obtained. The desired posterior marginals can be approximated by the empirical distribution of the \( N \) values \( \theta_i^{(l)}, \theta_{2(l)}, \theta_{3(l)} \), \( i = M + 1, ..., M + N \), where \( M \) is large enough to give sufficient precision to the empirical distribution of interest. However, the sample could be highly dependent, each realization being generated from the previous one. To monitor and to reduce this dependence, the sample autocorrelation of this chain is computed and a lag \( l \) is selected so that the sample autocorrelations at lag \( i \geq l \) are very small. Then \( \theta_1^{(l)}, \theta_2^{(l)}, \theta_3^{(l)} \), \( i = M + 1, ..., M + NL \), as random sample. A major advantage of using Gibbs is that there is no need to assume beforehand the number of outliers in the series and it can identify those observations which require further scrutinizing.

Some other useful methods for the detection of outliers in bilinear models are discussed by Franses and van Dijk (2000) and Tsay (2002).

### 4.3 Detection of Outliers in ARCH and GARCH Models

Outlier detection in ARCH and GARCH models was studied by van Dijk et al. (1999), Franses and van Dijk (2000), Charles and Darne (2004), Doornik and Ooms (2005) with particular attention to robust testing. Doula and Imon (2007) suggested using robust model fitting techniques such as the least median of squares (LMS) proposed by Rousseeuw (1984) or least trimmed squares (LTS) proposed by Rousseeuw and Leroy (1987) for the identification of multiple outliers in ARCH model. Charles and Darne (2004) proposed an outlier detection method for GARCH (1,1) consists of the following steps:

1. Estimate the GARCH (1,1) model for the observed series \( e_t \) and obtain the conditional variance \( \hat{h}_t \) and \( \hat{h} = e_t^2 - \hat{h}_t \).

2. Estimate \( \hat{e}_t (\tau) \), \( i = 1, 2 \) for all possible \( \tau = 1, 2, ..., n \) and compute \( \hat{\tau}_{\text{max}} \). The value of test statistic exceeds the prespecified critical value \( C \), an outlier is detected.

3. Replace \( e_t^2 \) with \( e_t^* = e_t^2 - \hat{\omega}_t \) for an AO, and \( e_{t+1}^* \) with \( e_{t+1}^* = e_{t+1}^2 - \hat{\omega}_2 \psi_j \) with \( j > 0 \) for IO.

4. Return to step (1) to estimate a GARCH (1,1) model for the series \( e_t^* \) and all the steps are repeated until no \( \hat{\tau}_{\text{max}} \) test-statistic exceed the critical value \( C \).

Doornik and Ooms (2005) present a new procedure for detecting additive outliers in GARCH(1,1) models. The outlier candidate is the observation with the largest standardized residual. First, a likelihood-ratio based tests determines the presence of an outlier. Next, a second LR test determines the type of outlier (volatility or level). This procedure is shown to be approximately independent from the GARCH parameters, with a null distribution that can be easily approximated.

For threshold models, Chan and Cheng (1994) modified the class of generalized M-estimates usually applied to linear models.
5. Discussion

Outliers do exist in linear and non-linear time series data. The increasing interest in detecting outliers in linear and non-linear time series is becoming vital in statistical analysis. It is difficult to say how frequent outliers occur, which is the best method to detect them and how serious their threat to the time series analysis.

Outliers are more critical in non-linear time series than in linear time series, because, their effects may be much longer persistent and they may have serious impact in parameter estimation. When outliers are very large, the detection method based on linear ARMA models or linear interpolators are relatively successful, but they are less useful when the anomalies are not so evident. It would be interesting, although difficult, to derive optimal nonlinear interpolators based on different nonlinear models and study their effectiveness in outlier identification.

The detected outliers are clearly important in the analysis of time series data, and they are responsible for a considerable amount of apparent nonlinearity in the series. The removal of the outliers does not remove all evidence of ARCH and bilinearity from the series. It seems therefore that the series should perhaps be modeled with an MA model with the outlier dummies and ARCH(1) errors.

It has been suggested that outliers can account for some detected nonlinearities in time series (see Balke and Fomby, 1997). Some kinds of nonlinearities are also closely related to ARCH, so it may be difficult to clearly identify which of these three (outliers, ARCH or nonlinearities) is responsible for causing deviations from linearity. There is also a possibility that even if all these three are present they may interact with one another. Different aspects of the interactions between ARCH and bilinear models are discussed by Weiss (1984, 1986), Higgins and Bera (1988) and others.

Detection of outliers at an early stage is important for any model since they can affect all stages of time series analysis: model identification, estimation and forecasting. Therefore it is essential to apply some modeling device to encounter the outlier effect. These can be robust methods, outlier modeling methods and/or outlier deletion of some of the observations. All these will be obviously applicable in some situations but none of them will probably applicable in every situation. So far there is not a fixed way to handle outliers and therefore caution has to be taken whenever handling outliers especially in nonlinear time series.

Reference


An Alternative Method for Selecting the Number of Components for Smooth Lack of Fit Tests

PPB Eggermont and VN LaRiccia
University of Delaware

Abstract
A common approach to testing the adequacy of a linear model in regression is as follows: A larger linear model, which contains the hypothesized model, is considered, and then using classical techniques one tests whether the betas associated with the added terms are zero. If the added terms are properly selected, then the power of such tests is determined by the number, M, of terms, or components, included in the larger model. For example, for such test to be consistent for a large class of alternatives one must have that $M \to \infty$ as $n \to \infty$. Common methods for selecting $M$ include differing deterministic choices, or one of the data driven procedures such as Aikie’s Information Criteria, Mallow’s Criteria or Schwartz’s Bayesian Criteria. All of these choices result in testing procedures which have certain good asymptotic properties. Here we compare some of the resulting tests for moderate sample sizes by simulation experiments.

The results of these experiments, from a power perspective, are somewhat disappointing. However, they lead to a new method for selecting not only $M$ but the specific components. The corresponding test procedure, in some sense, blends the good properties of the other procedures.

1 Introduction
In linear regression problems many times one wants to test the adequacy of an assumed parametric mean function. That is, one believes that the mean function is given by

$$H_0 : m(x) = \sum_{j=0}^{p} \beta_j h_j(x), \quad (1.1a)$$

where $p$ is a known integer, the $h_j$’s are known functions and the $\beta_j$’s are unknown. Further one wants to test this belief against the very general alternative hypothesis

$$H_A : m(x) \neq \sum_{j=0}^{p} \beta_j h_j(x), \text{ for some } x \text{ and for all } \beta. \quad (1.1b)$$

Typically this problem goes by the name of the Lack of Fit (LOF) problem. There exists an extensive literature on the LOF problem. One reason for this is that the alternative hypothesis is quite general, and one can always find a set of alternatives for which any given test has very low power. Here we are interested in component (or Neyman smooth) type tests. For a nice general discussion of such procedures one is referred to Hart(1997).

Here we shall restrict attention to the case where under both the null and alternative hypotheses

$$m \in L^2, \quad (1.2)$$

with

$$L^2 = \left\{ m : m \text{ and } m^1 \text{ are absolutely continuous and } \int |m^2(x)|^2 \, dx < \infty \right\},$$

where $m^1$ denotes the $i$th derivative of $m$. Further, it will be assumed that in the null hypothesis
the $h_j$’s are elements of a set of basis functions for $L^2$. (1.3)

In Section 4, the new test procedure for (1.1) under (1.2) and (1.3) is given. However, for clarity we begin the discussion by considering the following simple special case, studied in Eubank (2000). Specifically for a given sample size $n$, let $(T_1, Y_1), \ldots, (T_n, Y_n)$, denote the observations, with

$$T_i = (2i - 1)/2n.$$  

(i.e. a deterministic equally spaced design),

$$Y_i = m(T_i) + \varepsilon_i,$$  

with $m$ a smooth unknown function (i.e. $m \in L^2[0, 1]$), and the unobserved error terms, $\varepsilon_i$, independent identically distributed random variables, with

$$\varepsilon_i \sim N(0, 1).$$

Now the goal is to test the no-effect null hypothesis

$$H_0 : m(t) = \beta_0,$$  

versus the omnibus alternative

$$H_A : m(t) \neq \beta_0.$$

There exists quite an extensive literature on tests for the no-effect null. See Eubank (2000), and Hart (1997) for reviews of the literature. Note that many of the proposed tests for this problem start by considering a complete orthonormal system, $\text{CON}$, $\{\phi_j\}_{j=0}^\infty$, for $L^2[0, 1]$ with $\phi_0(x) = 1$. Remember for $m \in L^2[0, 1]$ we have that for almost all $t \in [0, 1]$ that

$$m(t) = \sum_{j=0}^{\infty} \beta_j \phi_j(t).$$

Given the CON the test statistic is then defined by

$$T(w) = \sum_{j=1}^{n} w_j \hat{\beta}_j^2,$$

where the $w_j$’s are either random or deterministic weights, and for all values of $j$,

$$\hat{\beta}_j = n^{-1} \sum_{i=1}^{n} \phi_j(T_i)Y_i,$$

is the estimator of $\beta_j$. Here we shall consider tests of this nature with

$$\phi_j(x) = \sqrt{2} \cos(j\pi x).$$

As stated before, this will be generalized in Section 4. Note that for this design and the cosine functions for any value of $p$, in the model

$$Y_i = \sum_{j=0}^{p} \beta_j \phi_j(x) + \varepsilon_i$$

written in matrix notation we have that

$$(X'X) = nI. \quad (1.5)$$

Hence the $\hat{\beta}_j$’s are the typical least squares estimators. Further the corresponding test statistic for $H_0 : \beta_1 = \beta_2 \ldots = \beta_p = 0$ is given by

$$T_p = \sum_{j=1}^{p} \hat{\beta}_j^2.$$  

(1.6)
Note that $T_p$ is in fact the basis of all Neyman type smooth tests. The difference between them being in how $p$ is selected.

In Section 2, the following 3 versions of Neyman’s smooth tests are considered. First define

$$T_M = n \sum_{j=1}^{M_n} \beta_j^2,$$

with $M_n$ taken as the solution to

$$\max_M \{n \sum_{j=1}^{M} \beta_j^2 - 2M\},$$

where the maximization is done over

$$M \in \{0, 1, \ldots n\},$$

with $\sum_{j=1}^{M} \beta_j^2$ taken to equal 0. (Note that $M_n$ is nothing more than the solution to Mallows criteria for selecting the order of a regression fit to the data.). The second statistic is given by

$$T_S = n \sum_{j=1}^{M_n} \beta_j^2,$$

where here $M_n$ is the solution to

$$\max_M \{n \sum_{j=1}^{M} \beta_j^2 - \log(n)M\},$$

where the maximization is now done over

$$M \in \{1, 2, \ldots n\},$$

note that if zero was also considered in the minimization, then $M_n$ is the solution to the Schwartz criteria for selecting the order of a regression fit to the data. Also for comparison purposes we also considered the test statistic,

$$T_D = n \sum_{j=1}^{M_n} \beta_j^2,$$

where $M_n$ is chosen in a deterministic fashion, as the closest integer to

$$M_n = \sqrt{n}.$$

Note that this is not the typical value taken. For classical smooth type tests $M_n$ is not selected this large. See for example Raynor and Best(1989). This choice was made to see how a test based on a large deterministic choice would compare with the tests based on common data driven procedures. Finally we also considered the Cramer von Mises type test statistic

$$T_C = n \sum_{j=1}^{n} \left(\frac{\beta_j^2}{(j\pi)^2}\right),$$

which is an approximation to a test proposed by Cox and Cohen(1989), and Buckley(1991).

It is well known that in this simple setting, there is no (even asymptotically) uniformly efficient lack of fit tests. For more on this, see the comments in Janssen(2000 and 2003). The original goal of this paper was thus to determine in which directions (i.e. types of alternatives) $T_M$ and $T_S$ are most powerful. A major problem in such considerations is that there exists at least three different asymptotic approaches to compare different LOF tests, and they lead to contradictory answers. See for example Eubank(2000), Inglol and Ledwina(2003) and Janssen(2003). In Section 2 we compare the four tests using Monte Carlo techniques. While the simulation study is limited, the results are
Quite revealing. The main conclusion being that if one desires an omnibus test, then $T_M$ is the best choice, with $T_D$ coming in a close second.

However, the price paid by the test procedure based on $T_M$ for being an omnibus test, is that it has relatively low power for many simple and realistic smooth alternatives. In section 3 a new smooth type test which compensates some what for this is developed. It is based on the fact that the three methods for selecting $M_n$ for the Neyman smooth tests come from results for non-parametric series regression estimators. The new test is motivated by the best properties of each of the three methods for estimating the mean function. Finally in Section 4 the new test procedure is generalized to other designs, and alternative CONS.

2 Power Comparison of the Four Tests.

In this section selected results from a simulation study comparing the four tests given in Section 1 are presented. The first part of the comparison is carried out in what is close to a principal components type analysis. That is for a given sample size, $n$, and constant $\beta$, the power of the four tests are computed for the single component models

$$Y_i = \beta_0 + \beta \phi_j(T_i) + \varepsilon_i,$$

for $j = 1, \ldots, n$. This sort of an approach was suggested by the remarks in Janssen (2003).

(2.2) Remark. Surprisingly (at least to the authors) even though the power of the tests for more realistic alternatives, like

$$m(t) = \beta_0 + \sum_{j=1}^{\infty} \beta_j \phi_j(t)$$

is not a linear function of the $\beta_j$'s, as you shall see, a table of the power of the tests for the different single component models does a good job of predicting the relative power of the four tests for such alternatives.

The specifics of the basic case considered are that $n = 100$ and $\beta = 0.80$. The reason for selecting this value of $\beta$ is that for the second component (i.e. $j = 2$ in (2.1)) it resulted in powers which ranged from 0.272 to 0.619, and it seemed to separate the powers of the four tests reasonably well. Further, as a test case this choice provides a problem which is quite challenging to the different test procedures. The reason being that the signal ($m(t)$) to noise ($\sigma$) ratio is quite small. Next for all four tests the critical values for $\alpha = 0.05$ were determined by a simulation experiment consisting of 5000 runs. Finally for each alternative model considered a simulation consisting of at least 1000 runs was performed.

The first four columns of Table 1 gives the power of the four tests for the single component models for $j \leq 20$. (Note that you should ignore column 5, for now.) Also, Beyond 20 all of the powers were more or less equal to $\alpha = 0.05$. Some of the implications of this table follow.

Conclusions for $T_C$ : As expected the test based on $T_C$ only has reasonable power for simple alternatives with $j = 1$ and 2. Note that theoretically the power of this test is greater than $\alpha$ for all values of $j$. However as Figure 1 illustrates even for moderate values of $j$ (i.e. $j = 5$) when $n = 100$, $\beta$ must be quite large before this test has reasonable power. (As one would expect the results are even worse for larger values of $j$.)

Conclusions for $T_S$ : The results for the test based on $T_S$ are more surprising. The power decreased in a monotone fashion, and relatively rapidly as $j$ increased. It was more or less equal to $\alpha$ for $j \geq 8$.

Conclusions for $T_D$ : The power of test based on $T_D$ is as theory predicts. (Remember that under our assumptions the exact distribution of this test is known) That is the power of the test is the same in the first $\sqrt{n}$ directions, and equals $\alpha$ for the remaining ones. The only surprising result here is that even with $M_n = \sqrt{n}$ (10) the power of the test, where it is not equal to $\alpha$, is very competitive.

So at least for the given problem all of the above three tests fail in one sense or another as an omnibus test. That is having reasonable power in many directions.
Conclusions for $T_M$: On the other hand the tests based on $T_M$, at least for the situation considered, comes the closest to being a true omnibus test. (That is it has a reasonable amount of power in many directions.) Note however that the power is not uniform in all directions. In fact it appears that the price one pays for the omnibus nature of this test is in its relatively lower power for the lower order single component alternatives. This is not what the authors had expected. The following remark is as close to a "deep statistical insight" for this phenomena as we can come up with.

(2.3) Remark. As we now illustrate, for the tests based on $T_M$ it is not the nature of the higher order components, but rather only their arbitrary location in the sequence, which results in their higher power for single component alternatives. In this remark we shall consider the alternative mean model,

$$m(t) = \beta_0 + \beta_3 \phi_3(t),$$

with $|\beta_3|$ reasonably larger than zero.

Also for this remark let $M_1$ denote the value of $M$ selected by Mallows procedure and let $T_1$ denote the corresponding test statistic (i.e. $T_M$). Next consider the following re-ordered sequence, $\{\psi_i\}_{i=1}^{\infty}$, of cosine components. Let

Table 1: Power of the Five Tests for Single Component Alternatives

<table>
<thead>
<tr>
<th>Component</th>
<th>$T_C$</th>
<th>$T_B$</th>
<th>$T_D$</th>
<th>$T_M$</th>
<th>$T_W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.705</td>
<td>0.582</td>
<td>0.352</td>
<td>0.223</td>
<td>0.437</td>
</tr>
<tr>
<td>2</td>
<td>0.223</td>
<td>0.619</td>
<td>0.352</td>
<td>0.272</td>
<td>0.458</td>
</tr>
<tr>
<td>3</td>
<td>0.091</td>
<td>0.361</td>
<td>0.352</td>
<td>0.277</td>
<td>0.429</td>
</tr>
<tr>
<td>4</td>
<td>0.072</td>
<td>0.248</td>
<td>0.352</td>
<td>0.335</td>
<td>0.457</td>
</tr>
<tr>
<td>5</td>
<td>0.052</td>
<td>0.123</td>
<td>0.352</td>
<td>0.348</td>
<td>0.415</td>
</tr>
<tr>
<td>6</td>
<td>0.065</td>
<td>0.081</td>
<td>0.352</td>
<td>0.377</td>
<td>0.386</td>
</tr>
<tr>
<td>7</td>
<td>0.058</td>
<td>0.066</td>
<td>0.352</td>
<td>0.398</td>
<td>0.363</td>
</tr>
<tr>
<td>8</td>
<td>0.056</td>
<td>0.057</td>
<td>0.352</td>
<td>0.348</td>
<td>0.313</td>
</tr>
<tr>
<td>9</td>
<td>0.056</td>
<td>0.056</td>
<td>0.352</td>
<td>0.346</td>
<td>0.314</td>
</tr>
<tr>
<td>10</td>
<td>0.055</td>
<td>0.056</td>
<td>0.352</td>
<td>0.247</td>
<td>0.224</td>
</tr>
<tr>
<td>11</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.245</td>
<td>0.050</td>
</tr>
<tr>
<td>12</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.188</td>
<td>0.050</td>
</tr>
<tr>
<td>13</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.193</td>
<td>0.050</td>
</tr>
<tr>
<td>14</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.164</td>
<td>0.050</td>
</tr>
<tr>
<td>15</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.126</td>
<td>0.050</td>
</tr>
<tr>
<td>16</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.129</td>
<td>0.050</td>
</tr>
<tr>
<td>17</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.109</td>
<td>0.050</td>
</tr>
<tr>
<td>18</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.111</td>
<td>0.050</td>
</tr>
<tr>
<td>19</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.100</td>
<td>0.050</td>
</tr>
<tr>
<td>20</td>
<td>0.055</td>
<td>0.050</td>
<td>0.050</td>
<td>0.066</td>
<td>0.050</td>
</tr>
</tbody>
</table>
Table 2: Power of the Five Tests for Mean Model (2.4)

<table>
<thead>
<tr>
<th>$K$</th>
<th>$T_C$</th>
<th>$T_S$</th>
<th>$T_D$</th>
<th>$T_M$</th>
<th>$T_W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.542</td>
<td>0.558</td>
<td>0.352</td>
<td>0.249</td>
<td>0.374</td>
</tr>
<tr>
<td>3</td>
<td>0.381</td>
<td>0.430</td>
<td>0.352</td>
<td>0.260</td>
<td>0.351</td>
</tr>
<tr>
<td>5</td>
<td>0.260</td>
<td>0.282</td>
<td>0.352</td>
<td>0.440</td>
<td>0.324</td>
</tr>
<tr>
<td>10</td>
<td>0.151</td>
<td>0.184</td>
<td>0.352</td>
<td>0.339</td>
<td>0.298</td>
</tr>
<tr>
<td>15</td>
<td>0.128</td>
<td>0.131</td>
<td>0.227</td>
<td>0.224</td>
<td>0.181</td>
</tr>
<tr>
<td>20</td>
<td>0.093</td>
<td>0.117</td>
<td>0.173</td>
<td>0.174</td>
<td>0.143</td>
</tr>
</tbody>
</table>

\[
\psi_1 = \phi_3, \\
\psi_2 = \phi_1, \\
\psi_3 = \phi_2,
\]

and for $i > 3$

\[
\psi_i = \phi_i.
\]

Also denote the Fourier coefficients for the new sequence by $\gamma_i$. For this ordering of the cosine components, let $M_2$ denote the value of $M$ selected by Mallows procedure and let $T_2$ denote the corresponding test statistic.

Now since the null distributions of $T_1$ and $T_2$ are the same, the rejection rule for the test based on either statistic is given by:

Reject $H_0$ if $T_i > C_\alpha$, for the same constant $C_\alpha$.

However unless $|\hat{\beta}_3|$ is much smaller than both $|\hat{\beta}_1|$ and $|\hat{\beta}_2|$ (which under the given alternative should not happen very often, and when it does both tests will fail to reject $H_0$), we shall have $M_2 \leq M_1$, and since for all $i$, $\beta_i \geq 0$ this in turn implies that $T_2 \leq T_1$. Hence for this alternative the test based on $T_2$ will have lower power.

Before proceeding it is important to see if the results for the single component alternatives actually reflect what happens for more complicated alternatives. (Hopefully they do, since then instead of having to simulate a large number of "reasonable" alternatives, all that is required is Table 1.) To check this out a few different types of alternatives are considered. For the first set we consider alternatives of the form

\[
m(t) = \beta_0 + \sum_{j=1}^{K} \beta_K \phi_j(t), \tag{2.4}
\]

where $\beta_K = \beta / \sqrt{K}$ for different values of $K$. The results for these alternatives are given in Table 2. Maybe we are being overly optimistic, but it seems (to us) that the relative ranking of the different tests for the alternatives in Table 2 could have been predicted by the results in Table 1.

For the last set we considered alternatives with more complicated $\beta_i$’s. Specifically the following 7 alternatives are considered.

(A.1) Linear Mean Model: Here is $m(t) = t$. For this model we have

\[
\beta_j = \begin{cases} 
-2^{3/2} \left( \frac{1}{j\pi} \right)^2 & \text{for } j \text{ odd}, \\
0 & \text{for } j \text{ even.}
\end{cases}
\]
(A.2) Quadratic Mean Model: For this case \( m(t) = 0.5t + t^2 \), and for all \( j \)

\[
\beta_j = \begin{cases} 
-\left[2^{3/2} + 2^{1/2}\right](1/j\pi)^2 & \text{for } j \text{ odd}, \\
2^{3/2}(1/j\pi)^2 & \text{for } j \text{ even}.
\end{cases}
\]

(A.3) Cubic Mean Model: Here \( m(t) = 0.5t + t^3 \), with for all \( j \)

\[
\beta_j = \begin{cases} 
\left((1/j\pi)^2\left[-4\sqrt{2}\right] + 12\sqrt{2}(1/j\pi)^4 \right) & \text{for } j \text{ odd}, \\
3\sqrt{2}(1/j\pi)^2 & \text{for } j \text{ even}.
\end{cases}
\]

(A.4) Decreasing \( \beta_j \)'s Model 1: Here letting \( z_j = (2j - 1)/2n \), we take for \( j = 1, \ldots, n \)

\[
\beta_j = 1/\left(700 \times z_j\right)
\]

(A.5) Rapidly Decreasing \( \beta_j \)'s: Again for \( j = 1, \ldots, n \), and letting \( z_j = (2j - 1)/2n \), we have

\[
\beta_j = 0.00001 \times z_j^{-2},
\]

(A.6) Increasing \( \beta_j \)'s: Taking \( z_j = (2j - 1)/2n \), we have

\[
\beta_j = 0.0003 \times z_j.
\]

(A.7) Increasing \( \beta_j \)'s, with the first three components zero: Taking \( z_j = (2j - 1)/2n \), we have

\[
\beta_1 = \beta_2 = \beta_3 = 0,
\]

and for \( j > 3 \) and \( j \leq n \)

\[
\beta_j = 0.0003 \times z_j.
\]

Figures 2 gives the graphs of \( m(t) \) for alternative models (A.4), (A.5), (A.6) and (A.7). The power results for these six alternatives are given in the first 4 columns of Table 3. Again it is our contention that the ranking of the tests is predictable from the values of the \( \beta_j \)'s and the results of Table 1.

**Remark.** The results given in Tables 2 and 3 seem to indicate that in comparing such tests all that is required is a Table similar to Table 1 for alternative models (A.1) through (A.7).

Note that the results in Table 3 for \( T_M \), the "best" omnibus test are somewhat troubling. The reason being that for many "smooth' alternatives to \( H_0 \) the Fourier coefficients, \( \beta_j \), tend to be larger.
Table 3: Power of the Five Tests for Mean Models (A.1) through (A.7)

<table>
<thead>
<tr>
<th>Model</th>
<th>TC</th>
<th>TR</th>
<th>TD</th>
<th>TM</th>
<th>TW</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A.1)</td>
<td>0.830</td>
<td>0.714</td>
<td>0.289</td>
<td>0.290</td>
<td>0.579</td>
</tr>
<tr>
<td>(A.2)</td>
<td>0.988</td>
<td>0.974</td>
<td>0.802</td>
<td>0.795</td>
<td>0.955</td>
</tr>
<tr>
<td>(A.3)</td>
<td>0.983</td>
<td>0.963</td>
<td>0.725</td>
<td>0.720</td>
<td>0.918</td>
</tr>
<tr>
<td>(A.4)</td>
<td>0.846</td>
<td>0.760</td>
<td>0.567</td>
<td>0.418</td>
<td>0.635</td>
</tr>
<tr>
<td>(A.5)</td>
<td>0.973</td>
<td>0.941</td>
<td>0.786</td>
<td>0.666</td>
<td>0.893</td>
</tr>
<tr>
<td>(A.6)</td>
<td>0.046</td>
<td>0.115</td>
<td>0.363</td>
<td>0.572</td>
<td>0.586</td>
</tr>
<tr>
<td>(A.7)</td>
<td>0.073</td>
<td>0.246</td>
<td>0.363</td>
<td>0.572</td>
<td>0.586</td>
</tr>
</tbody>
</table>

for low components and rapidly converge to zero as \( j \to \infty \). From the authors perspective we would be willing to give up some power for higher order alternatives (but not as much as is given up by the test associated with \( T_C \) or for that matter even \( T_S \)) in order to raise the power for lower order alternatives. How is this to be done? At first one might consider \( T_D \) with \( M_n = \sqrt{n} \). However we would like to do even better than this! (That is take some of the power associated with the higher order single component alternatives and move it to the lower components.) A test which appears to accomplish this is presented in the following Section.

3 A New Neyman Smooth Test.

In this section an alternative \( LOF \) test also based upon \( T_p \) given in (1.6) is developed. The basic idea is to try and capture the best properties of the three different ways of selecting \( M_n \). That is, by selecting \( M_n \) in a deterministic fashion, the power of the test is concentrated on a finite dimensional subspace. While Mallows selection criterion asymptotically selects, with very high probability, the correct value of \( M_n \) for estimating \( m(t) \) by a non-parametric series estimator. Finally given a finite dimensional model, say

\[
m(t) = \beta_0 + \sum_{j=1}^{M} \beta_j \varphi_j(t),
\]

Schwarz’s Bayesian criteria selects, with probability converging to one, the correct mean model from the collection of \( 2^M \) possible sub-models. Taking into account Remark (2.3), this last point should be useful.

The first step in defining the test statistic is to determine an upper limit on the number of components in a model like (3.1). The value, call it \( M_1 \), is taken to be the solution to the restricted Mallows problem,

\[
\max_M \left\{ n \sum_{j=1}^{M} \beta_j^2 - 2M \right\},
\]

where the maximization is now done over the set \( M \in \{0, 1, \ldots, \sqrt{n}\} \). Note in the first step we now restrict the minimization to integers \( \leq \sqrt{n} \). The next step is to use Schwarz’s criteria to select the ”best” sub-model from (3.1) with \( M \) now take to be \( M_1 \). For this problem, because of (1.5), this is easily done. Specifically define the reverse order statistics associated with \( \hat{\beta}_1^2, \ldots, \hat{\beta}_{M_1}^2 \). Call these \( \hat{\gamma}(1) \geq \hat{\gamma}(2) \geq \cdots \geq \hat{\gamma}(M_1) \).

Then determine the value \( M_2 \) which solves

\[
\hat{\gamma}(1) \geq \hat{\gamma}(2) \geq \cdots \geq \hat{\gamma}(M_1).
\]
\[ \max_M \left\{ n \sum_{j=0}^{M} \hat{\gamma}_{(j)} - \log(n)(M + 1) \right\}, \]

where \( M \in \{0, 1, 2, \ldots M_1 - 1\} \). Now using only the selected components the test statistic is then taken to be
\[ T_W = n \sum_{j=0}^{M_2} \hat{\gamma}_{(j)}. \]

To see how this test compares with the others go back to Tables 1 through 3 and consider the fifth column. The new test appears to behave as desired.

4 Tests for General LOF Problems.

In this section the test statistic, \( T_W \), is generalized to handle any collection of \( CONS \), \( \{h_i\} \), general designs, \( \{T_i\}_{i=1}^n \), and the case where \( Var(Y_i) = \sigma^2 \), with \( \sigma^2 \) unknown. Note for a fixed value of \( p \), the testing problem is given by (1.1). The basic difference between this case and the case in Section 3 is in the \( \sqrt{n} \ast \sqrt{n} \) design matrix, \( X^T X \), with \( (i,j) \)th element
\[ (X^T X)_{i,j} = n \sum_{k=1}^{n} h_i(T_k)h_j(T_k), \]
associated with the regression problem
\[ Y_i = \sum_{j=0}^{\sqrt{n}} \beta_j h_j(T_i) + \varepsilon_i. \]

Here it is not assumed that \( X^T X = nI \). The way around this is quite simple, we just reparameterize the problem to make it orthogonal. That is for each value of \( j \) we replace \( h_j \) by the function
\[ h_j^*(x) = \sum_{k=0}^{j} a_{k,j} h_j(x), \]
with the \( a_{k,j} \) selected so that the \( \sqrt{n} \ast \sqrt{n} \) design matrix, \( X^T X \), associated with the regression problem
\[ Y_i = \sum_{j=0}^{\sqrt{n}} \beta_j^* h_j^*(T_i) + \varepsilon_i \]
satisfies \( X^T X = nI \). Now the original testing problem, (1.1), is equivalent to
\[ H_0 : m(x) = \sum_{j=0}^{p} \beta_j^* h_j^*(x), \quad (4.1a) \]
versus
\[ H_A : m(x) \neq \sum_{j=0}^{p} \beta_j^* h_j^*(x), \quad \text{for some } x \text{ and for all } \beta^*. \quad (4.1b) \]
The test statistic is then given by
\[ T_W = n \hat{\sigma}^{-1} \sum_{j=p+1}^{M_2} \hat{\gamma}_{(j)}^*, \]
where \( \hat{\sigma} \) is the GSJS (for Gasser, Sroka, and Jennen-Steinmetz) estimator given on page 48 of Eubank (1999).

(4.2) Remark. It would be interesting to see which, if any, of the asymptotic results proven for \( T_C, T_M, T_S \) and \( T_D \) are valid for \( T_W \).
References


On The Asymptotic Variance of The Estimator of Attributable Risk and Testing Association In a $2 \times 2$ Contingency Table

Tanweer J Shapla
Department of Mathematics, Eastern Michigan University, Ypsilanti, MI 48197
Email: tshapla@emich.edu

Abstract

Attributable risk is one of the most important epidemiological indices to quantify the impact of the risk factor on disease. This paper considers derivation of the asymptotic variance for the estimator of attributable risk by using delta method under a cross-sectional sampling scheme. We also consider large sample approximate tests using attributable risk and log odds ratio for association of the risk factor and the disease outcome. It follows that tests based on attributable risk and log odds ratio are comparable with respect to testing size and power.

Key words: Attributable risk; Cross-sectional study; Asymptotic variance; Testing power.

1. Introduction

While the relative risk ($RR$) and the odds ratio ($OR$) have enjoyed widespread use in measuring the association between risk factor and disease, neither of them takes into account the actual number of cases which might play an important role in studying a disease with several risk factors with varying relative risk and prevalences. A risk factor of a disease with a high relative risk may have a low prevalence rate, whereas another risk factor with a relatively low relative risk may have a very high prevalence rate in the population, and hence is responsible for a sizeable fraction of cases. For example, exposure of industrial workers to various chemicals with high relative risk causes fewer cases of the carcinoma of the lung than the exposure to smoking with low relative risk because of the high prevalence rate of smoking in the population (Doll, 1959; Walter, 1976). Since the attributable risk ($AR$) takes into account both the prevalence of the risk factor and the relative risk (Benichou, 1991; Fleiss, 1979; Walter, 1976) and is estimable for all three designs, namely, case-control, cohort and cross-sectional sampling design, it has become popular to epidemiologists and public health practitioners in measuring the impact of a risk factor in developing a disease.

Introduced by Levin (1953), the attributable risk is defined as the proportion of the disease that could be avoided by eliminating the risk factor completely from the population, and is thus given by

$$AR = \frac{P(D) - P(D \mid E_0)}{P(D)}$$

(1.1) where $P(D)$ is the probability of disease in the population, $P(D \mid E_0)$ is the disease rate among the unexposed. It follows from equation (1.1) that

$$AR = \frac{P(E_1)(RR - 1)}{1 + P(E_1)(RR - 1)},$$

(1.2)

which establishes a relation among $AR$, exposure rate $P(E_1)$ and the relative risk $RR$ defined by

$$RR = \frac{P(D \mid E_1)}{P(D \mid E_0)}.$$

Walter (1976) has presented methods of estimation of $AR$ under case-control, cohort and cross-sectional sampling schemes and derived asymptotic variance for estimator of $AR$ using Fisher information matrix, which involves extensive computation, especially under the cross-sectional sampling scheme. In the present paper we derive an expression for the asymptotic variance of the estimator of $AR$ under cross-sectional sampling design by using delta method for a dichotomous risk factor with a dichotomous outcome variable. While the expressions
by the two methods are identical, the new method is computationally straightforward and can easily be extended for a risk factor with multiple exposure levels.

We also consider large sample approximate tests using attributable risk and log odds ratio for positive association of the risk factor and the disease outcome. It follows that tests based on attributable risk and log odds ratio are comparable with respect to size and power.

Section 2 reviews the estimation procedure of the asymptotic variance of $\hat{AR}$ using Fisher information matrix for a $2 \times 2$ contingency table under cross-sectional sampling design. Section 3.1 discusses derivation of the asymptotic variance of $\hat{AR}$ by using delta method. The distribution of $\hat{AR}$ and some useful results regarding independence between the risk factor and the disease outcome are provided in Section 3.2. An example illustrating the application of $AR$ in testing positive association between the respiratory disease and the locomotor disease is presented in Section 4. Section 5 discusses estimation of power and size for testing association using $\hat{AR}$ and $\log OR$ for large sample case.

2. Model Development

Given a random sample under a cross-sectional scheme, we wish to estimate attributable risk ($AR$) for a dichotomous risk factor and a dichotomous outcome variable. Let $E_i$ be the level of the risk factor under investigation with $E_0$ being the reference level. Let $n_{kd}$ be the random frequency of individuals falling into the cell at exposure level $k$ with disease status $d$, $d = 1$ (presence of disease, $D$), 2 (absence of disease, $\overline{D}$). Let $p_{kd}$, $0 < p_{kd} < 1$, be the probability of a subject falling into a cell having observed frequency $n_{kd}$. The table below summarizes the data structure given by the design.

<table>
<thead>
<tr>
<th>Exposure Status</th>
<th>Disease Status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_i$ (present)</td>
<td>$D$</td>
<td>$n_{11}$</td>
</tr>
<tr>
<td>$E_0$ (absent)</td>
<td>$\overline{D}$</td>
<td>$n_{01}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_1$</td>
<td>$n_2$</td>
</tr>
</tbody>
</table>

Let $\varphi_1 = P(D \mid E_i)$, $\varphi_0 = P(D \mid E_0)$ and $\psi = \frac{\varphi_1}{\varphi_0}$. Of course, $\varphi_1 = \frac{p_{11}}{(p_{11} + p_{12})}$ and $\varphi_0 = \frac{p_{01}}{(p_{01} + p_{02})}$. Following (1.1), Walter (1976) expressed $AR$ to be

$$AR = 1 - [\left(\frac{p_{11} + p_{12}}{\psi - 1} + 1\right)]^{-1}. \quad (2.1)$$

It is readily evident that given $n$, the vector $N = (n_{11}, n_{12}, n_{01}, n_{02})$ follows the multinomial distribution with parameters $n$ and $P = (p_{11}, p_{12}, p_{01}, p_{02})$, for which the log-likelihood is given by

$$L = \ln K + n_{11} \ln p_{11} + n_{12} \ln p_{12} + n_{01} \ln p_{01} + n_{02} \ln p_{02}.$$  

By the invariance property (Mood and Graybill, 1974) of the maximum likelihood estimator (MLE), the estimator of $AR$ is given by

$$\hat{AR} = 1 - [\left(\hat{p}_{11} + \hat{p}_{12}\right)(\psi - 1) + 1]^{-1}.$$  

Walter (1976) used the relationships $\hat{p}_{12} = 1 - p_{11} - p_{01} / f$ and $\hat{p}_{02} = -p_{01} + p_{01} / f$. 


where \( f = (p_{11} + p_{01})(1 - AR) \), to express the likelihood in terms of \( p_{11}, p_{01} \) and \( AR \). He used Fisher information matrix to derive the asymptotic variance of \( \hat{AR} \) where the distinct elements of the information matrix were given by

\[
\begin{align*}
\frac{n^{-1} E}{\frac{\partial^2 L}{\partial AR^2}} &= \frac{p_{01}^2 (p_{01} + p_{02})}{f^2 (1 - AR)^2 p_{12} p_{02}}, \\
\frac{n^{-1} E}{\frac{\partial^2 L}{\partial p_{11}^2}} &= \frac{1}{p_{11}} + \frac{1}{p_{12}} - \frac{2 p_{01}}{fAR p_{12}} + \frac{p_{11}^2 (p_{12} + p_{02})}{f^2 AR^2 p_{12} p_{02}}, \\
\frac{n^{-1} E}{\frac{\partial^2 L}{\partial AR \partial p_{11}}} &= \frac{1}{p_{01}} + \frac{1}{p_{02}} - \frac{2 p_{11}}{fAR p_{02}} - \frac{p_{11}^2 (p_{12} + p_{02})}{f^2 AR^2 p_{12} p_{02}}, \\
\frac{n^{-1} E}{\frac{\partial^2 L}{\partial AR \partial p_{01}}} &= \frac{p_{22}}{f (1 - AR) p_{12}} - \frac{p_{22}^2 (p_{12} + p_{02})}{f^2 AR^2 (1 - AR) p_{12} p_{02}}, \\
\frac{n^{-1} E}{\frac{\partial^2 L}{\partial AR \partial p_{01}}} &= \frac{- p_{01}}{f (1 - AR) p_{02}} + \frac{P_{11} p_{02}}{f p_{12} p_{02}} - \frac{P_{11} p_{02}^2 (p_{12} + p_{02})}{f^2 AR^2 p_{12} p_{02}}.
\end{align*}
\]

His expression of the variance of \( \hat{AR} \) is given by

\[
\begin{align*}
&\frac{V_{\hat{AR}}}{\hat{AR}} = \frac{1}{n} \left[ p_{11} (1 - p_{11}) - p_{11} p_{12} - p_{11} p_{01} - p_{11} p_{02} ight]^{-1} \left[ p_{11} p_{12} - p_{11} p_{01} - p_{11} p_{02} \right]. \\
&\text{Then, by the invariance property of the MLE, the estimator of } AR \text{ is given by } \\
&\hat{AR} = \frac{\hat{p}_{11} \hat{p}_{02} - \hat{p}_{12} \hat{p}_{01}}{\hat{p}_{11} \hat{p}_{02} - \hat{p}_{12} \hat{p}_{01} + \hat{p}_{01}}, \quad \text{where } \hat{p}_{k,d} = \frac{n_{k,d}}{n}, \quad k = 0,1; d = 1, 2 \text{ and } \\
&\hat{P} = (\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{01}, \hat{p}_{02}).
\end{align*}
\]

The covariance matrix of the estimate \( \hat{P} \) is given by

\[
\Sigma = \frac{1}{n} \left[ \begin{array}{cccc}
\hat{p}_{11} (1 - \hat{p}_{11}) & \hat{p}_{11} p_{12} & \hat{p}_{11} p_{01} & \hat{p}_{11} p_{02} \\
\hat{p}_{12} (1 - \hat{p}_{12}) & \hat{p}_{12} p_{12} & \hat{p}_{12} p_{01} & \hat{p}_{12} p_{02} \\
\hat{p}_{01} p_{11} & \hat{p}_{01} p_{01} & \hat{p}_{01} (1 - \hat{p}_{01}) & \hat{p}_{01} p_{02} \\
\hat{p}_{02} p_{11} & \hat{p}_{02} p_{01} & \hat{p}_{02} p_{01} & \hat{p}_{02} (1 - \hat{p}_{02})
\end{array} \right].
\]

Let \( \partial P \) be vector of the partial derivatives of \( g \) with respect to the components of \( \hat{P} \) evaluated at \( \hat{P} = P \). Then

\[
\partial P = \left( \frac{\partial g(\hat{P})}{\partial \hat{p}_{11}}, \frac{\partial g(\hat{P})}{\partial \hat{p}_{12}}, \frac{\partial g(\hat{P})}{\partial \hat{p}_{01}}, \frac{\partial g(\hat{P})}{\partial \hat{p}_{02}} \right)_{P=P},
\]

and by the use of delta method, the asymptotic variance of the estimator of \( AR \), \( g(\hat{P}) \) is given by

\[
\begin{align*}
&\frac{\partial}{\partial \hat{P}} g(\hat{P}) = \partial \Sigma \partial, \\
&\frac{\partial \Sigma}{\partial \hat{P}} g(\hat{P}) = \partial \Sigma \partial.
\end{align*}
\]
By (2.4), (2.2) and (2.3), it follows that
\[ \frac{\partial \bar{AR}}{\partial p_{11}} = \frac{p_{01} p_{02}}{(p_{01} + p_{11} p_{02} - p_{12} p_{01})^2}, \quad \frac{\partial \bar{AR}}{\partial p_{12}} = \frac{-p_{01}^2}{(p_{01} + p_{11} p_{02} - p_{12} p_{01})^2} \]
\[ \frac{\partial \bar{AR}}{\partial p_{01}} = \frac{-p_{11} p_{02}}{(p_{01} + p_{11} p_{02} - p_{12} p_{01})^2}, \quad \frac{\partial \bar{AR}}{\partial p_{02}} = \frac{p_{11} p_{01}}{(p_{01} + p_{11} p_{02} - p_{12} p_{01})^2} \]

Also it follows that,
\[ n\hat{\theta} \Sigma \hat{\theta} = \sum_{k=0}^{2} \sum_{d=1}^{2} \sum_{d=1}^{2} p_{k,d} \left( \frac{\partial \bar{AR}}{\partial \theta} \right)^2 = \left( \sum_{k=0}^{2} \sum_{d=1}^{2} p_{k,d} \frac{\partial \bar{AR}}{\partial \theta} \right)^2. \]

Note that
\[ \sum_{k=0}^{2} \sum_{d=1}^{2} p_{k,d} \left( \frac{\partial \bar{AR}}{\partial \theta} \right)^2 = \frac{p_{01} p_{02}^2 + p_{12} p_{01}^3 + p_{11} p_{02}^2 + p_{02} p_{11} p_{01}^2 - p_{01} (p_{11} p_{02} - p_{12} p_{01})^2}{(p_{01} + p_{11} p_{02} - p_{12} p_{01})^4}. \]

And,
\[ \left( \sum_{k=0}^{2} \sum_{d=1}^{2} p_{k,d} \frac{\partial \bar{AR}}{\partial \theta} \right)^2 = \frac{p_{01}^2 (p_{11} p_{02} - p_{12} p_{01})^2}{(p_{01} + p_{11} p_{02} - p_{12} p_{01})^2}. \]

Then,
\[ n\hat{\theta} \Sigma \hat{\theta} = \frac{p_{01} \left( (p_{11} p_{01} p_{02}^2 + p_{12} p_{01}^3 + p_{11} p_{02}^2 + p_{02} p_{11} p_{01}^2 - p_{01} (p_{11} p_{02} - p_{12} p_{01})^2) \right)}{(p_{01} + p_{11} p_{02} - p_{12} p_{01})^4}. \]

Therefore, the asymptotic variance of the estimator of \( \bar{AR} \) using delta method is given by
\[ v_D(\hat{\bar{AR}}) = (1-\bar{AR})^4 \left\{ \frac{p_{01}^3 (p_{11} p_{02} + p_{11} p_{01} + p_{01} p_{02}) - p_{01} (p_{11} p_{02} - p_{12} p_{01})^2 + p_{12} p_{01}^2}{n p_{01}} \right\}. \]

(2.3)

and an estimator of \( v_D(\hat{\bar{AR}}) \), \( v_D(\hat{\bar{AR}}) \) is obtained by replacing the parameters by the respective MLEs in (2.3).

**Corollary 1.** The expression for the variance obtained by Walter using Fisher information matrix and the one obtained by the proposed delta method are same, that is, \( v_f(\hat{\bar{AR}}) = v_D(\hat{\bar{AR}}) \).

**Proof:** Note that
\[
\begin{align*}
(p_{11} + p_{01}) (p_{01} + p_{02}) & \left\{ p_{01} (p_{12} p_{01} - p_{11} p_{02}) + p_{11} p_{02} \right\} \\
& = (p_{11} p_{01} + p_{11} p_{02} + p_{01}^2 + p_{01} p_{02}) (p_{12} p_{01}^2 - p_{11} p_{01} p_{02} + p_{11} p_{02}) \\
& = p_{11} p_{02} (p_{11} p_{01} + p_{11} p_{02} + p_{01} p_{02}) + p_{11} p_{02} p_{01}^2 - p_{12} p_{01} p_{02} - p_{11} p_{01}^3 + p_{12} p_{01} p_{02} + p_{12} p_{01}^3 + p_{11} p_{02}^2 \\
& \quad + p_{12} p_{01}^4 + p_{12} p_{01}^3 p_{02} - p_{11} p_{01}^2 p_{02} - p_{11} p_{01}^3 p_{02} - p_{11} p_{01}^2 p_{02} - p_{11} p_{01}^2 p_{02}^2 \\
& = p_{11} p_{02} (p_{11} p_{01} + p_{11} p_{02} + p_{01} p_{02}) + p_{11} p_{02} p_{01}^2 - p_{01} (p_{11} p_{02} - p_{12} p_{01})^2 + p_{12} p_{01}^3 (p_{11} + p_{12} \\
& \quad + p_{01} + p_{02}) - p_{11} p_{02} p_{01}^2 (p_{11} + p_{12} + p_{01} + p_{02}) \\
& = p_{11} p_{02} (p_{11} p_{01} + p_{11} p_{02} + p_{01} p_{02}) + p_{11} p_{02} p_{01}^2 - p_{01} (p_{11} p_{02} - p_{12} p_{01})^2 + p_{12} p_{01}^3 - p_{11} p_{02} p_{01} \\
& = p_{11} p_{02} (p_{11} p_{01} + p_{11} p_{02} + p_{01} p_{02}) - p_{01} (p_{11} p_{01} - p_{12} p_{01})^2 + p_{12} p_{01}^3 \quad (2.4)
\end{align*}
\]

By (2.4), (2.2) and (2.3), it follows that \( v_f(\hat{\bar{AR}}) = v_D(\hat{\bar{AR}}) \).

238
3.2. Asymptotic distribution of $\hat{AR}$

The following theorems will be useful to find the asymptotic distribution of $\hat{AR}$ to be considered for hypothesis testing procedure.

**Theorem 1.** The random vector $\hat{P} = (\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{01}, \hat{p}_{02})$ is asymptotically normally distributed with mean vector $P = (p_{11}, p_{12}, p_{01}, p_{02})$ and covariance matrix $\Sigma = \frac{1}{n} [\text{diag}(p_{11}, p_{12}, p_{01}, p_{02}) - P'P]$.

The proof of the theorem follows immediately from the direct application of the multivariate Lindeberg-Lévy CLT (Serfling, 2002).

**Theorem 2.** $\hat{AR}$ is asymptotically normally distributed with mean $AR$ and variance $\sqrt{\Delta}(\hat{AR})$.

The proof follows from the direct application of delta method.

Therefore, $Z = \frac{(\hat{AR} - AR)}{\sqrt{\Delta}(\hat{AR})}$ can be approximated by a standard normal distribution and hence we can employ $z$-test in testing for a specified value of $AR$ and for independence or association between the risk factor and the disease.

The propositions below establish relations between the attributable risk and the association of the risk factor and the disease outcome.

**Proposition 1.** Testing of independence or no association between the risk factor and the disease outcome in a $2 \times 2$ contingency table are equivalent to testing $AR = 0$.

**Proposition 2.** Testing of positive association is equivalent to testing $AR > 0$.

**Proposition 3.** $AR \geq 0$ if and only if $\frac{p_{11}}{p_1} \geq \frac{p_{01}}{p_0}$.

Propositions 1 and 2 follow immediately after some algebraic manipulations of the expression of $AR$ together with the conditions of independence and no association.

Also, by the fact that $AR = \frac{p_{11}p_{02} - p_{12}p_{01}}{p_{11}p_{12}p_{02} - p_{11}p_{02} + p_{01}} = \frac{p_{11}p_0 - p_{01}p_1}{p_1p_0}$, the proposition 3 follows immediately.

The theorems and the propositions above can be combined together for testing associations regarding the risk factor and the disease.

4. Application

In this section we provide an example to illustrate the application of $AR$ in testing association. The data for this example appears in Fleiss et al. (2003). A total of 2784 subjects in a community have been cross-classified by the presence or absence of the respiratory disease and the locomotor disease. Respiratory disease is thought of as a cause of the locomotor disease, a disease of bones and organs of movement. The following table summarizes the distribution of the subjects with respect to respiratory disease and locomotor disease status.

<table>
<thead>
<tr>
<th>Respiratory Disease</th>
<th>Locomotor Disease</th>
<th>Total</th>
<th>Proportion of Locomotor Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D$</td>
<td>$\bar{D}$</td>
<td></td>
</tr>
<tr>
<td>$E_1$ (present)</td>
<td>17</td>
<td>207</td>
<td>224</td>
</tr>
<tr>
<td>$E_0$ (absent)</td>
<td>184</td>
<td>2376</td>
<td>2560</td>
</tr>
<tr>
<td>Total</td>
<td>201</td>
<td>2583</td>
<td>2784</td>
</tr>
</tbody>
</table>

Since the rates of locomotor disease in people with and without respiratory disease (0.08 and 0.07, respectively) are virtually the same, we wish to test for no association between respiratory disease and locomotor disease. Because testing of no association between the exposure and disease outcome in a $2 \times 2$ contingency table is
equivalent to testing $H_0 : AR = 0$, we therefore can test $H_0 : no \, association$ between the risk factor and the disease outcome by means of test statistics involving $\hat{AR}$. In particular, we wish to test $H_0 : AR = 0$ versus $H_a : AR > 0$, where $H_a$ is equivalent to testing positive association. An equivalent test $H_0 : log OR = 0$ versus $H_a : log OR > 0$ has also been considered for comparison by means of test statistic involving $\hat{log OR}$. It follows that both statistics $\hat{AR}$ and $\hat{log OR}$ can be approximated by standard normal distributions given by

test statistics $Z = \frac{\hat{AR} - AR}{\sqrt{\nu D(\hat{AR})}}$ and $Z^* = \frac{\hat{log OR} - log OR}{\sqrt{\nu(\hat{log OR})}}$, respectively, where

$$\nu(\hat{log OR}) = \left( \frac{1}{n_{11} + 0.5} + \frac{1}{n_{12} + 0.5} + \frac{1}{n_{01} + 0.5} + \frac{1}{n_{02} + 0.5} \right)$$

(Haldane, 1956; Gart, 1966). Based on the above table, $\hat{AR} = 0.0045$ and $\nu D(\hat{AR}) = 0.0004212$. Then, under null hypothesis, the observed value of the test statistic $Z$ is found to be 0.2182. The value of $Z^*$ is found to be 0.2228. Therefore, at 5% significance level the data does not provide sufficient evidence to conclude that there is an effect of respiratory disease in developing locomotor disease.

**5. Testing Power and Size**

In this section we estimate testing power and size for test statistics $Z$ and $Z^*$ for testing $H_0 : no \, association$ versus $H_a : positive \, association$ by means of Monte Carlo simulation performed following the scheme given below.

1. Fix significance level $\alpha$, and the Monte Carlo sample size $M$.

2. Consider a $2 \times 2$ contingency table satisfying $\frac{p_{11}}{p_{1}} > \frac{p_{01}}{p_{0}}$.

3. Generate a random sample of size $n$ from the given multinomial distribution and find the value of the test statistics $Z$ and $Z^*$.

4. Compare the observed value of each test statistic with the critical value $z_\alpha$ and reject the null hypothesis if $Z \geq z_\alpha$, $Z^* \geq z_\alpha$, where $z_\alpha$ is the upper 100 ($\alpha$) th percentile of a standard normal distribution.

5. Repeat steps 3-4 $M$ times and count the number of rejections. The proportion of rejection over all the simulations gives the estimated power for both statistics, separately.

The simulation results for $M = 10000$ have been summarized in Table 3 for different values of $n$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$Z$</th>
<th>$Z^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.7386</td>
<td>0.7278</td>
</tr>
<tr>
<td>30</td>
<td>0.9063</td>
<td>0.9076</td>
</tr>
<tr>
<td>50</td>
<td>0.9882</td>
<td>0.9891</td>
</tr>
<tr>
<td>70</td>
<td>0.9981</td>
<td>0.9983</td>
</tr>
<tr>
<td>90</td>
<td>0.9996</td>
<td>0.9997</td>
</tr>
<tr>
<td>110</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>150</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>200</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
In order to estimate the size of the test we follow steps 1 through 5 above for a contingency table satisfying 
\[ \frac{p_{11}}{p_{1.}} = \frac{p_{01}}{p_{0.}} \] in step 2. The following table provides the estimated size for different values of \( n \).

| Table 4: Estimated size (\( \alpha = 0.05 \)) using \( Z \) and \( Z^* \) |
|---|---|---|
| \( n \) | \( Z \) | \( Z^* \) |
| 20 | 0.0179 | 0.0113 |
| 30 | 0.0195 | 0.0186 |
| 50 | 0.0325 | 0.0357 |
| 70 | 0.0454 | 0.0467 |
| 90 | 0.0570 | 0.0521 |
| 110 | 0.0540 | 0.0554 |
| 150 | 0.0594 | 0.0512 |
| 200 | 0.0567 | 0.0514 |

From simulation results it follows that both test statistics \( Z \) and \( Z^* \) are comparable in estimating testing power and size.

6. Discussion

While asymptotic variance of \( \hat{AR} \) by using Fisher information matrix and delta method are equivalent, the derivation of the former is computationally more extensive than that of the latter. The reason that both methods are equivalent may lie in the fact that both are based on the maximum likelihood estimators. Also the technique using delta method can easily be generalized to risk factors with multiple exposure levels. For testing purposes, both test statistics provide comparable results in terms of estimated size and power of the test of no association.

However, because of the fact that \( \hat{AR} \) provides the estimated proportion of the disease that could be avoided by eliminating the risk factor of interest, not estimable by the use of the \( \log \hat{OR} \), the attributable risk could be more appealing than that of the \( \log OR \) in measuring the effect of the risk factor in developing a disease.

References

Thoughts on Actuarial Science, Demography, Probability, Statistics and Stochastic Processes

John A. Beekman
Lincoln National Corp. Distinguished Professor Emeritus of Actuarial Science
Department of Mathematical Sciences
Ball State University
Muncie, Indiana 47306

Abstract. This paper is a collection of historical notes on eight topics. Section I is devoted to early probability and statistics courses, students, and research within our Department. Section II summarizes a little of the stochastic processes research performed over the years 1963-96 here. Section III is concerned with some important dates and events in actuarial science history in the United States, Canada, England, Scotland, Sweden, and other countries. Section IV briefly discusses the actuarial science programs at Ball State University. Section V and VI are short essays on Demography, and Risk Theory, from a global perspective. Section VII describes some of the changes over the years in actuarial texts and programs, here and at other colleges and universities. Section VIII tries to capture the exciting ideas brought to actuarial employment and education by the Actuarial Research Conferences, held annually since 1966. Section IX records eighty-eight References which contain ideas relating to the earlier eight sections.

Dedication. It is a pleasure to dedicate this paper to my friend Dr. Mir Masoom Ali, the George and Frances Ball Distinguished Professor of Statistics. Dr. Ali has been a leader in bringing new statistics courses and ideas in his outstanding teaching manner to hundreds of our students. Moreover, he has done excellent research in mathematical statistics, and in the process he has inspired members of the Department of Mathematical Sciences to vigorously do research in their areas of study. Professor Ali has performed exemplary service for the Department, the College of Sciences and Humanities, and Ball State University. Especially noteworthy is his co-creation with Dr. Charles Sampson of Eli Lilly Corp. of the annual Midwest Biopharmaceutical Statistics Workshops, and his continuous support of the series.

I. Early Probability and Statistics Courses, Students, Research

Ball State University was on the Quarter System when I arrived in September, 1963. There were four quarter hour classes on Probability and Statistics (calculus based) which I taught in the Winter and Spring Terms. The numbers completing those courses were 23, and 11. Alan Huckleberry studied for 12 quarter hours of Independent Study. The topics covered were Lebesgue integration, stochastic processes, and statistical estimation. Alan completed a good portion of (Wilks, 1962). He also wrote an 8 quarter hour M.S. thesis entitled “Cramer – von Mises Nonparametric Statistics” (Huckleberry, 1964). His thesis, other M.S. theses, B.S. Honors Theses, and my authored and co-authored books and research papers all are in Special Collections and Archives, Bracken Library, Ball State University. Alan earned a Ph.D. in mathematics at Stanford University, became a Professor of Mathematics at Notre Dame University, South Bend, Indiana, and for sometime now has been a Professor of Mathematics in Germany.

During 1964 – 65, two sections of undergraduate probability were offered with 25, and 14 completing the course. At the graduate level, James L. Allen studied 9 quarter hours of Special Studies in Statistics with me. This included extensive study of (Wilks, 1962), Lebesque Integration, Stieltjes Integration, and articles from Journal, American Statistical Association, and Annals of Mathematical Statistics. This intensive year led to Allen’s 8 quarter hour M.S. thesis entitled “A Nonparametric Statistical Test Involving a Random Number of Random Variables”, which was completed during the summer of 1965 (Allen, 1965). Allen and I went on to co-author two papers, (Allen, Beekman, 1966, 1967). The subject of these papers was initiated by M. Kac in 1949 (Kac, 1949). Kac defined a statistic which appeared useful for statistical problems arising in insurance, biology, and telephone engineering. In such fields, a random
number of observations would be observed in a fixed time period. A Poisson distribution could serve for the distribution of number of observations. Distributions of the M. Kac statistic can be used to determine upper and lower confidence contours for an unknown distribution, or in testing a distribution hypothesis. Our research was partially supported by the National Science Foundation. Allen worked as a statistician for Miles Laboratories, Elkhart, Indiana, and earned a Ph.D. in Econometrics from Notre Dame University.

During 1965 – 66, 31 students completed the undergraduate probability course, and 26 students finished the undergraduate statistics course. A 12 quarter hour sequence in graduate level mathematical statistics was started during that year. The number of students completing the 3 quarters were 16, 10, and 6 respectively. Two 8 quarter hour M.S. theses were completed during this year, and following summer. Thomas J. Morgan’s thesis was entitled “Numerical Solution of Schroedinger Equation” (Morgan, 1966). Tom has spent his professional years at the Enrico Fermi Laboratory, and Argonne National Laboratory in Chicago. His work there helped in the writing of Section 1.1 “The Probability of Controlled Fission” of (Beekman, 1974). Kenneth G. Hummel’s thesis was devoted to “Error Analysis of Morgan’s Method and a New Numerical Solution of Schroedinger’s Equation”, (Hummel, 1966). A portion of Section 7.7 “Finite Difference Approach to Generalized Schroedinger Equations” of (Beekman, 1974) contained brief summaries of Morgan’s and Hummel’s M.S. theses. Hummel went on to earn a Ph.D. in Mathematics. Karen K. Sullivan wrote a B.S. Honors Thesis entitled “Linear Programming (Sullivan, 1966).

During 1966 – 67, the 3 quarter sequence in graduate mathematical statistics was offered again, with 14, 10, and 7 completing the courses, respectively. Paul S. Bell studied an independent readings course in actuarial science, passed Actuarial Exam 1, and was offered a tax-free fellowship to study actuarial science at the M.S. level at The University of Iowa, Iowa City. Paul has had a distinguished actuarial career, starting with Connecticut General Life Insurance Company, Hartford, Connecticut. Paul is a Fellow, Society of Actuaries (FSA). Presently he is the Senior Vice President and Regional Certified Financial Officer, American International Group Inc., in Hong Kong. Jerry R. Beehler was one of the students in the graduate statistics courses. He earned a Ph.D. in Mathematics and is a Vice President, Tri State University, Angola, Indiana. During the summer of 1967, Carol Spilker completed her B.S. Honors Thesis on Monte Carlo Techniques (Spilker, 1967).

Two B.S. Honors Theses were completed during 1967 – 68. Jacqueline Lehman’s thesis was on Semi-groups in Probability Theory (Lehman, 1968). Jackie went on to earn a M.S. in Mathematics at Indiana University. Virginia Gray’s thesis was entitled Markov Chains (Gray, 1968). During this year, 25 students completed the undergraduate probability course, and 11 students finished the undergraduate statistics course. The numbers of students completing the 3 quarters of graduate statistics were 10, 7, and 6 respectively.

During 1968-69, the 3 quarter sequence in graduate probability and stochastic processes was initiated. The number of students completing the 3 quarters were 17, 10, and 5 respectively. The undergraduate classes in probability and statistics were completed by 24, and 19 students. Four students took the actuarial exam on calculus on November 13, 1968. The students traveled to Indianapolis to take the exam. Mel McFall, and Tim Alford passed. The first actuarial exams administered at Ball State University were given on May 14, 1969. Tim Alford passed the probability and statistics exam.

II. Stochastic Processes Research

During the eight summers of 1965, 66, 67, 68, 69, 70, 71, and 72 I had support from the National Science Foundation for stochastic processes research. Brief summaries of some of that research will now be given.

In the papers (Cameron, 1960, 1962), R.H. Cameron defined several special integrals connected with the Wiener integral, and studied the existence and properties of those integrals. Under certain hypotheses on the integrand, these integrals are connected with R. P. Feynman’s function space integral (Feynman, 1948), in the one dimensional case. Feynman integrals of certain classes of functionals serve as solutions to the Schroedinger equation. The paper (Beekman, 1965) extended those ideas from the Wiener stochastic process to Gaussian Markov stochastic processes. It considered a partial differential equation which plays the role for Gaussian Markov processes that the generalized heat flow equation (Cameron, 1954) does for the Wiener process.
In (Beekman, 1967) an integral representation for a solution of a certain boundary value problem is obtained. The solution involved an expectation over the sample functions of a Gaussian Markov process. The three processes considered are the Wiener process, the Doob-Kac process, and the Ornstein-Uhlenbeck process. M. Kac studied the Wiener version of the boundary value problem in (Kac, 1949), as did R.H. Cameron in (Cameron, 1954).

The paper (Beekman, 1968) derived some formulas for obtaining the ruin function and the distribution of total claims of collective risk theory. Five examples illustrated the use of the ruin function in studying retention levels, determining the size of "acceptable" adverse fluctuation, and calculating net premiums for stop-loss reinsurance treaties.

Operator valued function space integrals for the Wiener, Ornstein-Uhlenbeck, and Doob-Kac stochastic processes are defined in (Beekman, Kallman, 1971). For functionals of the form $F(x) = \exp\left\{ \int_s^t \theta(t-w, x(w))dw \right\}$ where $\theta(t,u)$ is bounded and almost everywhere continuous, the special integrals satisfy integral equations related to the generalized Schrödinger equation studied by the first author. R.H. Cameron and D.A. Storvick in (Cameron, Storvick, 1968) defined an operator valued function space integral based on the Wiener stochastic process which, for an appropriate functional, solved an integral equation related to the Schrödinger equation.

An approximation to the finite time ruin function was obtained in (Beekman, Bowers, 1972). That research was conducted jointly at The University of Iowa, and Drake University while this author was on sabbatical leave to the Department of Statistics and Actuarial Science, The University of Iowa. The academic year 1969-70, and the summers of 1970 and 1971 were spent there. Professor Robert V. Hogg was very encouraging to this research, and other writing done there. That time was also very beneficial to the inauguration of the actuarial programs at Ball State University in 1970.

During a number of summers, I was supported by Ball State University Summer Research Grants, and the Actuarial Education and Research Fund. One paper which resulted from such support was (Beekman, Fuelling, 1979). That paper developed a mathematical model for insurance company risks formed by a linear combination of four stochastic processes.

The first process modeled the evolution of claim patterns, recognizing both the random number of claims in a time period and the random nature of the claims. The other three processes modeled random deviations from assumptions about investment performance, operating expenses, and lapse expenses.

The paper (Beekman, Shiu, 1988) presents some diffusion models for bond prices. By treating the spot interest rate as a state variable and using the no-arbitrage principle, the price of a default-free and non-callable pure discount bond is expressed by a conditional expectation. The Ornstein-Uhlenbeck process is used to model the spot rate. That process is adjusted to exclude negative interest rates, and the associated bond-price partial differential equation is solved. Treating the yield rate as a state variable, and using the Brownian bridge process, a simpler bond price model results. Applications to immunization theory are given.

The paper (Beekman, Fuelling, 1990) presented a model which can be used when interest rates and future lifetimes are random, for certain annuities. The Ornstein-Uhlenbeck stochastic process was used to model interest randomness about a fixed level. The paper (Beekman, Fuelling, 1993) provided a comparable model for life insurance.

Two Research Projects, and two M.S. theses belong to this section. Mel McFall wrote a Research Project on net stop-loss premiums (McFall, 1972). N. Rajaram wrote a M.S. Thesis on the connection between certain parabolic partial differential equations, stochastic processes, and function space integrals (Rajaram, 1972). Bart Frye wrote a Research Project on collective risk probabilities, random walk, and applications (Frye, 1974). Gary Lin wrote a M.S. Thesis on fast Fourier transform techniques applied to collective risk problems (Lin, 1977).

### III. Notes on Actuarial Science History

The Actuarial Society of America (ASA) was founded in 1889. The ASA was modeled after the Institute of Actuaries (formed in London in 1848) and the Faculty of Actuaries (formed in Edinburgh in 1856). The ASA published a professional journal, held periodic meetings, and sponsored
actuarial examinations. Actuarial students needed to pass the exams to become members of ASA.

In 1909 the American Institute of Actuaries (AIA) was founded. The AIA reflected western and smaller insurance company interests.

In 1949 the ASA and the AIA merged to become the Society of Actuaries (SOA).

The Casualty Actuarial Society (CAS) was formed in 1914 by actuaries working with the newly emerging workers’ compensation plans. The CAS is the professional body for actuaries working in property/casualty insurance.

Further notes on the history of the actuarial profession can be found in (Trowbridge, 1989). The above summary is drawn from that reference.

The American Mathematical Society conducted a Short Course entitled Actuarial Mathematics at The University of Wyoming, Laramie, on August 10 – 11, 1985. The Short Course was organized by Cecil J. Nesbitt, James C. Hickman, and Elias S.W. Shiu. Eight papers were presented, and their titles indicate the areas of actuarial science covered. The papers were:

“Introduction to Actuarial Mathematics” by J.C. Hickman;
“Updating Life Contingencies” by J.C. Hickman;
“Models in Risk Theory” by H.H. Panjer;
“Loss Distributions” by S.A. Klugman;
“Overview of Credibility Theory” by P.M. Kahn;
“A Survey of Graduation Theory” by E.S.W. Shiu;
“Actuarial Assumptions and Models for Social Security Projections” by J.A. Beekman;

This author’s paper was partially based on the monographs (Andrews, Beekman, 1987), and (Keyfitz, Beekman, 1984). The eight papers appeared as Volume 35 of Proceedings of Symposia in Applied Mathematics, American Mathematical Society, 1986. The Editor was Harry H. Panjer, and is listed in the References as (Panjer, 1986).


In 1999 the Society of Actuaries celebrated its 50th Anniversary. This led to the publication of two outstanding books which are listed in the References. The 50th Anniversary Monograph contains an excellent introductory essay “Milestone Ideas: Marking Fifty Years” by James C. Hickman, and the reprinting of seven major works which first appeared in the Transactions, Society of Actuaries. The first paper “A New Mortality Basis for Annuities” by Wilmer A. Jenkins and Edward A. Lew contained elaborate analyses of time trends in insurance and population mortality. Generation and year-of-exposure projection methods are presented and analyzed in considerable detail. The second paper “Fundamentals of Pension Funding” by C.L. Trowbridge provided a framework for discussing funding methods for pensions. The seventh paper is “An Actuarial Layman’s Guide to Building Stochastic Interest Rate Generators” by J.A. Tilley. The construction of stochastic interest rate generators is an important component of an actuarial model of a financial security system.

The book “The First 50 Years: Society of Actuaries 1949 – 1999” demonstrates how actuaries helped their companies and institutions fulfill the promises made through insurance policies, retirement plans, Social Security and other social programs, employee benefits, and investment plans. The book brings to life many events which influenced financial security throughout the world. It contains a remarkable timeline from 1649 through 1998, emphasizing events which were significant in the evolution of countries, actuarial science, demography, economics, insurance, probability, Social Security, and statistics. This wonderful
timeline and introductory remarks are the comprehensive work of James C. Hickman.

IV. Actuarial Science Programs at Ball State University

The M.A. program was started in 1970; the B.S. program was begun in 1974. The monograph (Beekman, 2004) contains in Chapter 2 an Actuarial Alumni list which was initially prepared by myself, up through the classes of 1997. Bart Frye, F.S.A., long-time professor of actuarial science added some names that I missed, since I never had those students in a class, or as advisees. Mr. Frye updates the list each year. In the summary of these alumni, the following abbreviations are used:

FSA: Fellow, Society of Actuaries; ASA: Associate, Society of Actuaries; FCAS: Fellow, Casualty Actuarial Society; ACAS: Associate, Casualty Actuarial Society.

Those designations are earned by passing rigorous examinations. Average times to become an Associate, and a Fellow are three years past a B.S. or M.A. degree in actuarial science, and three years beyond the Associateship. In a few cases, the listed student completed almost all of the B.S. or M.A. program, without being awarded the degree.

Summary of Lists of Actuarial Alumni

M.A. Class of 1972 1
M.A. Class of 1973 6

M.A. and B.S. Classes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of students</td>
<td>8</td>
<td>10</td>
<td>15</td>
<td>17</td>
<td>10</td>
<td>18</td>
<td>15</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of students</td>
<td>16</td>
<td>18</td>
<td>13</td>
<td>9</td>
<td>18</td>
<td>15</td>
<td>12</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of students</td>
<td>19</td>
<td>23</td>
<td>31</td>
<td>15</td>
<td>32</td>
<td>31</td>
<td>20</td>
<td>17</td>
</tr>
</tbody>
</table>

The total number of students was 423. Of these, 183 were in the M.A. program, and 240 were in the B.S. program. As of June 30, 2002, 99 were FSAs, 86 were ASAs, 4 were FCASs, and 7 were ACASs. Thus, 196 of the 423 had earned an actuarial designation. As of December, 2005, there were 18,179 Fellows and Associates, Society of Actuaries, and 4,120 Fellows and Associates, Casualty Actuarial Society. Thus, 1% of the Society of Actuaries, and ¼% of the Casualty Actuarial Society received a significant part of their actuarial training at Ball State University.

Many of the undergraduate majors in actuarial science earned a minor in the Principles of Business. A few students also earned a minor in Econometrics.

The need for actuaries is large. Actuaries are trained to develop and administer programs which insure people against the random events which lead to serious losses. Those risks are to a person’s life, health, home, car, retirement pensions, travel plans, and investments. Actuaries are trained in probability, statistics, and the mathematics needed to develop premiums for insurance coverage, the reserves needed to provide long-time coverage, and to make financial projections. More and more actuaries are involved with the ever-changing health insurance scene, including Medicare, and private health insurance plans. Long-term care coverage is requiring a growing amount of actuarial effort. Mergers of insurance companies necessitates the careful planning that actuaries can provide. Changes in the Social Security Old Age, Survivors, and Disability Program require actuarial expertise. Thus, graduates of the B.S. and M.A. programs in actuarial science will have exciting and rewarding professional careers. That has been true since the programs started in the 1970’s.

Starting in 1998, there has been an Actuarial Science Advisory Board. It consists of sixteen practicing actuaries, and five professors of actuarial science. The Board has a day long meeting in Muncie on the last Friday in November before Thanksgiving. In addition, many conference calls help the ideas move forward. The Board members plus many alumni have been very helpful in marketing the actuarial programs to prospective students, providing useful advice for possible revisions of the programs, and raising contributions to the various actuarial scholarships. The advisors for the Actuarial Science Advisory Board and the Insurance Advisory Board have been Dr. Jack
Foley, FSA, and Dr. Steve Avila, respectively. Mr. C. Gary Dean, FCAS, Lincoln Financial Group Distinguished Professor of Actuarial Science was the advisor for the Actuarial Science Advisory Board during Fall 2004 – December, 2005. Dr. Avila established the Insurance Advisory Board somewhat earlier than the establishment of the Actuarial Science Advisory Board.

Alumni and friends of the Actuarial Science Program have been generous in their contributions to four scholarship accounts for undergraduates, two scholarship and assistantship accounts for graduate students, and two endowment accounts for the Lincoln National Corporation Distinguished Professorship of Actuarial Science.

V. Demography

Over the years, the Office of the Chief Actuary in Baltimore, Maryland has done very important demographic work for the United States. The demographers and actuaries there have produced many studies which have reported on masses of data projections with respect to life expectancies at ages 0 and 65, population numbers by age and sex, disabled populations, causes of death, and other subjects. The projections have been up to 75 years in the future, so that the Office of the Chief Actuary can provide insightful guidance to the Trustees of the Social Security Trust Funds, and to the U.S. Congress. The original Social Security Act was passed by the U.S. Congress in 1935. Over the intervening years, demographers and actuaries have used their projections and cost analyses to advise the U.S. Congress as changes have been made to the original program, including the Disabled Workers Program, Dependents Program, Medicare, and Medicaid. The crucial 1983 Amendments to the Social Security Act permitted the Trust Funds to move away from their perilously low amounts into safer monetary levels.

The Trustees Reports provide short term (10 years) and long term (75 years) financial projections for

1. Federal Old-Age and Survivors Insurance Trust Fund;
2. Federal Disability Insurance Trust Fund;
3. Federal Hospital Insurance Trust Fund;

These Trust Funds receive income and make payments for the Old-Age Survivors and Disability Income programs, and the Medicare program. The two components of the Medicare program help pay for hospital, home health, skilled nursing facility, hospice care, physician, outpatient hospital, and, prescription drug expenses. The lengthy Trust Reports (more than 200 pages on Trust Funds (1) and (2), and Trust Funds (3) and (4) are submitted to the Speaker of the House of Representatives, and the President of the Senate. The References provide the titles of these Reports. Luckily for readers, there is a short (about 15 pages) Summary of the Trustee Reports. Each year this Summary can be obtained at www.socialsecurity.gov/OACT/TRSU/trsummary.html. Also see References.

A leading book on Social Security is (Robert J. Myers, 1993). Mr. Myers is a Fellow, Society of Actuaries, Fellow, Casualty Actuarial Society, and Fellow, Conference of Actuaries in Public Practice. He served as Chief Actuary, Social Security Administration from 1947 to 1970, as Deputy Commissioner of Social Security from 1981 to 1982, and as Executive Director, National Commission on Social Security Reform during 1982-83.

As the 1983 Amendments proved, there can be considerable changes of programs of Social Security. There can be delays in income benefit increases, acceleration in contribution rates, taxation of income benefits, and gradual increases in the age of qualification for full retirement benefits. The paper (Beekman, 1986) provides a lengthy exposition of the actuarial assumptions and models for Social Security projections. The monograph (Andrews, Beekman, 1987) provides an overview of the enormous amount of actuarial analysis performed by the Office of the Actuary, Social Security Administration, in regard to the Old-Age, Survivors, and Disability Insurance program. Since the publication of that monograph, many comprehensive Actuarial Studies have been published. The titles of Studies 117, 118, 119, and 120 are given in the References, and indicate their subject matter. These studies are very complete and are of length 92 pages, 90 pages, 300 pages, and 186 pages, respectfully.

The Social Security Administration also administers the Supplemental Security Income
(SSI) program. Although this program is less known than the OASDI program, it is a very large program. To quote from the Executive Summary of the 2005 Annual Report of the Supplemental Security Income Program,

- The SSI program is a nationwide Federal assistance program administered by the Social Security Administration that guarantees a minimum level of income for needy aged, blind, or disabled individuals.

- In January 2005, 6.8 million individuals received monthly Federal SSI payments averaging $404.

- Federal expenditures for case payments under the SSI program during calendar year 2004 increased 3.8 percent to $34.2 billion, and the cost of administering the SSI program in fiscal year 2004 increased 5.6 percent to $2.8 billion.

The paper (Beekman, Donkar, 2001) discussed the relationships between the SSI program and the OASDI programs during the 1990s. The paper focuses on the persons concurrently receiving benefits from both programs, but analyzed from the alternative perspectives of the SSI and OASDI programs separately.

There are two Ball State University graduates whose work in demography should be described in detail. Md. Humayun Kabir earned a Master of Science degree in 1975. His thesis (Kabir, 1975) used some techniques from mathematical demography to create thirteen projections of the Bangladesh female population at 5-year intervals from 1966 through 2026. Original data gathered in Bangladesh by the Institute of Statistical Research and Training, Dacca University, and by the Bangladesh Institute of Development Economics was used. Mr. Kabir went on to earn a M.S. in Demography at Harvard University, and a Ph.D. in Demography at Brown University. Dr. Kabir is now a Professor in the Institute of Statistical Research and Training, University of Dhaka, Bangladesh. He has been a leader in offering advice on population growth to his country. Dr. Kabir and I co-authored the paper (Beekman, Kabir, 1997). Its purpose was to describe a possible old-age social security program for Bangladesh, with actuarial projections through 2025. Such a program could be of considerable help in removing the large resistance to family planning.

John R. Wilmoth earned a B.S. in Actuarial Science in 1984. His Honors Thesis (Wilmoth, 1984) was largely developed in France where he was a visiting student for a year. John then went to Princeton University where he earned a Ph.D. in Demography.

Presently he is an Associate Professor, The University of California-Berkeley, where he is in charge of the graduate program in the Department of Demography. Actually, he is on leave from Berkeley to the United Nations in New York. Among his many publications, I would mention his authorship of 21 pages of a 90 page report from an eight member panel which reviewed the financial projection model used by the Social Security Administration to forecast the financial status of the OASDI system into the near and distant future. This report appears in the References.

For many years, actuarial students learned about demography from the book (Spiegelman, 1968). This is a very thorough reference, but does not have any problems to solve. I had developed some problems and a syllabus for a demography course which appeared in (Beekman, 1980). Professor Nathan Keyfitz of Harvard University had developed a longer set of problems which appeared as a new Chapter 20 in the revised printing of (Keyfitz, 1968, 1977). Through visits at Actuarial Research Conferences at Harvard University, and The University of Wisconsin, Professor Keyfitz and I agreed to co-author a demography book which emphasized problems and their solutions. This led to (Keyfitz, Beekman, 1984). This textbook was used as an official reference by the Society of Actuaries for its actuarial exam on demography for a number of years. The current Society of Actuaries reference on demography is (Brown, 1991).

The Society of Actuaries has sponsored symposia on the topic “Living to 100 and Beyond.” Recently, a call for papers was issued for the third such symposium. This conference will be held in January 2008, and invites authors to weigh-in on the implications of long life on social security, pensions, health care, long term care throughout the world. The first such symposium was held in January, 2002. Seven of the twenty papers which were presented appear in (North American Actuarial Journal, Volume 6, Number 3, July 2002). The same reference contains a Guest
Editorial “Advancing the Horizon for Centenarian Survival Studies” by Robert J. Johansen on pages iii, iv.

The paper (Kestenbaum, Ferguson, 2006) examines the number of centenarians in the United States. They discovered that records of enrollment in Medicare Part B provides a solid basis of studying the size, growth, and character of the centenarian population. They estimated the January 1, 1990 centenarian population to be 21,830, the January 1, 2000 such population to be 32,920, and projected the size of the centenarian population on January 1, 2010 to be 52,800. Their Table 4 provides the characteristics of persons age 100 and beyond on January 1, 2000 by age, sex, and race. The number of aged persons, circa 2000, for fourteen countries, is provided in Table 3.

The survival function is useful in analyzing the proportions of men and women who live to age 100 and beyond. This function is explained on page 52 of (Bowers et al, 1997). Let the random variable X denote the distribution of X, \( F_X(x) = \Pr(X \leq x) \), \( x \geq 0 \). The survival function \( s(x) = 1 - F_X(x) = \Pr(X > x) \), \( x \geq 0 \).

Thus \( s(x) \) is the probability a new-born will attain age \( x \). The monograph “Life Tables for the United States Social Security Area 1900 – 2100” (Actuarial Study No. 120, 2005) contains a wealth of material on living to age 100 and beyond. Figure 4b of that reference provides actual ages and projected ages at which \( s(x) = 0.00001 \). In 1900, such \( x \) was approximately 105 for both men and women. In 1980, \( x \approx 110 \) for men and \( x \approx 113 \) for women. It is projected that by 2060, \( x \approx 114 \) for men, and \( x \approx 116 \) for women.

Twenty-three students wrote B.S. Honors Theses in actuarial science in the years reported on in this essay (Beekman, 1996). Of the twenty-three students, five wrote theses which were devoted to demographic subjects. They were concerned with population projections for certain standard metropolitan statistical areas, actuarial projections for Social Security retirement benefits, and other demographic topics in an actuarial setting.

A rather new concept in demography is “active life expectancy”. This phrase first appeared in (Katz et al, 1983). I wrote a lengthy paper on that concept (Beekman, 1988) and how it could be used by actuaries. The last section of the paper presented a formula for a new premium for long-term health care for the elderly, using some of the mathematical foundations of active life expectancy. The formula is applied in an example which uses Medicare and medical data. These ideas were further pursued in the publication (Beekman, 1990). The research in the summers of 1987 and 1988 was supported by the Actuarial Education and Research Fund. A Ball State University Summer Grant during 1989 furthered the research on active life expectancies and their applications to long-term care coverages. The publication (Beekman, Frye, 1991) projected active life expectancies to the year 2080, under three sets of assumptions, for 65 year old men and women. Table 9 of that reference provided estimates of the expected numbers of years of dependency in activities of daily living (ADL) prior to death of 65 year old men and women, under three sets of assumptions on ordinary expectancy of life at age 65.

VI. Risk Theory

This section will sketch the history of collective risk theory. As explained in (Cramér, 1954), this theory was developed by the Swedish actuary Dr. F. Lundberg. Cramér provides the five Lundberg references, the first of which was published in 1909, in which the risk business of an insurance company can be regarded as a stochastic process with independent increments. The specific process is a generalized Poisson process. Cramér compares Lundberg’s work with the pioneering work on the Brownian movement process due to Bachelier in France, (Bachelier, 1900). The Brownian movement process is also a stochastic process with independent increments. Cramér authored an important monograph on the collective risk process in (Cramér, 1955). That work was a survey of collective risk theory from the point of view of the theory of stochastic processes. Harold Cramér was a giant in the areas of mathematical statistics, probability and stochastic processes, and actuarial science. A comprehensive article about his life and works is contained in (Blom, 1987). A complete reference to Cramér’s book Mathematical Methods of Statistics is repeated in our References. It is noted on page 1339 of (Blom, 1987) that Cramér was Chief Editor of Skandinavisk Aktuarietidsskrift.
(now Scandinavian Actuarial Journal) from 1940-1963. Because of my admiration for Cramér and his works, my first actuarial paper (Beekman, 1966) appeared in Skandinavisk Aktuarietidskrift.

The paper (Beekman, 1968) derived formulas for obtaining the ruin function and the distribution of total claims of collective risk theory. These results were illustrated by examples which show the use of the ruin function in studying retention limits for insurance policies and other uses. The derivation of the results was based on the mathematical theory of stochastic processes. By describing the risk business of an insurance company as a stochastic process, it was seen to relate to the theory of queues, dams, storage, statistics, physics, and other fields. A formula for approximating the ruin function is derived in (Beekman, 1969). Included in this paper is a discussion of how the ruin function can be used in deciding the amount of initial capital needed for a new line of insurance. The approximation was improved by Newton Bowers in (Bowers, 1969). The ruin function described in the above references involved an infinite time frame. Since this was bothersome to some readers, an approximation to the finite time ruin function was derived in the papers (Beekman, Bowers, 1972). Actuaries make assumptions about future claims, investment performance, operating expenses, and lapse expenses in developing premiums for new insurance policies. A mathematical model to study future deviations from those assumptions is presented in (Beekman, 1974). This model utilizes the compound Poisson process, and three other processes which are Gaussian Markov processes. Further study of that model led to the use of the Ornstein-Uhlenbeck stochastic process. The conditional mean function for that process reflects the stabilizing effects needed in a model for an economic process in which excessive movements are rare. Results of that study appear in (Beekman, 1975, 1976).

At about this time, the book (Beekman, 1974) appeared. It was devoted to the theory and applications of Gaussian Markov and collective risk stochastic processes. The book involved applications in insurance (collective risk), physics (quantum mechanics), electrical engineering, and statistics (limit laws for Kolmogorov and Kac statistics).

Clinton P. Fuelling brought his considerable expertise in computer science and numerical analysis in studying the multi-risk mathematical model. Our joint research culminated in the papers (Beekman, Fuelling, 1977, 1979, 1980).

VII. Changes in Actuarial Texts

For more than thirty years, actuarial students studied (Jordon, 1952, 1967) as their basic text for the actuarial mathematics of life insurance, and pensions. This changed with the publication of (Bowers, Gerber, Hickman, Jones, Nesbitt, 1986, 1997). This text used a stochastic approach to the mathematics of life contingencies, and built a structure which utilized risk theory, economics, and population theory. The last chapter of (Bowers et al, 1997) introduces models in which interest rates are random variables, discusses the management of interest rate risk, and the intersection of actuarial mathematics and financial economics.

The mathematics of finance is very much needed by actuaries. The text (Kellison, 1991) is the second edition of a familiar book. In addition to the classical subjects, it considers inflation, risk and uncertainty, yield curves, the tools available to quantify and manage the relationship between assets and liabilities, the techniques of immunization, scenario testing, and stochastic approaches to interest.

The subject of financial economics has numerous applications to investments, insurance, and pensions. The comprehensive volume (Panjer, 1998) is an important reference. Actuaries should have a good knowledge of the subject to manage risks on both the asset side and the liability side of the balance sheet. The referenced volume served a number of the actuarial and economics professors of Ball State University in a multi-year faculty seminar. We also studied topics from stochastic processes (Beekman, 1974), and stochastic differential equations (Oksendal, 2000).

The study of loss models is very important for actuaries. The text (Klugman, Panjer, Willmot, 2004) expresses this as follows on page xvii: “This textbook is organized around the principle that much of actuarial science consists of the construction and analysis of mathematical models which describe the process by which funds flow into and out of an insurance system.” The book concentrates on the outflow of cash due to the payment of benefits. The five parts of the text are: Introduction to Modeling, Actuarial models, Construction of empirical models, Parametric statistical methods, and Adjusted estimates and simulation.
A majority of our B.S. and M.A. actuarial science students used the excellent mathematical statistics textbooks (Hogg, Tanis, 2006), and (Hogg, McKean, Craig, 2005). As the years have gone by, these texts have been updated with revised editions. It should also be noted that these books are among the recommended texts of the Society of Actuaries and the Casualty Actuarial Society. The textbook (Klugman, Panjer, Willmot, 2004) had three goals, one of which was to update the distribution fitting material from (Hogg, Klugman, 1984). Some of our students have studied from the 1984 or 2004 text.

**VIII. Actuarial Research Conferences**

The first Actuarial Research Conference was held at The University of Michigan on November 14-16, 1966. The Conference had two objectives:

1. To review the current status of the various approaches to risk theory, to collect information on applications of risk theory to insurance problems, and to indicate directions for further research, particularly in regard to applications.

2. To delineate those aspects of multivariate analysis, which may have application to insurance problems, and to indicate actuarial research possibilities.

Cecil J. Nesbitt served as Chairman of the Organizing Committee and did an excellent job in making this first Actuarial Research Conference an outstanding success. There were 52 participants, of whom 34 were practicing actuaries, and 18 were professors of actuarial science, mathematics, or statistics. I gave a talk based on the paper (Beekman, 1966), and a preliminary version of (Beekman, 1968).

There was so much enthusiasm at this conference, that it became an annual event. A partial list of universities where Actuarial Research Conferences have been held would include: University of Iowa, University of Wisconsin, University of Waterloo, Harvard University, Brown University, Duke University, Ball State University, University of Manitoba, University of Western Ontario, Drake University, University of Calgary, Georgia State University, Laval University, Penn State University, Concordia University, and Oregon State University.

Ball State University hosted the Actuarial Research Conference in 1978, and 1996. At the 1978 Conference, Newton Bowers, Hans Gerber, James Hickman, Donald Jones, and Cecil Nesbitt announced their plans for the new actuarial textbook (Bowers, Gerber, Hickman, Jones, Nesbitt, 1986). There were 52 participants, of whom 23 were practicing actuaries, and 29 were professors of actuarial science.

The theme for the 1996 conference was Actuarial Applications of Financial Economic Theory. This was held on August 15-17, 1996. There were 99 participants, of whom 30 were practicing actuaries, and 69 were professors of actuarial science, insurance, mathematics, or statistics, and university administrators. Twenty seven papers were presented, including (Beekman, Kabir, 1997). As was true of other Actuarial Research Conferences, one of the papers was devoted to the research activities of the Society of Actuaries (Luckner, 1997). The papers of the Conference appear in Actuarial Research Clearing House, Issue 1997.1. This 428 page volume was packaged with a CD entitled “What’s New in Actuarial Education and Research, Highlights of the 31st Actuarial Research Conference”, (Shapiro, 1997). William Bart Frye was the Coordinator for the Conference. At the Conference, Harry Panjer discussed the forthcoming book on Financial Economics (Panjer, 1998). Of the eleven authors of this 670 page volume, ten were present at this Conference.

**IX. REFERENCES**


Actuarial Study No. 119, Short-Range Actuarial Projections of the Old-Age, Survivors, and Disability Insurance Program, 2005, Social Security Administration, Office of the Chief Actuary, August, 2005.

Actuarial Study No. 120, Life Tables for the United States Social Security Area 1900 – 2100, Social Security Administration, Office of the Chief Actuary, August, 2005.
Allen, James L., A Nonparametric Statistical Test Involving a Random Number of Random Variables, M.S. Thesis, Ball State University, Muncie, IN, 1965.


Beekman, John A., Contributions to the History of Ball State University, Archives and Special Collections, Bracken Library, Ball State University, Muncie, IN 2004.


Cameron, R. H. and D. A. Storvick, An operator valued function space integral and a related integral equation, J. Math. and Mech. 18 (1968), 517 – 552.


Hummel, Kenneth G., Error Analysis of Morgan’s Method and a New Numerical Solution of Schroedinger’s Equation, M.S. Thesis, Ball State University, Muncie, IN, 1966.


Lehman, Jacqueline, Semi-groups in Probability Theory, B.S. Honors Thesis, Ball State University, Muncie, IN 1968.


Mc Fall, M.C., Net stop-loss premiums, Actuarial Research Clearing House, Issue 1972.3.

Morgan, Thomas J., Numerical Solution of Schroedinger Equation, M.S. Thesis, Ball State University, Muncie, IN, 1966.


Spilker, Carol, Monte Carlo Techniques, B.S. Honors Thesis, Ball State University, Muncie, IN, 1967.


Estimating A Population Median With A Small Sample

Boris Shulkin and Shlomo Sawilowsky
College of Education
Wayne State University

Abstract

Various estimators of the population median are compared for small samples. The competitors included the sample median, Harrell-Davis estimate (HD), and newly introduced competitors: six distribution free estimates Circle, DblExp, Exp+, Exp-, AltExp, Fibon and one nearly distribution free statistic AllExp, for small sample sizes. A Fortran program was written to conduct a Monte Carlo investigation, where two criteria were used to benchmark the statistics, based on the (1) standard error (MSE) and (2) rank (RBE) for sample sizes of size N = 5, 10, 15, 20, 25, 30, and 50. Data were sampled from the normal, uniform, exponential, Chi squared (df – 2) and student (df = 3) distributions. AllExp performed extremely well under most sets of study conditions. For the distribution free case, it was shown that for sample sizes under 30, AltExp is a top performer for MSE criteria and Fibon emerges as the winner for RBE measure.

A frequently used measure of location is the percentile, defined as a value \( Q_p \) such that 100\( p \)% of the possible values fall below it. More formally, let \( F(\cdot) \) be the cumulative distribution function of a population. Then, \( Q_p \) can be defined as: \( Q_p = F^{-1}(p) \).

By far, the most studied class of measures of location is the family of measures of central tendency. Among them is sample mean, \( \bar{X} \),

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i ,
\]

where \( X = X_1, \ldots, X_n \) be the sample of size \( n \).

The median, the point on a scale below which 50 percent of the observations fall, is a venerable but also commonly used measure of central tendency. If a population distribution is known with cumulative distribution function \( F(\cdot) \), then the median can be formally defined as

\[
M = F^{-1}(0.5).
\]

Let \( X = X_1, \ldots, X_n \) be the sample of size \( n \) and \( X_{(1)} \leq \ldots \leq X_{(n)} \) are order statistics of this sample. Textbook authors (e.g. (Hogg & Tanis, 1993)) typically define the sample median (\( M_s \)) as follows:

\[
M_s = \begin{cases} 
X_{(m)}, & m = (n+1)/2, \text{ if } n \text{ is even} \\
\frac{(X_{(m)} + X_{(m+1)})}{2}, & m = n/2, \text{ if } n \text{ is odd}.
\end{cases}
\]  
(3)

For convenience, (3) can be rewritten as

\[
M_s = \frac{X_{(k)} + X_{(n+1-k)}}{2}, \text{ where } k = \lfloor (n+1)/2 \rfloor.
\]

Here \( \lfloor \cdot \rfloor \) is the integer value.

It is well known that the sample mean defined in (1) minimizes the Euclidian norm of deviations of the variate values from the value of the sample mean. More formally, the function

\[
\varepsilon_x(\xi) = \left( \sum_{i=1}^{n} \left| X_i - \xi \right|^2 \right)^{1/2}
\]

\( (5) \)
attains its minimum at $\xi = \bar{X}$, i.e.
\[
\min_{\xi} \varepsilon_\pi(\xi) = \varepsilon_\pi(\bar{X}). \tag{6}
\]
Similarly, the sample median defined as (4) minimizes absolute norm of deviations of the variate values from its value, i.e.
\[
\varepsilon_{M_s}(\xi) = \sum_{i=1}^{n} |X_i - \xi| \tag{7}
\]
reaches its minimum at $\xi = M_s$,
\[
\min_{\xi} \varepsilon_{M_s}(\xi) = \varepsilon_{M_s}(M_s). \tag{8}
\]
This approach can be generalized to define a class of statistics $C_p$ for measuring central tendency by applying the general Minkowski norm to the deviations of the sample variates from the value of that statistic. In other words, the function
\[
\varepsilon_C(\xi; p) = \left( \sum_{i=1}^{n} |X_i - \xi|^p \right)^{1/p} \tag{9}
\]
attains its minimum at the value of the statistic $C_p$:
\[
\min_{\xi} \varepsilon_C(\xi; p) = \varepsilon_C(C_p; p) \tag{10}
\]
Clearly, the sample mean and sample median are particular cases of (10) for $p=2$ and $p=1$ correspondently, i.e., $\bar{X} \equiv C_2$ and $M_s \equiv C_1$.

It has been established that (6) delivers a single solution that agrees with (1). In the case of (8), the matter is more complicated. It can be shown that in the case when sample size $n$ is an odd number, a single solution exists that agrees with definition (4). When $n$ is even, the value of $M_s$ is indeterminate, to be exact, any number between $X_{(k)}$ and $X_{(n+1-k)}$, $k = [(n+1)/2]$ minimizes expression in (7). This can be explained by the observation that function $\varepsilon_{M_s}(\xi)$ is not differentiable in the classical sense at the point $\xi = M_s$.

The inherent variability of sample values leads to the development of robust measures of central tendency. Many of these methods are based on the notion of a weighted mean:
\[
\overline{X}_w = \sum_{i=1}^{n} w_i X_i, \tag{11}
\]
where $w_i$, $i = 1, \ldots n$ are constants, which typically satisfy the condition: $\sum w_i = 1$. An example of a class of weighted means is the L-estimator, the general form of which is $\sum_{i=1}^{n} w_i X_{(i)}$, where the weights are applied to order statistics.

One of the most commonly used competitors of the sample median is the Harrell-Davis estimator (HD) (Harrell & Davis, 1982). This procedure is based on (Maritz & Jarrett, 1978). The process of computing this statistic involves computing a weighted sum of the order statistics of the sample at hand, with weights based on incomplete beta ratio. The values of the weights are distributed in such a way that the most weight is given to the values from the sample that are involved in computing sample median, and progressively smaller amounts are given to the rest of the order statistics as their values are further away from the center. More formally, let $X = (X_1, \ldots, X_n)$ be a random sample of size $n$ and $X = (X_{(1)}, \ldots, X_{(n)})$ be the order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$. 

257
The expected value of the k-th order statistic is given by
\[ E(X_{(k)}) = \frac{1}{\beta(k, n-k+1)} \int_{-\infty}^{\infty} x F(x)^{k-1} \{1 - F(x)\}^{n-k} dF(x), \]  \hspace{1cm} (12)
where \( \beta(\cdot, \cdot) \) is beta function and \( F(\cdot) \) is distribution function of \( X \). Through the limiting properties and convergence, the Harrell-Davis estimator for \( p \)-th population quantile takes the form of a weighted sum of order statistics with the weights based on incomplete beta function:
\[ M_{HD} = \sum_{i=1}^{n} W_{n,i}^{HD} X_{(i)}. \]  \hspace{1cm} (13)

The weights \( W_{n,i}^{HD} \) can be expressed as
\[ W_{n,i}^{HD} = \frac{1}{\beta((n+1)p, (n+1)(1-p))} \int_{(i-1)/n}^{i/n} y^{(n+1)p-1} (1-y)^{(n+1)(1-p)-1} \, dy, \quad i = 1, \ldots, n. \]  \hspace{1cm} (14)

The regularized beta function (also called the incomplete beta ratio) \( I_1(a, b) \) is defined as a ratio of the incomplete beta function \( \beta_1(a, b) = \int_{0}^{z} u^{a-1}(1-u)^{b-1} \, du \) and the complete beta function \( \beta(a, b) \equiv \beta_1(a, b) \).

Thus, for the case of population median estimate with \( p = 0.5 \), (14) can be rewritten as
\[ W_{n,i}^{HD} = I_{i/n}((n+1)/2, (n+1)/2) - I_{(i-1)/n}((n+1)/2, (n+1)/2), \quad i = 1, \ldots, n. \]  \hspace{1cm} (15)

An interesting property of (15) is that the resulting beta deviates represent the approximation of the probability that the i-th order statistic is the value of the population median. The approximation part of this statement entails replacing distribution function in (12) with a sample distribution resulting in (14). However, that observation is irrelevant to the task of finding the best estimate of the population median (or, for that matter, any specific quantile). In other words, this observation neither proves that the Harrell-Davis is the best estimator, nor does it preclude the possibility that there are other multipliers that may be substituted for (15) in (13) that produce a closer estimate to the population median.

The technique in (Harrell & Davis, 1982) can also be viewed as bootstrap estimator of \( E(X_{((n+1)p)}) \) (see (Efron, 1979), for an in-depth treatment of the general bootstrap procedure). In short, the estimate is built based on simulated resampling. Both the Harrell-Davis and a similar procedure introduced in (Kaigh & Lachenbrouch, 1982) are based on the same idea, except that (Harrell & Davis, 1982) is based on resampling with replacement in case \( k=n \), and the estimator in (Kaigh & Lachenbrouch, 1982) is based on resampling without replacement.

A measure of location is called resistant if deviations in the observation values produce little or no effect on the value of the estimator. The measure of the proportion of the observations in the sample that do not affect the value of the statistic (or have little effect on it) defines the measure of resistance of the statistic. The sample mean has a breakdown point of \( 1/n \), which means that a change in a single observation value can produce an arbitrarily large change in the value of the sample mean. The sample median’s finite breakdown point is slightly less than \( \frac{1}{2} \).

Another property of estimators that is often of interest in traditional applied statistics is the notion of an unbiased estimator. Typically, it is defined through mathematical expectation. Because the value of an estimator calculated on a sample can be viewed as a random variable, a mathematical expectation of it can be derived. Accordingly, an estimator \( \hat{\theta} \) is called an unbiased estimator of a population parameter \( \theta \) if the mathematical expectation of that estimator equals to the value of the parameter:
\[ E(\hat{\theta}(X)) = \theta. \]  \hspace{1cm} (16)
Additionally, various measures of scale for the estimator may be considered. There are several ways to estimate the goodness of the estimator in terms of scale. Least squares error, maximum likelihood, and minimum variance are examples. A member $\hat{\theta}_0$ of the family of unbiased estimators of population parameter $\theta$ with the smallest variance is referred to as the best unbiased estimator of that parameter:

$$\hat{\theta}_0: \text{var}(\hat{\theta}_0) = \min \{ \text{var}(\hat{\theta}) : E(\hat{\theta}) = \theta \}$$

A minimum variance unbiased estimator is a primary estimation criterion in the classical (non-Bayesian) approach to parameter estimation. For example, it can be shown that the sample mean is the best unbiased estimator of population mean.

Returning to the sample median, there are two limitations to its use:

1. Its sampling distribution is intractable. This issue deteriorates with odd sample sizes.
2. Unlike the sample mean, the sample median is not the best unbiased estimator of population median.

As a result, competitors were sought that better estimate population median.

The Harrell-Davis estimator was one such attempt, and it is considered by many (e.g. (Parrish, 1990); (Sheather & Marron, 1990); (Falk, 1985); (Wilcox, 1996)) to be the best. The Harrell-Davis estimator has many interesting theoretical properties. For example, it can be shown that it is asymptotically equivalent to sample median. Harrell-Davis estimator can be viewed as the limit of a bootstrap average as the number of bootstrap resamples tends to infinity. It has also been established that the Harrell-Davis estimator tends to be more efficient as compared with the sample median, especially for small samples.

Even though the Harrell-Davis statistic possesses these, as well as other, impressive theoretical properties, there is no proof that demonstrates it is the best in its class. In other words, based on currently existing theory and practice, there is no reason to believe that it is impossible to find another procedure that will perform better (e.g. more efficiently in some sense) for a narrow and/or wide class of conditions (e.g. for small samples).

Statement of the problem.

The purpose of this study is to compare the sample median, Harrell-Davis estimate (HD), and newly introduced competitors, for small sample sizes, in an effort to better estimate a population median.

A competitor is considered superior in estimating the population median regardless of distribution or sample size if it (a) has the smallest root mean square error from true value of the population median or (b) is the closest to the true value of the population median. The first approach is equivalent to the Euclidean metric; the second one is referred to as taxicab or the Manhattan metric.

Compaq Visual Fortran Professional Edition 6.6C was used, accessing the IMSL(R) Fortran 90 MP Library Version 4.01. A series of Monte Carlo repetitions for each of the combination of statistics/underlying distribution/sample size will be programmed in the following way: determine true median of the population; compute all the information necessary for using all the competing statistics, such as coefficients for weighted sum, etc.; generate a sample of given size from underlying distribution with replacement; compute the values of the statistics; calculate mean square distances to the true population median; rank the values of the statistics according to their absolute distance to the population median; record the results. After all the calculations are performed, the results will be presented in series of summarizing tables.

Sample sizes of $N = 5, 10, 15, 20, 25, 30$ and $50$ were used, with $10,000$ repetitions per experiment for the standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is a statistic distribution with probability density function

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Uniform distribution, a continuous distribution with probability density function that takes the following form:
for $x \in [a, b]$. The values of $a = 0$ and $b = 1$ are used in this study.

**Exponential distribution** (Exp). Consider a continuous random variable $X$ with a density function:

$$p(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{1}{b - a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x > b \end{cases}$$

for $x \in [a, b]$. The values of $a = 0$ and $b = 1$ are used in this study.

**Chi Squared distribution** (ChiSq(2)).

Let $Y_i, i = 1, \ldots, r$ be independent random variables having standard normal distribution (17). Consider a random variable $X$ such that:

$$X = \sum_{i=1}^{r} Y_i^2.$$ (20)

Then, $X$ has a $\chi^2$ (chi-squared) distribution with $r$ degrees of freedom. The density function for $\chi^2$ can be written as

$$p_r(x) = \frac{x^{r/2-1} e^{-x/2}}{\Gamma(\frac{1}{2} r) 2^{r/2}}$$

for $x \in [0, \infty)$. Here $\Gamma(\cdot)$ is a Gamma function defined as

$$\Gamma(x) = \int_0^\infty r^{x-1} e^{-r} dt.$$ (22)

For the study parameter $r = 2$ is used.

**Student distribution** (t(3)).

The Student distribution with $r = n - 1$ degrees of freedom is defined as the distribution of the random variable $t$. Its density function can be written as

$$p_r(x) = \frac{\Gamma\left(\frac{1}{2} (r + 1)\right)}{\sqrt{r\pi} \Gamma\left(\frac{1}{2} r\right) (1 + \frac{x^2}{r})^{\left(\frac{r+1}{2}\right)}}.$$ (23)

The Student distribution (23), under normality condition, has a property of being ‘the best in its class’ in a certain sense (i.e. UMPU – uniformly most powerful unbiased), providing that population variance is unknown. Parameter $r=3$ is used in the study.

**Competitors**

As earlier, let $X = (X_1, \ldots, X_n)$ be a random sample of size $n$ and $\bar{X} = (X_{(1)}, \ldots, X_{(n)})$ be the order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$.

**Circle**

Consider a random variable $X$ with the following pdf:
\[ p(x) = \begin{cases} 
0, & \text{if } x \leq -\sqrt{2/\pi} \\
\sqrt{2/\pi - x^2}, & \text{if } -\sqrt{2/\pi} \leq x \leq \sqrt{2/\pi} . \\
0, & \text{if } x \geq \sqrt{2/\pi} 
\end{cases} \quad (24) \]

Following the logic of (15), consider the following set of weights:

\[ W_{n,i}^C = \int_{-\infty}^{-\sqrt{2/\pi} + 2(i-1)\sqrt{2/\pi} n} p(u)du - \int_{-\infty}^{-\sqrt{2/\pi} + 2i\sqrt{2/\pi} n} p(u)du, \quad i = 1, \ldots, n . \quad (25) \]

The weights \( W_{n,i}^C \) can be interpreted as the probability that the value of a random variable with pdf (24) would fall between \( -\sqrt{2/\pi} + 2(i-1)\sqrt{2/\pi} n \) and \( -\sqrt{2/\pi} + 2i\sqrt{2/\pi} n \), \( i = 1, \ldots, n \). The result of weighting the order statistics \( \bar{X} = (X_1, \ldots, X_n) \) on weights \( W_{n,i}^C \) gives the following estimate of the population median:

\[ M_C = \sum_{i=1}^{n} W_{n,i}^C X(i) . \quad (26) \]

**DblExp**

A general formula for the density function for double exponential distribution (or Laplace distribution) is

\[ p(x) = \frac{e^{-|x-\mu|/\beta}}{2\beta} , \quad x \in \mathbb{R} , \quad (27) \]

where \( \mu \) is the location parameter and \( \beta \) is the scale parameter. The cdf can be written as

\[ F(x) = \frac{1}{2} (1 + \text{sgn}(x-\mu)(1-e^{-|x-\mu|/\beta})) . \quad (28) \]

When \( \mu = 0 \), which makes the pdf symmetric about \( x = 0 \), (28) can be written as

\[ F(x) = \begin{cases} 
1 - e^{-|x|/\beta} / 2, & \text{if } x > 0 \\
e^{-|x|/\beta} / 2, & \text{if } x < 0 \\
1/2, & \text{if } x = 0 
\end{cases} \quad (29) \]

In this study, \( \beta = 1 \) is used. First, \( n \) expected values (deviates) \( \Omega = (\omega_1, \ldots, \omega_n) \) are obtained using Monte Carlo simulation (10,000 repetitions). The weights are then calculated by the transformation

\[ W_{n,i}^{\text{DblExp}} = \omega_i / \left( n \sum_{i=1}^{n} |\omega_i| - \sum_{i=1}^{n} \omega_i \right) , \quad i = 1, \ldots, n . \quad (30) \]

The estimate of population median is calculated in the form

\[ M_{\text{DblExp}} = \sum_{i=1}^{n} W_{n,i}^{\text{DblExp}} X(i) . \quad (31) \]

**Exp+ and Exp-**

Consider a random variable with exponential distribution. The pdf has the form:
The distribution function for (32) can be written as

\[ F(x) = 1 - e^{-x/\mu} . \]  

Parameters \( \mu = 1 \) and \( \mu = -1 \) are used for \( \text{Exp}^+ \) and \( \text{Exp}^- \), respectively. Following the same procedure as in the case of double exponential distribution, \( n \) deviates \( \Omega = (\omega_1, \ldots, \omega_n) \) for each random variable are obtained using Monte Carlo simulation method (10,000 repetitions). For the weights, the following ‘normalizing’ transformation was used:

\[ W_{n,i}^{\text{Exp}} = \omega_i / \sum_{i=1}^{n} \omega_i, \quad i = 1, \ldots, n . \]  

The final form of population median estimate is a weighted sum of the order statistics

\[ M_{\text{Exp}} = \sum_{i=1}^{n} W_{n,i}^{\text{Exp}} X_{(i)} . \]  

\( \text{AltExp} \)

Consider a random variable with pdf (27) and distribution function (28). Fix parameter \( \mu = 0 \). Calculate the weights \( W_{n,i}^{\text{AltExp}} \) in the following form:

\[ W_{n,i}^{\text{AltExp}} = \frac{1}{2} \int_{-\infty}^{-n/3+2(-1)^{i}/3} (1 + \text{sgn}(x)(1-e^{-x^{3}/n}))dx - \int_{-\infty}^{-n/3+2(1)}/3 (1 + \text{sgn}(x)(1-e^{-x^{3}/n}))dx . \]  

The weights in (36) can be interpreted as the probability that a random variable with pdf (27) and parameters \( \mu = 0 \) and \( \beta = n/8 \) falls between \( -n/3+2(-1)^{i}/3 \) and \( -n/3+2i/3 \). The values for these parameters were obtained through series of Monte Carlo minimization studies. The estimate is calculated as a weighted sum

\[ M_{\text{AltExp}} = \sum_{i=1}^{n} W_{n,i}^{\text{AltExp}} X_{(i)} . \]  

\( \text{Fibon} \)

Consider a Fibonacci sequence: \( f_1, f_2, f_3, \ldots \) where \( f_1 = 1, \ f_2 = 2, \) and \( f_i = f_{i-2} + f_{i-1} \) for \( i = 3, 4, 5, \ldots \). Further, consider the set

\[ \omega_i \equiv \omega_{n-i+1} = f_i, \quad \text{for } i = 1, \ldots, n/2, \text{ if } n \text{ is even} \]

\[ \omega_i \equiv \omega_{n-i+1} = f_i, \quad \text{and } \omega_{i+n/2} = \omega_{n/2} + \omega_{n/2-1}, \text{ for } i = 1, \ldots, (n-1)/2, \text{ if } n \text{ is odd} \]  

Take a Fibonacci sequence half way through the sample size, and mirror it over to the other half. The following transformation normalizes (38)

\[ W_{n,i}^{\text{Fib}} = \omega_i / \sum_{i=1}^{n} \omega_i, \quad i = 1, \ldots, n . \]  

The weighted sum of order statistics gives the estimate

\[ M_{\text{Fib}} = \sum_{i=1}^{n} W_{n,i}^{\text{Fib}} X_{(i)} . \]  

\( \text{AllExp} \)

The remaining statistic used in the study is somewhat distribution dependent. Assume that certain, very ‘general’ properties of underlying distribution are known. In particular, suppose that the shape of probability distribution function can be estimated. In particular, the following loosely defined characteristics of p.d.f. are of the interest: Growth, Decay, Symmetry Type 1 (with high kurtosis) and Symmetry Type 2 (with low kurtosis). Symmetry Type 1 and Type 2 may also include distributions that are not necessarily symmetrical per se, but
exhibit certain properties that make them a better fit for a particular statistic. For example, a distribution with some skewness (either positive or negative) and large kurtosis fit better with Symmetric Type 1

Combine statistics AltExp, DblExp, Exp- and Exp+ using the following rule:

1. Use Exp- for Growth Distributions (such as Asymmetric Growth and Bimodal)
2. Use Exp+ Decay Distributions (such as Asymmetric Decay and Mass G)
3. Use AltExp Symmetric Type 1 Distributions (such as Normal, Exp, ChiSq(2), t(3), MultiModal)
4. Use DblExp for Symmetric Type 2 Distributions (such as Uniform, Sm Sym, Mass 0 and Digit preference)

The resulting aggregate statistic will be referred to as AllExp.

**Benchmarks for success**

**Mean Square Error (MSE) and Rank Based Error (RBE)**

There are several ways to judge which competitor is superior in estimating the population median regardless of distribution or sample size. Two of these will be considered. One benchmark is the smallest root mean square error from the population median (MSE). Another is the closeness to the population median (RBE).

Let $P_M$ be population median. Let $N_R$ be a number of Monte Carlo repetitions and $M_i^j$ be the median estimate by $j^{th}$ method in $i^{th}$ repetition, $j = 1,\ldots,N_M$. Here $N_M$ is number of methods. Then Mean Square Error (MSE) can be defined as follows:

$$
\varepsilon_{MSE}^j = \sqrt{\frac{1}{N_R} \sum_{i=1}^{N_R} (M_P^i - M_i^j)^2}.
$$

Further, calculate deviation of each estimate from population median

$$
\Delta M_i^j = |M_P^i - M_i^j|, j = 1,\ldots,N_M, i = 1,\ldots N_R.
$$

For each $i = 1,\ldots,N_R$ find a set of indexes $I(j), j = 1,\ldots,N_M$ such that

$$
\Delta M_i^{(1)} \leq \Delta M_i^{(2)} \leq \ldots \leq \Delta M_i^{(N_M)}.
$$

The Rank Based Error (RBE) can now be defined as follows:

$$
\varepsilon_{RBE}^j = \frac{\sum_{i=1}^{N_R} (I(j) - 1) / N_M}{N_R}.
$$

Table 1. Mean Square Error (MSE × 10⁻⁴) of Competing Sample Statistics from the Population Median Rank for Sample of Size N = 5 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>53\6</td>
<td>46\5</td>
<td>44\3</td>
<td>44\2</td>
<td>75\8</td>
<td>74\7</td>
<td>43\1</td>
<td>45\4</td>
</tr>
<tr>
<td>Unif</td>
<td>18\6</td>
<td>15\5</td>
<td>13\2</td>
<td>13\1</td>
<td>21\8</td>
<td>21\7</td>
<td>14\4</td>
<td>14\3</td>
</tr>
<tr>
<td>Exp</td>
<td>46\4</td>
<td>45\3</td>
<td>50\6</td>
<td>51\7</td>
<td>111\8</td>
<td>31\1</td>
<td>41\2</td>
<td>48\5</td>
</tr>
<tr>
<td>ChiSq2</td>
<td>92\4</td>
<td>90\3</td>
<td>99\6</td>
<td>101\7</td>
<td>220\8</td>
<td>62\1</td>
<td>81\2</td>
<td>95\5</td>
</tr>
<tr>
<td>t(3)</td>
<td>63\3</td>
<td>61\2</td>
<td>69\5</td>
<td>71\6</td>
<td>131\8</td>
<td>129\7</td>
<td>57\1</td>
<td>65\4</td>
</tr>
</tbody>
</table>
Table 2. Rank Based Error (RBE \times 10^{-4}) of Competing Sample Statistics from the Population Median\Rank for Sample of Size N = 5 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>3734\6 3378\5 2939\3 2872\2 4871\8 4839\7 2268\1 3095\4 2268\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unif</td>
<td>4303\6 3478\5 2472\2 2042\1 4812\7 4921\8 3052\4 2916\3 2042\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp</td>
<td>3559\7 2934\2 3243\4 3545\6 5749\8 3435\5 2471\1 3061\3 2471\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChiSq2</td>
<td>3484\6 2942\2 3277\4 3540\7 5725\8 3473\5 2495\1 3060\3 2495\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t(3)</td>
<td>3268\5 3105\2 3258\4 3440\6 4829\7 4840\8 2124\1 3132\3 2124\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Mean Square Error (MSE \times 10^{-4}) of Competing Sample Statistics from the Population Median\Rank for Sample of Size N = 10 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>37\6 34\5 31\3 31\2 78\7 79\8 30\1 32\4 30\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unif</td>
<td>13\6 12\5 10\2 9\1 22\7 22\7 11\3 11\4 9\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp</td>
<td>31\4 30\3 37\6 41\7 116\8 29\2 28\1 32\5 28\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChiSq2</td>
<td>62\4 61\3 75\6 83\7 233\8 58\2 56\1 66\5 56\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t(3)</td>
<td>42\4 40\2 45\5 51\6 135\8 130\7 38\1 41\3 38\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Rank Based Error (RBE \times 10^{-4}) of Competing Sample Statistics from the Population Median\Rank for Sample of Size N = 10 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>3422\6 3200\5 2625\3 2559\2 5625\7 5705\8 3076\4 3076\4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unif</td>
<td>3924\6 3444\5 2245\2 3304\5 4049\7 6584\8 3817\6 3817\6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp</td>
<td>2883\4 2548\2 3304\5 4049\7 6584\8 3817\6 3817\6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChiSq2</td>
<td>2873\4 2528\2 3319\5 4071\7 6577\8 3800\6 3800\6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t(3)</td>
<td>2867\3 2658\2 3012\4 3359\6 5630\7 5639\8 1802\1 1802\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7. Mean Square Error (MSE \times 10^{-4}) of Competing Sample Statistics from the Population Median\Rank for Sample of Size N = 20 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>27\6 25\5 22\3 22\2 83\8 83\7 21\1 24\4 21\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unif</td>
<td>10\6 9\5 7\2 6\1 23\7 23\8 8\3 8\4 6\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp</td>
<td>21\4 20\2 29\5 36\7 120\8 31\6 19\1 21\3 19\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChiSq2</td>
<td>44\4 42\2 60\5 73\7 242\8 62\6 40\1 43\3 40\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t(3)</td>
<td>30\4 28\3 31\5 37\6 139\8 137\7 27\1 28\2 27\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Rank Based Error (RBE \times 10^{-4}) of Competing Sample Statistics from the Population Median\Rank for Sample of Size N = 20 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>3245\6 2972\5 2400\3 2343\2 6228\8 6227\7 1832\1 2748\4 1832\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unif</td>
<td>3555\6 3220\5 2061\2 1332\1 6149\7 6192\8 2857\4 2630\3 1332\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp</td>
<td>2566\4 2153\2 3322\5 4496\7 6927\8 4387\6 1830\1 2314\3 1830\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChiSq2</td>
<td>2556\4 2168\2 3323\5 4495\7 6931\8 4341\6 1840\1 2341\3 1840\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t(3)</td>
<td>2764\5 2437\3 2750\4 3225\6 6205\7 6238\8 1955\1 2421\2 1955\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11. Mean Square Error (MSE \times 10^{-4}) of Competing Sample Statistics from the Population Median\Rank for Sample of Size N = 30 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>22\6 20\5 18\3 18\2 84\7 85\8 17\1 20\4 17\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unif</td>
<td>8\6 7\5 5\2 5\1 23\8 23\7 7\3 7\4 5\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp</td>
<td>18\4 17\2 26\5 34\7 123\8 31\6 16\1 17\3 16\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChiSq2</td>
<td>36\4 34\2 53\5 68\7 245\8 63\6 33\1 34\3 33\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t(3)</td>
<td>24\4 23\2 25\5 30\6 139\8 137\7 21\1 23\3 21\1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

264
Table 12. Rank Based Error (RBE × 10⁻⁴) of Competing Sample Statistics from the Population Median\Rank for Sample of Size N = 30 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>3173\6</td>
<td>2835\5</td>
<td>2302\3</td>
<td>2262\1</td>
<td>6347\7</td>
<td>6332\7</td>
<td>2801\3</td>
<td>2852\4</td>
<td>1262\1</td>
</tr>
<tr>
<td>Unif</td>
<td>3407\6</td>
<td>2976\5</td>
<td>2019\2</td>
<td>1262\1</td>
<td>6347\8</td>
<td>6332\7</td>
<td>2801\3</td>
<td>2852\4</td>
<td>1262\1</td>
</tr>
<tr>
<td>Exp</td>
<td>2353\4</td>
<td>1988\3</td>
<td>3430\5</td>
<td>4799\7</td>
<td>6985\8</td>
<td>6569\6</td>
<td>1797\1</td>
<td>1984\2</td>
<td>1797\1</td>
</tr>
<tr>
<td>ChiSq2</td>
<td>2376\4</td>
<td>2013\2</td>
<td>3384\5</td>
<td>4750\7</td>
<td>6986\8</td>
<td>6656\6</td>
<td>1806\1</td>
<td>2015\3</td>
<td>1806\1</td>
</tr>
<tr>
<td>t(3)</td>
<td>2719\5</td>
<td>2380\2</td>
<td>2615\4</td>
<td>3149\6</td>
<td>6418\8</td>
<td>6398\7</td>
<td>1923\1</td>
<td>2392\3</td>
<td>1923\1</td>
</tr>
</tbody>
</table>

Table 13. Mean Square Error (MSE × 10⁻⁴) of Competing Sample Statistics from the Population Median\Rank for Sample of Size N = 50 for Various Theoretical Distributions.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>17\6</td>
<td>16\4</td>
<td>14\3</td>
<td>14\2</td>
<td>86\8</td>
<td>86\7</td>
<td>13\1</td>
<td>16\5</td>
<td>13\1</td>
</tr>
<tr>
<td>Unif</td>
<td>6\6</td>
<td>6\4</td>
<td>4\2</td>
<td>4\1</td>
<td>12\8</td>
<td>24\7</td>
<td>5\3</td>
<td>6\5</td>
<td>4\1</td>
</tr>
<tr>
<td>Exp</td>
<td>14\4</td>
<td>13\2</td>
<td>23\5</td>
<td>33\7</td>
<td>125\8</td>
<td>32\6</td>
<td>13\1</td>
<td>13\3</td>
<td>13\1</td>
</tr>
<tr>
<td>ChiSq2</td>
<td>28\4</td>
<td>26\1</td>
<td>48\5</td>
<td>66\7</td>
<td>251\8</td>
<td>64\6</td>
<td>27\2</td>
<td>27\3</td>
<td>27\2</td>
</tr>
<tr>
<td>t(3)</td>
<td>19\4</td>
<td>18\2</td>
<td>19\5</td>
<td>24\6</td>
<td>139\8</td>
<td>139\7</td>
<td>17\1</td>
<td>18\3</td>
<td>17\1</td>
</tr>
</tbody>
</table>

The next 2 tables, Table 15 and Table 16, contain mean ranks for each competing estimator and each population distribution averaged over all sample sizes for MSE and RBE correspondently. As earlier, rows of the tables correspond to the 13 studied population distributions and columns correspond to the competing estimators.

Table 15. Mean Ranks of Mean Square Error (MSE) Estimation of Various Estimators of the Population Median by Distribution, 10,000 Repetitions, n=5,10,15,20,25,30&50.

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>6.00</td>
<td>4.86</td>
<td>3.00</td>
<td>2.00</td>
<td>7.57</td>
<td>7.43</td>
<td>1.00</td>
<td>4.14</td>
<td>1.00</td>
</tr>
<tr>
<td>Unif</td>
<td>6.00</td>
<td>4.86</td>
<td>2.00</td>
<td>1.00</td>
<td>7.71</td>
<td>7.29</td>
<td>3.14</td>
<td>4.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Exp</td>
<td>4.00</td>
<td>2.29</td>
<td>5.43</td>
<td>7.00</td>
<td>8.00</td>
<td>4.57</td>
<td>1.14</td>
<td>3.57</td>
<td>1.14</td>
</tr>
<tr>
<td>ChiSq2</td>
<td>4.00</td>
<td>2.14</td>
<td>5.43</td>
<td>7.00</td>
<td>8.00</td>
<td>4.57</td>
<td>1.29</td>
<td>3.57</td>
<td>1.29</td>
</tr>
<tr>
<td>t(3)</td>
<td>3.86</td>
<td>2.29</td>
<td>5.00</td>
<td>6.00</td>
<td>8.00</td>
<td>7.00</td>
<td>1.00</td>
<td>2.86</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 16. Mean Ranks of Rank Based Error (RBE) Estimation of Various Estimators of the Population Median by Distribution, 10,000 Repetitions, n=5,10,15,20,25,30&50

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>6.00</td>
<td>4.86</td>
<td>3.00</td>
<td>2.00</td>
<td>7.57</td>
<td>7.43</td>
<td>1.00</td>
<td>4.14</td>
<td>1.00</td>
</tr>
<tr>
<td>Unif</td>
<td>6.00</td>
<td>4.71</td>
<td>2.00</td>
<td>1.00</td>
<td>7.29</td>
<td>7.71</td>
<td>3.71</td>
<td>3.57</td>
<td>1.00</td>
</tr>
<tr>
<td>Exp</td>
<td>4.43</td>
<td>2.00</td>
<td>4.86</td>
<td>6.86</td>
<td>8.00</td>
<td>5.86</td>
<td>1.14</td>
<td>2.86</td>
<td>1.14</td>
</tr>
<tr>
<td>ChiSq2</td>
<td>4.29</td>
<td>1.86</td>
<td>4.86</td>
<td>7.00</td>
<td>8.00</td>
<td>5.86</td>
<td>1.14</td>
<td>3.00</td>
<td>1.14</td>
</tr>
<tr>
<td>t(3)</td>
<td>4.71</td>
<td>2.29</td>
<td>4.00</td>
<td>6.00</td>
<td>7.43</td>
<td>7.57</td>
<td>1.00</td>
<td>3.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The next 2 tables, Table 17 and Table 18, contain mean ranks for each competing estimator and each sample size averaged over five studied theoretical distributions for MSE and RBE correspondently. Rows of the tables correspond to the 7 studied sample sizes and columns correspond to the competing estimators. For example, row 2 column 7 in Table 17 reads 1.40. This means that using MSE criteria, for samples size 10, the statistic AltExp ranks 1.40 on average (for the five theoretical distributions studied) among all competing estimators.

Table 17. Mean Ranks of Mean Square Error (MSE) Estimation of Various Estimators of the Population Median by Sample size for Theoretical Distributions, 10,000 Repetitions

<table>
<thead>
<tr>
<th>Distr</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>6.00</td>
<td>4.86</td>
<td>3.00</td>
<td>2.00</td>
<td>7.57</td>
<td>7.43</td>
<td>1.00</td>
<td>4.14</td>
<td>1.00</td>
</tr>
<tr>
<td>Unif</td>
<td>6.00</td>
<td>4.86</td>
<td>2.00</td>
<td>1.00</td>
<td>7.71</td>
<td>7.29</td>
<td>3.14</td>
<td>4.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Exp</td>
<td>4.00</td>
<td>2.29</td>
<td>5.43</td>
<td>7.00</td>
<td>8.00</td>
<td>4.57</td>
<td>1.14</td>
<td>3.57</td>
<td>1.14</td>
</tr>
<tr>
<td>ChiSq2</td>
<td>4.00</td>
<td>2.14</td>
<td>5.43</td>
<td>7.00</td>
<td>8.00</td>
<td>4.57</td>
<td>1.29</td>
<td>3.57</td>
<td>1.29</td>
</tr>
<tr>
<td>t(3)</td>
<td>3.86</td>
<td>2.29</td>
<td>5.00</td>
<td>6.00</td>
<td>8.00</td>
<td>7.00</td>
<td>1.00</td>
<td>2.86</td>
<td>1.00</td>
</tr>
</tbody>
</table>
266

Table 18. Mean Ranks of Rank Based Error (RBE) Estimation of Various Estimators of the Population Median by Sample size for Theoretical Distributions, 10,000 Repetitions

<table>
<thead>
<tr>
<th>N</th>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4.60</td>
<td>3.60</td>
<td>4.40</td>
<td>4.60</td>
<td>8.00</td>
<td>4.60</td>
<td>2.00</td>
<td>4.20</td>
<td>1.40</td>
</tr>
<tr>
<td>10</td>
<td>4.80</td>
<td>3.60</td>
<td>4.40</td>
<td>4.60</td>
<td>7.80</td>
<td>5.20</td>
<td>1.40</td>
<td>4.20</td>
<td>1.00</td>
</tr>
<tr>
<td>15</td>
<td>4.80</td>
<td>3.20</td>
<td>4.40</td>
<td>4.60</td>
<td>7.80</td>
<td>6.40</td>
<td>1.40</td>
<td>3.40</td>
<td>1.00</td>
</tr>
<tr>
<td>20</td>
<td>4.80</td>
<td>3.40</td>
<td>4.00</td>
<td>4.60</td>
<td>7.80</td>
<td>6.80</td>
<td>1.40</td>
<td>3.20</td>
<td>1.00</td>
</tr>
<tr>
<td>25</td>
<td>4.80</td>
<td>3.40</td>
<td>4.00</td>
<td>4.60</td>
<td>7.80</td>
<td>6.80</td>
<td>1.40</td>
<td>3.20</td>
<td>1.00</td>
</tr>
<tr>
<td>30</td>
<td>4.80</td>
<td>3.20</td>
<td>4.00</td>
<td>4.60</td>
<td>7.80</td>
<td>6.80</td>
<td>1.40</td>
<td>3.40</td>
<td>1.00</td>
</tr>
<tr>
<td>50</td>
<td>4.80</td>
<td>2.60</td>
<td>4.00</td>
<td>4.60</td>
<td>8.00</td>
<td>6.60</td>
<td>1.60</td>
<td>3.80</td>
<td>1.20</td>
</tr>
</tbody>
</table>

The next 2 tables, Table 21 and Table 22, contain mean ranks for each competing estimator averaged over five studied theoretical distributions and seven sample sizes for MSE and RBE correspondently. Columns in the tables correspond to the competing estimators.

Table 21. Mean Ranks of Mean Square Error (MSE) Estimation of Various Estimators of the Population Median, for Theoretical Distributions, 10,000 Repetitions

<table>
<thead>
<tr>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.77</td>
<td>3.29</td>
<td>4.17</td>
<td>4.60</td>
<td>7.86</td>
<td>6.17</td>
<td>1.51</td>
<td>3.63</td>
<td>1.09</td>
</tr>
</tbody>
</table>

Table 22. Mean Ranks of Rank Based Error (RBE) Estimation of Various Estimators of the Population Median, for Theoretical Distributions, 10,000 Repetitions

<table>
<thead>
<tr>
<th>SMedian</th>
<th>HDavis</th>
<th>Circle</th>
<th>DblExp</th>
<th>Exp+</th>
<th>Exp-</th>
<th>AltExp</th>
<th>Fib</th>
<th>AllExp</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.09</td>
<td>3.14</td>
<td>3.74</td>
<td>4.57</td>
<td>7.66</td>
<td>6.89</td>
<td>1.60</td>
<td>3.31</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Conclusions and Recommendations

It is important to first make few remarks about interpretation of the rankings. Tables 1 through 14 are fairly straightforward to read: the errors are tabulated along with corresponding ranks (integer numbers) for each distribution free statistic, ranging from 1 through 8. The values that are tabulated in Tables 15 through 24, however, are cumulative ranks. This results in a fractional rank value for each competing estimate, with a minimum and maximum of 1 and 8.

For each set of assumptions, there may or may not be a statistic that is a clear winner. If such statistic does emerge, it has the rank value less than 2. However, if such an estimate does not exist, the best choice of statistic in such a case would be the one with the smallest rank value among all competing statistics. Consider, for example, the results summarized in Table 20. For samples size 5 the statistic DblExp ranks 3.25 on average. Note, however, that this is the lowest average rank for this sample size among all distribution free statistics considered in this study, which makes DblExp estimate the best choice under mentioned above assumptions.

According to the data in Tables 17 through 20, AllExp estimate’s average ranking is under 2 in every set of assumptions (generalized over more than one population distribution type) included with the study, with exception of RBE error for non-theoretical distribution with sample sizes 25 and 50. Even in these 2 cases, AllExp is still the best statistic among the ones included in the present study, with average rankings of 2.0 and 2.12 correspondently. This makes the AllExp estimate a clear winner among all the competitors under assumptions of this study. The obvious disadvantage of AllExp, of course, is the fact that a certain à priori information (albeit only sketchy) about the population distribution is necessary for its effective use.

Overall, for both type of errors MSE and RBE considered, irregardless of the sample size, AltExp comes out to be the obvious winner. However, the second best varies depending on the set of assumptions. For the case of MSE criteria, HD and Fibon are tied for second place. HD clearly dominates the second place for both small and large sample size, although Fibon takes over for middle sample size (20 and 25).
Having established this, it is important to mention that HD and Fibon demonstrate extremely similar levels of performance, with the exception at the large sample size (50). The situation changes dramatically for the case of RBE criteria, in which Fibon statistic controls a strong second place.

Several important points are worth being mentioned. It appears that with a little a priori knowledge about the population distribution, as is necessary for the use of AllExp estimate, the results can be improved dramatically, as compared to a pure distribution free case. In every set of assumptions, AllExp comes out to be the best statistic for estimating the population median.

Unexpectedly, to some extent, Fibon shows a strong and persistent third and fourth place in averaged rankings over all distributions for all sample sizes, while first and second places do not appear to have a single strong contender. This can possibly make it a very good alternative as a solo statistic in distribution free situation. Another interesting observation is that for every set of assumptions, with exception of a single case (MSE criteria for theoretical distributions), the least likely challenger, the Circle estimate, shows results that closely match those of HD.

**Recommendations**

General recommendations, based on the results of this study, are summarized as follows:

1. As long as some basic information about the population distribution is known, such as (1) growth, (2) decay, (3) symmetry with high kurtosis, or (4) symmetry with low kurtosis, a ‘nearly’ distribution free statistic AllExp should be used for estimating the population median for all sample sizes up to 50. This is the best statistic choice (among the ones that were part of this study) in either the case of the standard error (MSE) or the case of the closest to the population median (RBE).
2. If there is concern regarding the standard error of the estimate, but there is no à priori information regarding the population (i.e., the shape of the distribution), then for sample sizes up to 30 AltExp is the best choice of estimator.
3. If there is concern with how close the estimate has to be to the population median, and no à priori information regarding the population distribution, then, for sample sizes up to 30, Fibon is the best choice of estimator.
4. In both cases 2 and 3 above, HD should be the statistic of choice for sample sizes 50 and above.

**References**


SUM OF HYPERGEOMETRIC SERIES FUNCTIONS USING PROBABILITY DISTRIBUTIONS

Munir Ahmad
National College of Business Administration & Economics
Lahore, Pakistan
Email: drmunir@brain.net.pk

and

Ayesha Roohi
Statistics Department
Lahore College for Women University
Lahore, Pakistan
Email: ayesha_roohi_2004@yahoo.co.uk

ABSTRACT

In this paper we use Dacey’s (1972) probability function to obtain some new results of sums of hypergeometric series functions.

1. INTRODUCTION

Ahmad and Roohi (2004, 2005) have obtained the sum of hypergeometric series function using discrete probability functions. In this paper, sum of a few new hypergeometric series functions have derived using Dacey (1972) generalized discrete probability function.

2. SUM OF HYPERGEOMETRIC SERIES FUNCTIONS

Theorem (1):

Let \( h \) \( F_m \left[ (a) ; (b) ; \theta \right] \) be defined as [See Rainville (1960)]:

\[
\begin{align*}
\sum_{s=0}^{m} (1)^{s} \binom{k}{s} b_{s} \left( \begin{array}{c}
\gamma(a) + \theta s + (b) \\
\gamma(a + s + b + \theta)
\end{array} \right)
\end{align*}
\]

Then

\[
\sum_{s=0}^{m} (1)^{s} \binom{k}{s} b_{s} \left( \begin{array}{c}
\gamma(a) + \theta s + (b) \\
\gamma(a + s + b + \theta)
\end{array} \right)
\]

where \( n = 1, 2, \ldots, k \geq 1 \) and \( b_j \neq 0, -1, -2, \ldots \) \( j = 1, 2, \ldots, m \) and \( \theta \in R \)

Proof:

Suppose \( X \) is a discrete random variable with probability function defined by Dacey (1972) as:

\[
P(X = x) = C^n \left( \begin{array}{c}
x \\
\gamma(a); (b)
\end{array} \right) \theta^x, \quad x = 0, 1, \ldots, n
\]

where \( C^{x} = h_{+1} F_m \left[ -n, (a); (b); \theta \right] \), \( \gamma_s [(a); (b)] = \frac{\Gamma[(a + s); (b + x)]}{\Gamma[(a); (b)]} \)
and \( \Gamma[(a);(b)] = \frac{\Gamma a_1 \Gamma a_2 \ldots \Gamma a_h}{\Gamma b_1 \Gamma b_2 \ldots \Gamma b_m} \).

It is known that
\[
\prod_{s=1}^{k} \left( \frac{1}{x+s} \right) = \sum_{s=1}^{k} \frac{(-1)^{s+1}}{(x+s)(s-1)!} \frac{1}{x+s}, \quad x \geq 0 \quad k = 1,2,3,\ldots \tag{4}
\]

Now
\[
E \left[ \prod_{s=1}^{k} \left( \frac{1}{X+s} \right) \right] = \sum_{s=1}^{k} \frac{(-1)^{s+1}}{(k-s)!s!} E \left[ \frac{1}{X+s} \right] \tag{5}
\]

where
\[
E \left[ \frac{1}{X+s} \right] = C \sum_{s=0}^{n} \frac{1}{x+s} \left( n \right)^{S} \gamma_{S}[(a);(b)]
\]

\[
= C n^{-1} \left\{ \frac{n^S}{s+1} \left[ a_{1}a_{2}a_{h}+ \frac{n^S}{s+1} \left( \frac{\theta}{b_{1}b_{2}b_{m}}(b_{m}+1) \right) \left[ a_{1}(a_{1}+1)\ldots a_{h}(a_{h}+1) \right] \right] \right\} \]

\[
= C n^{-1} \left\{ \frac{n^S}{s+1} \left[ a_{1}a_{2}a_{h}(\theta)+ \frac{n^S}{s+1} \left( \frac{\theta}{b_{1}b_{2}b_{m}}(b_{m}+1) \right) \left[ a_{1}(a_{1}+1)\ldots a_{h}(a_{h}+1) \right] \right] \right\} \]

Substituting (6) in (5) we get:
\[
E \left[ \prod_{s=1}^{k} \left( \frac{1}{X+s} \right) \right] = C n^{-1} \left\{ \frac{n^S}{s+1} \left[ a_{1}a_{2}a_{h}+ \frac{n^S}{s+1} \left( \frac{\theta}{b_{1}b_{2}b_{m}}(b_{m}+1) \right) \left[ a_{1}(a_{1}+1)\ldots a_{h}(a_{h}+1) \right] \right] \right\} \tag{7}
\]

Also
\[
E \left[ \prod_{s=1}^{k} \left( \frac{1}{X+s} \right) \right] = C \sum_{s=0}^{n} \left( \frac{n^S}{x+s} \right) \gamma_{S}[(a);(b)] \prod_{s=1}^{k} \left( \frac{1}{x+s} \right)
\]

\[
= C (k-1)^{-1} \left\{ \frac{(-1)^{k}}{k+1} \left[ a_{1}a_{2}a_{h}+ \frac{(-1)^{k}}{k+1} \left( \frac{\theta}{b_{1}b_{2}b_{m}}(b_{m}+1) \right) \left[ a_{1}(a_{1}+1)\ldots a_{h}(a_{h}+1) \right] \right] \right\} \]

\[
= C (k-1)^{-1} \left\{ \frac{(-1)^{k}}{k+1} \left[ a_{1}a_{2}a_{h}+ \frac{(-1)^{k}}{k+1} \left( \frac{\theta}{b_{1}b_{2}b_{m}}(b_{m}+1) \right) \left[ a_{1}(a_{1}+1)\ldots a_{h}(a_{h}+1) \right] \right] \right\} \tag{8}
\]
Equating (7) and (8) we get (2).

Corollary (1):

If we put \((a_i) = 1 = b_j; \ i = 1, 2, \ldots h \) and \( j = 1, 2, \ldots m \) and \( \theta = \frac{p}{1-p} \) where \( 0 < p < 1 \) in (2) we get:

\[
\sum_{s=0}^{k} (-1)^{s+1} \binom{k}{s} \left[ \sum_{j=1}^{h} F_{m+1}[s, a_j; s+1, b_j; \theta] \right] = \sum_{s=0}^{k} (-1)^{s+1} \binom{k}{s} \left[ \sum_{j=1}^{h} F_{m+1}[s, a_j; s+1, b_j; \theta] \right]
\]

Theorem (2):

\[
\sum_{s=0}^{k} (-1)^{s+1} \binom{k}{s} F_{h+1} F_{m+1}[s, (a); s+1, (b); \theta] = \sum_{s=0}^{k} (-1)^{s+1} \binom{k}{s} F_{h+1} F_{m+1}[s, (a); s+1, (b); \theta]
\]

where \( k \geq 1, a_i \in R, i = 1, 2, \ldots h, b_j \neq 0, -1, -2, \ldots j = 1, 2, \ldots m \).

Proof:

Suppose \( X \) is a discrete random variable with probability function defined at (3):

Using the identity (4) and the definition of expectation we have:

\[
E \left[ \frac{1}{X+s} \right] = C \sum_{x=0}^{\infty} \frac{\theta^x}{x!} \gamma_x[(a); (b)] \frac{1}{x+s}
\]

\[
= s^{-1} C \left[ 1 + \frac{\theta s \ a_1 a_2 \ldots a_h}{s+1} \frac{a_1 (a_1+1) \ldots a_h (a_h+1)}{s^2} \theta^2 + \ldots \right],
\]

\[
= s^{-1} C \ h_{h+1} F_{m+1}[s, (a); s+1, (b); \theta].
\]

Using (11) in (10), we get:

\[
E \left[ \prod_{i=1}^{k} \frac{1}{X+s} \right] = C \sum_{x=0}^{\infty} \frac{\theta^x}{x!} \gamma_x[(a); (b)] \prod_{x=1}^{k} \frac{1}{x+s}
\]

Also

\[
E \left[ \prod_{x=0}^{k} \frac{1}{X+s} \right] = C \sum_{x=0}^{\infty} \frac{\theta^x}{x!} \gamma_x[(a); (b)] \prod_{x=1}^{k} \frac{1}{x+s}
\]

\[
E \left[ \prod_{x=0}^{k} \frac{1}{X+s} \right] = C(k)^{-1} \left[ 1 + \frac{1}{k+1} \frac{\theta s \ a_1 a_2 \ldots a_h}{s+1} \frac{a_1 (a_1+1) \ldots a_h (a_h+1)}{s^2} \theta^2 + \ldots \right],
\]

\[
= C(k)^{-1} h_{h+1} F_{m+1}[1, (a); s+1, (b); \theta]
\]

Equating (12) and (13) give the required result.

Corollary (2):

Substituting \( (a) = 1, (b) = \lambda, \lambda > 0, \theta > 0 \) in (13), we get
\[
\sum_{s=1}^{k} (-1)^{s+1} \binom{k}{s} F_2\left[1, s; \lambda, s+1; \theta\right] = F_2\left[1, k+1; \lambda; \theta\right]
\]  
(14)

For \( \lambda = 1 \) in (14) we have
\[
\sum_{s=1}^{k} (-1)^{s+1} \binom{k}{s} F_1\left[s, s+1; \theta\right] = F_1\left[1, k+1; \theta\right]
\]  
(15)

**Corollary (3):**

For \((a) = r > 0\), no denominator parameter \((b)\) and \(\theta = q, 0 < q < 1\) in (13) we get:
\[
\sum_{s=1}^{k} (-1)^{s+1} \binom{k}{s} F_2\left[s, r; s+1; q\right] = F_2\left[1, r; k+1; q\right]
\]  
(16)

Further substituting \(r = 1\) in (16) we have
\[
\sum_{s=1}^{k} (-1)^{s+1} \binom{k}{s} F_1\left[s, 1; s+1; q\right] = F_1\left[1, 1; k+1; q\right].
\]  
(17)

**Corollary (4):**

If \((a)\) consists of two values \((1, a)\) and \((b) = a_1 + \rho + 1\) where \(a_1, \rho > 0\) and \(\theta = 1\), then (10) becomes:
\[
\sum_{s=1}^{k} (-1)^{s+1} \binom{k}{s} F_2\left[s, 1, a; s+1, a_1 + \rho + 1; 1\right] = F_2\left[1, 1, a; k+1, a_1 + \rho + 1; 1\right]
\]  
(18)

Substituting \(a_1 = 1\) in (18) gives:
\[
\sum_{s=1}^{k} (-1)^{s+1} \binom{k}{s} F_2\left[s, 1, 1, s+1, \rho + 2; 1\right] = F_2\left[1, 1, 1; k+1, \rho + 2; 1\right]
\]  
(19)

**Theorem (3):**

\[
\sum_{s=1}^{k} (-1)^{s+1} \binom{k+d}{s+d} \frac{(s+d-1) \cdots s}{d!} h+iF_{m+1}\left[(a), s+d; (b), s+d+1; \theta\right] = h+iF_{m+1}\left[(a), d+1; (b), d+k+1; \theta\right], k \geq 1.
\]  
(20)

**Proof:**

Suppose \(X\) is a discrete random variable with probability function defined in (3), we have
\[
E\left[\frac{1}{X+s}\right] = C \sum_{s=1}^{\infty} \frac{\theta^{s-d}}{(s-d)!} h+iF_{m+1}\left[(a); (b)\right] \frac{1}{s+s}
\]

\[
= C(s+d)^{-1} \left[1+ s+d \frac{a_1, a_2, \ldots, a_n \theta^2}{b_1 b_2 \cdots b_m} + \frac{s+d+1}{b_1 (b_1+1) \cdots (b_m+1) 2^n} + \cdots\right]
\]

\[
= C(s+d)^{-1} h+iF_{m+1}\left[(a), s+d; (b), s+d+1; \theta\right].
\]  
(21)

Using the result (5) we have:
\[
E \left[ \prod_{i=1}^{k} \left( \frac{1}{X+s} \right) \right] = C \sum_{s=1}^{k} \frac{(-1)^{s+1}}{s!(d+1)(d+2)\cdots(d+k)} \frac{1}{(k-s)!} \frac{a_{s+1}a_{s+2}\cdots a_{k}}{b_{s}b_{s+1}b_{s+2}\cdots b_{k}} \left[ (a), s+d; (b), s+d+1; \theta \right] \]  

Also

\[
E \left[ \prod_{i=1}^{k} \left( \frac{1}{X+s} \right) \right] = C \sum_{s=1}^{k} \frac{\theta^{s-d}}{(s+1)!} \frac{1}{(1+1)(1+2)\cdots(1+k)} \frac{a_{s+1}a_{s+2}\cdots a_{k}}{b_{s}b_{s+1}b_{s+2}\cdots b_{k}} \left[ (a); (b) \right] \prod_{i=1}^{k} \left( \frac{1}{x+s} \right)
\]

\[
= C \sum_{s=1}^{k} \frac{1}{(d+1)(d+2)\cdots(k+d)(k+1)!!} \left[ 1 + \frac{a_{s+1}a_{s+2}\cdots a_{k}}{b_{s}b_{s+1}b_{s+2}\cdots b_{k}} \frac{d+1}{k+d+1} \theta \right] 
\]

\[
+ \frac{a_{1}(a_{1}+1)\cdots a_{k}(a_{k}+1)}{b_{1}(b_{1}+1)\cdots b_{k}(b_{k}+1)} \frac{1}{(d+1)(d+2)\cdots(2k+1)} \frac{\theta^{2}}{2!} \cdots
\]

\[
E \left[ \prod_{i=1}^{k} \left( \frac{1}{X+s} \right) \right] = C \left[ (d+1)(d+2)\cdots(k+d) \right]^{-1} \frac{a_{s+1}a_{s+2}\cdots a_{k}}{b_{s}b_{s+1}b_{s+2}\cdots b_{k}} \left[ (a), d+1; (b), k+d+1; \theta \right]
\]

Equations (21) and (22) give the required result (20).

**Corollary (5):**

For \((a) = (1, 1), d = 1\) and \((b) = 2\) in (23) we get:

\[
\sum_{s=1}^{k} (-1)^{s+1} \frac{(k+1)}{s+1} s \gamma_{s+1} F_{s+2} \left[ 1, 1, s+1; 2, s+2; \theta \right] = \gamma_{1} F_{1} \left[ 1, 1; k+2; \theta \right], \quad 0 < \theta < 1
\]

**Theorem (4):**

For \(\alpha \geq 0\):

\[
\sum_{s=1}^{k} (-1)^{s+1} \frac{(d+1)(d+2)\cdots(d+k)}{s+d} \frac{h_{s+1}F_{s+1}}{F_{s+1}} \left[ (a), s+d; (b), s+d+1; \theta \right]
\]

\[
= h_{1} \left[ (a), \quad d+1; \quad (b), \quad k+d+1; \quad \theta \right],
\]

\[
a_{j} \in R, i = 2, \ldots, \theta \in R \text{ and } b_{j} \neq 0, -1, -2, \ldots
\]

**Proof:**

Let \(X\) be a discrete random variable with probability function defined in (3), then:

\[
E \left[ \prod_{i=1}^{k} \left( \frac{1}{X+s} \right) \right] = C \sum_{s=1}^{k} \frac{\theta^{s-d}}{(s+1)!} \frac{1}{(1+1)(1+2)\cdots(1+k)} \frac{a_{s+1}a_{s+2}\cdots a_{k}}{b_{s}b_{s+1}b_{s+2}\cdots b_{k}} \left[ (a); (b) \right] \prod_{i=1}^{k} \left( \frac{1}{x+s} \right)
\]

\[
= C \left[ (d+1)(d+2)\cdots(d+k-1) \right]^{-1} \left[ 1 + \frac{a_{s+1}a_{s+2}\cdots a_{k}}{b_{s}b_{s+1}b_{s+2}\cdots b_{k}} \frac{d+1}{k+d} \theta \right]
\]

\[
+ \frac{a_{1}(a_{1}+1)\cdots a_{k}(a_{k}+1)}{b_{1}(b_{1}+1)\cdots b_{k}(b_{k}+1)} \frac{1}{(d+1)(d+2)\cdots(2k)} \frac{\theta^{2}}{2!} \cdots
\]

\[
E \left[ \prod_{i=1}^{k} \left( \frac{1}{X+s} \right) \right] = C \left[ (d+1)(d+2)\cdots(d+k-1) \right]^{-1} h_{1} F_{1+1} \left[ (a), d+1; (b), k+d; \theta \right]
\]
Corollary (6):

For \((a) = (1, 1), (b) = 2, d = 1\) and \(0 < \theta < 1\), we get from (25):

\[
\sum_{s=1}^{k} (-1)^{s+1} \binom{k+1}{s+1} F_2 \left[ 1,1, s+1; 2, 2; \theta \right] = \frac{k}{\theta} F_1 \left[ 1,1; k+2; \theta \right].
\] (27)

REFERENCES


Fuzzy Set Representation of a Prior Distribution

Glen Meeden*
School of Statistics
University of Minnesota
Minneapolis, MN 55455
glen@stat.umn.edu

For a subjective Bayesian uncertainty about the unknown parameter or state of nature can often be expressed through a prior distribution. If \( \theta \) denotes a typical parameter value and \( \Theta \) the set of all possible parameter values then the prior distribution over \( \Theta \) summarizes their knowledge and beliefs about the parameter.

Fuzzy set theory is another approach to representing uncertainty. A fuzzy set \( A \), a subset of \( \Theta \), is characterized by its membership function. This is a function defined on \( \Theta \) whose range is contained in the unit interval. At a point \( \theta \) the value of the membership function is a measure of how much we think \( \theta \) belongs to the set \( A \). Statisticians, both Bayesian and frequentist, have been slow to embrace fuzzy set theory. Recently Singpurwala and Booker, (*Journal of the American Statistical Association*, 2004) have proposed a model which incorporates fuzzy membership functions into a subjective Bayesian setup. However the fuzzy membership functions do not have a probabilistic interpretation.

The concept of confidence intervals is a frequentist approach to expressing uncertain about an unknown parameter. It has long been recognized that naive users have difficulty interpreting confidence intervals. They have a

*Research supported in part by NSF Grant DMS 0406169
tendency to give a probabilistic interpretation to the observed confidence interval.

It has also long been known that for discrete data conventional confidence intervals, which we will also call “crisp” confidence intervals, using a term from fuzzy set theory, can perform poorly. Because of the inherent flaws in crisp confidence intervals for discrete problems Geyer and Meeden, (Statistical Science, 2005) suggested a new confidence interval notion called fuzzy confidence intervals. Given the data a fuzzy confidence interval is just the membership function of the set of plausible or reasonable values for $\theta$. One way to think about such membership functions is that they are generalizations of randomized intervals where no randomization is ever implemented. They argued that fuzzy confidence intervals overcome the difficulties of the usual crisp intervals for discrete probability models.

In terms of frequency of coverage discrete data Bayesian credible intervals will suffer from the same problem that conventional intervals do. This suggests that it could be of interest to some Bayesians to be able to use their prior or posterior to get a sensible fuzzy interval instead of the usual Bayesian credible interval.

Here we consider a no data statistical decision problem where the set of possible decisions is the class of all membership functions defined on $\Theta$. We then define a family of loss functions. Such a loss function measures the loss incurred when the prior distribution is replaced by a fuzzy membership function. For any loss function in the family and a given prior distribution we solve the resulting no data decision problem. This gives a method for converting a prior or posterior distribution into a fuzzy membership function. Conversely, for a fuzzy membership function we find the family of prior distributions whose solutions to the no data decision problem is this function. This sets up a formal relationship between the two theories which allows one to move back and forth between them.
The Impact of Different Math Prerequisites on the Performance of Business Statistics Students

Jeffrey J. Green  
Associate Professor of Economics  
Phone: (765) 285-5206  
Fax: (765) 285-4313  
jgreen@bsu.edu

Courtenay C. Stone  
Professor of Economics  
Phone: (765) 285-2857  
Fax: (765) 285-4313  
o0ccstone@bsu.edu

Abera Zegeye  
Associate Professor of Economics  
Phone: (765) 285-1117  
Fax: (765) 285-4313  
azegeye@bsu.edu

Thomas A. Charles  
Senior Research Analyst  
Phone: (765) 285-5152  
Fax: (765) 295-8024  
tcharles@bsu.edu

Ball State University  
Muncie, IN 47306

Preliminary Working Paper: Please do not cite without authors’ permission
Version: 4/2/07

Jeffrey J. Green, Courtenay C. Stone, Abera Zegeye and Thomas A. Charles

“…[S]tudents are generally screened by the use of prerequisite course requirements … before enrollment into a course or program is granted. An additional concern is whether the screening process correctly identifies students with inadequate math skills and prevents their enrollment in courses they are poorly prepared for.”

1 The authors are Associate Professor, Professor and Associate Professor of Economics and Senior Research Analyst, respectively, at Ball State University, Muncie IN 47306
2 Ely and Hittle (1990), p. 59
The Association to Advance Collegiate Schools of Business (hereafter AACSB) standards require accredited business school curriculums to provide their students with “analytic skills.” While AACSB does not mandate any specific method for achieving this result, the vast majority of accredited business schools require their students to take one or more courses in business statistics. Because of low and declining levels of math skills of matriculating U.S. college students, most introductory statistics courses have prerequisite math courses that are ostensibly intended to provide the mathematical background necessary to make statistics less difficult for students to learn and, therefore, for instructors to teach.  

While AACSB does not mandate any specific method for achieving this result, the vast majority of accredited business schools require their students to take one or more courses in business statistics. Because of low and declining levels of math skills of matriculating U.S. college students, most introductory statistics courses have prerequisite math courses that are ostensibly intended to provide the mathematical background necessary to make statistics less difficult for students to learn and, therefore, for instructors to teach.  

How can we determine exactly how effective the selected prerequisite math skills and courses are in aiding business statistics students to succeed? Once in a great while, there are substantial change in math prerequisites that provide a unique opportunity to examine their impact on student performance in business statistics. In this paper, we examine the consequences of recent substantial changes in math prerequisites on student performance in the business statistics course at Ball State University. 

The Change in Math Prerequisites

ECON 221: Business Statistics is one of nine courses that students are required to complete before they are admitted to the business college and permitted to pursue their desired business degrees. Until a few years ago, two three-unit mathematics courses—MATHS 131: Finite Mathematics for Business and MATHS 132: Brief Calculus for Business—were the standard math prerequisites for ECON 221.  

Course descriptions of these courses, their prerequisites and the other courses discussed in the following sections are shown in Table 1.

During the 2003-2004 academic year, the business college’s Undergraduate Curriculum Committee began to consider reducing the math prerequisite for ECON 221 to a single math course. Initially, their discussion focused on dropping MATHS 132 and simply making MATHS 131 the prerequisite math course. The final decision, however, was to change the math prerequisite to a newly-proposed four-unit course, MATHS 135: Mathematics for Business. Accordingly, the Mathematical Sciences (MATHS) Department removed MATHS 131 and added MATHS 135 to the university’s 2004-2006 Undergraduate Catalog and began teaching MATHS 135 in Fall 2004.

During the transition period from the old math prerequisite regime to the new one, the college had to decide what to do about a sizeable number of prospective business students who had completed MATHS 131 but were caught in midst of the math prerequisite change and the timing of the MATHS department’s adjustment to the course changes—MATHS 132 was not offered after Spring 2004 and MATHS 135 was not offered until Fall 2004. The college decided to permit these students to enroll in ECON 221 in Fall 2004—in effect providing them with a two-semester window of opportunity to take ECON 221 after having taken only MATH 131.

In response to this limited opportunity, students enrolled in ECON 221 in record numbers. For example, Fall semester enrollment in ECON 221 in 2000-2003 had averaged about 200 students (standard deviation of 28 and a maximum of 237 students), while Spring enrollment in 2000-2003 had averaged about 270 students (standard deviation of 28 and a maximum of 304 students). However, ECON 221 enrollment soared to 340 students in Fall 2004 and 355 students in Spring 2005. In contrast, ECON 221 enrollment in the Fall 2005 and Spring 2006 semesters were 216 and 293 students, respectively.

---

1 There is considerable controversy about whether this requirement is being met. Business schools have come increasingly under criticism for failing to provide their graduates with the analytic skills necessary for success in the business world. See, for example, Bennis and O’Toole (2005), DeAngelo, DeAngelo and Zimmerman (2005) and Stern and Tseng (2002).

2 For comments on the decline in U.S. math skills, see, for example, (Begley (2004), Kronholz (2004) and Prystay (2004)).

3 Two other math courses, MATHS 161 or MATHS 165, were considered as acceptable alternatives to the MATHS 131/132 sequence and, therefore, would also satisfy the math prerequisite for ECON 221. We discuss these alternatives in more detail later in this paper.

4 There were 347 students enrolled in MATHS 131 in Spring 2004. The MATHS department, inadvertently, no doubt, exacerbated the problem by offering MATHS 131 (44 students were enrolled), but not MATHS 132 during the 2004 summer sessions.

5 Although the intent was to permit students who had completed MATHS 131 to enroll in ECON 221 only in Fall 2004, students who had enrolled in the statistics course in Fall 2004 and had received any grade in the class (including W or F) were permitted to re-enroll in ECON 221 in Spring 2005 to improve their grade.
The Data
We acquired selected academic and demographic data on all BSU students who had earned any grade (including W, which indicates "withdrawal" during the semester) in ECON 221 from Fall 2001 through Summer 2005. There were 1,694 students who had received a grade ranging from A through W in ECON 221 after taking their math prerequisite course(s) at the university. This data enabled us to determine the impact of reducing the math prerequisite to MATHS 131 alone compared to the previous math prerequisites. Because this change was the one that was initially considered, our results show how ECON 221 student performance in general would have been affected had this change actually been adopted.

The ECON 221 information listed in Table 1 indicates that, under the former math prerequisites (those prior to Spring 2005), students had two alternatives to the MATHS 131/132 sequence. They could take either MATHS 161 or 165, both of which have additional MATHS courses as prerequisites. We are able to assess the performance of those students who used these MATHS courses as a prerequisite for ECON 221 compared to those who took MATHS 132/132 or MATHS 131 alone.

Assessing Student Performance in Business Statistics: The Variables and The Model
The focus of our study is the impact of differential math prerequisite courses on student performance in ECON 221. The symbols and definitions of the variables used in our study are shown in Table 2. We used two alternative measures of student performance in ECON 221 as our dependent variable. For our OLS model, we used E221 Grade, the numeric grade that students received the first time they enrolled in ECON 221. For our ordered probit model, we used DE221 Grade, a dummy variable for the grade that the students received the first time they enrolled in ECON 221.

We also used several explanatory variables that have been used in other studies of student performance or are related to the math prerequisites for ECON 221. We used two 0-1 dummy variables to determine whether participation in intercollegiate athletics, Athlete or the student’s gender, Female, had any influence on student performance in ECON 221. MATHS Grade is the grade (F = 0, D- = .7, … A- = 3.7, A = 4) that the students received in their prerequisite math class and Lag GPA is the students’ cumulative GPA in the semester prior to their enrollment in ECON 221. D131, D161 and D165 are the usual 0-1 dummy variables identifying whether the student’s math prerequisite course was MATHS 131, MATHS 161 or MATHS 165 instead of the MATHS 131/132 sequence.

Our general model for student performance in ECON 221 is:

\[
E221 \text{ Grade} = \beta_0 + \beta_1 \text{ Athlete} + \beta_2 \text{ Gender} + \beta_3 \text{ MATHS Grade} + \beta_4 \text{ Lag GPA} + \beta_5 D131 + \beta_6 D161 + \beta_7 D165.
\]

We expect that \( \beta_3, \beta_4, \beta_6 \) and \( \beta_7 > 0 \) and \( \beta_5 < 0 \). MATHS Grade and Lag GPA are measures of the student’s success in acquiring the prerequisite math skills and the student’s overall academic accomplishment—both of which should have positive effects on the student’s performance in ECON 221. D161 and D165 identify students who took math courses that are more analytic in scope and have additional math prerequisites when compared to the MATHS 131/132 sequence. Consequently, we expect students who have taken MATHS 161 or MATHS 165 will perform better in ECON 221 than those who took MATHS 131/132. In contrast, because D131 indicates students who took MATHS 131 only, we expect that these students to perform at a lower level in ECON 221.

---

8 Our data set excludes transfer students, “guest” students, students who had not taken their math prerequisites at the university and any others who were not first-time matriculating students.

9 There are 327 students (out of 1694 students in our data set) who used MATHS 131 as their math prerequisite to enroll on ECON 221.

10 We had hoped to also show the impact of the new math prerequisite, MATHS 135, on student performance in ECON 221 in this paper. Unfortunately, only 47 of the 1,694 students in our data set had taken ECON 221 by Summer 2005 after completing MATHS 135 earlier. In addition, about 40% of these students had enrolled in ECON 221 during the first Summer Session 2005.

11 There were 267 students who had taken MATHS 161, and 53 students who had taken MATHS 165, prior to enrolling in ECON 221.

12 Students can take ECON 221 up to three times. Those who receive a W or F the first time will retake ECON 221 again unless they decide either to abandon the pursuit of a business or Economics degree or to take statistics at another university and transfer the requirement and unit credit (but not the grade) to BSU.
We are agnostic about the expected impact of Athlete or Female on ECON 221 student performance. Participation in intercollegiate athletics imposes considerable physical, mental and time demands on the athlete—thus, perhaps, $\beta_1 < 0$. But, at least one study has found that the extracurricular activities have no statistically significant effect on student performance, so, perhaps, $\beta_1 = 0$. And, finally, because the latest NCAA Academic Progress Report rates BSU’s athletic programs as among the best in the country for setting and achieving high academic standards, perhaps, at least at BSU, $\beta_1 > 0$.

Studies that include gender as an explanatory variable have shown that females perform better than males or no better (no differently) than males. More recently, Wall Street Journal articles have reported that women score well on the quantitative portion of the GMAT and, as a result, are underrepresented in graduate business programs. So, because Athlete and Female have appeared in several studies with varying results, we have no maintained hypotheses about the likely signs or statistical significance of their estimated coefficients on student performance in ECON 221.

**The Results**

“Given modern computers, from a practical perspective the most difficult aspect of logit and probit models is presenting and interpreting the results. The coefficient estimates, their standard errors, and the value of the log-likelihood function should be reported in any application. The coefficients give the signs of the partial effects of each $x_j$ on the response probability.”

We estimated equation (1) using two different statistical models: an OLS model (Table 2) and an ordered probit model (Tables 3 and 4). Although probit (or logit) model estimation is frequently used in studies like this, we have included the OLS results for two reasons. First, and most importantly, the results are very similar to the probit results in terms of the expected signs and statistical significance of the estimated coefficients for the math prerequisite variables. Second, as Wooldridge notes, OLS results are easier to explain and interpret when compared to interpreting probit estimation. For these reasons, we have included both the OLS and probit estimation results here.

The full and reduced model estimations in Tables 3 - 5 strongly support our maintained hypotheses about the effects of the alternative math prerequisites on student performance in ECON 221. The estimated coefficients for $D_{161}$ and $D_{165}$ are both positive and statistically significant. ECON 221 students who took either MATHS 161 or 165 had higher (or the probability of higher) grades than those who took the MATHS 131/132 sequence. Similarly, as we anticipated, the estimated coefficient for $D_{131}$ is negative and statistically significant. ECON 221 students who took only MATHS 131 received lower grades than those who took the MATHS 131/132 sequence.

As expected, the grade received in the prerequisite math course, MATHS Grade had a positive and significant effect on ECON 221 student performance. Students with better math skills, as measured by their course grade in whatever math course they used as the prerequisite, performed better in ECON 221. And, again as expected, the estimated impact of Lag GPA (the previous semester’s cumulative grade point average) was positive and significant. This result shows that better students in general, as measured by their overall academic performance, also perform better in ECON 221.

The estimated impact of Female indicates that women earn higher grades in ECON 221. Our study also indicates that participation in collegiate sports, Athlete, has no statistically significant impact on student performance in ECON 221.

The results in Tables 2-4 support our general view that the choice of math prerequisites matters significantly for student performance in business statistics. Consequently, suggested changes in these math prerequisites should be carefully assessed to make certain that they have no detrimental effects on student performance. The well-known Hippocratic Oath

---

13 Brasfield et al.
15 Pritchard et al.
16 Brasfield et al.; Von Allmen
17 Alsop.
18 Wooldridge (2003), pp. 559-60.
19 We also estimated our model using binomial probit and logit and ordered logit estimation. The results of these model estimations are closely similar to those reported in the text.

Please contact one of the authors if you are interested in seeing these results.
that medical doctors take should also apply to faculty committees who are considering changes in prerequisite courses for any course or program: “First, do no harm.”

Previous Studies of the Impact of Math Prerequisites/Courses on Student Performance

How do our results compare with those of other studies? Although the literature on factors that influence student performance in various business courses is extensive, only a few of these studies, unfortunately, explicitly examine the impact of specific math courses and/or prerequisites.

Brasfield et al. (1992) use an ordered probit model to examine the influence of a number of variables, including 0-1 dummy variables for completion of two math courses, MA120 and MA220, Mathematics with Applications to Business I and II, respectively, on student performance in the microeconomics and macroeconomics courses. The content of MA120 and MA 220 are similar to MATHS 131 and 132. Business students are required to complete MA220 and the two economics courses for their Bachelor’s degree. However, there are no math prerequisites for either economics course. Consequently, Brasfield et al. were able to examine the differing impacts of completing neither math course, MA120 only or MA220 prior to enrolling in either or both economics courses on student performance in the economics courses. They found that

“…MA220 was positive and significant in both … regression(s). … However, results … indicate that students who had completed only MA120 did not do significantly better than those students who had not yet completed that course. Combined, these results lead to the conclusion that completion of the math sequence is desirable before students register for either economics class.”

Brasfield et al.’s results on the differential impact of MA120 vs. MA220 on student performance in economics courses closely matches our results for the relative impact of MATHS 131 vs. MATHS 132 on student performance in ECON 221.

Von Allmen (1992) used an ordered probit model to assess the impact of selected variables on student performance in intermediate microeconomics. The intermediate microeconomics course has a calculus prerequisite which students can meet by taking MA170 (taken by students who score better on the math test given to matriculating students) or the MA 106/166 two-course sequence. Von Allen uses the grade received in the prerequisite math course (MA 166 or MA 170) and a 0-1 math dummy (DUM166). He finds that

“The grade in calculus … has a positive significant effect on performance in microeconomics. … The lack of significance of the DUM166 … supports the hypothesis that students with weak skills can be brought up to the level of their better prepared or more able counterparts.”

Von Allmen’s finding that the grade earned in their prerequisite math course has a positive significant on student performance in economics courses is the same result that we found for students in our business statistics course. However, in contrast to Von Allmen’s study, we found that students who took the higher level math classes (MATHS 161 and 165) received higher grades than those who took the two-course MATHS 131/132 sequence. We plan to extend our research to see whether his results or ours hold for student performance in ECON 201 and 202.

Summary

Math prerequisites for business courses are intended to provide students with the necessary skills to more easily learn the material covered in the course. However, it appears that very little analysis has been applied to assess which math prerequisites are appropriate. Consequently, there is virtually no way to know if the screening process is working as anticipated or, for that matter, could be improved.

Recently, the MCOB decided to reduce the math prerequisites for business statistics, thus providing one of those rare opportunities that enable researchers to actually see how student performance is affected by such changes. Even more interestingly, during the transition period from the previous regime (two three-unit courses) to the new one (a four-unit course), students were offered a brief two-semester period in which the math prerequisite was reduced to a single three-unit course.

In this study, we show that the sharp, albeit temporary, reduction in math prerequisites from two three-unit courses to one three-unit

---

20 Brasfield et al., pp. 243-4

21 Von Allmen, p. 20.
course had a significant negative impact on student performance in business statistics. Because this specific change was, at one time, under consideration to be the one that the MCOB would adopt, our findings have a more general application than this brief episode might suggest. Apparently, some changes in math prerequisites are actually hazardous to students and faculty of business statistics courses. It would be nice to know which ones they were before they were inadvertently adopted.

References


Wooldridge, Jeffrey M. (2003), Introductory Econometrics: A Modern Approach, 2e (Thomson/South-Western)

Table 1. Business Statistics: Past and Present Mathematics Prerequisites

<table>
<thead>
<tr>
<th>ECON 221 Business Statistics</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to various statistical and probabilistic concepts and techniques with application to business problems including random variables and probability distributions, measures of central tendency and dispersion, testing of hypotheses, simple linear regression, and correlation</td>
<td></td>
</tr>
<tr>
<td><strong>Current Prerequisites:</strong> (Beginning in Spring 2005) MATHS 135; sophomore standing; demonstrated proficiency in computer skills.</td>
<td></td>
</tr>
<tr>
<td><strong>Previous Prerequisites</strong> (Fall 2004 only): MATHS 131 or its equivalent. Because MATHS 132 was last offered in Spring 2004, students who had passed MATH 131, but had not passed</td>
<td></td>
</tr>
</tbody>
</table>
MATH 132, were permitted to enroll in ECON 221 in Fall 2004.

Previous Prerequisites (Up through Summer 2004): MATHS 132 or its equivalent (MATHS 161 or 165). Not open to students who have credit in MATHS 181.

MATHS 131 Finite Mathematics for Business. (3)
Topics in mathematics particularly suited to the needs of business majors, including mathematics of finance, probability, matrix algebra, and linear programming. This course, which does not appear in the 2004-2006 university catalog, was last taught in Summer 2004.

Previous Prerequisite: MATHS 108 or an appropriate score on the mathematics placement test.

MATHS 132 Brief Calculus for Business. (3)
Brief survey of differential and integral calculus. Emphasizes applications to business. This course, although listed in the 2004-2006 university catalog, was last taught in Spring 2004.

Current Prerequisites: MATHS 135
Previous Prerequisites: MATHS 131

MATHS 135 (131) Mathematics for Business. (4)
Topics in mathematics particularly suited to the needs of business majors, including mathematics of finance, probability, and differential calculus. This course, which covers selected topics previously included in MATHS 131 AND 132, is a new course that was first taught in Fall 2004.

Prerequisites: MATHS 108 or an appropriate score on the mathematics placement test.

MATHS 161 Applied Calculus 1. (3)
Discussion of limits, derivatives, differentials, and appropriate applications. The definite integral, area, fundamental theorem of calculus, indefinite and improper integrals.

Prerequisites: MATHS 112 or permission of the department chairperson. Not open to students who have credit in MATHS 165.

MATHS 165 Calculus 1. (4)
Differential calculus of algebraic and transcendental functions and applications, antidifferentiation and the Riemann integral. Includes the use of graphing calculators and computer software.

Prerequisites: MATHS 111, 112; or sufficient background in algebra and trigonometry as evidenced by the student’s high school record, SAT/ACT scores, and/or score on the mathematics placement test.

MATHS 108 Intermediate Algebra (3)
Review of factoring, quadratic equations and inequalities, relations and functions, rational exponents, systems of linear equations, and exponential and logarithmic functions.

TABLE 2: Variable Symbols and Definitions

<table>
<thead>
<tr>
<th>Dependent Variables:</th>
</tr>
</thead>
<tbody>
<tr>
<td>E221 Grade: The numeric grade (F = 0, D- = .7, D = 1, …, A- = 3.7, A = 4) earned by the student the first time he/she enrolled in ECON 221. [OLS regression analysis].</td>
</tr>
<tr>
<td>DE221 Grade: A dummy variable (F = 0, D- = 1, D = 2, …, A = 12) for the grade that the student received the first time that he/she enrolled in ECON 221. [Ordered Probit analysis]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Independent Variables:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athlete: A 0-1 dummy variable identifying whether a student competes in intercollegiate athletics (Athlete = 1) or not (Athlete = 0).</td>
</tr>
<tr>
<td>Female: A 0-1 dummy gender variable identifying whether a student is a female (Female = 1) or not (Female = 0).</td>
</tr>
<tr>
<td>MATHS Grade: The numeric grade (F = 0, D- = .7, … A- = 3.7, A = 4) that the student received in his/her prerequisite math class (Either MATHS 132, 161 or 165. MATHS 131 only for the “special exemption” students starting in Fall 2004).</td>
</tr>
<tr>
<td>Lag GPA: The student’s cumulative GPA in the semester prior to his/her first enrollment in ECON 221.</td>
</tr>
<tr>
<td>D131: A 0-1 dummy variable identifying whether the student took MATHS 131 as the prerequisite math course for ECON 221 (D131 = 1) or not (D131 = 0). This applies only for students who were caught in the math prerequisite changes starting in Fall 2004.</td>
</tr>
<tr>
<td>D161: A 0-1 dummy variable identifying whether the student took MATHS 161 as the</td>
</tr>
</tbody>
</table>
prerequisite math course for ECON 221 (D161 = 1) or not (D161 = 0).

D165: A 0-1 dummy variable identifying whether the student took MATHS 165 as the prerequisite math course for ECON 221 (D165 = 1) or not (D165 = 0).

**TABLE 2: OLS Model Estimation**

**Dependent Variable:** E221 Grade  
**Method:** Least Squares  
**Included observations:** 1576

<table>
<thead>
<tr>
<th>Variable</th>
<th>General Model</th>
<th>Reduced Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATHLETE</td>
<td>-0.117569</td>
<td>1.290228</td>
</tr>
<tr>
<td>FEMALE</td>
<td>0.107327</td>
<td>2.009306</td>
</tr>
<tr>
<td>MATH Grade</td>
<td>0.231735</td>
<td>6.809915</td>
</tr>
<tr>
<td>LAG GPA</td>
<td>1.147823</td>
<td>17.36005</td>
</tr>
<tr>
<td>D131</td>
<td>-0.152638</td>
<td>-2.345935</td>
</tr>
<tr>
<td>D135</td>
<td>0.164545</td>
<td>1.125509</td>
</tr>
<tr>
<td>D161</td>
<td>0.244054</td>
<td>3.515296</td>
</tr>
<tr>
<td>D165</td>
<td>0.332301</td>
<td>2.358162</td>
</tr>
<tr>
<td>C</td>
<td>-1.646351</td>
<td>-10.92447</td>
</tr>
</tbody>
</table>

Ad. R-sq 0.393802 0.393412  
Log likelihood -2195.197 -2196.708  
Schwarz crit. 2.827828 2.820402  
DW stat 2.048386 2.046897
Table 3. Ordered Probit Estimation: Full Model

Dependent Variable: DE221 Grade
Method: ML - Ordered Probit
(Quadratic hill climbing)
Number of ordered indicator values: 12
Convergence achieved after 5 iterations
Covariance matrix computed using second derivatives

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATHLETE</td>
<td>-0.136006</td>
<td>0.098516</td>
<td>-1.380539</td>
</tr>
<tr>
<td>FEMALE</td>
<td>0.111451</td>
<td>0.057914</td>
<td>1.924424</td>
</tr>
<tr>
<td>MATH Grade</td>
<td>0.260387</td>
<td>0.036707</td>
<td>7.093604</td>
</tr>
<tr>
<td>LAG GPA</td>
<td>1.271849</td>
<td>0.074589</td>
<td>17.05133</td>
</tr>
<tr>
<td>D131</td>
<td>-0.147498</td>
<td>0.070399</td>
<td>-2.095175</td>
</tr>
<tr>
<td>D135</td>
<td>0.181698</td>
<td>0.15803</td>
<td>1.149772</td>
</tr>
<tr>
<td>D161</td>
<td>0.295855</td>
<td>0.076372</td>
<td>3.873875</td>
</tr>
<tr>
<td>D165</td>
<td>0.322375</td>
<td>0.15226</td>
<td>2.117268</td>
</tr>
</tbody>
</table>

Limit Points

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LIMIT_1:C(9)</td>
<td>2.823677</td>
<td>0.172594</td>
<td>16.36018</td>
</tr>
<tr>
<td>LIMIT_2:C(10)</td>
<td>2.835927</td>
<td>0.172601</td>
<td>16.43051</td>
</tr>
<tr>
<td>LIMIT_3:C(11)</td>
<td>3.431125</td>
<td>0.174714</td>
<td>19.63856</td>
</tr>
<tr>
<td>LIMIT_4:C(12)</td>
<td>3.522349</td>
<td>0.175248</td>
<td>20.09918</td>
</tr>
<tr>
<td>LIMIT_5:C(13)</td>
<td>3.616703</td>
<td>0.175836</td>
<td>20.56863</td>
</tr>
<tr>
<td>LIMIT_6:C(14)</td>
<td>4.29843</td>
<td>0.18108</td>
<td>23.73769</td>
</tr>
<tr>
<td>LIMIT_7:C(15)</td>
<td>4.426172</td>
<td>0.182081</td>
<td>24.30886</td>
</tr>
<tr>
<td>LIMIT_8:C(16)</td>
<td>4.552175</td>
<td>0.183194</td>
<td>24.849</td>
</tr>
<tr>
<td>LIMIT_9:C(17)</td>
<td>5.298604</td>
<td>0.191186</td>
<td>27.71447</td>
</tr>
<tr>
<td>LIMIT_10:C(18)</td>
<td>5.48822</td>
<td>0.193189</td>
<td>28.4086</td>
</tr>
<tr>
<td>LIMIT_11:C(19)</td>
<td>5.675345</td>
<td>0.195528</td>
<td>29.02576</td>
</tr>
</tbody>
</table>

Akaike info criterion | 3.744004 | Schwarz criterion | 3.808655
Hannan-Quinn criter. | 3.768029 |
Log likelihood | -2931.275 | Avg. log likelihood | -1.85995
Restr. log likelihood | -350.773 | LR index (PseudoR2) | 0.125194
LR statistic (8 df) | 838.9964 | Prob. (LR stat) | 0
Table 4. Ordered Probit Estimation: Reduced Model

Dependent Variable: DE221 Grade  
Method: ML - Ordered Probit (Quadratic hill climbing)  
Included observations: 1576  
Number of ordered indicator values: 12  
Convergence achieved after 5 iterations  
Covariance matrix computed using second derivatives

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE</td>
<td>0.117813</td>
<td>0.057768</td>
<td>2.039429</td>
<td>0.0414</td>
</tr>
<tr>
<td>MATH Grade</td>
<td>0.264132</td>
<td>0.036594</td>
<td>7.217856</td>
<td>0</td>
</tr>
<tr>
<td>LAG GPA</td>
<td>1.259455</td>
<td>0.074145</td>
<td>16.9864</td>
<td>0</td>
</tr>
<tr>
<td>D131</td>
<td>-0.15329</td>
<td>0.069944</td>
<td>-2.19163</td>
<td>0.0284</td>
</tr>
<tr>
<td>D161</td>
<td>0.287185</td>
<td>0.076004</td>
<td>3.77857</td>
<td>0.0002</td>
</tr>
<tr>
<td>D165</td>
<td>0.317696</td>
<td>0.151995</td>
<td>2.090181</td>
<td>0.0366</td>
</tr>
</tbody>
</table>

Limit Points

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIMIT_1:C(7)</td>
<td>2.803659</td>
<td>0.172133</td>
<td>16.28772</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_2:C(8)</td>
<td>2.815877</td>
<td>0.172139</td>
<td>16.35817</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_3:C(9)</td>
<td>3.409861</td>
<td>0.174205</td>
<td>19.57383</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_4:C(10)</td>
<td>3.500904</td>
<td>0.174733</td>
<td>20.03578</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_5:C(11)</td>
<td>3.595102</td>
<td>0.175315</td>
<td>20.50654</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_6:C(12)</td>
<td>4.276418</td>
<td>0.180556</td>
<td>23.68472</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_7:C(13)</td>
<td>4.404236</td>
<td>0.181564</td>
<td>24.25727</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_8:C(14)</td>
<td>4.530377</td>
<td>0.182687</td>
<td>24.79857</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_9:C(15)</td>
<td>5.276775</td>
<td>0.190701</td>
<td>27.67037</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_10:C(16)</td>
<td>5.466053</td>
<td>0.192696</td>
<td>28.36615</td>
<td>0</td>
</tr>
<tr>
<td>LIMIT_11:C(17)</td>
<td>5.652707</td>
<td>0.195019</td>
<td>28.98545</td>
<td>0</td>
</tr>
</tbody>
</table>

Akaike info criterion 3.74357 Schwarz criterion 3.801416  
Log likelihood -2932.93 Hannan-Quinn criter. 3.765066  
Restr. log likelihood -3350.77 Avg. log likelihood -1.861  
LR statistic (6 df) 835.6805 LR index (Pseud-R2) 0.1247  
Prob. (LR stat) 0
Training Environmental Statisticians – Tomorrows Problem Solvers

William F. Hunt, Jr., Dr. Kimberly Weems, Dr. William Swallow and Andrew Moore
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203

Dr. Nagambal Shah and Dr. Monica Stephens
Spelman College
Atlanta, GA 30314-4399

ABSTRACT

How could a win-win strategy be used to train young people in environmental statistics and at the same time analyze environmental data for Federal, State and local agencies, that have not been analyzed until now? This paper will discuss a course that has been developed to train undergraduate students in environmental statistics and the impact the course has had on the students, the clients and the university. This training comes in support of a National Science Foundation Grant, Collaborative Research: Training Environmental Statisticians Using Complicated Data Sets to Make More Informed Environmental Decisions. Currently, a collaborative effort is being undertaken with Spelman College, a historically black college for women in Atlanta, Georgia. This collaborative effort will demonstrate that this approach is portable to other universities and colleges with an undergraduate statistics program and at those without, as long as there are some courses in statistics and a statistician with an interest in environmental statistics. The intent of this collaborative effort is to adapt, modify and enhance the Environmental Statistics Practicum Program, which was developed at NC State University. The collaborators are currently implementing and adapting the environmental statistics program at Spelman College, which represents those colleges without a formal undergraduate statistics program. In summary, the classes have created a win-win situation for the students, the clients and the university and provide an alternative way to complete environmental data analysis. Examples of the students work will be presented in the paper.

INTRODUCTION

For many years, the issues of how best to develop statistical partners in academe, industry and government have been raised (see Lynne Hare’s bibliography: Statistical Partners in Academe, Industry and Government – Reference Literature on the web site: http://web.utk.edu/~wparr/hareexpect.html). These 77 references emphasize the importance of training statisticians to meet the needs of industry and government. All levels of government tend to collect large quantities of data, which largely go unanalyzed. This is particularly true with the collection of environmental data.

Cobb (1993) has proposed that educators should (a) place less emphasis on mathematics and more emphasis on data analysis, (b) increase the use of data sets from domains recognizable to students and (c) to learn by doing. At North Carolina State University (NCSU), the benefits of coupling education and research have been described by anecdotes such as "...seemingly mediocre students caught fire after being involved in meaningful research.” More authoritative support for research-based learning at the undergraduate level has been set forth in the Boyer Report (see Boyer, 1996). The BMS/NRC study reported in Educating Mathematical Scientists: Doctoral Study and the Postdoctoral Experience in the United States (1992) cite the benefits of early involvement in research. In recent years, research universities have been criticized for ignoring their educational mission in favor of their research mission. The coupling of education and research, at least in the major field, turns a disadvantage into an advantage.

How can a win-win-win situation be created that would benefit students, faculty and government agencies and in the process encourage undergraduate students to pursue advanced degrees...
or careers in environmental statistics? A proof of concept study has been undertaken at North Carolina State University with the introduction of two new courses entitled Environmental Statistics Practicum and Special Topics in Environmental Statistics (see Hunt, 2000', Hunt, 2001', Hunt, 2002" and Hunt, 2004'). The objectives of the environmental statistics courses are: (1) to provide a consulting opportunity for the students with Federal, State or local environmental agencies; (2) focus on the application of the student’s technical skills to a real problem; (3) have the students gain consulting experience; and (4) develop their oral and written communication skills.

The students learn how to prepare a final report, brief clients at the client’s office, present poster papers at technical conferences, and write papers for publication. Students have done work for seventeen client organizations: (1) the Southern Oxidant Study at North Carolina State University (NCSU); (2) the U. S. Environmental Protection Agency’s (USEPA) National Exposure Research Laboratory; (3) the USEPA’s Office of Air Quality Planning and Standards; (4) the USEPA’s Office of Environmental Information in Washington, DC; (5) the Air Division of the North Carolina Department of Environment and Natural Resources (NCDENR); (6) the Forsyth County Environmental Affairs Department in Winston Salem, NC; (7) the U. S. Department of State; (8) Environment Canada; (9) the University of Texas; (10) the Texas Commission on Environmental Quality; (11) the Water Division of the North Carolina Department of Environment and Natural Resources (NCDENR); (12) the North Carolina State Climate Office; (13) the USEPA Region 4 Office in Atlanta; the (14) The Georgia Department of Natural Resources; (15) Mid-Atlantic Regional Air Management Association; (16) the New Jersey Department of Environmental Protection; and (17) RTP Environmental Inc. In addition to briefing their clients and providing the client’s with final reports, they have presented papers at 44 professional meetings, research symposia, etc.

This work has addressed several critical areas:

1. The need to train undergraduates in analyzing important complicated and messy data sets.
2. The National, State and international need to analyze environmental data to make better environmental policy decisions.
3. The need to encourage students to pursue graduate degrees in statistics, keeping people in the pipeline to pursue PhDs.
4. The need to analyze real data for real clients in the workplace and make the student a desirable candidate for employment upon graduation.
5. The need to foster collaboration between majority and minority institutions and contribute to diversity in environmental decision making.
6. The need to develop statistical partners in academia, government, and industry.

**BACKGROUND**

There is strong evidence that information obtained through problem solving is better learned than information simply learned by rote (Adams, Kasserman, Yearwood, Perfetto, Bransford & Franks, 1988); Lockhart, Lamon & Gick, 1988). Conway, Cohen and Stanhope (1991) showed that memory for materials learned in a research methods course was enhanced by having students design and conduct experiments rather than passively learn about course content. The objective of this course is to have the students engage in active communication with their clients, understand the client’s needs and then conduct an exploratory analysis of the client’s data to answer the client’s questions. The students visit an air monitoring site location to see how air monitoring data are collected and an air monitoring laboratory to see how the data are analyzed in the laboratory, as well. After the site and laboratory visits, the students focus on possible causes of uncertainty in the data collection process. The students learn how to apply their statistical and data analytic skills to the data and to present their results by developing their speaking and writing skills. They interact several times over the course of the semester with their clients before they make their final presentations to their clients. The students have accomplished a great deal in this course. In addition to briefing their clients and providing the client’s with final reports, they have participated in 22 professional meetings, research symposia, etc. The meetings they participated in are:

1. **The Southern Oxidant Study Data Analysis Workshop**, Research Triangle Park, NC, March 9, 2000;
2. **NCSU Undergraduate Research Symposium**, McKimmon Center, Raleigh, NC, April 27, 2000;
3. USEPA Technical Workshop on PM \(2.5\) Monitoring, Quality Assurance, and Data Analysis, Cary, NC, May 22-25, 2000;


5. Future Directions in Air Quality Research, Ecological, Atmospheric, Regulatory/Policy and Educational Issues, Research Triangle Park, NC February 12, 2001;


11. NCSU Undergraduate Research Symposium, McKimmon Center, Raleigh, NC, April 18, 2002.


23. OPT-ED Alliance Day, Raleigh, NC, Sept. 24, 2004

24. South Atlantic States Section (SASS) of the AWMA Annual Meeting, Virginia Beach, VA Nov. 4-5, 2004

25. Triangle Undergraduate Research Symposium, NC State University, Raleigh, NC, Nov. 6, 2004.


27. Meredith College: Mathematical Association of America, Southeastern Section, 84th Annual Meeting, Raleigh, NC, March 11-12, 2005


29. Capital Research Day, North Carolina State Legislature, Raleigh, NC, April 12, 2005

30. USEPA Earth Day Celebration, Research Triangle Park, NC, April 21, 2005

31. 14th Annual NC State Undergraduate Research Symposium, McKimmon Center, April 28, 2005.


33. Fourth Annual NC State Undergraduate Summer Research Symposium, Raleigh, NC, August 4, 2005.

34. Alumni and Friends Weekend, College of Physical and Mathematical Sciences, NC State University, Raleigh, NC, September 23, 2005.

35. AWMA Specialty Conference, Environmental Data Analysis – Assess Health and Environmental Impacts, Developing Policy, and Achieving
After seven years, twelve students have graduated with a master’s degree in statistics and seven are continuing on for a Ph.D. Twenty-four students have gone onto graduate school programs in statistics. Eight students are employed at the Research Triangle Institute as environmental statisticians and fourteen students have worked part time at the USEPA as statisticians. The students have given 163 professional presentations and have written almost as many papers and reports.

### COURSE DESIGN

#### Lecture Segments

Lectures occupy the majority of class time in college courses (Mathie, et al., 1993). The approach used in this course is to provide lecture material for the students to increase their understanding of the environmental issues that they will encounter when working on the problems for their clients. Team training is also an important component of the course material. Peters and Waterman (1982) state that one of the key principles practiced, by “excellent” companies, was strong employee participation. Bradford and Raines (1993) show that today’s young workers are learning new management models in college that give more power to front-line employees and flatten the organizational chart. Team training is an important aspect of the new workplace. Kelly (1996) defines a team as a small number of people with complimentary skills who are committed to a common purpose, set of performance goals, and an approach for which they hold themselves mutually accountable. The lessons with team training make use of material provided by Whitney (1996), whose work was done for the USEPA’s Office of Air Quality Planning and Standards. Statistical methods and the types of presentations used in environmental data analysis are an important part of the lecture material. The availability of written material in both handout form and its availability on the web page allows students to concentrate on spoken lectures without being troubled by the need to take notes. The lectures occupy approximately half of the class time. Field trips to air monitoring sites and laboratories are incorporated into the course so that the student can see how the data are collected and to better understand the possible sources of error associated in data collection. The students meet their clients at the client’s place of business. This is done to reinforce the importance of the analysis that they are undertaking for the client. The students’ briefings and reports examine data that have not been analyzed before. The students present their findings in a number of ways – formal briefings (30 minutes) with questions and answers, abridged briefings (10 to 15 minutes) with questions and answers, and poster presentation briefings (2 minutes) with questions and answers. The students are trained using traditional approaches - overheads and handouts. There has been some use of the multimedia projectors with presentations written in PowerPoint. They also learn to give briefings using handouts in a conversational mode. For long distance clients, such as Environment Canada, the
students learn how to give briefings using teleconferencing or via conference call.

Several initial homework assignments are used for the students’ first problem. The initial homework assignment could include an examination of lead air monitoring data. Here the students examine the policy implications associated with interpreting the results of an analysis of variance of lead monitoring data (Hunt, 1984\textsuperscript{xvi}). Alternatively, a simple data collection effort could be undertaken to test a hypothesis. Both types of problems foster teamwork, dividing parts of the problem among team members and providing a baseline measure of the students’ ability to conduct a briefing. The second type of problem is illustrated with the fall 2001 class, which was asked to examine the question (a positive environmental effect): “Are the students, faculty and alumni of NCSU more likely to drive a red car, because they are associated with NCSU, than is the general public?” This sample exercise went very well with the students dividing themselves into three teams. Each team addressed the question for the students, faculty and staff or the alumni. A student reporter for the student newspaper, The Technician, was invited to the student presentations. Their work was so well received that it was reported in the student newspaper, The Technician\textsuperscript{xxvii} and carried on the university wire service. Additional examples will be developed as part of the course materials.

**Statistical Methods**

The students employ classical statistical methods to their analysis of the data. Smith (1998\textsuperscript{xviii}) found very positive effects of incorporating active-learning strategies in his classes. The activities allowed students to learn about statistics by getting first-hand experience in conducting statistical analysis. By designing studies, collecting and analyzing data, and preparing written reports, students come to see the importance of statistics and gain interest in and excitement about examining whether the results support their predictions and what they reveal about the data that the students have analyzed. The Environmental Statistics Practicum takes this one step further by having the students work with real clients, analyze the client’s data, interpret it, brief the client on the results of their work and write a written report. Often their work is of such a high quality that the students prepare poster presentations for the NCSU Undergraduate Research Symposium and make presentations before professional society and technical meetings. The students employ many different statistical methods depending upon the nature of the question being raised. The problems could involve forecasting (Harrington, 2000\textsuperscript{xxix}, Woodside, 2001\textsuperscript{xv}, Donaghy and Sorrell, 2003\textsuperscript{xxv}, Lawhorn and Ridenhour 2004\textsuperscript{xxvii} and Gracien and Hare, 2005\textsuperscript{xxxviii}). Three of these papers by students - Daric Harrington, Karen Donaghy and Courtney Sorrell- received cash awards for best papers at the NCSU Undergraduate Research Symposium. Kathy Woodside received the award for best paper at The Mathematical Association of America and Pi Mu Epsilon Mathfest 2001 Meeting in Madison, Wisconsin. Karen Donaghy and Courtney Sorrell extended their research by responding to a question raised by Environment Canada. Instead of using a 24-hour measurement of PM fine, they used a daily 3-hour maximum average. Environment Canada wanted to know if it would be easier to predict a daily 3-hour maximum average. They greatly improved the ability to forecast PM fine by developing models to forecast both weekend and weekday by season. They improved the accuracy to 81 percent, lowered the false alarm rate to 19 percent, increased the critical success index to 57 percent and increased the probability of detection to 70 percent. Each student received a $2000 Undergraduate Research Award from North Carolina State University and continued their work in the spring semester of 2003.

The students examine data using scatter plots, correlation analysis, regression analysis, analysis of variance, etc. (McMichael and DeFrancis, 2000\textsuperscript{xxxv}, Madsen, Copeland and Crotty, 2000\textsuperscript{xxv}; Cason, Clarke and Ness, 2001\textsuperscript{xxvi}, Bartz and Woodside, 2001\textsuperscript{xxvii}, Hayden, Riley and Dail, 2003\textsuperscript{xxvii}, Madsen 2005\textsuperscript{xxxiii}, Holland, 2006\textsuperscript{xxxv}). Other statistical methods examining spatial patterns and trends are also being used (Copeland, 2001\textsuperscript{xxviii}, Crotty, 2001\textsuperscript{xxviii}, Thomas, Brooker and Cheng, 2001\textsuperscript{xxviii}, Ogorek, 2003\textsuperscript{xxiv}, Darlington and Currier, 2004\textsuperscript{xxv} and Stines, 2005\textsuperscript{xxxvii}). Brian Copeland in a project for the USEPA demonstrated that standard conditions of temperature and pressure resulted in a positive bias in air toxics measurements in the Western States when compared with using local conditions of temperature and pressure. Based upon his work the USEPA changed its requirements so that air toxic data is now reported in local conditions of temperature and pressure.

Another analysis conducted by Schnell, Gabig and Spruell (2002\textsuperscript{xxvii}) provided a basis for changing the fine particulate matter standard to a more
A question rose in a workshop held by Environment Canada, "Towards a Canadian Air Quality Index" that the students answered for Environment Canada: “Could a three hour standard be used instead of a 24-hour standard?” In order to do this there must be a correlation between different particulate matter indicators – the daily maximum 3 hour average and the daily 24 hour average. The students answered this question in the affirmative in the analysis they did for Environment Canada. (See Figure 1).

Figure 1. Comparison of PM$_{2.5}$ daily maximum 3 hour average with the 24-hour average in Kitchener, Ontario, Canada.

Often the students produce new and novel approaches to solving environmental problems. Three student projects involving novel approaches have been published (Camalier, Yoshimoto and Stines, 2004; White, 2005 and Hornsby and Jackson, 2006). One technique was adapted by Brian Stines under the direction of Dr. David Dickey. Brian was able to create a three dimensional video showing how the number of malfunctions at a chemical plant changed over time. The data are displayed in Figure 2. From the graph, you can see that the numbers of malfunctions are a result of facility downwind of the Clinton site. Over time the malfunctions are being reduced along with their magnitude. The video takes 10 seconds to show what’s happening and is a distillation of tens of thousands of ambient 1,3 butadiene measurements. A picture is worth a thousand words!

Three students – Jeffrey Thomas, Darious Brooker and Ho Ling Cheng went on to win an undergraduate research award for their work. Each student received a $2000 scholarship in the Spring 2002 semester from NCSU to continue with his or her statistical analysis of the Toxic Release Inventory data. Quality control techniques are also used to examine problems (Gallins, Stidham and Bartz, 2001). The report prepared by students – Paul Gallins, Sam Stidham and Janet Bartz was used to change the quality control procedures used in the chemical analysis of volatile organic compound data under by the NCDENR. In 2002-2003, Karen Donaghy and Courtney Sorrell each won the $2000 scholarship. In 2003 and 2004, fifteen students each won the $500 Undergraduate Research Award for their projects. In 2005, eight students each received a $750 Undergraduate Research Award. Since 2000, the students have won over $33,000 for their projects.

Prototype

A prototype of the materials that have been developed can be viewed at: http://www.stat.ncsu.edu/~st495a_info/.

SPELMAN COLLEGE COLLABORATION

Spelman College has both a course, taught during the Spring Semester, and a summer research institute, called the Spelman Environmental Statistics Summer Institute (SSESI) that exposes students to work with environmental data. Spelman has taken a slightly different approach from NCSU. The Spelman approach seeks to expose students to a highly interdisciplinary research environment in which students use statistical analysis to analyze environmental data, but also explore issues such as...
environmental justice, impacts of environmental air quality on communities and health, and the often differing perspectives of academia, industry, and the community on environmental air quality.

Spelman does not have a formal department in statistics although statistics is taught in the mathematics department and several other departments including psychology and economics. Hence, the students that participate in the SESSI program are required to have either credit for an undergraduate-level course in statistics or to have participated in the spring course, called the Environmental Statistics Practicum, the semester prior to their participation in SESSI. During the spring course students learn how to use the SAS statistics software package, gain an understanding of environmental air quality, air quality measurement, environmental justice, and health impacts. They gain practical use of the skills they have learned by writing and presenting a report of their analysis of an environmental dataset.

SESSI is a six-week intensive research experience where students from varying academic backgrounds work on teams to analyze environmental data provided by the Georgia Department of Natural Resources (Georgia DNR) and the Environmental Protection Agency (EPA), Region IV who serve as clients for the students.

The objectives of SESSI are to:

• Involve students in interdisciplinary research on the undergraduate level and have them make a valuable contribution.
• Introduce students to statistics and to the application of statistical methods to “real world” problems.
• Introduce students to the analysis of “real world” environmental problems.
• Encourage students to pursue advanced degrees in either mathematics/statistics or in some other area of environmental research.

During the SESSI program, the students meet and consult with a client(s) and evaluate the needs of the client. At the end of the summer the participants present their results to the client and at appropriate symposia and technical and professional meetings. The students are paid a stipend of $2000.00 for their participation in the institute and are provided room and board from SESSI at Spelman College.

The SESSI program had its first session during the summer 2004 with six students. The SESSI participants included students with majors in mathematics, political science, and economics and ranged from freshman- to junior-level. Two groups of three students each were chosen (dependent on their academic background and experience) to work on two datasets, one provided by the DNR involving an industrialized section of Louisville Kentucky called “Rubbertown” and the other provided by the EPA involving air quality data for the Atlanta Metro area. The students have made several presentations to both the EPA and the Georgia DNR. The students have participated in the National Association of Mathematicians, Undergraduate MathFest and will present at other regional and national conferences. Che’ Smith, a student who participated in the spring course and who served as a mentor and teaching assistant for the SESSI program is currently completing an Honors Thesis on air quality data in the Atlanta Metro area. Her thesis is involves analyzing the air quality before and after September 11, 2001 (see Figure 3). The data were collected at the DeKalb air monitoring site downwind of the Atlanta Airport (Smith, 2005). Clearly, there is a significant reduction in benzene levels as a result of closing the airport on September 11, 12 and 13, 2001. The week before was not a normal week because of the Labor Day Holiday. When the average benzene for September 11, 12 and 13 is compared with the average benzene for September 18, 19 and 20 the impact of the airport closure can really be seen. Che’ has won several awards for her research and has made presentations at regional and national conferences.

Figure 3. Comparison of diurnal patterns in benzene at the DeKalb air monitoring site before, during and after September 11, 2001.
Spelman’s approach is to reach students who may not have a formal statistics background or a background in environmental studies. Hence, institutions that do not have statistics departments may use this approach to introduce its students to environmental data analysis and research. The materials that Spelman has developed can be easily disseminated, via the web, to these institutions. Much of the success of the Spelman program, however, relies on the expertise of the professionals at DNR and EPA who give their time to speak and interact with the students on a regular basis. The goal is to expose the students to research and statistics and ultimately to encourage them to pursue graduate work in environmental research or some other related environmental field.

Summer Workshop: a Winning Partnership for Training Environmental Problem Solvers

A joint North Carolina State University – Spelman College Workshop: A Winning Partnership for Training Environmental Problem Solvers was held at NC State University in Raleigh, NC on July 6-7, 2005 (Hunt, Shah, Stephens, and Weems, 2005). The workshop addressed several questions: Is there someone out there who has an interest to analyze the environmental data that are frequently collected by Federal, State and Local agencies? Are there some environmental problems and data sets out there that young enquiring minds would love to solve? Can we develop a win-win-win partnership among the environmental agencies and universities? Can such a partnership encourage the number of undergraduate students going to advanced degrees in statistics?

This workshop and the training partnerships are funded by a grant from the National Science Foundation. Sixty two people attended the workshop from environmental agencies, industries and other academic institutions interested in developing partnerships similar to the ones developed by NC State and Spelman. As a part of the grant from NSF, undergraduate students from NC State and Spelman are selected to work on various environmental problems. These students worked closely with scientists from several agencies who provided the relevant problems and data sets to work with. They worked in teams that included graduate students and faculty, in addition to the scientists at these agencies.

Dr. Ellis Cowling, University Distinguished Professor, College of Natural Resources, NCSU, gave the keynote address on the importance of statistical analysis. He made several important points pertaining to the work of Hunt, Shah, Weems and Stephens:

- While much money and professional time and energy are currently being spent by many different federal, state, and private sector organizations in the US and other countries collecting environmental data, much too little professional time, energy, and money is invested in analysis, interpretation, and dissemination of environmental data and information.

- He believes, together with Bill Hunt, Nagambal Shah, Kimberly Weems and Monica Stephens in both the programs at NCSU and Spelman College, that wider implementation of the Environmental Statistics Practicum idea could lead to a renaissance in what can be learned from environmental data that has already been collected and will be collected in the future.

- The Environmental Statistics Practicum idea developed under an NSF grant provides a win-win-win opportunity for all parties involved: undergraduate and graduate students, university faculty, and various federal and state government agencies and private-sector clients.

RECOMMENDATIONS AND CONCLUSIONS

This course demonstrates that a win-win-win situation can be created that benefits students, faculty and government agencies and in the process encourages undergraduate students to pursue advanced degrees and/or careers in environmental statistics. This prototype works. We believe that this course can be duplicated at other universities in partnership with government agencies. The work at Spelman College strongly suggests that a similar environmental statistics practicum can be developed at colleges without a formal undergraduate program in statistics. Similar courses could be constructed with other government agencies or private corporations. We are looking for interested partners that we can work with to implement this course. Government agencies tend to collect vast quantities of data that are rarely turned into information. This course is designed to turn data into information for the benefit of students, faculty and their clients and the general public!

All aspects of the course, lectures, data bases and student reports and papers will be freely available on the Internet.
REFERENCES

ii The Boyer Commission on Educating Undergraduates in the Research University, Reinventing Undergraduate Education: A Blueprint for America’s Research Universities, 1996.


xviii Technician (2001) “ Seeing Red: NCSU community more likely to drive red cars.”


xxiv Katina Gracien and Brian Hare, “A Comparison of AirNow and AQS Particulate Matter Databases,” AWMA Specialty Conference, Environmental Data Analysis – Assess Health and Environmental Impacts, Developing Policy, and Achieving Regulatory Compliance, Oak Brook, IL, October 4-5, 2005.


Copeland, B. (2001) “Standard Conditions of Temperature and Pressure vs. Local Conditions - What does it mean for air pollution control?” NCSU Undergraduate Research Symposium, McKimmon Center, Raleigh, NC.


http://www.ncsu.edu/undergrad-research/urj/


http://www.ncsu.edu/undergrad-research/urj/


Joint Distribution of Some Random Variables in Crossing of Two-Stage Erlang Processes

Mir Ghulam Hyder Talpur* and Iffat Zameer*

*Visiting Scholar Department of Mathematical Sciences, Ball State University, Muncie, IN, USA
*Department of Statistics, University of Sindh, Jamshoro, Pakistan.

ABSTRACT

In this study, the methodological techniques of z and L transform are used to obtain the joint probability generating function, joint probability density function and joint cumulative probability distribution function for three random variables namely $N_\xi$ (number of interarrivals (customers or units)), $N_\eta$ (number of vacations taken by servers) and $T_\xi$ (time taken by the servers during the vacation) by constructing an absorbing Markov process, where two sequences of the random variables $X_i$ and $Y_j$ are associated with 2-stages Erlang processes.

Key Words: Erlang process, Renewal process, Absorbing Markov process.

1 INTRODUCTION

Aki, Balakrishnan and Mohanty (1996), analyzed the waiting time for the first consecutive successes of a specified length and for sooner or later between successes run and a failure run with specified length respectively. There are two standard approaches to these problems. One is to solve a system of equations of conditional probability generating function. Then some characteristics such as probability functions and moments are derived from an expansion of the solution. The other is to use a typical sequence and split it into sub-sequences which could be interpreted to obtain the total characteristics of the distribution. They used the former approach to unify various approaches which had been attempted in order to extend the study of waiting time problems from the first order Markov dependent trials to the second order Markov dependent trials. The latter approach is used by Balakrishnan et al. (1993). Viveros et al. (1993) derived the binomial and negative binomial analogues under correlated Bernoulli trials.

This study was extended by Uchida (1998) by deriving the generalized probability generating function of the distributions of the waiting times until the rth occurrence among the events $\{E_i\}_{i=0}^{\infty}$. Koutras (1997) studied waiting time problem $W_{r,k}^{(a)}$ (a =I,II,III) by means of the method of finite Markov Chain Imbedding developed by Fu and Koutras (1993) and subsequently refined by Koutras and Alexandrou (1995), Fu (1996) extended this method for multistage trials and obtained the waiting time distribution of general compound pattern. Antzoulakos (1999) developed general technique for the study of waiting time problems $W_{r,k}^{(a)}$, $W_s$, $W_l$. Methodologically his technique was equivalent to the method of p.g.f. and by employing simple concepts of Markov chain theory, some general results are established which provide tools for the study of waiting time distributions. The statistical analyses extended the ideas to include waiting times for the occurrence of event. This is done by replacing the Laplace transform with mgf’s and incorporating probabilities into branches. Thus nodes and branches represented events and the waiting times for the occurrence of such events. Some authors provide an interdiction to flow graphs as used in engineering. The flow graph method bypass the traditional computations done for stochastic process, see (Taylor and Karlin 1994). The standard approach to analyze continuous time Markov chain involved solving the Chapman Kolmogrov equations for the generator matrix of the process. The generator was used to compute either the Laplace transform of the transition probabilities or the
probability generating functions of the process; it is also used by Talpur and Shi (1994). The main tool by all the aforementioned authors are combinatorial and/or suitable Markov chain embeddings. The applied nature of these problems is apparent. A routine derivation of explicit closed-form expression for joint moment generating functions associated with sooner and later waiting time in general discrete and continuous time model are performed. Stefanov (2000) used either finite – state Markov chain models or more generally finite-state semi-Markov process models with either discrete or continuous-time parameter.

### 1.1 PROBLEM DESCRIPTION

After going through the survey of the literature it has been found that renewal processes are widely used in reliability theory and models of queuing theory. The two theories are based on counting processes. It is in common practice that one has to deal with the situations where the difference between two or more counting processes is examined. The stochastic processes are found very helpful in analyzing such type of situations. Kroese (1992) showed the difference process of the two counting processes as

\[ D(t) = N_1(t) - N_2(t) \]

where \( N_1(t) \) and \( N_2(t) \) are two counting processes associated with corresponding renewal sequence of \( \{X_i\} \) and \( \{Y_j\} \). The problem considered in this study is the dual problem of the work of Kroese (1992) and the extension of the work of Talpur and Shi (1994). It is based upon the renewal sequence of two variables \( \{X_i\} \) and \( \{Y_j\} \)as shown in the fig-1

![Figure 1](image)

\[ \xi_N = \min \{n/T_n \geq S_N \} \]

where \( \xi_N \) is random variable. \( N \) is a constant and

\[
\begin{align*}
S_0 &= 0, S_N = X_1 + X_2 + \ldots + X_n \\
T_0 &= 0, T_n = Y_1 + Y_2 + \ldots + Y_n \\
T_{\xi_N} &= \sum_{j=1}^{\xi_N} Y_j 
\end{align*}
\]

\( X \) is representing the inter arrival, \( Y \) is the number of vacations performed by the server. Both variables are discrete having renewal processes at each occurrence. The level of absorption was achieved at nth arrival of \( X_n \). After nth arrival the nth vacation \( Y_n \) of the server would happen. The difference of the time at which the nth vacation happened and the nth customer arrived is the crossing time of the server. The probability generating
function. Probability density function, cumulative probability distribution function for joint distribution of the three random variables, $T_{x_N}, T_{z_N}$ and $\eta_N$ are obtained.

1.2 ASSUMPTIONS.

Let $N$ be a constant, $X_i$ and $Y_j$ be two sequences of random variables. Suppose that $X_i$, $i = 1, 2, 3, \ldots$; independently and identically distributed with finite mean $\lambda^1$ and $Y_j$, $j = 1, 2, 3, \ldots$; are independently and identically distributed with finite mean $\mu^1$.

$N_1(t)$ is the Erlang process associated with $X_i$ in which the distribution of $X_i$ is 2-stage Erlang distribution.

$N_2(t)$ is the Erlang process associated with $Y_j$ in which the distribution of $Y_j$ is 2-stage Erlang distribution.

$X_i$ and $Y_j$ are mutually independent.

2 ABSORBING MARKOV PROCESS AND ABSORBING TIME DISTRIBUTION

We consider a Markov process $\{X(t), t \geq 0\}$ on the state space $E = \{0,1,2,\ldots\}$. If $E_0$ and $E_1$ are two non-null subset of $E$ and they satisfy:

1) $E_0 \supset E_1 = E$.
2) $E_0$ is the absorbing state set and $E_1$ is the transient state set.
3) $\alpha_{E_0}$ is the given initial condition.

The absorbing Markov process (A.M.P) is constructed to analyze the problem consider the A.M.P $\{N_1(t), N_2(t), I(t), J(t)\}$ in which $N_1(t)$, and $N_2(t)$ are the counting process associated with $X_i$ and $Y_j$, respectively. $I(t)$ and $J(t)$ represent the phases of $X_i$ and $Y_j$ at time $t$ respectively. Its state space $E = \{(i,k,j,l), (i',j')| i, j = 0,1,\ldots; k,l = 1,2; \ i' = N', N+1',\ldots; \ j' = 1', 2',\ldots\}$. The absorbing states are shown in Figure 2.
Let
\[ P_{ij}(k,l,t) = p(N_1(t) = i, N_2(t) = j, I(t) = k, J(t) = l) \]
and
\[ P_{ij}(t) = [p_{ij}(1,1,t) \ldots p_{ij}(1,n,t) \ldots p_{ij}(m,1,t) \ldots p_{ij}(m,n,t)] \]

By the transition-rate diagram we can get the system of differential equations as follows
\[
P_{ij}'(t) = p_{ij}(t) - \left\{ \left( \begin{array}{c}
\lambda & -\lambda \\
0 & \mu
\end{array} \right) + \left( \begin{array}{c}
\mu & -\mu \\
0 & \mu
\end{array} \right) \right\} + p_{i-1,j}(t) \left( \begin{array}{c}
0 \\
\lambda
\end{array} \right) + p_{ij-1}(t) \left( \begin{array}{c}
0 \\
0
\end{array} \right) \]
\[ i = 0,1, \ldots N-1; \; j = 0,1,2, \ldots; \] (2.1)
\[
P_{ij}'(t) = p_{ij}(t) - \left\{ \left( \begin{array}{c}
\lambda & -\lambda \\
0 & \mu
\end{array} \right) + \left( \begin{array}{c}
\mu & -\mu \\
0 & \mu
\end{array} \right) \right\} + p_{i-1,j}(t) \left( \begin{array}{c}
0 \\
\lambda
\end{array} \right) \]
\[ i = N, \; j = 0,1,2, \ldots; \] (2.2)

The initial conditions are \( p_{i0}(0) = (1,0) \) and all others are equal to zero. We use \( z \) and \( L \) transforms to solve this system of equations.

Generating function with respect to number of service vacations, i.e., \( u \) can be written as
\[
\begin{align*}
\sum_{i=0}^{\infty} p_{i0}(t) u^i &= 0, \\
\sum_{i=0}^{\infty} p_{ij}'(t) u^i &= 1
\end{align*}
\]
\[ |u| \leq 1 \] (2.3)

Putting the values of \( p_{ij}'(t) \) from eq. (2.1)
\[
P_{i}(t,u) = \sum_{j=0}^{\infty} \left\{ p_{ij}(t) - \left( \begin{array}{c}
\lambda + \mu & -\lambda \\
0 & \lambda + \mu
\end{array} \right) \right\} + p_{i-1,j}(t) \left( \begin{array}{c}
0 \\
\lambda
\end{array} \right) + p_{ij-1}(t) \left( \begin{array}{c}
0 \\
0
\end{array} \right) \] \[ i = 0,1, \ldots N-1; \; j = 0,1,2, \ldots; \] (2.4)
\[
P_{i}(t,u) = - p_{i}(t,u) \left( \begin{array}{c}
\lambda + \mu & -\lambda \\
0 & \lambda + \mu
\end{array} \right) + p_{i-1}(t,u) \left( \begin{array}{c}
0 \\
\lambda
\end{array} \right) + u p_{j}(t,u) \left( \begin{array}{c}
0 \\
0
\end{array} \right) \]
\[ i = 0,1, \ldots N-1; \; j = 0,1,2, \ldots; \] (2.5)

The equation (2.2) can be generalized on the same pattern as
\[
P_{i}(t,u) = p_{ij}(t) \left( \begin{array}{c}
\lambda + \mu & -\lambda \\
0 & \lambda + \mu
\end{array} \right) + p_{i-1}(t,u) \left( \begin{array}{c}
0 \\
\lambda
\end{array} \right) \]
\[ i = N, \; j = 0,1,2, \ldots; \] (2.6)

Now the initial condition is \( p_{i0}(0,u) = (1,0) \) and all others are equal to zero. The probability generating function for number of customers, i.e, \( z \) can be written as
\[
P(t,u,z) = \sum_{i=0}^{\infty} p_{i}(t,u) z^i \quad \quad |z| \leq 1
\]
\[ P'(t,u,z) = \sum_{i=0}^{\infty} p_i(t,u)z^i \]

By putting the values of \( p_i(t,u) \)

\[ P'(t,u,z) = \sum_{i=0}^{\infty} \left[ -p_i(t,u) \left( \begin{array}{cc} \lambda + \mu & - (\lambda + \mu) \\ 0 & \lambda + \mu \end{array} \right) + p_{i-1}(t,u) \left( \begin{array}{cc} 0 & 0 \\ 0 & \lambda \end{array} \right) + up_i(t,u) \left( \begin{array}{cc} 0 & 0 \\ u & 0 \end{array} \right) \right] z^i \]

\[ P'(t,u,z) = -p_i(t,u, z) \left( \begin{array}{cc} \lambda + \mu & - (\lambda + \mu) \\ 0 & \lambda + \mu \end{array} \right) + p_{i-1}(t,u, z) \left( \begin{array}{cc} 0 & 0 \\ z\lambda & \lambda \end{array} \right) + \sum_{i=0}^{N-1} p_i(t,u) \left( \begin{array}{cc} 0 & 0 \\ u^\mu & 0 \end{array} \right) z^i \]

\[ i = 0, 1, ..., N - 1; \quad j = 0, 1, ...; \]  \hfill (2.7)

The Laplace transform can be defined as (Pipes & Harwill (1970)).

\[ p^*(s,u,z) = \int_0^\infty \exp(-st)p(t,u,z)dt. \quad \text{Re}(s) > 0. \]

The initial conditions for this stage are \( p(0,u,z) = (1,0) \) and all others are equal to zero. By applying L-transform on equation (2.7) we get

\[ p^*(s,u,z) = \left( \begin{array}{cc} 1 & 0 \\ 0 & u^\mu \end{array} \right) \left( \begin{array}{cc} s + \lambda + \mu & - (\lambda + \mu) \\ - z\lambda & s + \lambda + \mu \end{array} \right) \]

\[ (2.8) \]

Now taking the L-transform of equation (2.5)

\[ sp_i^*(s,u) - p_i(0,u) = - p_i^*(s,u) \left( \begin{array}{cc} \lambda + \mu & - (\lambda + \mu) \\ 0 & \lambda + \mu \end{array} \right) + p_{i-1}^*(s,u) \left( \begin{array}{cc} 0 & 0 \\ 0 & \lambda \end{array} \right) + p_i^*(s,u) \left( \begin{array}{cc} 0 & 0 \\ u^\mu & 0 \end{array} \right) \]

\[ p_0^*(s,u) = (1,0) \left( \begin{array}{cc} s + \lambda + \mu & - (\lambda + \mu) \\ - u^\mu & s + \lambda + \mu \end{array} \right)^{-1}. \]

\[ i = 0. \]  \hfill (2.9)

And

\[ p_i^*(s,u) = p_{i-1}^*(s,u) \left( \begin{array}{cc} s + \lambda + \mu & - (\lambda + \mu) \\ - u^\mu & s + \lambda + \mu \end{array} \right)^{-1} \left( \begin{array}{cc} 0 & 0 \\ 0 & \lambda \end{array} \right) \]

\[ i = 1, 2, ..., N - 1. \]

Therefore

\[ \sum_{i=0}^{N-1} p_i^*(s,u)z^i = p_0^*(s,u)z^0 + p_1^*(s,u)z^1 + ... + p_{N-1}^*(s,u)z^{N-1} \]  \hfill (2.10)

The values of \( p_0^*(s,u)z^0 + ... + p_{N-1}^*(s,u)z^{N-1} \) are substituted respectively in equation (2.10). After solving the sequence of equations stepwise and the equation is formulated in terms of \( p_0^*(s,u)z^0 \) we get

\[ \sum_{i=0}^{N-1} p_i^*(s,u)z^i = (1 \left( \begin{array}{cc} s + \lambda + \mu & - (\lambda + \mu) \\ - u^\mu - \lambda z & s + \lambda + \mu \end{array} \right)^{-1} \left( \begin{array}{cc} 0 & 0 \\ 0 & \lambda z \end{array} \right) \left( \begin{array}{cc} 0 & 0 \\ - u^\mu - \lambda z & s + \lambda + \mu \end{array} \right)^{-1} \left( \begin{array}{cc} 0 & 0 \\ 0 & \lambda z \end{array} \right) \right)^N \]  \hfill (2.11)
This position gives one step back to the absorption state. The Markovian property specifies that the future position is dependent of present position. In our study the absorption state is accomplished for $i = N$. The time taken during the vacations $T_{\xi_N}$, number of vacations $\xi_N$ and number of arrivals $\eta_N$ are organized for the absorption state of the system and are evaluated by considering the (N-1)th stage.

3 Joint Probability Distribution Functions

Theorem (3.1). The Joint probability generating (transform) function of three random variables $T_{\xi_N}, \xi_N$ and $\eta_N$ are given by

$$f^{*}(s,u,z) = u\left(1 - \frac{1}{s + \lambda + \mu} \right) \left( \frac{0}{s + \lambda + \mu} \right)^N \left( \frac{s + \lambda + \mu}{s + \lambda + \mu} \right) \left( \frac{-\lambda}{s + \lambda + \mu} \right)^{-1} \left( \frac{\mu}{s + \lambda + \mu} \right)$$

Proof. The $z$ transform while considering the number of arrivals can be written as

$$p^{*}(t,u,z) = \sum_{i=0}^{\infty} p^{*}_{i}(t,u)z^{i}$$

$$= \sum_{i=0}^{N-1} p^{*}_{i}(t,u)z^{i} + \sum_{i=N}^{\infty} p^{*}_{i}(t,u)z^{i}$$

$$= \sum_{i=N}^{\infty} \sum_{j=1}^{\infty} p^{*}_{i}(t,u)z^{i} = p^{*}(t,u,z) \cdot \sum_{i=0}^{N-1} p^{*}_{i}(t,u)z^{i}$$

Both sides of the equation are multiplied by $u\left( \frac{\mu}{\mu} \right)$.

$$\sum_{i=N}^{\infty} \sum_{j=1}^{\infty} p^{*}_{i}(t,u)z^{i} = u \left[ p^{*}(t,u,z) - \sum_{i=0}^{N-1} p^{*}_{i}(t,u)z^{i} \right] \left( \frac{\mu}{\mu} \right)$$

The Laplace transform is applied to solve the service vacation time with number of vacations made for number of customers arrived and by putting the value of $p^{*}(s,u,z)$ from equation (2.8) it yields

$$f^{*}(s,u,z) = u \left\{ (1 - \frac{1}{s + \lambda + \mu}) \left( \frac{0}{s + \lambda + \mu} \right)^N \left( \frac{s + \lambda + \mu}{s + \lambda + \mu} \right) \left( \frac{-\lambda}{s + \lambda + \mu} \right)^{-1} \left( \frac{\mu}{s + \lambda + \mu} \right) \right\}$$

putting the value of $\sum_{i=0}^{N-1} p^{*}_{i}(s,u)z^{i}$ from equation (2.11) and simplifying we get

$$f^{*}(s,u,z) =$$

$$u \left\{ (1 - (1 - 0) + (1 - 0)) \left( \frac{s + \lambda + \mu}{s + \lambda + \mu} \right) \left( \frac{-\lambda}{s + \lambda + \mu} \right)^{-1} \left( \frac{\mu}{s + \lambda + \mu} \right) \right\}$$

The joint probability transform function (generating function) for three random variables $T_{\xi_N}, \xi_N$ and $\eta_N$ is then


\[ f^*(s,u,z) = u(1) \left\{ \begin{pmatrix} s + \lambda + \mu & -(\lambda + \mu) \\ -u \mu & s + \lambda + \mu \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \lambda z \end{pmatrix} \right\}^N \left\{ \begin{pmatrix} s + \lambda + \mu & -(\lambda + \mu) \\ -\lambda z & s + \lambda + \mu \end{pmatrix}^{-1} \begin{pmatrix} \mu \end{pmatrix} \right\} (2.12) \]

**Theorem 3.2:** The joint probability density function for three random variables \( T_{\xi_N}, \xi_N \) and \( \eta_N \) that is time taken by service vacations, number of vacations taken by the server and number of inter arrivals is given by

\[
p(T_{\xi_N} \leq t, \xi_N = j, \eta_N = i) = \binom{N + j - 2}{j - 1} \left[ \lambda^i (\lambda + \mu)^{j+i-1} \frac{t^{2j+2i-2}}{(2j + 2i - 2)!} e^{-(\lambda + \mu)t} + \lambda^i (\lambda + \mu)^{j+i} \frac{t^{2j+2i-1}}{(2j + 2i - 1)!} e^{-(\lambda + \mu)t} \right]
\]

**Proof:** By the definition of the \( z \) and L transforms it can be written as

\[
f^*(s,u,z) = \sum_{i=N}^\infty \sum_{j=1}^\infty \int_0^\infty \exp(-st) dp \{ T_{\xi_N} \leq t, \xi_N = j, \eta_N = i \} u^j z^i
\]

By putting the value of \( f^*(s,u,z) \)

\[
f^*(s,u,z) = u(1) \left\{ \begin{pmatrix} s + \lambda + \mu & -(\lambda + \mu) \\ -u \mu & s + \lambda + \mu \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \lambda z \end{pmatrix} \right\}^N \left\{ \begin{pmatrix} s + \lambda + \mu & -(\lambda + \mu) \\ -\lambda z & s + \lambda + \mu \end{pmatrix}^{-1} \begin{pmatrix} \mu \end{pmatrix} \right\}
\]

Let \( a = s + \lambda + \mu \) and using some algebraic skills, applying the rule of geometric series as given by Finney, Weir and Giordano (2001) we obtain

\[
f^*(s,u,z) = \frac{u}{a} \left( \frac{\lambda z}{a} \right)^N \left( \frac{\lambda + \mu}{a} \right) \sum_{k=0}^\infty \left( -N \right) \left( \frac{\lambda + \mu}{a} \right)^k \left( \frac{u \mu}{a} \right)^k \mu \sum_{l=0}^\infty \left( \frac{\lambda + \mu}{a} \right)^l \left( \frac{\lambda z}{a} \right)^l \left\{ 1 + \frac{\lambda + \mu}{a} \right\}.
\]

The negative binomial theorem is applied to simplify the series as done by Hogg and Craig (1995) and putting \( j = k + 1, \ i = N + l \) we get

\[
f^*(s,u,z) = \sum_{j=1}^\infty \sum_{i=j}^\infty \left( \begin{pmatrix} N + j - 2 \\ j - 1 \end{pmatrix} \frac{\lambda}{a} \left( \frac{\mu}{a} \right)^i \left( \frac{\lambda + \mu}{a} \right)^{j+i-1} \left\{ 1 + \frac{\lambda + \mu}{a} \right\} z^i u^j.
\]

After comparing the coefficients of \( z^i, u^j \) and the inverse of the Laplace transform was taken to obtain the joint probability density function of three random variables \( T_{\xi_N}, \xi_N \) and \( \eta_N \) as done by Talpur and Shi (1994). Putting the value of \( a \) it then be written as

\[
p(T_{\xi_N} \leq t, \xi_N = j, \eta_N = i) = \binom{N + j - 2}{j - 1} \left[ \lambda^i (\lambda + \mu)^{j+i-1} \frac{t^{2j+2i-2}}{(2j + 2i - 2)!} e^{-(\lambda + \mu)t} + \lambda^i (\lambda + \mu)^{j+i} \frac{t^{2j+2i-1}}{(2j + 2i - 1)!} e^{-(\lambda + \mu)t} \right]
\]

Hence it is proved.

### 3.3 Joint Cumulative probability Distribution Function.

**Theorem 3.3.1** The joint cumulative probability distribution function of three random variables \( T_{\xi_N}, \xi_N \) and \( \eta_N \) is given by

\[
p(T_{\xi_N} \leq t, \xi_N = j, \eta_N = i) = \binom{N + j - 2}{j - 1} \left[ \lambda^i (\lambda + \mu)^{j+i-1} \frac{t^{2j+2i-2}}{(2j + 2i - 2)!} e^{-(\lambda + \mu)t} + \lambda^i (\lambda + \mu)^{j+i} \frac{t^{2j+2i-1}}{(2j + 2i - 1)!} e^{-(\lambda + \mu)t} \right]
\]

302
\[
\left( N + j - 2 \right) \left( \frac{\lambda}{\lambda + \mu} \right)^{j-1} \left( \frac{\mu}{\lambda + \mu} \right)^j \left[ \sum_{r=0}^{2j+2i-1} \frac{[(\lambda + \mu)^r]}{r!} e^{-(\lambda + \mu)t} \right] + \left[ \sum_{r=0}^{2j+2i-1} \frac{[(\lambda + \mu)^r]}{r!} e^{-(\lambda + \mu)t} \right].
\]

(2.14)

**Proof.** The joint cumulative distribution function for three random variables \( T_{sN}, \xi_N \) and \( \eta_N \) can be obtained by taking the integration of their joint probability density function as done by Medhi (1982) and Ifat (2004). Hence the proof is obvious.

## 4 RESULTS AND DISCUSSION

### 4.1 RESULTS

Results of the joint probability distribution functions of three random variables namely \( \eta_N \) (number of customers arrive in the system), \( T_{sN} \) (Vacation time) and \( \xi_N \) (Number of vacations done by the servers)

| Probability Generating Function \( f^+(s,u,z) \) | \( u(1 \ 0) \left\{ \begin{array}{c}
\left( s + \lambda + \mu - (\lambda + \mu)^{-1} \left[ \begin{array}{c}
0 \\
0 \\
\mu
\end{array} \right] \right)^{N} \\
\left( \begin{array}{c}
0 \\
0 \\
\mu
\end{array} \right) \\
\left( \begin{array}{c}
0 \\
0 \\
\mu
\end{array} \right)
\end{array} \right\} \\
\left( \begin{array}{c}
s + \lambda + \mu - (\lambda + \mu)^{-1} \\
-\lambda \\
s + \lambda + \mu
\end{array} \right) \left( \begin{array}{c}
\mu
\end{array} \right)
\) |
| --- | --- |

| Probability Density Function \( p_T \leq t, \xi_N = j, \eta_N = i \) | \( \left( N + j - 2 \right) \left( \frac{\lambda}{\lambda + \mu} \right)^{j-1} \left( \frac{\mu}{\lambda + \mu} \right)^j \left[ \sum_{r=0}^{2j+2i-1} \frac{[(\lambda + \mu)^r]}{r!} e^{-(\lambda + \mu)t} \right] + \left[ \sum_{r=0}^{2j+2i-1} \frac{[(\lambda + \mu)^r]}{r!} e^{-(\lambda + \mu)t} \right] \) |

| Cumulative Probability Distribution Function \( p_T \leq t, \xi_N = j, \eta_N = i \) | \( \left( N + j - 2 \right) \left( \frac{\lambda}{\lambda + \mu} \right)^{j-1} \left( \frac{\mu}{\lambda + \mu} \right)^j \left[ \sum_{r=0}^{2j+2i-1} \frac{[(\lambda + \mu)^r]}{r!} e^{-(\lambda + \mu)t} \right] + \left[ \sum_{r=0}^{2j+2i-1} \frac{[(\lambda + \mu)^r]}{r!} e^{-(\lambda + \mu)t} \right] \) |

Table 1: Joint Distribution for three random variables \( T_{sN}, \xi_N \) and \( \eta_N \)

The probability density function satisfies the 2-stage Erlang distribution evaluated for the joint probability distribution of random variables relating to the crossing time spent for reasonable number of vacations \( T_{sN} \) recorded during the number of vacations \( \xi_N \) receiving reasonable number of arriving customers \( \eta_N \) for the achievement of absorption state.

### 4.2 DISCUSSION

In this study the crossing of the renewal sequences of number of arriving customers and number of vacations made by servers at service channels is analyzed. The number of arriving customers as well as the number of server vacations provided discrete variables while the relevant period taken by number of vacations is associated with continuous variable. The bulk arrival and multi-channel service establishes the phases for two-stage Erlang processes. These phases are solved and the resulting joint probability distribution function is the convolution of the probability density function of the respective variables showing a gamma distribution as shown by Chaudhry and Temleton (1983).
REFERENCES


Ifat, Z., (2004) Joint Distribution of some random variables of crossing time and renewal numbers related with two Erlang processes, Thesis of Master of Philosophy, Department of Statistics University of Sindh Jamshoro

Koutras, M.V. (1997). Waiting time distributions associated with runs of fixed length in two state Markov chains,


Sampling Plans Excluding Certain Neighboring Units

Kyoungah See* and A. John Bailer
Department of Mathematics and Statistics, Center for Environmental Toxicology and Statistics, Miami University, Oxford, Ohio 45056, U.S.A.
*e-mail: seek@muohio.edu

SUMMARY
Ecological and environmental studies frequently involve work in spatially constrained settings, where some physical condition influences the effectiveness of standard sampling plans. Additionally, some studies involve expensive and time-consuming sampling processes. These conditions motivate researchers to find sampling plans that provide the highest precision at the lowest cost. Our application involved studies of forest-canopy insects, where sampling plots in close, physical proximity to each other might exhibit greater similarity than distant plots. Several sampling plans that exclude some types of “neighboring” units were developed and this paper illustrates how two, spatially constrained sampling plans can be applied to studies of forest-canopy insects. Sampling of this type is called Sampling Excluding Neighboring Units (SENU) and is a two-dimensional adaptation of Balanced Sampling Excluding Contiguous Units (BSEC; Hedayat, Rao, and Stufken, 1988). Key features of this paper are: (i) the construction and illustration of sampling plans designed to yield representative samples by avoiding the simultaneous selection of units that are, in some sense, neighbors; (ii) an investigation into the adequacy of these sampling plans when used with small sample sizes; and (iii) a comparison of the relative efficiencies of the two SENU type designs and simple random sampling.

Key Words: Contiguous units, relative efficiency, Spatial correlation,

1. INTRODUCTION
Ecological and environmental sampling problems occur in spatial contexts. For example, consider estimating the number of distinct insect species (abundance) in a forest ecosystem, in which organism distribution is affected by many factors over a large region, including environmental conditions and spatial patterns. Various sampling plans have been considered for situations, with the purpose of providing better estimates of population parameters, such as mean or total abundance (e.g., Christman, 1997; Hedayat, Rao, and Stufken, 1988; Hedayat and Stufken, 1998; See et al., 2000; Thompson, 1990; Thompson and Seber, 1996). For the example mentioned above, one could designate approximately rectangular sampling areas within a forest, referred to as strata, and superimpose a grid within each stratum. In such situations, neighboring grid units could be expected to yield similar results (Hedayat and Stufken, 1998). Therefore, with a fixed sample size, the inclusion of neighboring units may result in a less representative sample, particularly where the species composition differs widely over the entire geographical range.

Excluding neighboring units has been shown to provide an estimator of the population total with a smaller variance when the sample units are in a one-dimensional ordering, called Balanced Sampling Excluding Contiguous Units (BSEC) (Hedayat, Rao and Stufken, 1988).

While considerable research has been conducted on plans that avoid contiguous units in a one-dimensional setting (Hedayat and Sinha, 1991; Hedayat and Stufken, 1998; Stufken, 1993; Stufken et al., 1999), there are few results for two-dimensional settings (See et al., 2000, 2007a, 2007b). We refer to this type of sampling plan as Sampling Excluding Neighboring Units (SENU). These are designs in which the entire sample is chosen such that no two neighboring units are included, and which have been generally shown to provide a more precise estimator of population means when neighboring units are expected to provide similar responses.

In this paper, two SENU-type plans, in addition to simple random sampling, are applied to the actual data from a study that investigated the species abundance and diversity of beetles in the canopy of a beech-maple forest. In this particular
study, we found added justification for the use of SENU-type plans as the sampling method used in one plot could alter the responses from neighboring plots. A previous study showed that certain tree species (e.g. beech) had a low insect diversity and abundance, while other trees (e.g. oaks) supported a diverse and abundant insect assemblage. Therefore, in this study, we used sampling designs that avoided simultaneous selection of neighboring plots that shared the same tree crowns or the same admixture of tree species. Having acquired these data, we were able to investigate the accuracy of the subject sampling plans using smaller sample sizes, i.e., subsets of observations selected from the entire set that was measured.

This paper is organized as follows: Section 2 describes each sampling design that was used in this study in general context. Section 3 describes how proposed sampling designs were applied to our insect data. Sections 4 and 5 present results of the data analysis and a small simulation study, respectively, and these are discussed further in Section 6.

2. SAMPLING DESIGNS

We consider a rectangular study region consisting of \( N = LM \) units of equal size, laid out in \( L \) rows and \( M \) columns with \( L \geq 2 \) and \( M \geq 2 \). The rows and columns might be conceptualized as a series of transects through a forest where the \( M \) column transects are perpendicular to the \( L \) row transects. The (fixed) sample size is denoted by \( n(<N) \). The units will be labeled by row and column membership \((i, j)\), where \( i = 1, 2, \cdots, L \) and \( j = 1, 2, \cdots, M \). A sample \( \{(i_1, j_1), (i_2, j_2), \cdots, (i_n, j_n)\} \) of \( n \) distinct units may include multiple samples from each row and column. In other words, some \( \{i_1, i_2, \cdots, i_n\} \) of \( \{1, 2, \cdots, L\} \) or \( \{j_1, j_2, \cdots, j_n\} \) of \( \{1, 2, \cdots, M\} \) may be the same. We note that the case where the region is subdivided into many cells with a single row (or column) can be treated as a one-dimensional BSEC design (Hedayat et al., 1988).

In any sampling design, with or without replacement, the probability \( \pi_u \) that unit \( u \) (a cell \((i, j)\) in the figure above) is included, and the probability \( \pi_{uv} \) that both unit \( u \) and unit \( v \) are included in the sample, are called the first order and second-order inclusion probabilities, respectively. For any sampling design, an unbiased estimator of the population total \( \tau \) introduced by Horvitz-Thompson (1952) is

\[
\hat{\tau}_H = \sum_{k=1}^{n} \frac{y_{uk}}{\pi_{uk}} = \sum_{k=1}^{n} \frac{y_{uk}}{\pi_{uk}}
\]

where \( n \) is the effective sample size and the summation is over the distinct units \( u_k = (i_k, j_k) \), \( k = 1, 2, \cdots, n \); \( y_{uk} \) is the value of the variable obtained from unit \( u_k \); and \( I_{uk} \) is the indicator variable such that \( I_{uk} = 1 \) with probability \( \pi_{uk} \), and

\[
I_{uk} = \begin{cases} 
1, & \text{if the unit } u_k \text{ was included in the sample} \\
0, & \text{Otherwise}
\end{cases}
\]

If the first-order inclusion probabilities are positive, the Horvitz-Thompson estimator for the population total achieves unbiasedness by dividing the value for each unit in the sample by the probability that the unit is included in the sample. We note that the Horvitz-Thompson estimator of population mean \( \mu \) would be computed by

\[
\hat{\mu}_H = \frac{1}{N} \sum_{k=1}^{n} \frac{y_{uk}}{\pi_{uk}} = \frac{\hat{\tau}_H}{N}.
\]

The variance of Horvitz-Thompson estimator for population mean \( \hat{\mu}_H \) is given by

\[
\text{Var}(\hat{\mu}_H) = \frac{1}{N^2} \left[ \sum_{k=1}^{n} \sum_{m=1}^{n} \frac{y_{uk} y_{um}}{\pi_{uk} \pi_{um}} \left( \frac{1}{\pi_{uk} \pi_{um}} - 1 \right) \right] \]

(2.3)

provided all of the joint inclusion probabilities \( \pi_{uv} \) are positive. We note that two of the sampling plans described below, Random Sampling Excluding Neighboring Units and Random Sampling Excluding Contiguous Units, have a zero second-order inclusion probability for the neighboring pairs of units. As a result, the variance of the Horvitz-Thompson estimator of the population mean cannot be estimated unbiasedly for these designs.

I. Simple Random Sampling without replacement (SRS). For SRS, we randomly select \( n \) units, without replacement, from the \( N \) points defined by the grid points in the \( L \times M \) rectangular region. For this plan, the probability of selecting any particular sampling point \((i_u, j_u)\), and the probability of selecting any two points, \((i_u, j_u)\) and \((i_v, j_v)\), do not depend on the choice of units \( u \) and \( v \), and are given by \( \pi_u = n/N \) and

\[
\pi_{uv} = \frac{n}{N^2}
\]
respectively. Thus, from (2.2), the Horvitz-Thompson estimator of the population mean is simply estimated by the sample mean of the \( n \) sample points, \( \hat{\mu}_{HT} = \bar{y} \). Substituting the inclusion probabilities \( \pi_{ii} \) and \( \pi_{ij} \) in (2.3), the variance of the Horvitz-Thompson estimator \( \hat{\mu}_{HT} \) can be reduced to

\[
\text{var}(\hat{\mu}_{HT}) = \text{var}(\bar{y}) = \frac{\sigma^2 (N-n)}{n N (N-1)},
\]

where

\[
\sigma^2 = \sum_{x=1}^{N} (x - \mu)^2 / (N-1).
\]

Further, an unbiased estimator of the \( \text{var}(\hat{\mu}_{HT}) \) can be obtained by replacing \( \sigma^2 \) by the sample variance \( s^2 \).

II. Random Sampling Excluding Neighboring Units (RSENU). For this plan, we randomly select \( n \) distinct units, where each unit has the same selection probability, is admissible if it has not yet been selected, and is not the neighbor of a previously selected unit. Two units \((i_1, j_1)\) and \((i_2, j_2)\) are said to be neighbors if \( |i_1 - i_2| \leq 1 \) and \( |j_1 - j_2| \leq 1 \). A corner unit like \((1,1)\) has three neighbors, other edge units like \((1,2)\) have five neighbors, and interior units like \((2,2)\) have eight neighbors. Therefore, even first-order inclusion probabilities are not constant, so that the Horvitz-Thompson estimator for the population mean is now not given by the sample mean. Although biased, we decided to use the sample mean to estimate the population mean, and a mean squared error (MSE) criterion rather than a variance criterion to judge the efficiency of this strategy. In addition, we can examine the size of the bias in this situation.

III. Random Sampling Excluding Contiguous Units (RSECU). In this plan, two units are neighbors, if and only if, they share a common border; i.e., two units \((i_1, j_1)\) and \((i_2, j_2)\) are contiguous units if either “\(|i_1 - i_2| \leq 1 \) and \( j_1 = j_2 \)” or “\(i_1 = i_2\) and \(|j_1 - j_2| \leq 1\)”. Thus a corner unit like \((1,1)\) has two neighbors, other edge units like \((1,2)\) have three neighbors, and interior units like \((2,2)\) have four neighbors. In sampling, we randomly select \( n \) distinct units, one at a time, as in RSENU, keeping in mind that a unit is admissible if it is not one of the units sharing a common border with a selected unit. Once again, since the first order inclusion probabilities are not constant, MSE was used in evaluating the efficiency of the sampling plan.

IV. Random Sampling Excluding Neighboring Units and Identical Rows and Columns (RSENU&IRC). This plan combines features of RSENU and Random Sampling Excluding Identical Rows and Columns (RSEIRC, refer to page 277 of See et al., 2000), which randomly selects \( n \) distinct units with the restriction that no pair of units can be selected from the same row or column. Thus, a corner unit like \((1,1)\) has \( L + M - 1 \) neighbors, which include all points in the same row, \( M \), or same column, \( L \), plus the diagonally adjacent point, \((2,2)\). Other edge units, like \((1,2)\), have \( L + M \) neighbors, and interior units like \((2,2)\) have \( L + M + 2 \) neighbors. So, the first-order inclusion probabilities are not constant. Therefore, as with RSENU above, we used MSE criterion to evaluate the efficiency of this sampling plan.

3. SAMPLING PLANS FOR INSECT ABUNDANCE STUDY

In this section, we describe how the sampling plans listed in Section 2, were utilized in the field study at with colleagues D. Golden and T. Crist at Bachelor Reserve at Miami University, Oxford, Ohio. Ten plots, each 25m×25m in size and having from 3 to 10 tree species, were established with 50-100m between plots. Apart from differences in tree species richness, tree size and density were generally similar among plots. Each plot was subdivided into 25 subplots, 5m×5m in size, and we wished to select smaller number of units (e.g., 5 units) from the subplots. Insects were collected from the forest canopy after fogging with an insecticide. The object of the experiment was to obtain the total beetle density sampled in collecting trays.

3.1 Sampling procedure for subplots

For the purpose of selecting 5 of the 25 subplots, RSENU&IRC was applied such that the following two conditions were satisfied: (i) all subplots surrounding the selected subplot were avoided, and (ii) only one subplot from each row and column was selected. If we do not impose the two restrictions above, and if we randomly pick 5 units (or subplots) from 25, then the number of possible selections
must be \( \binom{25}{5} \), which is the number of ways to select 5 units from 25. However, if we impose the second condition (ii), then each sample of 5 will contain exactly 1 unit from each row and each column. In order to see this, view the region as a 5 by 5 grid of subplots, labeled as follows:

\[
\begin{array}{ccccc}
(1,1) & (1,2) & (1,3) & (1,4) & (1,5) \\
(2,1) & (2,2) & (2,3) & (2,4) & (2,5) \\
(3,1) & (3,2) & (3,3) & (3,4) & (3,5) \\
(4,1) & (4,2) & (4,3) & (4,4) & (4,5) \\
(5,1) & (5,2) & (5,3) & (5,4) & (5,5)
\end{array}
\]

From the table, the unit from the first row could be (1,1), (1,2), (1,3), (1,4) or (1,5). Now, suppose we choose (1,1) from the first row, then from the second row we cannot choose (2,1) or (2,2), because each is a neighbor of (1,1). So, we must choose (2,3), (2,4) or (2,5) from the second row.

**Case 1**: Assuming we choose (2,3), then (3,5) is the only possible selection from the third row because of the first condition (i), and we must choose (4,2) from the fourth row and (5,4) from the fifth row. (We refer to this as plan 1).

**Case 2**: Assuming we choose (2,4) from the second row, then (3,2) is the only possible selection from the third row, and (4,5) and (5,3) must be chosen from the fourth and fifth rows respectively. (We refer to this as plan 2).

**Case 3**: Assuming we select (2,5) from row 2, then (3,2) and (3,3) are possible selections from row 3. However, neither of these selections allows us to meet the first condition (i) in our selections from rows 4 and 5.

Similarly, we can show that there are exactly 14 ways to choose five subplots by utilizing RSENU & IRC in this setting. For example, with \( i = 1 \), we use the subplots (1,1), (2,3), (3,5), (4,2) and (5,4); with \( j = 3 \), we select (1,3), (2,1), (3,4), (4,2), and (5,5).

### 3.2 Sampling within subplots

As described in Section 3, each selected subplot (of 25 m\(^2\) area) is subdivided further into a 3×3 grid of cells. These cells are the basic units from which the population data are collected and recorded. That is, a set of nine insect counts, one for each cell, is obtained as data, as shown in Table 1. In our efforts to investigate the adequacy of reduced sampling, we employed the sampling plans RSENU and RSEC as proposed in Section 2, in addition to simple random sampling, to each of the 50 subplots. Thus, each of the 3×3 subplots defines a population of N=9 elements from which we sampled n=2 or n=3 units using the designs described below. Here, we will illustrate our example using sample size n=3 only. For a 3×3 grid population size N=9, with sample size n=3, let the units be labeled as 1=(1,1), 2=(1,2), ..., 9=(3,3) for convenience as follows:

\[
\begin{array}{ccc}
(1,1) & (1,2) & (1,3) \\
(2,1) & (2,2) & (2,3) \\
(3,1) & (3,2) & (3,3)
\end{array}
\Rightarrow
\begin{array}{ccc}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{array}
\]

**Simple Random Sampling without replacement (SRS):** Utilizing SRS, we randomly select 3 sampling units from each subplot which contains the units \( \{1,2,\ldots,9\} \) without replacement. The first-order and second-order inclusion probabilities do not depend on the choice of units, and are given by \( \pi_i = 1/9 \) and \( \pi_{ij} = (3\cdot2)/(9\cdot8) = 1/12 \), respectively. The Horvitz-Thompson estimator of the population mean is simply the sample mean: \( \hat{\mu}_{HT} = \bar{y} \), and the variance of the sample mean can be computed by substituting the inclusion probabilities into equation (2.3).

**Random Sampling Excluding Neighboring Units (RSENU):** Strictly avoiding all surrounding units, there are five permissible samples of size n=3 containing “unit 1”, namely, \( \{1,3,7\} \), \( \{1,3,8\} \), \( \{1,3,9\} \), \( \{1,6,7\} \), and \( \{1,7,9\} \). However, there is no possible sample of size n=3 containing “unit 5”. Therefore, the first-order inclusion probability of selecting the “unit 5” is zero, but that of “unit 1” is positive. This plan allows 8 possible sample combinations: \( \{2,7,9\} \), \( \{3,4,9\} \), \( \{3,7,9\} \) in addition to the five listed above.

**Random Sampling Excluding Contiguous Units (RSECU):** Excluding only the units that share a common border with the selected unit, we pick a sample of size n=3. In this plan, “unit 5” has a positive first-order inclusion probability unlike the RSENU case. A corner unit (e.g., “unit 1”) has two neighbors (“unit 2” and “unit 4”) to avoid, while an interior unit (”unit 5”) has four units (units 2, 4, 6, 9).
8) to avoid. Thus, the inclusion probability of a unit varies according to its location in the plot. This plan allows 22 possible sample combinations of the units: \{135, 137, 138, 139, 157, 159, 167, 168, 179, 246, 248, 249, 267, 268, 279, 348, 349, 357, 359, 379, 468, and 579\}.

4. ANALYSIS AND RESULTS IN INSECT ABUNDANCE STUDY

One of the main goals of this study was to determine whether it is necessary to sample each of the nine units of the selected subplot or whether a smaller number of cells will give an adequate estimation of the true population response. With a \(3 \times 3\) grid from which to select, we were able to study sample sizes of \(n=2\) and \(n=3\) only, due to the restrictions defined by the sampling plans. The mean of the N=9 cells defines the population mean for these data.

The sample means were compared to the population means. Both bias, the difference between the sample mean and the population mean, and mean squared error (MSE), the average squared difference between the sample and population mean, were assessed by averaging these quantities over all possible samples of size \(n=2\) or size \(n=3\) from the N=9 size population. For example, 36 simple random samples of \(n=2\) are possible for each N=9 population defined by the \(3 \times 3\) grid of observations. When calculating MSE, we examined the average squared difference between the sample mean, for each of the 36 possible samples, and the subplot (population) mean. Note that we were essentially conducting a data-based simulation experiment in which we examined all samples from a series of 50 observed.

Table 1 presents an overall summary of the results. For sample sizes \(n=2\) and \(n=3\), the mean squared errors of these two sampling plans were compared to those of simple random sampling. Both SENU-type sampling plans produced better results than simple random sampling. RSENU produced the best results, yielding the minimum MSE and thus, the highest relative efficiency of all designs considered. Figure 1 displays MSE plots based on some of the results of the insect abundance data analysis. Parts (a) and (c), in Figure 1, show that the MSE based on \(n=3\) is uniformly smaller than that of \(n=2\), for both RSENU and RSECU sampling designs. Parts (b) and (d) compare both of the SENU-type sampling designs to SRS for the \(n=3\) case. SENU-type designs yield slightly smaller MSEs, with RSENU showing a slight advantage over RSECU in their comparisons to SRS, for all 50 population subplots.

<table>
<thead>
<tr>
<th>Sample Size of N=9</th>
<th>Sampling Design</th>
<th>k(^a)</th>
<th>(d)</th>
<th>MSE(^b) Mean (empirical SD)</th>
<th>RE (d/\min)^c</th>
<th>Bias index(d) ((100\times \frac{\phi}{\mu})) (empirical SD)</th>
<th>% [95%CI that includes (\mu)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=2&lt;br&gt; SRS&lt;br&gt; 36</td>
<td>2.0911 (2.54)</td>
<td>1.13</td>
<td>0.00</td>
<td>68.0 (10.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSENU&lt;br&gt; 16</td>
<td>1.8439 (2.03)</td>
<td>1.00</td>
<td>0.61</td>
<td>70.4 (14.3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSECU&lt;br&gt; 24</td>
<td>1.9881 (2.37)</td>
<td>1.08</td>
<td>0.16</td>
<td>69.8 (11.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=3&lt;br&gt; SRS&lt;br&gt; 84</td>
<td>1.1949 (1.45)</td>
<td>1.65</td>
<td>0.00</td>
<td>82.3 (7.9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSENU&lt;br&gt; 8</td>
<td>0.7235 (0.73)</td>
<td>1.00</td>
<td>0.71</td>
<td>88.0 (14.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSECU&lt;br&gt; 22</td>
<td>1.0537 (1.33)</td>
<td>1.46</td>
<td>0.14</td>
<td>87.2 (10.3)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. \(k\) = the number of ways to obtain all possible sampling plans of size \(n=(2,3)\) from \(N=9\) population units under the particular sampling design.
- b. The average mean squared errors are obtained over 50 population subplots.
- c. \(RE(\min)=\frac{MSE(Plan \ d)}{MSE(\min)}\).
- d. Bias index\(=100\times \frac{\phi}{\mu}\), where \(\phi = |\hat{\mu} - \mu| / \mu\).

The relative efficiency (RE) of a sampling plan is defined as the ratio of its MSE to that of the sampling plan with the minimum MSE. As RSENU had the minimum MSE, all relative efficiencies were defined in relation to RSENU. By definition, then, the relative efficiency of RSENU was unity, as shown in Table 1. For example, the relative efficiency of simple random sampling for \(n=3\) is listed as 1.6516, which implies that the simple random sampling plan has an MSE which is 65% higher than that of RSENU.

The Bias Index \((\phi)\) in Table 1 is defined as the absolute value of the percentage difference between the estimated mean and the true mean, divided by the true mean. Although RSENU and RSECU had greater bias than SRS (which is unbiased), the bias we found was considered relatively small. For example, the bias index of RSENU for \(n=3\) is .71%, which shows that our estimation of the mean abundance varied by less than 1% from the true mean abundance. The average percentages of the computed 95% confidence intervals that contained the true mean are displayed in the last column. For
n=3, RSENU and RSECU sampling plans that contain the true mean abundance display higher percentages of the confidence intervals than the simple random sampling plan.

Figure 1. Mean Squared Error Plots: Insect Abundance Data

5. SIMULATION STUDY

Utilizing the real data set, we have shown that SENU-type sampling plans produce better results than SRS, as seen in the comparison of relative efficiencies listed in Table 2. However, in viewing the MSE plots in Figure 1, it is not visibly obvious that RSENU provided better results than SRS. Therefore, we conducted a small simulation study to compare these sampling designs with respect to MSE and bias. In particular, we generated $3 \times 3$ plots according to correlated multivariate normal distributions, where the correlations between units decreased as a function of the distance between units. We adopted a stationary spatial correlation model based on a power function, that is, we set $\text{corr}(\text{unit } u, \text{ unit } v) = \rho^{d_{uv}}$, where $d_{uv}$ is the Euclidean distance between sampling units $u$ and $v$ with $\rho = .4, .6, .8, \text{ and } .9$. For example, in the simulation, the correlation between neighboring (or adjacent) units would be $\rho$ (e.g., .6), while the correlation between the most distant units would be $\rho^{\sqrt{3}}$ (e.g., $.6^{\sqrt{3}} = .24$). Therefore, the covariance is assumed to be a function of the distance between the locations $u$ and $v$. Three different cell mean configurations were considered as simulation conditions:

\[
\begin{pmatrix}
5 & 5 & 5 \\
5 & 5 & 5 \\
5 & 5 & 5 \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
2 & 3 & 4 \\
3 & 4 & 5 \\
4 & 5 & 6 \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
5 & 7 & 11 \\
8 & 12 & 14 \\
\end{pmatrix}
\]

From the generated data, samples of size 2 and 3 were drawn according to SRS, RSENU and RSECU, and MSE and Bias were calculated. Each mean-correlated configuration was simulated 4,000 times. As described earlier, the MSE for a sampling design, applied to a simulated “population” $3 \times 3$ grid, is based upon average squared differences between estimates and the corresponding parameters of the mean, over all possible samples. These MSE estimates are based upon different numbers of possible samples, depending upon the sampling design (see Table 1). Note that responses in plots are correlated in this simulation unlike See et al. (2000).

In Table 2, the advantage of RSENU over SRS is clearly seen for sample size n=3. The relative efficiencies of SRS with respect to RSENU range from 1.76 to 2.51. (Note that the relative efficiencies of SRS with respect to RSECU (not included in Table 2) range from 1.29 to 2.12.) For sample size n=2, we obtain similar results, though the advantage is less pronounced as seen in the following:

\[
\begin{pmatrix}
5 & 5 & 5 \\
2 & 3 & 4 \\
2 & 5 & 8 \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
5 & 5 & 5 \\
3 & 4 & 5 \\
5 & 7 & 11 \\
\end{pmatrix}
\]

\[
\begin{pmatrix}
5 & 5 & 5 \\
4 & 5 & 6 \\
8 & 12 & 14 \\
\end{pmatrix}
\]

For both n=2 and 3, we observe that, as the
The mean function changes from a constant (not structured), as in M1, to the more structured configuration of M3, the advantage of RSENU (and also RSECU) becomes more obvious. In addition, the relative efficiency increases as the correlation between neighboring units increases. It is interesting to note that, in comparing RSENU to RSECU, RSECU has a slight advantage in terms of BIAS, while RSENU prevails in terms of MSE. As both are 'biased' techniques, we believe RSENU provided the better results over all.

### Table 2

Results of simulation study on $3 \times 3$ correlated multivariate normal data.

| Mean Configuration | $\rho$ | $\text{MSE}^a$ (empir. SD) | $\text{MSE}^b$ (empir. SD) | Bias index (100×$\phi$%) | $\text{MSE}^b$ (empir. SD) | Bias index (100×$\phi$%) | RE\(^b\)(SRS|RSENU) |
|--------------------|-------|---------------------------|---------------------------|--------------------------|---------------------------|--------------------------|--------------------------|
| $\{5 \ 5 \ 5\}$   | .4    | .1679 (.090)              | .0953 (.065)              | .05                      | .1299 (.071)              | .03                      | 1.76                     |
| $\{5 \ 5 \ 5\}$   | .6    | .1221 (.071)              | .0660 (.046)              | .13                      | .0885 (.048)              | .03                      | 1.85                     |
| $\{5 \ 5 \ 5\}$   | .8    | .0668 (.041)              | .0351 (.025)              | .09                      | .0458 (.026)              | .02                      | 1.90                     |
| $\{5 \ 5 \ 5\}$   | .9    | .0355 (.023)              | .0183 (.014)              | .01                      | .0232 (.013)              | .01                      | 1.94                     |
| $\{2 \ 3 \ 4\}$   | .4    | .4963 (.233)              | .2606 (.150)              | .03                      | .2939 (.132)              | .05                      | 1.90                     |
| $\{3 \ 4 \ 5\}$   | .6    | .4524 (.212)              | .2360 (.128)              | .23                      | .2522 (.110)              | .05                      | 1.92                     |
| $\{4 \ 5 \ 6\}$   | .8    | .3954 (.164)              | .2027 (.093)              | .18                      | .2078 (.082)              | .03                      | 1.95                     |
| $\{2 \ 5 \ 8\}$   | .4    | 3.3864 (.666)             | 1.5288 (.405)             | .56                      | 1.6482 (.343)             | .02                      | 2.22                     |
| $\{5 \ 7 \ 11\}$  | .6    | 3.3690 (.633)             | 1.4936 (.342)             | .48                      | 1.6073 (.301)             | .01                      | 2.26                     |
| $\{8 \ 12 \ 14\}$ | .8    | 3.3020 (.510)             | 1.4632 (.260)             | .52                      | 1.5602 (.237)             | .01                      | 2.26                     |
| $\{9 \ 13 \ 17\}$ | .9    | 3.2622 (.369)             | 1.4426 (.187)             | .55                      | 1.5388 (.176)             | <.01                     | 2.51                     |

a. The average mean squared errors are obtained over 4,000 simulations with associated empirical standard deviations.
b. RE\((SRS|RSENU)=\text{MSE}(SRS)\//\text{MSE}(RSENU)\).

Larger efficiencies were observed in the simulation, relative to our empirical study of the data. This difference is likely due to the greater correlation in the simulated data, compared to the empirical data. Real world data, such as our insect abundance data, is normally less structured in terms of trends in mean responses and may include some unusual and unpredictable observations. As decreasingly correlated data, with respect to the distances between plots, are an assumption upon which the success of our methods is based, the advantages are naturally pronounced when such conditions are imposed.

### 6. CONCLUSION

The chief advantage of Sampling Excluding Neighboring Units (SENU) type sampling plans is that specified neighboring units are deliberately excluded, based upon the assumption that they provide little new information to the sampling effort, in that they have similar responses and can be correlated with neighboring units. Utilizing the 1998 forest canopy insect abundance data, we demonstrated how to construct a sampling plan with an SENU-type design (Random Sampling Excluding Neighboring Units and Identical Rows and Columns) to generate samples that represent population abundance, while avoiding the simultaneous selection of units that are, in some sense, neighbors.

Next, we investigated the performance of SENU sampling plans using small sample sizes. In the field study, the use of sample sizes $n=2$ and 3 was necessary, as the sampling plan restrictions did not allow larger values of sample size from a square grid of $N=9$. However, as the results indicate, even the use of sample sizes $n=2$ and 3 provide relatively good estimations of population parameters in this empirical comparison. Note that the conclusions contained herein are based upon a study of a particular endpoint, namely species abundance. The
preferred sample design and sample size may vary with a different ecological endpoint, such as species diversity, the number of species present.

Our findings show that, (i) smaller sample size, such as 3, provides an adequate estimations of population abundance; and (ii) Random Sampling Excluding Neighboring Units (RSENU) is the most efficient sampling design investigated in this study. In fact, all SENU-type sampling plans appear to be more efficient than simple random sampling in our empirical study. These results were also supported by the simulation study described in Section 5. In cases where neighboring units are known to be similar, we therefore, believe SENU-type sampling designs to be of significant value.

While this discussion touches on just a few techniques that have been developed, and ideas that have been investigated in the two-dimensional case, much more work in this area lies ahead. Plans of this nature deserve more attention and further theoretical development is needed. Progress can take the form of developing better estimators, developing variance estimators, or finding similar plans that have simpler inclusion probabilities, yet possess all the merits of Random Sampling Excluding Neighboring Units and Random Sampling Excluding Contiguous Units.

REFERENCES


Difference Based Variance Estimators for Partially Linear Models
Karon Klipple and R. L. Eubank
San Diego City College and Arizona State University

Abstract. A difference based variance estimator is proposed for use in partially linear models. The estimator is shown to be asymptotically normal and to converge to the true variance at a parametric rate. Pseudo-residuals are defined for use with partially linear models that lead to development of diagnostic tests for heteroskedasticity. The ideas in the paper are illustrated with a data set concerning the relationship between a fetal liver antigen and subsequent incidences of cancer.

1. Introduction. A seminal paper by Gasser, Sroka and Jennen-Steinmetz (1985) (abbreviated as GSJS hereafter) prompted the development of new methods for estimation of the residual variance in nonparametric regression by demonstrating that the variance could be easily estimated with parametric efficiency without having to first estimate the underlying regression curve. In this paper we investigate how to extend the GSJS and related difference based estimators to semiparametric, partially linear, regression models.

The standard nonparametric regression model assumes that we observe responses \( y_1, \ldots, y_n \) at “time” points \( 0 < t_1 < \cdots < t_n < 1 \) with the \( t_i \) and \( y_i \) related via

\[
y_i = \mu(t_i) + \varepsilon_i, \quad i = 1, \ldots, n.
\]

Here \( \mu(\cdot) \) is an unknown function that is the inferential object and the \( \varepsilon_i \) are zero mean random errors with common variance \( \sigma^2 \). Provided that \( \mu(\cdot) \) is differentiable and the \( t \) ordinates are closely spaced, it is possible to (asymptotically) remove the effect of the mean function by differencing the data appropriately. The idea appears to have first been used by von Neumann, et al (1941) who proposed using successive differences to obtain the variance estimator

\[
\hat{\sigma}_N^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2
\]

that could be employed in situations where the mean was not constant. This estimator was used by Rice (1984) for estimation of the variance term in a bandwidth selection criterion for nonparametric kernel regression. Rice also suggested that one could similarly use residuals from straight line fits involving triples of points in lieu of successive differences. This latter approach was employed by GSJS to produce the estimator

\[
\hat{\sigma}_{GSJS}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n-2} \tilde{\varepsilon}_i^2
\]

where the \( \tilde{\varepsilon}_i \) are the so-called pseudo-residuals defined by

\[
\tilde{\varepsilon}_i = d_{i0}y_i + d_{i1}y_{i+1} + d_{i2}y_{i+2},
\]

with

\[
d_{i0} = -\frac{a_i}{(1 + a_i^2 + b_i^2)^{1/2}},
\]

\[
d_{i1} = \frac{1}{(1 + a_i^2 + b_i^2)^{1/2}},
\]

\[
d_{i2} = -\frac{b_i}{(1 + a_i^2 + b_i^2)^{1/2}}
\]

313
for
\[ a_i = \frac{t_{i+2} - t_{i+1}}{t_{i+2} - t_i} \quad \text{and} \quad b_i = \frac{t_{i+1} - t_i}{t_{i+2} - t_i} \] (8)

GSJS showed that when the \( \varepsilon_i \) are independent and identically distributed \( \sqrt{n}(\hat{\sigma}^2_{\text{GSJS}} - \sigma^2) \) has a limiting normal distribution. They also demonstrated the utility of pseudo-residuals for diagnosis of heteroskedasticity and lack-of-fit problems.

The \( d_{ij} \) in (4) are related to standard weights that arise in numerical differentiation formulae. This fact allows one to conclude that \( \tilde{\varepsilon}_i \approx d_{0i} \varepsilon_i + d_{11} \varepsilon_{i+1} + d_{12} \varepsilon_{i+2} \) with the accuracy of the approximation being a function of the differentiability of \( \mu(\cdot) \). A similar observation can be made concerning the successive differences that appear in \( \hat{\sigma}^2_N \). Thus, we can consider pseudo-residuals in a more general sense as weighted averages of the observations that are asymptotically free of the response means. Pseudo-residuals of this later form appear in Müller and Stadtmüller (1987) and Hall, Kay and Titterington (1990), for example.

In the present paper we will focus on an extension of model (1) wherein
\[ y_i = x_i^T \beta + \mu(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \] (9)
with \( \mu(\cdot) \) and the \( \varepsilon_i \) as before, \( x_i, i = 1, \ldots, n \), known \( p \)-vectors and \( \beta = (\beta_1, \ldots, \beta_p)^T \) a vector of unknown parameters. This formulation is appropriate for regression problems where some of the predictors can be assumed to have a linear relationship with the response. In the important special case where the \( x_i \) are composed of treatment indicator variables (9) provides a framework for nonparametric analysis of covariance and the inferential object is generally \( \beta \) rather than \( \mu(\cdot) \).

As a result of developments in Speckman (1988), it is known that the parameter vector \( \beta \) in (9) can be estimated with parametric efficiency. This entails that standard error type intervals can be employed for inference about \( \beta \) without the need for bias correction. Such intervals, in turn, require estimation of the error variance. We now show one way to obtain pseudo-residuals and difference based estimators that can be used for this purpose.

To motivate the form of our estimator let us first rewrite model (9) in vector/matrix notation as
\[ y = X\beta + \mu + \varepsilon, \] (10)
where \( y = (y_1, \ldots, y_n)^T, \mu = (\mu(t_1), \ldots, \mu(t_n))^T \), \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \) and \( X = [x_1, \ldots, x_n]^T \). Now let \( d = (d_0, \ldots, d_m)^T \) be a \((m+1)\)-vector whose elements satisfy
\[ \sum_{j=0}^m d_{ij} = 0 \quad \text{and} \quad \sum_{j=0}^m d_{ij}^2 = 1 \] (11)
and define the \((n - m) \times n\) matrix \( D \) to have first and last rows \([d^T, 0_{n-m-1}^T], [0_{n-m-1}^T, d^T]\), respectively, with \( i \)th row \([0_i, d_i^T, 0_{n-i-1}^T]^T \), \( i = 2, \ldots, (n - m - 1) \), where \( 0_r \) indicates a \( r \)-vector or all zero elements. Results in the next section then have the consequence that
\[ Dy \approx DX\beta + DE. \] (12)

Thus, standard linear models considerations suggest estimating \( \sigma^2 \) by
\[ \hat{\sigma}^2 = \frac{y^TD^T(I-P)Dy}{\text{tr}(D^T(I-P)D)} \] (13)
with \( \text{tr} \) the trace function for a square matrix and \( P \) the projection matrix
\[ P = DX(X^TD^TX)^{-1}X^TD^T. \] (14)
This estimator was first proposed in Eubank, et al (1998) and was shown to be a \( \sqrt{n} \)-consistent estimator of \( \sigma^2 \) in Eubank (1999) for the special case where the \( D \) matrix is constructed from the
GSJS weights in (5)–(8). Here we allow for general weights of the form (11) and develop asymptotic
distribution theory for the estimator as well as related statistics for testing for heteroskedasticity.
We will also establish that \( \tilde{\varepsilon} \approx D \varepsilon \) and, accordingly, define the pseudo-residuals for data from
model (10) to be
\[
\tilde{\varepsilon} = (I - P)Dy.
\]
(15)

As a result we can write the estimator in the more intuitive form \( \hat{\sigma}^2 = \tilde{\varepsilon}^T \tilde{\varepsilon} / \text{tr}(D^T (I - P) D) \).

The remainder of the paper is organized as follows. In the next section we establish asymptotic
normality and \( \sqrt{n} \)-consistency for (13) as an estimator of the error variance parameter. Then,
in Section 3, we explore the properties of the pseudo-residuals in (15) as a potential tool for the
detection of heteroskedasticity. The results from Section 3 are applied to a real data set in Section
4. All proofs are collected in the Appendix.

2. Large sample theory. In order to pursue an asymptotic development it is necessary to impose
some structure on the \( t_i \). One approach that is relatively general is to assume that the
ordinates are quantiles of some underlying design density: that is,
\[
\int_0^{t_i} f(t) dt = \frac{i - 1}{n - 1}, \quad i = 1, \ldots, n,
\]
(16)
for some differentiable, positive density \( f \) on \([0,1]\). A condition such as this ensures that the
successive differences \( t_i - t_{i-1} \), \( i = 2, \ldots, n \), are nonzero for all finite \( n \) yet decay to zero at the rate \( n^{-1} \). These latter two restrictions are actually the minimal conditions that are needed on the \( t_i \)
for Theorem 1 below to remain valid.

The main result of this section is provided by the following theorem.

THEOREM 1 Assume that the \( t_i \) satisfy (16), \( \mu(\cdot) \in C^1[0,1] \) and that \( \varepsilon_1, \ldots, \varepsilon_n \) are independent
and identically distributed with \( E \varepsilon_1 = E \varepsilon_1^2 = 0 \) and \( E \varepsilon_1^4 < \infty \). Then,
\[
\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = Z_n + o_p(1)
\]
(17)
with \( \tau^{-1/2}Z_n \) having a limiting standard normal distribution for
\[
\begin{align*}
\tau &= \sigma_0 + 2 \sum_{c=1}^{m} \frac{n - m - c}{n - m} \sigma_c, \\
\sigma_0 &= \sum_{i=0}^{m} d_i^4 + 6 \sigma^4 \sum_{i=0}^{m} \sum_{j=i+1}^{m} d_i^2 d_j^2 - \sigma^4, \\
\sigma_c &= (E \varepsilon_1^4 - \sigma^4) \sum_{i=0}^{m-c} d_i^2 d_{i+c}^2 + 4 \sigma^4 \sum_{i=0}^{m-c-1} \sum_{j=i+1}^{m-c} d_i d_j d_{i+c} d_{j+c}, \quad c = 1, \ldots, m - 1,
\end{align*}
\]
(18)
and \( \sigma_m = (E \varepsilon_1^4 - \sigma^4)d_m^2 d_0^2 \).

Theorem 1 has the implication that an interval with asymptotic probability \((1-\alpha)\) of containing
the true variance parameter is given by
\[
\hat{\sigma}^2 \pm z_{1-\alpha/2} \sqrt{\frac{\tau}{n}}.
\]
(21)
For purposes of confidence interval construction \( \tau \) must be estimated in (21) which requires estimation
of \( E \varepsilon_1^4 \). One possible approach is to use \( \sum_{i=1}^{n-m} \varepsilon_i^4 / (n - m) \) for this purpose. In the special case
of error distributions with zero kurtosis the problem becomes somewhat simpler in that \( \tau = \sigma^4 \tau^* \) with \( \tau^* \) a function only of the \( d_i \). Specifically,

\[
\tau^* = \sigma^*_0 + 2 \sum_{c=1}^{m} \frac{n-m-c}{n-m} \sigma^*_c, \tag{22}
\]
\[
\sigma^*_0 = 3 \sum_{i=0}^{m} d_i^4 + 6 \sum_{i=0}^{m} \sum_{j=i+1}^{m} d_i^2 d_j^2 - 1, \tag{23}
\]
\[
\sigma^*_c = 2 \sum_{i=0}^{m-c} d_i^2 d_{i+c}^2 + 4 \sum_{i=0}^{m-c-1} \sum_{j=i+1}^{m-c} d_i d_j d_{i+c} d_{j+c}, \quad c = 1, \ldots, m-1, \tag{24}
\]

with \( \sigma^*_m = 2d_m^2d_0^2 \). In practice it may often be satisfactory to simply use \( \hat{\sigma}^4 \tau^* \) to estimate \( \tau \).

In order to actually use \( \hat{\sigma}^2 \) with data one must select specific values for \( m \) and the \( d_i \). Optimal weights for this purpose in the context of nonparametric regression have been obtained by Hall, Kay and Titterington (1999). The arguments used to prove Theorem 1 have the implication that such weights are equally applicable to our setting. In particular, this means that optimal weights for \( \hat{\sigma}^2 \) are \( d_0 = .7071 = -d_1 \) for \( m = 1 \), \( d_0 = .809, d_1 = -.5, d_2 = -.309 \) for \( m = 2 \) and \( d_0 = 1.942, d_1 = .2809, d_2 = .3832, d_3 = -.8582 \) for \( m = 3 \).

The Hall, Kay and Titterington (1999) weights are optimal in a minimum variance sense. Dette, Monk and Wagner (1998) have found that using such weights can cause bias difficulties in smaller samples. As a result, they conclude it may be preferable to use weights that are more directly connected to numerical differentiation such as those employed by GSJS. Strictly speaking, the GSJS weights are not covered by Theorem 1 since they depend on \( n \). It is possible to extend the proof of Theorem 1 to account for such situations. Alternatively, under condition (16) the GSJS weights are asymptotically the same as using \( d_0 = d_2 = -1/\sqrt{8}, d_1 = 1/\sqrt{2} \) (Altman and Paulson 1993) so that their use with our estimator can be justified from that perspective.

3. Testing heteroskedasticity. A standard assumption in regression is that the error variance is constant. Violation of this assumption can lead to erroneous inference and, as a result, it is of interest to conduct diagnostic analysis to detect heteroskedasticity.

One approach to detection of heteroskedasticity for model (9) can be based on the pseudo-residuals defined in (15). This practice is motivated by the following result.

**Proposition 2** Assume that the elements of \( P \) are uniformly \( o(n^{-1/2}) \). Then, \( \hat{\varepsilon} = D\varepsilon + e_n \) where \( e_n \) is a random vector whose elements are uniformly \( o_P(1) \).

Proposition 2 justifies our designation of the elements of \( \hat{\varepsilon} \) as “pseudo-residuals” in that they behave asymptotically like the elements of \( D\varepsilon \) whose distribution does not involve \( \mu(\cdot) \). In addition, Proposition 2 suggests that the \( \hat{\varepsilon}_i^2 \) will behave like random variables with a constant mean when the error distribution has constant variance and will have nonconstant means otherwise. This latter feature indicates that examination of plots of the squared pseudo-residuals could be useful for detecting heteroskedasticity. We will discuss some specific plots that may be of interest subsequently.

The condition in Proposition 2 that \( P \) has elements decaying to zero faster than \( n^{-1/2} \) is not particularly restrictive. It is satisfied, for example, if the largest diagonal element of \( P \) is \( o(n^{-1/2}) \). In the special case that the diagonal elements of \( P \) are identical this convergence is actually of order \( O(n^{-1}) \). The following proposition suggests that one can expect the implications of Proposition 2 to apply in situations involving randomized experiments.

**Proposition 3** Let \( m = p = 1, d_0 = 1/\sqrt{2} = -d_1 \) and assume that the elements of the vector \( X \) in (9) have a Bernoulli distribution. Then, the elements of \( P \) are uniformly \( o_P(n^{-1}) \).
Diagnostic plots involving the pseudo-residuals can be obtained, for example, by plotting the \( \tilde{\varepsilon}_i \) against various functions of the \( t_i \) or other variables such as those that represent the columns of \( X \). Summary statistics along the lines of those proposed by Cook and Weisberg (1983) for ordinary regression analysis can aid in the interpretation of such plots. The Cook/Weisberg measures are basically test statistics for the significance of the slope coefficient when regressing a squared residual on the plotting variable. An application of this idea in our setting leads to summary statistics of the form

\[
T = \frac{\sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)(\tilde{\varepsilon}_i^2 - \sigma^2)}{\sqrt{\tau \sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)^2}} = \frac{\sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)\tilde{\varepsilon}_i^2}{\sqrt{\tau \sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)^2}}, \tag{25}
\]

where the \( w_{in} \) are weights that can be viewed as the values of a plotting variable and \( \bar{w}_n \) is their average. Thus, \( T \) provides a summary measure corresponding to a scatter plot of the points, \((w_1, \varepsilon_1^2), \ldots, (w_{(n-m)}n, \varepsilon_{n-m}^2)\). The value of \( \tau \) in this formula can be estimated as discussed in Section 2. To assess the significance of patterns that may appear in a plot \( T \) may be compared to percentiles of a standard normal distribution as a result of the next theorem.

**Theorem 4** Assume that the \( t_i \) satisfy (16), \( \mu(\cdot) \in C^1[0,1] \) and that \( \varepsilon_1, \ldots, \varepsilon_n \) are independent and identically distributed with \( E\varepsilon_1 = E\varepsilon_1^3 = 0 \) and \( E\varepsilon_1^4 < \infty \). Let \( \{w_{in}\}_{i=1}^{n-m} \) be a sequence of uniformly bounded positive weights with the property that

\[
\bar{w}_n = n^{-1} \sum_{i=1}^{n-m} w_{in} = B_1 + o(1) \tag{26}
\]

and

\[
\sum_{i=1}^{n-m-c} (w_{in} - B_1)(w_{(i+c)n} - B_1) = nB_2 + o(n), \quad c = 0, \ldots, m, \tag{27}
\]

for finite, positive constants \( B_1, B_2 \). Then, if the elements of \( P \) are uniformly \( o(n^{-1/2}) \), \( T \) in (25) converges in distribution to a standard normal random variable.

### 4. An example

In this final section we apply the results from Sections 2 and 3 to the analysis of data concerning alpha-feto protein levels in a group of women. The data is shown in Figure 1. The response in this case is the logarithm of alpha-feto protein measured in 857 different women on various days of their particular pregnancies. The subjects were monitored for a period of time subsequent to delivery over which incidences of cancer were recorded. An analysis of this data in Eubank and Li (2005) suggests that model (9) is appropriate with the \( t_i \) representing the day of pregnancy that the blood sample was taken and the \( x_i \) being scalar indicator variables that take on the values 1 or 0 depending on whether or not the subject eventually developed cancer. In contrast to the Eubank/Li development our focus here is on the issue of whether the data exhibits heteroskedasticity and whether there is a statistically significant cancer effect.

In accordance with the recommendation of Dette, Munk and Wagner (1998) we will use the GSJS weights from (5)–(8). However, we need to adjust this formulation to allow for the fact that there are repeated responses at some of the \( t \) ordinates in our data. In this respect we will follow the approach outlined in GSJS. If all three \( t \) ordinates are the same we use \( a_i = b_i = 1/2 \) in (8). In the case where only two ordinates are the same let \( t_{i+1} \) be the middle \( t \) ordinate and observe that there will be an integer \( k \) such that either \( t_{i-k} \leq t_{i-k+1} = \cdots = t_{i+1} < t_{i+2} \) or \( t_i < t_{i+1} = \cdots = t_{i+k+1} < t_{i+k+2} \). In the first instance we take

\[
a_i = \frac{t_{i+2} - t_{i+1}}{t_{i+2} - t_{i-k}} \quad \text{and} \quad b_i = \frac{t_{i+1} - t_{i-k}}{t_{i+2} - t_{i-k}}.
\]

317
and use

\[ a_i = \frac{t_{i+k+2} - t_{i+1}}{t_{i+k+2} - t_i} \quad \text{and} \quad b_i = \frac{t_{i+1} - t_i}{t_{i+k+2} - t_i} \]

for the second.

The squared pseudo-residuals obtained from the GSJS weights are plotted in Figures 2 and 3 against the (mean corrected) \( t_i \) and \( t_i^2 \). Neither of the plots exhibits a distinguished wedge shaped pattern that one would expect from heteroskedastic errors (e.g., Cook and Weisberg 1983). This visual examination is supported by the values of the corresponding test statistics (25) which are -1.011 and -0.832 for Figures 2 and 3, respectively. The parameter \( \tau \) that appears in the test statistics was estimated by \( \hat{\sigma}^4 \tau^* \) with \( \tau^* \) as defined in (22)-(24) and the \( \sigma_c^* \) were computed using the asymptotic form of the GSJS weights. The value of \( \hat{\sigma}^2 \) was found to be .266. Use of the alternative Hall, Kay and Titterington optimal weights for \( m = 2 \) produced a similar value of .275. We also computed \( \sum_{i=1}^{n-m} \varepsilon_i^4 / (n-m) \) to be .2384 \( \approx 3 \hat{\sigma}^4 \) = .2127. Thus, either of the two approaches for estimating \( \tau \) will lead to the same basic conclusions in this case. The largest diagonal element of \( P \) was .009.

If we now proceed under the assumption of constant error variance we can use nonparametric smoothing methods for partially linear models to test for the presence of a cancer effect: i.e., we can test the hypothesis that \( \beta = 0 \). If \( S \) denotes a symmetric smoother matrix with respect to the \( t \) ordinates, then the Speckman (1988) estimator of \( \beta \) is \( \hat{\beta} = (X^T(I - S)^2 X)^{-1} X^T(I - S)^2 y \) with
associated variance-covariance matrix obtained as the product of $\sigma^2$ with the matrix

$$V = (X^T(I - S)^2X)^{-1}X^T(I - S)^4X(I - S)^2X)^{-1}. $$

Using a cubic smoothing spline with unit smoothing parameter we obtained $\hat{\beta} = .011$ and $V = .0054$. Combining this with $\hat{\sigma}^2 = .266$ from before produces an approximate, estimated standard error of .038 for $\beta$. As a result we cannot conclude that $\hat{\beta}$ differs significantly from 0.
5. Appendix. In this section we provide details for the proofs of Theorems 1 and 4 and for Propositions 2 and 3. We begin with the proof of Theorem 1.

5.1. Proof of Theorem 1. The basic idea behind the proof stems from the decomposition

\[
\sqrt{n} \left( \frac{y^T D^T (I - P) D y}{\operatorname{tr}(D^T (I - P) D)} - \sigma^2 \right) = \sqrt{n} \left( \frac{\varepsilon^T D^T D \varepsilon}{\operatorname{tr}(D^T (I - P) D)} - \sigma^2 \right) \\
- \sqrt{n} \frac{\varepsilon^T D^T P D \varepsilon}{\operatorname{tr}(D^T (I - P) D)} + 2 \sqrt{n} \frac{\varepsilon^T D^T (I - P) D \mu}{\operatorname{tr}(D^T (I - P) D)} + \sqrt{n} \frac{\mu^T D^T (I - P) D \mu}{\operatorname{tr}(D^T (I - P) D)}. \tag{28}
\]

Lemmas 5–8 below show that all but the first term on the right hand side of (28) are asymptotically negligible and that this remaining term has the same limiting distribution as \( \sqrt{n}(\varepsilon^T D^T D \varepsilon / \operatorname{tr}(D^T D) - \sigma^2) \). But this quantity is the same as the one that arises when studying the limiting distribution of difference based variance estimators under the ordinary nonparametric regression model (1). The conclusion of the theorem is therefore a consequence of the central limit theorem for \( m \)-dependant random variables and arguments in GSJS.

**Lemma 5** Assume that the \( t_i \) satisfy (16) and that \( \mu \in C^1[0, 1] \). Then, \( \mu^T D^T D \mu = O(n^{-1}) \).
Proof. As a result of (11), (16) and a Taylor expansion we can write
\[
\sum_{j=0}^{m} d_j \mu(t_{i+j}) = \sum_{j=0}^{m} d_j \mu(t_{i+[m/2]}) + \sum_{j=0}^{m} d_j e_{ij} = \sum_{j=0}^{m} d_j e_{ij}
\]
where the \( e_{ij} \) are uniformly of order \( n^{-1} \). The result is then a consequence of the fact that
\[
\mu^T D^T D \mu = \sum_{i=1}^{n-m} (\sum_{j=0}^{m} d_j e_{ij})^2 \leq \sum_{i=1}^{n-m} \sum_{j=0}^{m} e_{ij}^2.
\]

**Lemma 6** \( \varepsilon^T D^T (I - P) D \mu = O_p(n^{-1/2}) \).

Proof. The quantity \( \varepsilon^T D^T (I - P) D \mu \) has mean zero and its variance is bounded above by the product of \( \mu^T D^T D \mu \) and the largest eigenvalue of \( DD^T \). Now \( DD^T \) is \((m+1)\)-banded matrix with non-zero elements of the form \( \sum_{i=j}^{m} e_{ij} \). Gershgorin’s theorem (e.g., Franklin 1968, Section 6.8) therefore implies that the largest eigenvalue of \( DD^T \) is bounded independent of \( n \) which completes the proof.

**Lemma 7** \( \text{tr}[D^T (I - P) D] = n - m + O(1) \).

Proof. Condition (11) entails that \( \text{tr}[D^T (I - P) D] = n - m - \text{tr}[D^T P D] \). Since \( P \) is a projection matrix we can write \( \text{tr}[D^T P D] = \sum_{i=1}^{\text{rank}(P)} e_i^T D D^T e_i \) with the \( e_i \) being the eigenvectors corresponding to the \( \text{rank}(P) \) unit roots of \( P \). But, the \( e_i^T D D^T e_i \) are uniformly bounded by the largest eigenvalue of \( DD^T \) which, in turn, is uniformly bounded independent of \( n \).

**Lemma 8** \( \varepsilon^T D^T P D \varepsilon = O_p(1) \).

Proof. We have \( E \varepsilon^T D^T P D \varepsilon = \sigma^2 \text{tr}[D^T P D] \leq \sigma^2 \text{rank}(P) \lambda_{\text{max}}(DD^T) \) with \( \lambda_{\text{max}}(DD^T) \) the largest eigenvalue of \( DD^T \). An application of Markov’s Theorem completes the proof.

To finish the proof of Theorem 1 observe from Lemmas 7 and 8 that
\[
\sqrt{n} \frac{\varepsilon^T D^T P D \varepsilon}{\text{tr}(D^T (I - P) D)} = O_p(n^{-1/2})
\]
while Lemmas 5 and 6 in conjunction with Lemma 7 ensure that
\[
\sqrt{n} \frac{\varepsilon^T D^T (I - P) D \mu}{\text{tr}(D^T (I - P) D)} = O_p(n^{-1}) \quad \text{and} \quad \sqrt{n} \frac{\mu^T D^T (I - P) D \mu}{\text{tr}(D^T (I - P) D)} = O(n^{-3/2}).
\]

We now write
\[
\sqrt{n} \left( \frac{\varepsilon^T D^T D \varepsilon}{\text{tr}(D^T (I - P) D)} \right) = \sqrt{n} \left( \frac{\varepsilon^T D^T D \varepsilon}{\text{tr}(D^T D)} - \sigma^2 \right) + \sqrt{n} \varepsilon^T D \varepsilon \left[ \frac{1}{\text{tr}(D^T (I - P) D)} - \frac{1}{\text{tr}(D^T D)} \right]
\]
and use Lemma 7 along with the fact that \( \varepsilon^T D^T D \varepsilon/(n - m) \xrightarrow{p} \sigma^2 \) to finish the proof.

**5.2. Proofs of Propositions 2 and 3.** To verify Proposition 2 first write \( \tilde{\varepsilon} = (I - P)(D \mu + D \varepsilon) \). Now observe that the elements of the second term \( (I - P)D \mu \) are bounded in magnitude by the
square root of $\mu^T D^T (I - P) D \mu \leq \mu^T D^T D \mu$ which means they are uniformly $O(n^{-1/2})$ as a result of Lemma 5. Thus, it suffices to focus on the behavior of $(I - P) D \varepsilon$.

Let 

$$\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^T = D \varepsilon.$$ 

Then, $(I - P) D \varepsilon$ has typical element $\hat{\varepsilon}_i + \sum_{j=1}^{n-m} P_{ij} \hat{\varepsilon}_j$ with $P_{ij}$ denoting the elements of $P$. This quantity has mean zero and variance

$$\sum_{k=1}^{n-m} \sum_{r=1}^{n-m} P_{kr} \text{Var}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = 2 \sum_{k=1}^{n-m} \sum_{r=k}^{n-m} P_{kr} \sigma_{0(r-k)} \leq 2(n - m) \max \{\sigma_{ij} \mid P_{ij}^2 \}.$$

The desired conclusion is now a consequence of the stated assumptions on the elements of $P$.

For the proof of Proposition 3 let $X = (x_1, \ldots, x_n)^T$. Then, we have

$$P_{ij} = \frac{1}{2} \frac{(x_{i+1} - x_i)(x_{j+1} - x_j)}{\sum_{k=1}^{n-1} (x_{k+1} - x_k)^2}.$$ 

The numerator of this expression is at most 1 in magnitude and the denominator has mean $2(n - 1)\theta(1 - \theta)$ with $\theta$ the success probability associate with the $x_i$. Thus, $\max_{i,j} |P_{ij}| = O_P(n^{-1})$ from Markov’s inequality.

5.3. Proof of Theorem 4. First observe that $T$ in (25) can be written as

$$T = \frac{\sum_{i=1}^{n-m} (w_{in} - B_1)(\hat{\varepsilon}_i^2 - \sigma^2)}{\sqrt{\tau} \sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)^2} + (B_1 - \bar{w}_n) \frac{\sum_{i=1}^{n-m} (\hat{\varepsilon}_i^2 - \sigma^2)}{\sqrt{\tau} \sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)^2}.$$

As a result of Theorem 1 and assumptions (26)–(27) the second term on the right hand side of this expression is $o_P(1)$. Thus, it suffices to work with the first term on the right hand side of (30).

Now define

$$W = \text{diag}(w_{1n} - B_1, \ldots, w_{(n-m)n} - B_1).$$

and proceed in a similar fashion to the proof of Theorem 1 by working with the decomposition

$$\frac{\sum_{i=1}^{n-m} (w_{in} - B_1)(\hat{\varepsilon}_i^2 - \sigma^2)}{\sqrt{\tau} \sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)^2} = \frac{\sum_{i=1}^{n-m} (w_{in} - B_1)(\hat{\varepsilon}_i^2 - \sigma^2)}{\sqrt{\tau} \sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)^2} + \frac{1}{\sqrt{\tau} \sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)^2} [\hat{\varepsilon}^T PW \hat{\varepsilon}]$$

$$= \frac{\sum_{i=1}^{n-m} (w_{in} - B_1)(\hat{\varepsilon}_i^2 - \sigma^2)}{\sqrt{\tau} \sum_{i=1}^{n-m} (w_{in} - \bar{w}_n)^2}$$

$$- 2 \hat{\varepsilon}^T WP \hat{\varepsilon} + \mu^T D^T (I - P) W (I - P) \hat{\varepsilon} + \mu^T D^T (I - P) W (I - P) D \mu$$

with $\hat{\varepsilon}$ as defined in (29). Lemmas 5, 6 and 8 along with the boundedness of the $w_{in}$ insure that $\mu^T D^T (I - P) W (I - P) D \mu = O(n^{-1})$, $\mu^T D^T (I - P) W (I - P) \hat{\varepsilon} = O_P(n^{-1/2})$ and $\hat{\varepsilon}^T WP \hat{\varepsilon} = O_P(1)$.

Thus, the proof will follow from the central limit theorem for $m$-dependant random variables if we show that $\hat{\varepsilon}^T WP \hat{\varepsilon} = o_P(\sqrt{n})$ under our conditions on $P$.

Define $V = \{v_{ij}\} = D^T W$ and observe that $V$ is (lower) band limited with uniformly bounded elements. The typical element of $D^T WPD$ is of the form

$$\sum_{k=i-m-1}^{i} \sum_{r=j}^{j+m+1} v_{ik} d_{r-j} P_{kr}.$$
and therefore of uniform order \( o(n^{-1/2}) \).

Let \( H = D^T WP + D^T PW D \) and set \( Q = \varepsilon^T D^T WP D \varepsilon = \varepsilon^T H \varepsilon / 2 \). The matrix \( WP + PW \) is symmetric and has eigenvalues that are twice those of the positive semi-definite matrix \( W^{1/2} PW^{1/2} \) as a result of the Courant Minimax Theorem (Franklin 1968, Section 6.3). Thus, \( H \) is also positive semi-definite. Since \( EQ = \sigma^2 \text{tr} H / 2 = o(\sqrt{n}) \), Markov’s Theorem entails that \( \hat{\varepsilon}^T WP \hat{\varepsilon} = o_p(\sqrt{n}) \) and thereby completes the proof.

**Acknowledgements.** Eubank’s research was supported by NSF.

**References.**


Bootstrap Bias Reduction and Estimator Decomposition

Kristofer Jennings
Department of Statistics
Purdue University
West Lafayette, IN 47907

Abstract

Methods for bias reduction using the iterated bootstrap are presented for estimators with and without unbiased estimators. These techniques can be additionally be applied to problems of estimating variance and mean squared error without bias. Parameters which do not have unbiased estimators can be modified using a projection into a linear subspace so that a bias-reduced estimator and associated variance can be stably computed. Finally, a summary discussion of implementation techniques is given for producing these estimators efficiently.

1 Introduction

Efron’s (1979) bootstrap has been applied to numerous problems in statistics, both in the theoretical realm as well as in many important data analysis contexts. The original purpose for the development of the bootstrap was to generalize Quenouille’s jackknife (1949, 1956) for variance estimation and bias reduction. The jackknife’s effectiveness for bias reduction has been well-developed for a variety of problems.

To improve the jackknife’s asymptotic properties, Schucany, Gray, and Owen (1971) show that jackknife-based estimation could be used to reduce the bias of a broad class of estimators to an arbitrarily small level. The Schucany, Gray, and Owen (1971) approach achieves dramatic results by using a second level of resampling (that is, jackknifing the jackknifed estimator). Efron (1983) followed suit by showing how iterating the bootstrap (or, “bootstrapping the bootstrap”) could be used to sharpen the accuracy of many bootstrap methods. For the problem of bias estimation, Hall and Martin (1988) develop a technique that uses bootstrap iteration and show how it improves the asymptotic rate of convergence of the bias by orders of $n^{-1/2}$. (Hall (1992) and Shao and Tu (1995) present a fairly comprehensive discussion of this approach.)

Interest in the problem of bias reduction has perhaps waned since Hall and Martin’s (1989) seminal work codified an effective nonanalytic approach to this problem. Unbiased (or nearly unbiased) estimators are notoriously variable (Doss and Sethuraman, 1983); as a result, the application of bias reduction techniques rarely is considered important beyond theoretical discussions or difficult applications, such as density estimation or aspects of times series analysis. A renewed interest in using the bootstrap for point estimation can be found in research work on bagging (Friedman and Hall, 1999) and robust methodology (Ali and Imon, 2005). However, issues of point estimation such as bias represent a useful springboard to more “nonlinear” problems such as variance estimation or inference in highly nonlinear settings. The work presented here develops further the approach of Hall and Martin (1989) using the discussions of Chan and Lee (2001) to provide context.

This paper proceeds as follows. Section 2 develops the theory of bias reduction using the bootstrap for parameters that can be estimated without bias. Application to problems of variance and mean squared error estimation are discussed in Section 3. This theory is generalized for bias reduction in the context of estimators with no possible unbiased estimators available in Section 4. Finally, algorithms for implementation are summarized in Section 5.
2 Producing Unbiased Estimators with the Bootstrap

Hall and Martin (1988) consider a real-valued functional parameter \( \theta(F) \) that has a “plug-in” estimator \( \hat{\theta}(F_n) \), where \( F_n \) is the empirical cumulative distribution function estimator based on the i.i.d. sample \( \{X_1, \ldots, X_n\} \). If the expectation of any plug-in function \( \theta_{\ell-1}(F_n) \) is given by the expression \( \theta(F) + c(F)n^{-j} + O_p(n^{-m}) \) for some \( j \leq m - 1 \) and sufficiently smooth function \( c(\cdot) \) that does not depend on \( n \), then the function \( \theta_{\ell}(F_n) \), given by

\[
\theta_{\ell}(F_n) = \theta(F) + \theta_{\ell-1}(F_n) - E[\theta_{\ell-1}(F_n^*|F_n)],
\]

where \( E[\theta_{\ell-1}(F_n^*|F_n)] \) is the bootstrap expected value of \( \theta_{\ell-1}(F_n) \), has expectation \( \theta(F) + R_n \), where \( R_n \) converges to zero at rate \( O_p(n^{-j+1}) \). In general, if \( E[\theta(F_n)] - \theta(F) \) is a sufficiently smooth function converging to zero at rate \( O_p(n^{-j}) \), the estimator

\[
\theta_{\ell}(F_n) = \sum_{j=1}^{\ell+1} \binom{\ell + 1}{j} (-1)^{j+1} E[\theta(F_n,j)|F_n],
\]

where \( E[\theta(F_n,j)|F_n] \) is the \( j \)th bootstrap iteration mean of \( \theta(F_n) \), converges in expected value to \( \theta(F) \) at rate \( O_p(n^{-j+1}) \).

Chan and Lee (2001) develop a discrete Markov chain process approach to describe bootstrap iteration based on functions of the vector of completely enumerated resampled statistics. As a result, the limiting bias-reduced estimator \( \theta_\infty = \lim_{n \to \infty} \theta_{\ell}(F_n) \) can be computed exactly with a finite (if often prohibitive) amount of calculation. Jennings (2002) shows how \( \theta_\infty \) is an unbiased estimator for \( \theta(F) \), as long as an unbiased estimator exists. This condition is equivalent to being able to express \( E[\theta(F_n)] \) as a finite polynomial of \( n \) with coefficient functionals of \( F \). An important condition for this unbiasedness property requires that the sample size \( n \) is large enough. In particular, if the plug-in estimator \( \theta(F_n) \) can be expressed as a \( V \)-statistic \( \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq n} \psi_k(X_{i_1}, \ldots, X_{i_k}) \) with kernel \( k \) no greater than \( n \), then an unbiased estimator can be derived; however, if no such \( V \)-statistic representation is possible, then any estimator calculated via resampling the plug-in estimator will necessarily be biased.

Since the bias of the estimator with infinite \( V \)-statistic expansions (Shieh, 1994) depends on the size of the associated infinite U-statistic (Kohatsuiga, 1991), establishing the asymptotic properties of the bias of these limiting estimators is very difficult analytically. Finite sample simulations show that the \( \theta_\infty \) statistic has impractically large biases; an empirical method of “projecting” the estimator into small-dimensional U-statistic space will be presented in Section 4.

3 Variance and Mean Squared Error Estimation

A useful corollary result to the construction of unbiased estimators \( \theta_\infty \) from Section 2 is that if a \( k \)-degree construction of an unbiased estimator exists and \( n \geq 2k \), then the Chan and Lee (2001) construction can be used to find an unbiased estimator of the variance of the \( \theta_\infty \). This estimator is the sign-flipped limit of the bootstrap variance (calculated via a matrix formulation) as the Hall and Martin (1988) process reaches its limit. (The value of 0 is substituted as the “bootstrap variance” of the original estimator.)

The formulation of the mean squared error (MSE) involves both the estimation of the (squared) bias and the variance. In machine learning and other statistics applications, having an unbiased estimator of MSE is an important step to model selection. Jennings (2002) develops an argument which shows that if \( \hat{\theta}_0 \) is an unbiased estimate of \( \theta_0 \), then the MSE of an alternative estimator \( \hat{\theta}_\gamma \) (which is unbiased for the parameter \( \theta_\gamma \)) is expressible as

\[
E[\hat{\theta}_\gamma - \hat{\theta}_0]^2 + 2cov(\hat{\theta}_\gamma, \hat{\theta}_0) - var(\hat{\theta}_0),
\]

which can be estimated with no bias using the formulation from the previous section. A simulation study of this estimation technique for ridge regression is given in Jennings (2002).
4 An Empirical Hoeffding Decomposition with Remainder

As mentioned in previous sections, bias elimination is possible only for those statistics for which an unbiased estimator exists at that sample size. At the same time, there is a great deal to be gained (via variance and MSE estimation) by analyzing this general class of estimators. In this section, an algorithm is presented for decomposing resampled statistics into a U-statistic space and an orthogonal remainder space. This decomposition is an empirically-based generalization of the U-statistic decomposition and closely resembles the ANOVA decomposition of Efron and Stein (1981).

To see this, consider the vector of completely enumerated versions of the resampled statistic \( \theta^*(F_n) \). A linear combination of these statistics produces the bias-reduced estimators \( \theta_l \) from (1) in Section 2 as well as the limiting estimator \( \theta_\infty \). The space of \( k \)-dimensional V-statistics can be expressed as a linear subspace within this vector space of resampled statistics. As a result, any statistic for which all resampled versions of this statistic are finite can be projected into this linear subspace. For true V-statistics with kernel \( \psi \) of degree no greater than \( k \), this projection reproduces the original resampled statistics; for general finite-valued statistics, there is a (typically small) remainder. Thus, any finite resampled statistic \( \theta^*(F_n) \) can be decomposed into a V-statistic, for which a \( k \)-dimensional U-statistic \( (k \leq n) \) and its unbiased variance estimator can be constructed (as long as \( k \leq n/2 \)). For the remainder, the usual sample variance calculation in the orthogonal projection usually serves well as a bootstrap variance estimator that, together with the unbiased variance estimator, can be used for confidence interval construction or MSE evaluation. The properties of this estimator have not been completely developed, but simulation results show very favorable finite-sample bias properties.

5 Implementation

Extensive discussions of implementation issues of bootstrap bias reduction can be found in Hall and Martin (1988), Hall (1992), Chan and Lee (2001), and Jennings (2002). We summarize the major approaches here.

The initial approach to bootstrap bias reduction was Hall and Martin’s (1988) iterated bootstrap form, as formulated in (1), was a linear combination of bootstrap means at different iteration levels. It is important to note that unlike many other applications of the iterated bootstrap, sampling at each iteration can be done independently; as a result, the usually prescribed nesting sampling structure for resampling iteration is not required. Thus, this formula is computationally easy to take out to an arbitrary level of iteration.

Chan and Lee (2001) suggest a method for bootstrap iteration evaluation based on matrix algebra and complete enumeration of the bootstrap space. Each resampled statistic takes a place in a vector, upon which the bootstrap acts as a matrix multiplier; the iterated bootstrap value is the transformed value in the entry for the original estimator. Beyond the difficulty of evaluating the completely enumerated bootstrap space (for a sample size of 7, the space already contains 1726 values to compute), the matrix algebra quickly becomes prohibitively difficult. Jennings (2002) suggests a simplification of the bootstrap algebra calculation wherein each entry in the new vector is the sum of all resampled statistics with the same permuted weight vector. Other algorithmic considerations, such as the projection on the U-statistic space described in Section 4, can also be accomplished easily in this space.

Another approach to the calculation of bias-reduced estimators uses the independent iterated bootstrap calculations from the initial Hall and Martin (1988) algorithm in this section. However, instead of a linear combination of the means, the bias-reduced is a projection of a linear regression of the resampled values along the “iteration” axis and projected to “iteration -1”. It can be shown that the iterated bootstrap distribution has expected value that is a linear combination of functions of the form \( \sum_{i=1}^n a_i \lambda_i^\ell \), where \( \ell \) is the iteration index, and the \( \lambda_i^\ell \)'s are eigenvalue of the bootstrap transition matrix. For a particular value of \( n \), values of \( \lambda_i \) can be shown to be of the form \( \lambda_i = n^{-1} \frac{i^\ell}{\binom{n}{\ell}} \) for \( i = 1, \ldots, n \). In practice, this process of regression and projection is not stable unless the coefficients on \( \lambda_i \) are severely shrunk or zeroed out for moderate to large values of \( i \).
Using this last method, estimates of other values can be improved as well. For instance, the variance can be regressed on the $\lambda_i^j$ through the value of zero (at the “zeroth” iteration) to the negative of a bias-reduced estimate of the variance of the new estimator. Similarly, regressing on just a few $\lambda_i$’s can dramatically improve estimation of bootstrap quantiles and thus be used to improve the coverage properties of the percentile bootstrap.

6 Conclusion

Using the iterated bootstrap to improve bias has not received much attention recently. However, the approaches described in this paper demonstrate that the properties of these techniques can continue to be improved. Further development of algorithms represent a promising area of study for the improvement of point estimates and related inferential results.

References


Bayesian Prediction for the Linear Model With Equi-correlated Responses: An Application of the Generalized Multivariate Modified Bessel Distribution

Lehana Thabane  
Department of Clinical Epidemiology and Biostatistics  
McMaster University  
Hamilton, ON L8N 1G6, Canada  
Email: ThabaneL@mcmaster.ca

and

B. M. Golam Kibria  
Department of Statistics  
Florida International University  
University Park  
Miami FL 33199, USA  
Email: kibriag@fiu.edu

Summary

This paper provides the Bayesian prediction for the linear regression model with equi-correlated responses. Assuming the generalized inverse Gaussian prior density for the scale or variance parameter of the general linear model, we obtain the predictive distribution to be the generalized multivariate modified Bessel-type distribution. This result is in contrast with the Student-t distribution obtained using inverted chi-square prior distribution for variance. The predictive distribution depends on the intra-class correlation and the possible estimation strategies of the parameters are discussed.

1 Introduction

A $p$-dimensional random variable $Y = (Y_1, ..., Y_p)$ is said to have a generalized modified Bessel $(GMBD_p)$ with location parameter $\mu$ and scale matrix $\Sigma$ if it’s density function is given by

$$ p(y) = \frac{|\Sigma|^{-\frac{1}{2}} \left( \frac{\lambda}{2} \right)^{\frac{\nu}{2}}}{(2\pi)^{\frac{p}{2}} K_{\nu} \left( \sqrt{\psi} \right)} \left\{ 1 + \frac{1}{\psi} (y - \mu) \Sigma^{-1} (y - \mu) \right\}^{\frac{\nu-\frac{1}{2}}{2}} \times K_{\frac{\nu-\frac{1}{2}}{2}} \left( \sqrt{\lambda \psi} \left[ 1 + \frac{1}{\psi} (y - \mu) \Sigma^{-1} (y - \mu) \right] \right),$$

(1.1)

where $K_{\nu}(z)$ is the modified Bessel function of the third kind (see Gradshteyn and Ryzhik (1980, p. 970)). The domain of the parameters $(\nu, \psi, \lambda)$ is given by

$$\psi > 0, \; \lambda \geq 0 \quad \text{for} \quad \nu < 0$$

$$\psi > 0, \; \lambda > 0 \quad \text{for} \quad \nu = 0$$

$$\psi \geq 0, \; \lambda > 0 \quad \text{for} \quad \nu > 0,$$

(1.2)
Figure 1: Joint pdf $p(\epsilon_1, \epsilon_2)$ for different values of $(\psi, \lambda, \nu)$: $A=(1,1,-1); B=(1,1,1); C=(1,2,-1); D=(1,2,1)$.

Figure 1 provides some plots of the distribution (1.1) for $p = 2$ and for different values of $(\psi, \lambda, \nu)$. Thabane and Haq (2003) investigated some statistical properties and Bayesian applications of the distribution (1.1). Thabane and Drejic (2002, 2003) studied some hypothesis testing problems under model (1.1). We note here that pdf (1.1) is a member of the spherically symmetric class of distributions (cf. Ng et al., 1990), a special case of the symmetric multivariate hyperbolic distributions of Barndorff-Nielsen (1978) (cf. Thabane and Kibria 1991 and Thabane and Haq 2003), and member of the elliptically symmetric class of distributions (see Fang et al. (1990)). The multivariate modified Bessel distribution proposed by Bhattacharya (1987) is a special case of this distribution for $\Sigma = I, \lambda > 0, \psi > 0$ and $\nu > p/2$. For more details on other special cases of distribution (1.1), we refer the reader to Thabane and Haq (2003).

Following Thabane and Haq (2003), a $p$-dimensional variable $Y$ with distribution (1.1) will be denoted by

$$Y \sim GMBD_p(\mu, \Sigma, \lambda, \psi, \nu).$$

In this paper, we consider the normal linear model with correlated errors. We assume an exchangeable correlation structure between the errors (i.e. the errors are equi-correlated). The objective is to adopt the
Bayesian framework using informative priors to develop procedures for making inferences about future unobserved responses. In particular, we assume a generalized inverse Gaussian prior for the scale parameter and demonstrate that the results lead to Bessel-type distribution with pdf of the form given by equation (1.1), which is in contrast to the Student-t distribution usually obtained with the use of an inverted chi-squared prior.

The rest of the paper is as follows: In Section 2, we introduce the linear model and derive the posterior distribution for the model coefficients. Section 3 deals with the predictive inferences for futures responses. Some concluding remarks are given in Section 4.

2 The Model

Consider the following linear regression model

$$y = X\beta + \epsilon,$$  

(2.3)

where \( y \) is an \( n \times 1 \) vector of responses, \( X \) is an \( n \times p \) matrix of rank \( p \) of non-stochastic elements, \( \beta \) is an \( p \times 1 \) vector of regression parameters, \( \epsilon \) is an \( n \times 1 \) error vector with \( \epsilon \sim N_n \left(0, \tau^2 \Omega \right) \), where \( \Omega \) is an \( n \) dimensional positive definite matrix and has the following intra-class correlation structure

$$\Omega = \begin{pmatrix}
1 & \rho & \rho & \cdots & \rho \\
\rho & 1 & \rho & \cdots & \rho \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\rho & \rho & \rho & \cdots & 1
\end{pmatrix}$$

$$= (1 - \rho)I_n + \rho 1_n 1_n' - \frac{1}{n-1} < \rho < 1.$$  

(2.4)

Here \( I_n \) as an \( n \)-dimensional identity matrix and \( 1_n \) is \( n \times 1 \) vector of ones. The intra class correlation coefficient \( \rho \) is often used to measure the degree of intra-family resemblance with respect to biological attributes such as height, weight, blood pressure etc. It also arises in education, genetics and psychology, where the population may be divided into clusters. For example, in sampling from a biological population, it is advantageous to select a sample of population of clusters and then to select a sample of organisms within the clusters. Such kind of sampling procedure is that sample observations within a cluster may exhibit a residual covariance of intra-class correlation structure rather than diagonal form (Wiorkowski, 1975). Normal regression theory with intra-class correlation structure for the error variables has been considered by various researchers. To mention a few, Halperin (1951), Han (1969), Rao and Wang (1995), Haq and Ng (1979), Jelenkowska (1999), Donner and Koval (1980a, b) and Ashworth and Armstrong (2003). Haq (1978) considered the location-scale model with equi-correlated responses to derive the predictive distribution for a Gaussian regression model. Kibria and Haq (1999) considered the multivariate linear model with error having a multivariate t distribution and intra-class correlation structure. They considered the structure relation of the model and derived the predictive distribution as a multivariate t for known \( \rho \). The Bayesian estimation of
the intra-class correlation coefficients in multivariate mixed linear model is considered by Jelenkowska (1999).

### 2.1 Prior Distribution

The likelihood function is given by

$$L(\beta, \tau^2 | \Omega, y) = \left( \frac{1}{2\pi \tau^2} \right)^{\frac{n}{2}} |\Omega|^{-1/2} \exp \left\{ -\frac{1}{2\tau^2} (y - X\beta)' \Omega^{-1} (y - X\beta) \right\}, \quad (2.5)$$

We assume that we have some prior information on the parameter space of \((\beta, \tau^2)\) summarized by the densities

$$p(\beta | \tau^2, \Omega) = \frac{|X'\Omega^{-1}X|^{-\frac{1}{2}}}{(2\pi \tau^2)^\frac{n}{2}} \exp \left\{ -\frac{1}{2} (\beta - \beta_0)' X' \Omega^{-1} X (\beta - \beta_0) \right\},$$

$$p(\tau^2) \propto (\tau^2)^{\nu-1} \exp \left\{ -\frac{1}{2} \left( \frac{\psi}{\tau^2} + \lambda \tau^2 \right) \right\}, \quad (2.6)$$

where \((\psi, \lambda, \nu)\) are as defined in (1.2). The distribution (2.6) is the generalized inverse Gaussian distribution (see Barndorff-Nielsen (1978) and Barndorff-Nielsen et al. (1978)).

Several authors (Raiffa and Schlaifer (1961), Aitchison and Sculthorpe (1965), Guttman (1970), Aitchison and Dunsmore (1975) among others) have performed the Bayesian analysis of the normal model assuming the normal-inverted chi-squared conjugate prior. For a long time, the chi-squared prior was often used to simplify the analysis by avoiding tedious algebraic manipulations. However, with advances in computer technology and developments of Bayesian software such as WINBUGS (Spiegelhalter et al. 2004), use of conjugate priors for the sake of mathematical tractability is no longer necessary. Following Thabane and Haq (1999), we propose to use the generalized inverse Gaussian prior for the variance. The structure of the normal-generalized inverse Gaussian prior provides a great flexibility of choice of the hyper-parameter values \((\beta_0, \psi, \lambda, \nu)\) such that the assumed prior still reflects the true picture of the investigator’s prior opinions. Further, assuming a wider class of distributions such the generalized inverse Gaussian distribution for the prior allows one to perform sensitivity analysis to assess the impact of different prior views or opinions on predictive or posterior inferences.

### 2.2 The Posterior Distribution of \(\beta\)

In this section, we derive the posterior distribution of \(\beta\) assuming that the intra-class correlation coefficient \(\rho\) is known. Combining the likelihood function with the prior, we obtain the joint posterior density of \((\beta, \tau^2)\) as

$$p(\beta, \tau^2 | y, \Omega) = C_4 \left( \frac{\psi}{2} \right)^{\frac{n}{2} - \frac{1}{2}} \exp \left\{ -\frac{1}{2\tau^2} \left[ d + (\beta - \hat{\beta})' X' \Omega^{-1} X (\beta - \hat{\beta}) \right] \right\} \times \exp \left\{ -\frac{\lambda \tau^2}{2} \right\}, \quad (2.7)$$

331
where
\[
\hat{\beta} = \frac{\beta_0 + \hat{\beta}}{2},
\]
\[
\hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y,
\]
\[
d = (n - p)s^2 + (\beta_0 - \hat{\beta})' X' \Omega^{-1} X (\beta_0 - \hat{\beta}) + \psi
\]
\[
(n - p)s^2 = (y - X\hat{\beta})' \Omega^{-1}(y - X\hat{\beta}).
\]
Also,
\[
X' \Omega^{-1} X = \frac{1}{1 - \rho} X' X - \frac{\rho}{1 - \rho} \frac{(X'1)(1'X)}{1 + \rho(n - 1)}
\]
\[
(X' \Omega^{-1} X)^{-1} = (1 - \rho)(X' X)^{-1} + \frac{\rho(1 - \rho)(X' X)^{-1} X' 1' X(X' X)^{-1}}{[1 + \rho(n - 1) - \rho^2 X(X' X)^{-1} X' 1']}
\]
\[
(X' \Omega^{-1} y) = \frac{X'y}{1 - \rho} - \frac{\rho}{1 - \rho} \frac{X 1'y}{X' 1'y},
\]
and \(C_4\) is the normalizing constant.

The marginal posterior density of \(\beta\) is obtained integrating (2.7) with respect \(\tau^2\):
\[
p(\beta|\Omega, y) = \int_0^\infty C_4 (\tau^2)^{\frac{n - p - 1}{2}} \exp\left\{ -\frac{1}{2\tau^2} \left[ d + (\beta - \hat{\beta})' X' \Omega^{-1} X (\beta - \hat{\beta}) \right] \right\} \times \exp\left\{ -\frac{\lambda \tau^2}{2} \right\} d\tau^2.
\]
(2.8)

To evaluate the integral (2.8), we use the following result (see Gradshteyn and Ryzhik (1980, p. 340, formula 9)),
\[
\int_0^\infty z^{\nu - 1} \exp\left\{ -\frac{a}{z} - bq \right\} dz = 2 \left( \frac{a}{b} \right)^{\frac{\nu}{2}} K\nu(2\sqrt{ab}), \quad (Re a > 0, Re b > 0).
\]
(2.9)

Therefore,
\[
p(\beta|\Omega, y) = C_5 \left\{ d + (\beta - \hat{\beta})' X' \Omega^{-1} X (\beta - \hat{\beta}) \right\}^{\frac{n - p - \nu}{2}}
\]
\[
\times K\nu(2\sqrt{d\lambda})(\lambda)^{\frac{n - p}{2}},
\]
(2.10)

where
\[
C_5 = \frac{|X' \Omega^{-1} X|^{-\frac{\nu}{2}}}{2(2\pi)^{\frac{\nu}{2}} \times K\nu(\sqrt{d\lambda})(\frac{\lambda}{d})^{\frac{n - p}{2}}}
\]
Thus, for known \(\rho\), the posterior density (2.7) is the generalized multivariate modified Bessel distribution. That is,
\[
\beta|\Omega, y \sim GMBDp(\hat{\beta}, (X' \Omega^{-1} X)^{-1}, d, \lambda, \nu).
\]

Using the results by Thabane and haq (2003), the posterior mean and variance of \(\beta|\Omega, y\) are respectively given by
\[
E(\beta|y) = \hat{\beta} \quad \text{and}
\]
\[
Var(\beta|\Omega, y) = \sqrt{\frac{d}{\lambda}} \times \frac{K\nu(\sqrt{d\lambda})}{K\nu(\sqrt{d\lambda})} \times (X' \Omega^{-1} X)^{-1}.
\]
\section{Prediction of Future Responses}

In this section, we consider the derivation of the joint prediction distribution of \( m \) future responses. Let \( y_f = (y_{f1}, \ldots, y_{fm})' \) be a set of future responses generated by the model

\[ y_f = X_f \beta + \epsilon_f, \tag{3.11} \]

where \( X_f \) is an \( m \times p \) matrix of known elements and \( \epsilon_f \) is an \( m \times 1 \) error vector with \( \epsilon_f \sim N_m(0, \tau^2 \Omega) \). In the Bayesian context, the predictive density of \( Y_f \) given the set of observed responses \( y \) is given by

\[ p(y_f | \Omega, y) = \int \int p(y_f | \beta, \tau^2, \Omega)p(\beta, \tau^2 | \Omega, y) d\beta \ d\tau^2, \]

where

\[ p(y_f | \Omega, \beta, \tau^2) = \left( \frac{1}{2\pi \tau^2} \right)^\frac{p}{2} |\Omega|^{-1/2} \exp \left\{ \frac{1}{2\tau^2} (y_f - X_f \beta)' \Omega^{-1} (y_f - X_f \beta) \right\}, \]

and \( p(\beta, \tau^2 | y) \) is as given in (2.7).

Therefore,

\[ p(y_f | \Omega, y) = C_6 \left\{ d + (y_f - X_f \hat{\beta})' S_f^{-1}(y_f - X_f \hat{\beta}) \right\}^{\frac{n_p - \nu}{2}} \times K_{\frac{n_p - \nu}{2}} \left( \sqrt{\lambda \left[ d + (y_f - X_f \hat{\beta})' S_f^{-1}(y_f - X_f \hat{\beta}) \right]} \right), \tag{3.12} \]

where

\[ C_6^{-1} = (2\pi)^{\frac{p}{2}} |S_f|^{-\frac{p}{2}} \left( \frac{d}{\lambda} \right)^{\frac{n_p - \nu}{2}} \times K_{\frac{n_p - \nu}{2}} \left( \sqrt{d\lambda} \right), \]

and

\[ S_f = I_m + (1 - \rho) X_f (X'X)^{-1} X_f' + \frac{\rho(1 - \rho) X_f (X'X)^{-1} X'X (X'X)^{-1} X_f}{[1 + \rho(n - 1) - \rho 1' X (X'X)^{-1} X' 1]}, \]

Again, we note that the density (3.12) has the same form as distribution (1.1) and therefore for known \( \rho \),

\[ y_f | y \sim GMBD_m \left( X_f \hat{\beta}, S_f, d, \lambda, \nu \right). \]

The mean and variance are respectively given by

\[ E(y_f | y) = X_f \hat{\beta} \quad \text{and} \]

\[ \text{Var}(y_f | y) = \left( \frac{d}{\lambda} \right)^{\frac{\nu}{2}} \times \frac{K_{\frac{n_p - \nu}{2}} \left( \sqrt{d\lambda} \right)}{K_{\frac{n_p - \nu}{2}} \left( \sqrt{d\lambda} \right)} S_f. \]

We observed that both posterior distribution of \( \beta \) and predictive distribution of \( y_f \) depend on \( \rho \). Quite often \( \rho \) is also an unknown, but can be estimated from the data. There are several options that can be used: the first option would to adopt an empirical Bayes approach and estimate \( \rho \) directly from the data using the method of moments or maximum likelihood estimate, and the substitute the estimate \( \hat{\rho} \) for \( \rho \) in

333
(2.7) and (3.12) to obtain approximate posterior distribution \( p(\beta | y, \hat{\rho}) \) and predictive distribution \( p(y_f | y, \hat{\rho}) \), respectively. Alternatively one could assume an informative prior \( p(\rho) \) and then obtain the posterior and predictive distributions as

\[
p(\beta | y) = \frac{1}{\pi^2} \int p(\beta | y, \rho) p(\rho) \, d\rho \\
p(y_f | y) = \frac{1}{\pi^2} \int p(y_f | y, \rho) p(\rho) \, d\rho
\]

However, the above integrals can not be evaluated in closed form, but numerical integration using MCMC methods in WinBUGS can be employed to sample from the respective distributions.

## 4 Concluding Remarks

The application of the GMBD is considered in the Bayesian analysis of the normal linear model under the assumption of the generalized inverse Gaussian prior density for the scale parameter of the model. The analysis showed that both posterior density of parameters and predictive density of future response(s) lead to the generalized multivariate modified Bessel distribution. The estimation of the intra-class correlation has been discussed. Finally, the findings of this paper include some of the earlier results as special cases obtained under the normal-inverted chi-squared and non-informative prior densities. The application of the findings of the paper are under the current investigations.

ACKNOWLEDGMENT

The research was supported in part by the Natural Sciences and Engineering Research Council of Canada for the first author.

References


Middle Censoring for Circular Data

S. RAO JAMMALAMADAKA
University of California, Santa Barbara, USA,

and

VASUDEVAN MANGALAM
Universiti Brunei Darussalam, Brunei

Abstract

A generalized censoring scheme in the survival analysis context was introduced by the authors in Jammalamadaka and Mangalam (2003). In this article we discuss how such a censoring scheme applies to circular data and in particular when the original data is assumed to come from a parametric model such as the von Mises. Maximum likelihood estimation of the parameters as well as their large sample properties are considered under this censoring scheme.

1 Introduction

Let \( \alpha_1, \alpha_2, \ldots, \alpha_n \) be a set of independent and identically distributed (i.i.d.) measurements on two-dimensional directions. Such measurements, called angular or circular data, can be represented as points on the circumference of a circle with unit radius. They may represent wind directions, the vanishing angles at the horizon for a group of birds, or the times of arrival at a hospital emergency room where the 24 hour cycle is represented as a circle. In assigning numerical values to such directions, one has to keep in mind the arbitrary choice of the zero-direction, as well as the sense of rotation. For definiteness, all throughout this paper, we measure angles in the range \((0, 2\pi)\) and use anti-clockwise direction as positive. However, the statistical measures as well as methods should be independent of these choices.

We consider the censoring problem where one may not be able to observe all the data points. For instance, a bird’s vanishing angle at the horizon might be obscured by a passing cloud or a (fixed) hill range so that one sees some actual observations \( \alpha_i \) while the others may be simply noted as falling inside a random interval \((L_i, R_i)\). A similar thing happens when the registration counter of a hospital emergency room is closed for a temporary period and all patients who arrive during that period are registered as having come during that interval. We consider both parametric and nonparametric estimation problems in this context and evaluate the loss in efficiency because of such censoring. In the next section, we consider the Circular Normal or von Mises model for the original data with different distributions that generate the intervals of censoring, \((L_i, R_i)\), and consider the Maximum Likelihood Estimates (MLEs) of the parameters of interest. In the final section, we establish the consistency and asymptotic normality of these estimates, allowing one to find large-sample confidence intervals etc.

2 Censoring in von Mises Model

Let \( \alpha_1, \ldots, \alpha_n \) be a set of angular measurements and suppose they follow a von Mises distribution. Recall that a random angle \( A \) is said to follow a von Mises distribution with mean direction \( \mu \) and concentration parameter \( \kappa \), to be denoted by a \( \text{vM}(\mu, \kappa) \), if it has the pdf

\[
f(\alpha; \theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\alpha - \mu)}, \quad 0 \leq \alpha < 2\pi,
\]

where \( \theta = (\mu, \kappa) \in \Theta = [0, 2\pi) \times [0, \infty) \) and \( I_\nu \) is the modified Bessel function of the first kind and order \( \nu \) (see for instance, Jammalamadaka and SenGupta, 2001). Our goal is to estimate \( \theta \) when some of the observations are censored by intervals of the type \((l, r)\) where \( l \) and \( r \) forms an arc on the circumference. Denoting the observed arcs as \((l_i, r_i)\), the likelihood function takes the form
\[ L_n (\alpha; \theta) = \frac{1}{(2\pi I_0(\kappa))} \exp \left[ \kappa \sum_{i=1}^{n} \delta_i \cos (\alpha_i - \mu) \right] \prod_{i=1}^{n} \left[ \int_{l_i}^{r_i} \exp [\kappa \cos (t - \mu)] \, dt \right]^{1-\delta_i} \]

where \( \delta_i \) takes the value “0” if the observation is censored and the value “1” if it is uncensored. If the arc straddles the origin \( 0 = 2\pi \) so that \( r < 0 < l \), the integral over the arc starting from \( l \) and ending at \( r \) is interpreted as the sum of integrals from \( l \) to \( 2\pi \) and 0 to \( r \). When convenient, we indicate this arc by \( A_{l,r} \) and its complement by \( \bar{A}_{l,r} \). Thus the log-likelihood function is given by

\[ l_n (\alpha; \theta) = -n \log 2\pi - n \log I_0(\kappa) + \kappa \sum_{i=1}^{n} \delta_i \cos (\alpha_i - \mu) \]

\[ + \sum_{i=1}^{n} (1 - \delta_i) \log \left[ \int_{A_{l_i,r_i}} \exp [\kappa \cos (t - \mu)] \, dt \right]. \tag{2.1} \]

The MLE of \( \theta \) is \( \hat{\theta} = \hat{\theta}_n = (\hat{\mu}_n, \hat{\kappa}_n) \) which maximizes this likelihood function. Explicit form is of course, not available and computational aspects of the MLE are discussed below.

For computing the MLEs and the information matrix, we need the following notations

\[ A_0 (\kappa) = \frac{I_1 (k)}{I_0 (k)}, \quad B_{0i} (\mu, \kappa) = \int_{l_i}^{r_i} \exp [\kappa \cos (t - \mu)] \, dt, \quad B_{1i} (\mu, \kappa) = \int_{l_i}^{r_i} \sin (t - \mu) \exp [\kappa \cos (t - \mu)] \, dt, \]

\[ B_{2i} (\mu, \kappa) = \int_{l_i}^{r_i} \cos (t - \mu) \exp [\kappa \cos (t - \mu)] \, dt, \quad B_{3i} (\mu, \kappa) = \int_{l_i}^{r_i} \sin^2 (t - \mu) \exp [\kappa \cos (t - \mu)] \, dt, \]

\[ B_{4i} (\mu, \kappa) = \int_{l_i}^{r_i} \cos^2 (t - \mu) \exp [\kappa \cos (t - \mu)] \, dt, \]

\[ B_{5i} (\mu, \kappa) = \int_{l_i}^{r_i} \sin (t - \mu) \cos (t - \mu) \exp [\kappa \cos (t - \mu)] \, dt, \]

and note

\[ \frac{\partial B_{0i}}{\partial \mu} = \kappa B_{1i}, \quad \frac{\partial B_{0i}}{\partial \kappa} = B_{2i}, \quad \frac{\partial B_{4i}}{\partial \mu} = \kappa B_{3i} - B_{2i}, \quad \frac{\partial B_{4i}}{\partial \kappa} = B_{5i}, \quad \text{and} \quad \frac{\partial B_{2i}}{\partial \kappa} = B_{4i}. \]

The derivatives of the log likelihood function are given by

\[ \frac{\partial l_n}{\partial \mu} = \kappa \sum_{i=1}^{n} \delta_i \sin (\alpha_i - \mu) + \kappa \sum_{i=1}^{n} (1 - \delta_i) \frac{B_{1i} (\mu, \kappa)}{B_{0i} (\mu, \kappa)} \]

\[ (2.2) \]

and

\[ \frac{\partial l_n}{\partial \kappa} = -n A_0 (\kappa) + \sum_{i=1}^{n} \delta_i \cos (\alpha_i - \mu) + \sum_{i=1}^{n} (1 - \delta_i) \frac{B_{2i} (\mu, \kappa)}{B_{0i} (\mu, \kappa)} \]

\[ (2.3) \]

The second derivatives are given by

\[ \frac{\partial^2 l_n}{\partial \mu^2} = -n \sum_{i=1}^{n} \delta_i \cos (\alpha_i - \mu) + \kappa \sum_{i=1}^{n} (1 - \delta_i) \left[ \frac{\kappa B_{3i} (\mu, \kappa) - B_{2i} (\mu, \kappa)}{B_{0i} (\mu, \kappa)} \right] - \kappa \left( \frac{B_{1i} (\mu, \kappa)}{B_{0i} (\mu, \kappa)} \right)^2 \]

\[ \frac{\partial^2 l_n}{\partial \kappa^2} = -n A'_0 (\kappa) + \sum_{i=1}^{n} (1 - \delta_i) \left[ \frac{B_{4i} (\mu, \kappa)}{B_{0i} (\mu, \kappa)} - \left( \frac{B_{2i} (\mu, \kappa)}{B_{0i} (\mu, \kappa)} \right)^2 \right] \]

\[ \frac{\partial^2 l_n}{\partial \kappa \partial \mu} = \sum_{i=1}^{n} \delta_i \sin (\alpha_i - \mu) + \sum_{i=1}^{n} (1 - \delta_i) \left[ \frac{B_{1i} (\mu, \kappa) + B_{5i} (\mu, \kappa)}{B_{0i} (\mu, \kappa)} - \frac{B_{1i} (\mu, \kappa) B_{2i} (\mu, \kappa)}{B_{0i} (\mu, \kappa)} \right] \]

337
Expressions 2.2 and 2.3 are equated to zero and solved numerically. By substituting these solutions into the information matrix, one obtains the "observed information" matrix viz.

\[ \hat{I} = \begin{bmatrix} \frac{\partial^2 t_n}{\partial \theta_1^2} & \frac{\partial^2 t_n}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 t_n}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 t_n}{\partial \theta_2^2} \end{bmatrix} \bigg|_{\theta = \hat{\theta}}. \]

Then, provided censoring is not too strong, \( \sqrt{n} (\hat{\theta}_n - \theta) \) will be asymptotically normal with mean zero and covariance \( \hat{I}^{-1} \) as shown in the next section.

3 Large Sample Properties of the MLEs under censoring

We assume that the random censoring mechanism is independent of the variable of interest and does not involve \( \theta \). Let \( \theta_0 \) denote the true value of the parameter and let

\[ p(\theta, l, r) = P_\theta [A \in (l, r)] = \int_l^r f(t; \theta) dt. \]

Let

\[ g_1(\theta, l, r) = -\log (2\pi I_0(\kappa)) + \kappa \int_{A_{l,r}} \cos(t - \mu) f(t; \theta_0) dt \]

\[ + p(\theta_0, l, r) \log \left[ \int_{A_{l,r}} \exp[\kappa \cos(t - \mu)] dt \right] \]

and define a function \( g \) on the parameter space as \( g(\theta) = \int g_1(\theta, l, r) dF_{LR} \)

**Theorem 3.1** Assume that the identifiability condition

\[ p(\theta_0) = P_{\theta_0} \{A \in (L, R)\} < 1 \]

is satisfied. Let \( \Sigma_1(\theta) \) be the dispersion of \( X_1(\theta) = \left( \frac{\partial X_1(\theta)}{\partial \theta_1}, \frac{\partial X_1(\theta)}{\partial \theta_2} \right) \) where

\[ X_1(\theta) = -\log [2\pi I_0(\kappa)] + \kappa \delta \cos(A_i - \mu) + (1 - \delta) i \log \left[ \int_{L_i} \exp[\kappa \cos(t - \mu)] dt \right] \]

Let \( \Sigma(\theta) = \left[ g''(\theta) \right]^{-1} \Sigma_1(\theta) \left[ g''(\theta) \right]^{-1} \). Then (i) \( \hat{\theta}_n \rightarrow \theta_0 \) a.s.(\( P_{\theta_0} \)) and (ii) \( \sqrt{n} (\hat{\theta}_n - \theta_0) \Rightarrow N_2(0, \Sigma(\theta_0)) \).

**Proof:** For brevity, we skip the proof of consistency, which is shown elsewhere. Let \( I_n(\theta) = \left( \frac{\partial^2 t_n}{\partial \theta_1^2}, \frac{\partial^2 t_n}{\partial \theta_1 \partial \theta_2}, \frac{\partial^2 t_n}{\partial \theta_1 \partial \theta_2}, \frac{\partial^2 t_n}{\partial \theta_2^2} \right) \) and

\[ I''(\theta) = \begin{bmatrix} \frac{\partial^2 t_n}{\partial \theta_1^2} & \frac{\partial^2 t_n}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 t_n}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 t_n}{\partial \theta_2^2} \end{bmatrix}. \]

By the standard multivariate Taylor expansion (see (8.19), Vol.2 of Apostol (1967)), we get

\[ I'(\hat{\theta}_n) = I_n(\theta_0) + (\hat{\theta}_n - \theta_0) \left( I''(\theta_0) \right) + \| \hat{\theta}_n - \theta_0 \| E \left( \hat{\theta}_n, \theta_0 \right) \]

where the function \( E \) is such that \( \lim_{x \rightarrow y} E(x, y) = 0 \). As \( \hat{\theta}_n \) is the maximizer of the likelihood function, \( I_n(\hat{\theta}_n) = 0 \). From this it follows that

\[ (\hat{\theta}_n - \theta_0) I''(\theta_0) = -I'(\theta_0) - \| \hat{\theta}_n - \theta_0 \| E \left( \hat{\theta}_n, \theta_0 \right) \]

338
\[
\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) = \frac{-f'(\theta_0)}{\sqrt{n}} \left[ f''(\theta_0) \right]^{-1} - \frac{1}{\sqrt{n}} \left\| \hat{\theta}_n - \theta_0 \right\| E \left( \hat{\theta}_n, \theta_0 \right) \left[ f''(\theta_0) \right]^{-1} \quad (3.5)
\]

By the strong law of large numbers, \( \frac{f'(\theta_0)}{n} \) converges to a constant matrix given by \( E_{\theta_0} \left[ X''_1(\theta_0) \right] = g''(\theta_0) \). By the consistency result, \( \frac{1}{\sqrt{n}} \left\| \hat{\theta}_n - \theta_0 \right\| E \left( \hat{\theta}_n, \theta_0 \right) \) goes to zero a.s., and hence so does the second term on the RHS of (3.5). As a consequence of the standard multivariate central limit theorem for i.i.d. random variables, \( \frac{f'(\theta_0)}{\sqrt{n}} \) goes to a multivariate normal distribution with mean zero and dispersion \( \Sigma_1 \) and the desired result follows.

References


Abstract. This paper presents a robust alternative to least squares for estimating parameters of a general nonlinear model. We consider an estimator which minimizes the Wilcoxon dispersion function and establish its asymptotic properties under mild regularity conditions similar to those used in least squares and least absolute deviations estimation. As in linear models, these rank-based procedures provide estimators that are highly efficient.

Key words and phrases: nonlinear regression, dispersion function, order statistics.

1 Introduction

In this paper, we consider the following general nonlinear model,

$$y_i = f_i(\theta_0) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $f_i$ are known real valued functions defined on a compact space $\Theta$ and $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random errors with probability density function $h(t)$. There has been considerable work on least squares (LS) estimation of $\theta_0$. The asymptotic properties and conditions needed for the numerical stability of the LS estimation procedure were investigated in Jennrich (1969). Malinvaud (1970) and Wu (1981) have further investigated large sample properties of the LS estimator. LS estimation in nonlinear models is a direct extension of its estimation in linear models. The same norm (Euclidean) is minimized to obtain the LS estimate of $\theta_0$; i.e., the geometry stays the same.

Linear regression based on ranks was first proposed by Jurečková (1971) and Jaeckel (1972). McKean and Schrader (1980) showed that these R estimates are based on minimizing a norm based on a score function. Hence, the geometry of these estimates is similar to that of LS in the sense that one norm has been substituted for another. Unlike the Euclidean norm, the norm associated with R estimates leads to highly efficient, robust estimates.

Based on geometry, R estimates can naturally be extended to nonlinear models. As in the case of LS estimates, the same norm can be used to obtain R estimates for nonlinear models as for linear models. Thus the linear model interpretation of the estimates carries over to nonlinear models. These R estimates and the associated norm depend on the score function chosen. The two most popular score functions are the sign score function ($L_1$) and the Wilcoxon score function (linear score function). In simple location models, sign scores result in medians as the location estimates, while the Wilcoxon score function results in Hodges-Lehmann estimates; see the monograph by Hettmansperger and McKean (1998) for discussion. For normal errors, the sign and Wilcoxon estimates have ARE’s (relative to LS) of 63% and 95%, respectively. Further, these efficiencies carry over to the linear model. The high efficiency of the Wilcoxon procedures relative to LS makes them attractive alternatives to LS procedures.

For the nonlinear model (1), Oberhofer (1982) obtained the weak consistency for R estimates based on the sign scores, i.e., the $L_1$ estimate. In Section 2 of this paper, we obtain the consistency and asymptotic normality for the nonlinear R estimates of Model (1) based on the Wilcoxon score function. We show that
the ARE for these Wilcoxon estimates relative to LS is the same as in the linear model. Hence, as in the linear model, the Wilcoxon nonlinear estimates offer the user a highly efficient alternative to LS nonlinear estimates. In Section 3, we present examples and a small simulation study which demonstrate the finite sample robustness of the unweighted Wilcoxon estimates. Proofs are provided in Section 5.

2 WILCOXON ESTIMATION

2.1 Definition and Existence

Consider the nonlinear model (1) and let \( \mathbf{y} = (y_1, \ldots, y_n)^T \) and \( \mathbf{f}(\mathbf{\theta}) = (f_1(\mathbf{\theta}), \ldots, f_n(\mathbf{\theta}))^T \). Given a norm \( \| \cdot \| \) on n-space, a natural estimator of \( \mathbf{\theta} \) is a value \( \hat{\mathbf{\theta}} \) which minimizes the distance between the response vector \( \mathbf{y} \) and \( \mathbf{f}(\mathbf{\theta}); \) i.e., \( \hat{\mathbf{\theta}} = \text{Argmin}_{\mathbf{\theta} \in \Theta} \| \mathbf{y} - \mathbf{f}(\mathbf{\theta}) \| \). If the norm is the Euclidean norm then \( \hat{\mathbf{\theta}} \) is the LS estimate.

In this paper, we consider the Wilcoxon norm given by,

\[
\| \mathbf{u} \|_W \equiv (2n(n+1))^{-\frac{1}{2}} \sum_{i<j} |u_i - u_j| ,
\]  

(2)

where \( \mathbf{u} \) is a point in \( \mathbb{R}^n \). The quantity given in (2) may be represented as a linear function of the order statistics of \( \mathbf{u} \) (see, for example, Hettmansperger and McKean (1998) page 73) as,

\[
\| \mathbf{u} \|_W \equiv n^{-\frac{1}{2}} \sum_{i=1}^{n} a_{W,n}(i) u_{(i)} ,
\]  

(3)

where \( a_{W,n}(i) = \varphi_W(i/(n+1)) \) where \( \varphi_W(u) \) is the Wilcoxon (linear) score function given by \( \varphi_W(u) = u-1/2 \). This representation is the one considered by Jaeckel (1972) in estimating linear regression parameters.

Technically the function (2) is a pseudo-norm; that is, it satisfies all the properties of a norm except, the property \( \| \mathbf{u} \|_W = 0 \Leftrightarrow \mathbf{u} = \mathbf{0} \) is replaced by \( \| \mathbf{u} \|_W = 0 \Leftrightarrow u_1 = u_2 = \cdots = u_n \).

We define the Wilcoxon estimator of \( \mathbf{\theta}_0 \), denoted hereafter by \( \hat{\mathbf{\theta}}_{W,n} \), as

\[
\hat{\mathbf{\theta}}_{W,n} = \text{Argmin}_{\mathbf{\theta} \in \Theta} \| \mathbf{y} - \mathbf{f}(\mathbf{\theta}) \|_W .
\]  

(4)

It will be convenient to use the notation, \( D_n(\mathbf{y}, \mathbf{\theta}) \equiv \| \mathbf{y} - \mathbf{f}(\mathbf{\theta}) \|_W \). When there is no confusion we will drop the \( \mathbf{y} \) from the notation. Furthermore, in the remainder of this paper we let \( A^o \) represent the interior of the space \( A \) and \( \partial A \) represent the boundary of \( A \). The set subtraction of \( B \) from \( A \) is denoted by \( A \setminus B \).

The following assumption suffices for the existence of the Wilcoxon nonlinear estimate:

**A1:** For all \( i, f_i(\mathbf{\theta}) \) is defined and continuous for all \( \mathbf{\theta} \in \Theta \).

This implies that the dispersion function is a continuous function of \( \mathbf{\theta} \) and, hence, since \( \Theta \) is compact, that the Wilcoxon estimate \( \hat{\mathbf{\theta}} \) exists. We state this as a theorem.

**Theorem 2.1.** Under Model (1) and Assumption A1, \( \hat{\mathbf{\theta}}_{W,n} \) exists.

The sign score function given by \( \varphi_S(u) = \text{sgn}(u-(1/2)) \) is another popular score function. The norm associated with this score function is (3) but with the sign scores \( a_{S,n}(i) = \varphi_S(i/(n+1)) \). Let \( \hat{\mathbf{\theta}}_S \) denote the estimate based on this norm. Its existence follows in the same way as the existence of the Wilcoxon. In order to see the relationship between \( \hat{\mathbf{\theta}}_S \) and the \( L_1 \) estimate of \( \mathbf{\theta} \), denote the \( L_1 \) norm by, \( \| \mathbf{u} \|_{L_1} \equiv n^{-\frac{1}{2}} \sum_{i=1}^{n} |u_i| \). Let \( \hat{\mathbf{\theta}}_{L_1} \) denote the estimate based on this norm. The following lemma is the nonlinear analogue of Theorem 3.8.1 of Hettmansperger and McKean (1998). Its proof is quite similar to the linear model result.

**Lemma 2.1.** If \( \Theta \) is a compact subset of \( \mathbb{R}^p \), then

\[
f(\hat{\mathbf{\theta}}_{L_1}) = f(\hat{\mathbf{\theta}}_S) + \text{med}\{y_i - f_i(\hat{\mathbf{\theta}}_S)\} \mathbf{1} ,
\]  

where, \( \mathbf{1} \) is a vector of \( n \) ones.

341
Oberhofer (1982) obtained asymptotic theory for the $L_1$ estimate $\hat{\theta}_{L_1}$. As in the linear case, the ARE of the $L_1$ estimate, relative to LS at normal errors, is low, 63%. In the next two sections, we derive the asymptotic theory (consistency and asymptotic normality) of the Wilcoxon nonlinear estimator $\hat{\theta}_{W,n}$. As in the linear model situation, we show that the Wilcoxon nonlinear estimate has an ARE of 95%, relative to LS at normal errors. Thus the Wilcoxon estimate provides a highly efficient, nonlinear estimate of $\theta_0$.

2.2 Consistency

Before we establish the consistency of $\hat{\theta}_{W,n}$ we introduce some helpful notation. Let $\theta$ and $\theta^*$ be points in $\Theta$. We denote the residuals at $\theta$ by $e_i(\theta) = y_i - f_i(\theta)$ for $1 \leq i \leq n$. For $1 \leq i, j \leq n$, we define $W_{ij}(\theta, \theta^*) \equiv |e_i(\theta) - e_j(\theta)| - |e_i(\theta^*) - e_j(\theta^*)|$. Further, let

$$h_i^*(\theta, \theta^*) \equiv f_i(\theta^*) - f_i(\theta),$$

$$h_{ij}(\theta, \theta^*) \equiv h_i^*(\theta, \theta^*) - h_j^*(\theta, \theta^*),$$

and

$$\Delta_n(\theta, \theta^*) \equiv n^{-1} \sum_{i=1}^{n} [h_i^*(\theta, \theta^*)]^2.$$

Let $G$ denote the distribution function of $\varepsilon_i - \varepsilon_j$.

We need the following assumptions:

A2: $\theta_0 \in \Theta^\circ$.

A3: $\lim_{n \to \infty} n^{-1} \Delta_n(\theta, \theta_0) = 0$ for all $\theta \in \Theta$.

A4: $G(0) = 1/2$.

A5: There exist $\eta > 0$ and $n_0$ such that for all $n \geq n_0$ and all $\theta \in \Theta^*$, where $\Theta^*$ is a closed subset of $\Theta \setminus \{\theta_0\}$,

$$\inf_{\theta \in \Theta^*} n^{-2} \sum_{i<j} |h_{ij}(\theta_0)| \times \min \{G(|h_{ij}(\theta, \theta_0)|/2) - 1/2, 1/2 - G(-|h_{ij}(\theta, \theta_0)|/2)\} \geq \eta.$$

Assumption A3 is the same as Oberhofer’s (1982) assumption A4. Jennrich (1969) assumes that $\Delta_n(\theta, \theta^*)$ converges uniformly to a continuous function $\Delta(\theta, \theta^*)$ for all $\theta, \theta^*$ in $\Theta$ and $\Delta(\theta, \theta_0) = 0$ if and only if $\theta = \theta_0$. This of course implies A3. Wu (1981) assumes that $n \Delta_n(\theta, \theta^*)$ diverges as $n$ approaches infinity. This is weaker than Jennrich’s assumption and A3 since it does not restrict $\Delta_n(\theta, \theta_0)$ to converge at the rate of $n$. Assuming the existence of the dot product $\sum f_i \varepsilon_i$, in light of Lemma 4.2, one can easily observe that Wu’s condition is sufficient for the consistency of the LS estimator. This, however, is not true in our case as $\Delta_n$ only comes into play as part of an upper bound on $D_n(\theta)$. Discussion on Assumption A5 follows Theorem 2.2.

Now let $\Theta^*$ be a closed subset of $\Theta$ not containing $\theta_0$. The weak consistency of $\hat{\theta}_{W,n}$ will then follow if for all such $\Theta^*$ and every $\theta \in \Theta^*$,

$$\lim_{n \to \infty} P(\inf_{\theta \in \Theta^*} |D_n(\theta) - D_n(\theta_0)| > 0) = 1. \tag{5}$$

**Lemma 2.2.** Under A3,

$$\{D_n(\theta) - D_n(\theta_0)\} - E\{D_n(\theta) - D_n(\theta_0)\} \to 0,$$

in probability.

**Lemma 2.3.** Under A4,

$$E\{D_n(\theta) - D_n(\theta_0)\} \geq [2n(n + 1)]^{-1} \sum_{i<j} |h_{ij}(\theta, \theta_0)| \times \min\{G(|h_{ij}(\theta, \theta_0)|/2) - 1/2, 1/2 - G(-|h_{ij}(\theta, \theta_0)|/2)\}.$$

342
We now state and prove the main theorem of this section.

**Theorem 2.2.** Under A1-A5, $\hat{\theta}_{W,n}$ is weakly consistent for $\theta_0$.

Assumption A5 is similar to that assumed in Oberhofer (1982). As the following lemma shows assumption A5 is an identifiability assumption. We will give the following definition of identifiability of measurable functions. A similar definition can be found in Seber and Wild (1989).

**Definition 2.1.** Let $f$ be $\lambda$-measurable, where $\lambda$ is a $\sigma$-finite measure. The parameters of the nonlinear regression problem $y = f(\theta) + \varepsilon$ are said to be unidentifiable if there exist two distinct points $\theta_1$ and $\theta_2$ such that $\lambda\{f(\theta_1) \neq f(\theta_2)\} = 0$.

**Lemma 2.4.** If for each $\theta$, $f(\theta)$ is a real measurable function on a measure space $(\omega, \mathcal{F}, \lambda)$, where $\lambda$ is a $\sigma$-finite measure, then a necessary condition for A5 to hold is,

$$\lambda\{f(\theta) \neq f(\theta_0)\} > 0,$$

for any $\theta \neq \theta_0$.

If $f$ is a known function that depends on a set of random predictors, in addition to the parameters, then we obtain a natural extension of this. The following corollary gives the result.

**Corollary 2.1.** Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space and $f_i(\theta) = f(z_i, \theta)$. Assume $z_i$ are $n$ independent identically distributed $m$ dimensional random vectors with range $Z \subset \mathbb{R}^m$. Then a necessary condition for A5 is that for all $z_i \in Z$,

$$P\{\omega \in \Omega : (f(z_i(\omega), \theta) = f(z_i(\omega), \theta_0))\} < 1,$$

for $\theta \neq \theta_0$.

The following lemma gives sufficient conditions for A5 to hold.

**Lemma 2.5.** If

A5.1: $\varepsilon_i - \varepsilon_j$ have density $g$ continuous at 0 with $g(0) > 0$, and

A5.2: There exist $\eta > 0$ and $n_0$ such that for all $n \geq n_0$ and all $\theta \in \Theta^*$

$$\inf_{\theta \in \Theta^*} n^{-2}\{h_{ij}(\theta, \theta_0)\}^2 \geq \eta,$$

then A5 is true.

For most practical purposes A5.1 and A5.2 are easier to verify than A5. As discussed in the remark immediately following the definition of assumption A5, A5.2 is a double-indexed version of Jennrich’s (1969) assumption on $\Delta_n(\theta, \theta_0)$.

### 2.3 Asymptotic Normality

In this section we will investigate the asymptotic distribution of $\hat{\theta}_{W,n}$. The theory for the asymptotic normality uses tangent planes. A local linearity of the expectation surface, $\mathcal{S}(\theta) = \{f(\theta) : \theta \in \Theta\}$, is furnished by the consistency of $\hat{\theta}_{W,n}$ in the locality of $\theta_0$. Hence the asymptotic theory of linear models plays a very important role in showing the asymptotic normality of $\hat{\theta}_{W,n}$. Details concerning the asymptotic properties of rank estimators of linear model parameters can be found in Hettmansperger and McKean (1998).

In addition to A1 - A5, we will assume that the following conditions are satisfied.

N1: For $i = 1, \ldots, n$, $f_i(\theta)$ is continuously differentiable at $\theta_0$ with respect to $\theta$.

N2: The sequence of matrices

$$n^{-1}\sum_{i=1}^n\{\nabla f_i(\theta_0)\}\{\nabla f_i(\theta_0)\}^T$$

converges to a positive definite matrix $\Sigma(\theta_0)$ where $\nabla f_i(\theta)$ is the $p \times 1$ derivative vector of $f_i(\theta)$ with respect to $\theta$. 

343
N3: The error density $g$ has a finite Fisher information.

For $i = 1, \ldots, n$, let

$$e_i^*(\theta) \equiv y_i - f_i(\theta_0) - \{\nabla f_i(\theta_0)\}^T(\theta - \theta_0).$$

Note that $e_i^*(\theta)$ are the error terms of the linear regression,

$$y_i^* = x_i^T \theta_0 + \varepsilon_i,$$

where $y_i^* = y_i - f_i(\theta_0) + \{\nabla f_i(\theta_0)\}^T \theta_0$ and $x_i^* = \nabla f_i(\theta_0)$. Note that the probability density function of the errors of Model 6 is $h(t)$, i.e., the density function of $\varepsilon_i$. Define the corresponding Wilcoxon dispersion function as,

$$D_n^*(\theta) \equiv [2n(n+1)]^{-1} \sum_{i<j} |e_i^*(\theta) - e_j^*(\theta)|.$$

Furthermore, let,

$$\tilde{\theta}_n = \arg\min_{\theta \in \Theta} D_n^*(\theta).$$

Under assumption N2, we apply Lemma 4.4 to obtain Noether’s condition,

$$\max_{1 \leq i \leq n} \{\nabla f_i(\theta_0)\}^T \left[ \sum_{i=1}^n \{\nabla f_i(\theta_0)\} \{\nabla f_i(\theta_0)\}^T \right]^{-1} \{\nabla f_i(\theta_0)\} \to 0,$$

as $n \to \infty$. The following theorem uses this condition to establish the asymptotic normality of $\tilde{\theta}_n$. A rigorous derivation of the result may be found in Hettmansperger and McKean (1998).

Theorem 2.3. Under model (6) and assumptions N2, N3 we have,

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \overset{D}{\to} N_p(0, \tau^2 \Sigma(\theta_0)),$$

where

$$\tau^{-1} = \sqrt{\frac{12}{12}} \int h^2(t) \, dt,$$

and $\Sigma(\theta_0)$ is given in assumption N2.

Let $\nabla f_0$ be the $n \times p$ matrix with the $i$th row given by $\{\nabla f_i(\theta_0)\}^T$. The projection operator onto the tangent plane at $\theta_0$ of the expectation surface, $S$, is given by

$$P_n = \nabla f_0 (\nabla f_0^T \nabla f_0)^{-1} \nabla f_0^T.$$

Assume

H1: $\lim_{n \to \infty} \max_{1 \leq i \leq n} p_{ii} = 0$, where $p_{ii}$ is the $i$th diagonal entry of $P_n$.

The following corollary of Theorem 2.3 shows that H1 may be used to prove the asymptotic normality of $\tilde{\theta}_n$ whenever it is convenient. The proof follows by Lemma 4.5 and Theorem 2.3.

Corollary 2.2. Under model (6) and assumptions H1, N3 we have,

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \overset{D}{\to} N_p(0, \tau^2 \Sigma(\theta_0)),$$

where

$$\tau^{-1} = \sqrt{\frac{12}{12}} \int h^2(t) \, dt,$$

and $\Sigma(\theta_0)$ is given in assumption N2.
The approach we follow to prove the asymptotic normality is via Slutsky’s Theorem (see Serfling (1980)). Recall that Lemma 4.3 gives a probabilistic bound on the asymptotic distance between the minimizers of $A_n$, a convex process with a possibly flat bottom, and $B_n$, a process whose minimizer is asymptotically unique in a neighborhood which shrinks at the rate of $1/\sqrt{n}$. In light of Lemma 2.3 and the consistency of $\hat{\theta}_{W,n}$, we have a neighborhood where the minimum of $D_n$ is unique as $n \to \infty$. Moreover, the process $D_n^*$ is convex as shown in Theorem 1 of Jaeckel (1972). Thus, $D_n$ and $D_n^*$ may be treated as $B_n$ and $A_n$, respectively, in Lemma 4.3.

The following lemma in addition to (10) gives the main result of this section which is given in Theorem 2.4 below.

**Lemma 2.6.** Under A1 - A5, N1 - N3,

$$\sqrt{n}(\hat{\theta}_{W,n} - \hat{\theta}_n) \xrightarrow{D} 0.$$  

The proof of the following theorem follows immediately from Lemma 2.6 via an application of Slutsky’s Theorem.

**Theorem 2.4.** Under A1 - A5 and N1 - N3,

$$\sqrt{n}(\hat{\theta}_{W,n} - \theta_0) \xrightarrow{D} N_p(0, \tau^2 \Sigma(\theta_0)),$$  

where $\tau$ is as given in (12).

In the next two subsections, we briefly discuss two results based on this theorem. The first concerns asymptotic relative efficiency and the second concerns applications of the asymptotic representation of the estimate.

### 2.4 Asymptotic Relative Efficiency

Let $\hat{\theta}_{LS,n}$ denote the LS estimator of $\theta_0$. Under suitable regularity conditions, the asymptotic distribution of the LS estimator is given by

$$\sqrt{n}(\hat{\theta}_{LS,n} - \theta_0) \xrightarrow{D} N_p(0, \sigma^2 \Sigma(\theta_0)),$$  

where $\sigma^2$ is the variance of the random error $\varepsilon_i$. It follows immediately, from expressions (13) and (14), that, for any component of $\theta_0$, the asymptotic relative efficiency (ARE) between the Wilcoxon estimator and the LS estimator of the component is given by the ratio $\tau^2/\sigma^2$. This, of course, is the ARE between the Wilcoxon and LS estimators in linear models. If the error distribution is normal, then this ratio is the well known number 0.955. Hence, there is only a loss of 5% efficiency, if one uses the Wilcoxon estimator instead of the LS estimator when the errors are normally distributed. In contrast, the $L_1$ estimator has the asymptotic relative efficiency of 63% relative to the LS estimator. The ARE between the Wilcoxon and $L_1$ estimators at normal errors is 150%. Hence, as in the linear model case, the Wilcoxon estimator is a highly efficient estimator for nonlinear models. For heavier tailed error distributions the Wilcoxon estimator is generally much more efficient than the LS estimator. We discuss such results for a family of contaminated normal error distributions in Section 3.

### 2.5 Asymptotic Representation for $\hat{\theta}_{W,n}$

As a simple corollary, a useful asymptotic representation of the Wilcoxon estimate is obtained. By (10), it follows, as in Hettmansperger and McKean (1998), that,

$$\sqrt{n}(\bar{\theta}_n - \theta_0) = \tau(n^{-1}X^TX*)^{-1/2}X^T\{G(y^* - X^*\theta_0) - 1/2\} + o_p(1),$$

where $X^*$ is the $n \times p$ matrix with the $i$th row given by $(\nabla f_i(\theta_0))^T$ and $y^*$ is an $n \times 1$ vector with the $i$th component $y_i - f_i(\theta_0) + (\nabla f_i(\theta_0))^T \theta_0$. Now applying Lemma 2.6 we have the same asymptotic representation for the Wilcoxon estimate, i.e.,

$$\sqrt{n}(\hat{\theta}_{W,n} - \theta_0) = \tau(n^{-1}X^TX*)^{-1/2}X^T\{G(y^* - X^*\theta_0) - 1/2\} + o_p(1).$$
Based on (16), we can obtain the influence function of the Wilcoxon estimate. Assume \( f_i \) depends on a set of predictors \( z_i \in \mathcal{Z} \subset \mathbb{R}^m \) as \( f_i(\theta) = f(z_i, \theta) \). Assume also that \( f \) is a continuous function of \( \theta \) for each \( z \in \mathcal{Z} \) and is a measurable function of \( z \) for each \( \theta \in \Theta \) with respect to a \( \sigma \)-finite measure. Under these assumptions, the representation above gives us the local influence function of the Wilcoxon estimate at the point \((z_0, y_0)\),

\[
IF(z_0, y_0; \hat{\theta}_{W,n}) = \tau \{ \Sigma(\theta_0) \}^{-1} \{ G(y_0) - 1/2 \} \nabla f(z_0, \theta_0).
\]

Note that the influence function is unbounded if the tangent plane \( S \) at \( \theta_0 \) is unbounded. This phenomenon corresponds to the existence of high leverage points in linear regression.

### 2.6 Implementation

To implement the asymptotic inference based on the Wilcoxon estimate we need a consistent estimator of the variance-covariance matrix. Define the statistic \( \Sigma(\hat{\theta}_{W,n}) \) to be \( \Sigma(\theta_0) \) of expression (N2) with \( \theta_0 \) replaced by \( \hat{\theta}_{W,n} \). By Assumption (N2) and the consistency of \( \hat{\theta}_{W,n} \) to \( \theta_0 \), \( \Sigma(\hat{\theta}_{W,n}) \) converges in probability to \( \Sigma(\theta_0) \). Next, it follows from the asymptotic representation (16) that the estimator of \( \tau \) proposed by Koul et al. (1987) for linear models is also a consistent estimator of \( \tau \) for our nonlinear model. We denote this estimator by \( \hat{\tau} \). Thus \( \hat{\tau}^2 \Sigma(\hat{\theta}_{W,n}) \) is a consistent estimator of the asymptotic variance-covariance matrix of \( \hat{\theta}_{W,n} \).

#### 2.6.1 Estimation Algorithm

Similar to the LS estimates for nonlinear models, a Gauss-Newton type of algorithm can be used to obtain the Wilcoxon fit. Recall that this is an iterated algorithm which uses the Taylor Series expansion of the function \( f(\theta) \) evaluated at the current estimate to obtain the estimate at the next iteration. Thus each iteration consists of fitting a linear model. Abebe and McKean (2007) show that this algorithm for obtaining the Wilcoxon fit converges in a probability sense. Using this algorithm, all that is required to compute the Wilcoxon nonlinear estimate is a computational procedure for Wilcoxon linear model estimates. There are several ways of computing this estimate for a linear model, including: the package RGLM of Kapenga et al. (1995) and the R routines developed by Terpstra and McKean (2005). These are discussed further in Abebe and McKean (2007).

### 3 NUMERICAL EXAMPLES AND A SIMULATION STUDY

#### 3.1 Numerical Examples

Although our focus was in developing the asymptotic theory of rank regression for nonlinear models, we consider a few examples that demonstrate the robustness and efficiency properties of the rank estimators in comparison to the least squares (LS) estimator in practical situations. Since most data contain contamination, due to either the faulty nature of the mechanism which produces the data or human error in handling the data, the use of procedures such as the ones developed in this study becomes one of the ways of making sensible inference.

Through our examples we will illustrate that the Wilcoxon estimator is a robust alternative to LS. All our estimates are computed using the package RGLM of Kapenga et al. (1995).

**Example 3.1 (Chwirut’s data).** These data are taken from the ultrasonic block reference study by Chwirut (1979). The response variable is ultrasonic response and the predictor variable is metal distance. The study involved 214 observations. The model under consideration is,

\[
f(x_i; \theta_1, \theta_2, \theta_3) \equiv \frac{\exp[-\theta_1 x_i]}{\theta_2 + \theta_3 x}, \quad i = 1, \ldots, 214.
\]

Using the Wilcoxon and LS fitting procedures, we fit the (original) data and then a data set with one observation replaced by an outlier. Figure 1 displays the results of the fits.

For the original data, as shown in the figure and by the estimates given in Table 3.1, the LS and Wilcoxon fits are quite similar. As shown in the residual plots of Figure 1, there are several moderate outliers in the
Figure 1: Analysis of Chwirut's data
Table 1: Wilcoxon and LS estimates based on the original data with standard errors (SE) and the Wilcoxon estimates based on the data with substituted gross outlier.

<table>
<thead>
<tr>
<th></th>
<th>Original Data Set</th>
<th>Outlier Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wil. Est.</td>
<td>SE</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.1902</td>
<td>0.0161</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>0.0061</td>
<td>0.0002</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>0.0197</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

original data set. These outliers have an impact on the LS estimate of scale, the square-root of $MSE$, which has the value $\hat{\sigma}_{LS} = 3.36$. In contrast, the Wilcoxon estimate of $\tau$ is $\hat{\tau} = 2.45$ which explains the Wilcoxon’s smaller standard errors than those of LS in the table of estimates.

For robustness considerations, we introduced a gross outlier in the response space, (observation 17 was changed from 8.025 to 5000). The Wilcoxon and LS fits were obtained. As shown in Figure 1, the LS estimate essentially did not converge. From the plot of the fitted models and residual plots, it is clear that the Wilcoxon fit performs dramatically better than its LS counterpart. In Table 3.1 the Wilcoxon estimates are displayed with their standard error. There is basically little difference between the Wilcoxon fits for the original set and the data set with the gross outlier.

Example 3.2 (Lanczos’ data). In this example we consider a generated data set given in Lanczos (1956). Twenty four observations were generated to 5-digits of accuracy using $f(x) = 0.0951 \exp(-x)+0.8607 \exp(-3x)+1.5576 \exp(-5x)$. Naturally the model we consider is, $f_i(\theta) = \theta_1 \exp(-\theta_2 x_i) + \theta_3 \exp(-\theta_4 x_i) + \theta_5 \exp(-\theta_6 x_i), \ i = 1, \ldots, 24$.

Just as in Example 3.1 we fitted both models with and without an outlier present. This time the outlier introduced does not deviate much from the form of the model as can be seen in Figure 2. The effect of the outlier is clearly seen in the LS residual plot which developed a wave-like pattern. The LS fit follows the outlier; but since the estimation is done under smoothness and shape restrictions imposed by the model, points in the neighborhood of the outlier will also be affected to some degree. This produces the wave in the residual plot.

3.2 A Simulation Study

In Section 2.4, we showed that the efficacy of the nonlinear Wilcoxon estimator is the same as its efficacy in the linear model case. In particular, at normal errors the ARE between the Wilcoxon and LS nonlinear estimators is 0.955. When the distribution of the errors has a heavier tail than the tails of the standard normal distribution the value of the ARE rises substantially. In this subsection, we discuss this ARE for contaminated normal error distributions and compare the results to an empirical finite sample study.

We start by defining $CN(\gamma, \eta) \equiv (1-\gamma)N(0,1) + \gamma N(0, \eta), \ 0 \leq \gamma \leq 1$.

where $\eta > 0$, to be the contaminated normal distribution. In this case the contaminating distribution is also normal but with a variance different from 1.

Taking the ratio of the asymptotic variances of the LS estimator and the Wilcoxon estimator and applying simple algebra shows that the asymptotic relative efficiency of the Wilcoxon estimator relative to the LS estimator when the errors come from the $CN(\gamma, \eta)$ distribution is given by,

$$ARE(\gamma, \eta) = 12 \left[ \frac{(1-\gamma)^2}{2\sqrt{\pi}} + \frac{\gamma^2}{2\sqrt{\eta \pi}} + \frac{\sqrt{2\gamma(1-\gamma)}}{\sqrt{\pi(\eta+1)}} \right]^2 \left[ (1-\gamma) + \gamma \eta \right].$$

It is easy to see that $ARE(0, \eta) = ARE(1, \eta) = ARE(\gamma, 1) = 3/\pi$. Furthermore, $ARE(\gamma, \eta)$ is increasing in both its arguments. So, either an increase in contamination or an increase Table 2 displays values of this ARE for specified values of $\gamma$ and $\eta$. 348
Figure 2: Analysis of Lanczos’ data
In order to investigate the finite sample efficiency of the Wilcoxon estimator relative to the LS estimator, we consider the function,

\[ f_i(\theta) = \exp(x_i \theta), \quad i = 1, \ldots, n. \]

This functional form is then used to generate a vector of response by fixing \( \theta = \log(2) \) and adding random errors as,

\[ y_i = \exp(x_i \log(2)) + \varepsilon_i, \quad x_i = 1, \ldots, n, \]

where \( x_i \) are uniformly distributed over the interval \((0, 5)\) and \( \varepsilon_i \) are sampled from, \( \mathcal{CN}(\gamma, \eta) \).

We performed 1000 repetitions at \( n = 20 \) and obtained LS and Wilcoxon fits using the algorithm given by Sievers and Abebe (2004). The finite sample relative efficiency (RE) is then taken to be the ratio of the bootstrap variance of the LS fit to that of the Wilcoxon fit. The estimated values of the relative efficiency are given in Table 2. One can observe that the estimated values of RE are in a close proximity of the true ARE values.

### Table 2: Estimated relative efficiencies of Wilcoxon relative to LS

<table>
<thead>
<tr>
<th>( \gamma )</th>
<th>RE(( \eta = 3 ))</th>
<th>ARE(( \gamma, \eta = 3 ))</th>
<th>RE(( \eta = 10 ))</th>
<th>ARE(( \gamma, \eta = 10 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma = 0.00 )</td>
<td>0.957</td>
<td>0.955</td>
<td>0.960</td>
<td>0.955</td>
</tr>
<tr>
<td>( \gamma = 0.01 )</td>
<td>1.019</td>
<td>1.009</td>
<td>1.826</td>
<td>1.836</td>
</tr>
<tr>
<td>( \gamma = 0.05 )</td>
<td>1.193</td>
<td>1.196</td>
<td>4.796</td>
<td>4.769</td>
</tr>
<tr>
<td>( \gamma = 0.10 )</td>
<td>1.363</td>
<td>1.373</td>
<td>7.399</td>
<td>7.280</td>
</tr>
<tr>
<td>( \gamma = 0.15 )</td>
<td>1.479</td>
<td>1.497</td>
<td>8.695</td>
<td>8.757</td>
</tr>
<tr>
<td>( \gamma = 0.20 )</td>
<td>1.558</td>
<td>1.575</td>
<td>9.193</td>
<td>9.430</td>
</tr>
</tbody>
</table>

4 AUXILIARY ASYMPTOTIC RESULTS

Let \((\Omega, \mathcal{F}, P)\) be a probability space. We shall write \( h(\theta) \) instead of \( h(\theta)(\omega) \) where \( \omega \in \Omega \) and \( \theta \in \Theta \) when there is no confusion regarding the stochastic nature of the function \( h \) (i.e. \( h(\theta) \) is \( P \)-measurable).

#### 4.1 Consistency

Lemma 4.1 is a generalization of the results of Oberhofer (1982) and Bhattacharyya et. al. (1992), where it is used in establishing the consistency of \( L_1 \) estimators of nonlinear regression parameters.

**Lemma 4.1.** Let \( \{\Gamma_n : n \geq 1\} \) be a real valued sequence of functions defined on \( \Omega \times \Theta \) where \( \Theta \) is a compact space. Let \( \theta_0 \in \Theta^* \) and \( \Theta^* \) be an arbitrary compact subset of \( \Theta \setminus \{\theta_0\} \). If

(i) \( \Gamma_n \) is uniformly continuous on \( \Theta \) for each \( \omega \in \Omega \), uniformly in \( n \),

(ii) there exist a sequence of real valued functions \( \mu_n \) defined on \( \Theta \) such that for each \( \omega \in \Omega \), \( \Gamma_n(\theta) - \mu_n(\theta) \to 0 \) in probability for all \( \theta \in \Theta \) as \( n \to \infty \), and

(iii) there exist a \( \beta = \beta(\Theta^*) > 0 \) and a \( n_0 = n_0(\Theta^*) \) such that for all \( n \geq n_0 \),

\[ \inf_{\theta \in \Theta^*} \mu_n(\theta) \geq \beta, \]

then

\[ \lim_{n \to \infty} P[ \inf_{\theta \in \Theta^*} \Gamma_n(\theta) > 0 ] = 1. \]

**Proof.** Let \( \theta^* \) be an arbitrary point in \( \Theta^* \). Since by (iii) of the theorem we have a \( \beta > 0 \) and a \( n_0 \) such that

\[ \inf_{\theta \in \Theta^*} \mu_n(\theta) \geq \beta, \]

350
whenever \( n \geq n_0 \), (ii) implies that
\[
\lim_{n \to \infty} P \left[ \Gamma_n(\theta^*) \geq \beta/2 \right] = 1 .
\] (17)
Because \( \Gamma_n \) is uniformly continuous in \( \theta \) on \( \Theta \), uniformly in \( n \), there exist an open set \( K^* \) and a \( n^* \geq n_0 \) such that for \( \theta^* \in K^* \) and for all \( \theta \in K^* \), \( |\Gamma_n(\theta) - \Gamma_n(\theta^*)| < \beta/4 \) for \( n \geq n^* \). Hence, by (17), with high probability,
\[
\Gamma_n(\theta) > \beta/4 ,
\]
for all \( \theta \in K^* \) and sufficiently large \( n \). So, \( \inf_{K^*} \Gamma_n(\theta) \geq \beta/4 \) with high probability, i.e.,
\[
\lim_{n \to \infty} P \left[ \inf_{K^*} \Gamma_n(\theta) \geq \beta/4 \right] = 1 .
\] (18)
This is true for all \( \theta^* \) in \( \Theta^* \). Since \( \Theta^* \) is compact, this produces a finite subcover of such sets \( (K^*_1, \ldots, K^*_7) \) covering \( \Theta^* \). Therefore, \( K^* \) in (18) can be replaced by \( \Theta^* \).

In the discussions of Battacharyya et al. (1992) and Oberhofer (1982) the function \( \mu \) is assumed to be the expectation of \( \Gamma_n \) and independent of \( n \). Lemma 4.1 requires the existence of a function, \( \mu_n \), so that the stochastic function \( \Gamma_n - \mu_n \) converges pointwise to 0 and it makes no assumption concerning the existence of the expectation of \( \Gamma_n \).

The above lemma plays a very important role in establishing the consistency of minimizers of dispersion functions. If \( D_n(\theta) \) is a dispersion function, (i) - (iii) of Lemma 4.1 for \( \Gamma_n(\theta) \equiv D_n(\theta) - D_n(\theta_0) \) establish the weak consistency of the minimizer of \( D_n \) as the following lemma of Wu (1981) shows. The proof is similar to the proof of Lemma 1 of Wu (1981).

**Lemma 4.2.** Assume
\[
\hat{\theta}_n \equiv \text{Argmin}_{\theta \in \Theta} D_n(\theta)
\]
exists. Suppose, for any arbitrary compact subset \( \Theta^* \) of \( \Theta \setminus \{ \theta_0 \} \),
\[
\liminf_{n \to \infty} \inf_{\theta \in \Theta^*} \left[ D_n(\theta) - D_n(\theta_0) \right] > 0 ,
\]
a.s. (or in probability). Then, \( \hat{\theta}_n \to \theta_0 \) a.s. (or in probability) as \( n \to \infty \).

Lemma 4.2 is very general in that it only requires the existence of the limit infimum of the process as opposed to the limit which we have assumed in Lemma 4.1. When the limit exists and is finite, the result of Lemma 4.1 is equivalent to the sufficient condition given in Lemma 4.2 via an application of the Dominated Continuity of Measure Theorem (see, for example, Fristedt and Gray (1997)).

### 4.2 Asymptotic Distance Between Minimizers

The following result concerns the asymptotic distance between minimizers of dispersion functions. A version of the result was used by Jaeckel (1972). The version given here is the one found in Hjort and Pollard (1993) generalized to metric spaces.

Assume that \( (\Theta, \rho) \) is a compact metric space. Let \( A_n \) be a real valued convex random function defined on \( \Omega \times \Theta \) and let \( B_n \) be an approximation of \( A_n \) in some compact subspace, \( \Theta_n \) of \( \Theta \). For \( \omega \in \Omega \), we assume that the minimizer, \( \beta_n(\omega) \), of \( B_n \) is unique on \( \Theta_n \); however, we make no assumption as to the uniqueness of the minimizer, \( \alpha_n(\omega) \), of \( A_n \). Furthermore, let \( C_n(\eta) = \{ \theta \in \Theta : \rho(\theta, \beta_n) \leq \eta \} \).

The proof of the following lemma may be constructed along the lines of Hjort and Pollard (1993) in a straightforward manner.

**Lemma 4.3.** For \( \theta \in \Theta \) and \( \eta > 0 \),
\[
P \left[ \rho(\alpha_n, \beta_n) \geq \eta \right] \leq P \left[ \Delta_n(\eta) \geq h_n(\eta) \right],
\]
where
\[
\Delta_n(\eta) = \sup_{\theta \in C_n(\eta)} |A_n(\theta) - B_n(\theta)| ,
\]
and
\[
h_n(\eta) = \inf_{\theta \in \partial C_n(\eta)} B_n(\theta) - B_n(\beta_n) .
\]
To show the asymptotic equivalence of the two minimizers, \( \alpha_n \) and \( \beta_n \), we may apply Theorem 4.3 with \( \eta = \delta/\sqrt{n} \), \( \delta > 0 \). Sufficient conditions are \( h_n(\delta/\sqrt{n}) \) is stochastically bounded above zero and \( \Delta_n(\delta/\sqrt{n}) \) converges to zero in probability as \( n \to \infty \).

This just says the minimizer of \( B_n \) is unique in a shrinking ball as \( n \to \infty \). The process \( A_n \) is allowed to have a flat bottom as in most rank dispersion functions. In most cases, \( A_n \) is taken to be the local convex approximation of \( B_n \) via a Taylor series expansion.

4.3 Conditions for Normality

Noether’s condition is one of the sufficient conditions for asymptotic normality of an estimator. For \( i = 1, \ldots, n \), let \( x_i = (x_{i1}, \ldots, x_{ip})' \) denote \( p \times 1 \) vectors. Then Noether’s condition is given by,

\[
\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \frac{x_{ik}^2}{\sum_{j=1}^{p} x_{jk}^2} \to 0 \quad \text{for all } k = 1, \ldots, p,
\]

as \( n \to \infty \). The result of this section gives sufficient conditions needed for Noether’s condition. The following lemma along with a proof can be found in Wu (1981).

**Lemma 4.4.** Let \( x_i, i = 1, \ldots, n \), be \( p \times 1 \) vectors such that there exist \( \alpha_n \uparrow \infty \) and \( \lim_{n \to \infty} \alpha_{n-1}/\alpha_n = 1 \) with \( \alpha_n^{-1} \sum_{j=1}^{n} x_j x_j' \) converging to a positive definite matrix \( \Sigma \). Then

\[
\max_{1 \leq i \leq n} \frac{\max_{1 \leq j \leq n} (\sum_{j=1}^{n} x_j x_j')^{-1} x_i}{\max_{1 \leq i \leq n} h_{ii}} \to 0
\]

as \( n \to \infty \).

Define \( X \) to be the \( n \times p \) matrix with the \( i \)th row given by \( x_i' \). Now \( H_n = X(X'X)^{-1}X' \) is the projection matrix onto the column space of \( X \). Another condition often used in proving asymptotic normality (known as Huber’s condition) is

\[
\lim_{n \to \infty} \max_{1 \leq i \leq n} h_{ii} = 0,
\]

where \( h_{ii} \) is the \( i \)th diagonal entry of \( H_n \). The following lemma given by Hettmansperger and McKean (1998) shows that Huber’s condition is sufficient for Noether’s condition.

**Lemma 4.5.** Under the above assumptions,

\[
\max_{1 \leq k \leq p} \max_{1 \leq i \leq n} \frac{x_{ik}^2}{\sum_{j=1}^{p} x_{jk}^2} \leq \max_{1 \leq i \leq n} h_{ii}.
\]

5 PROOFS

5.1 Results of Section 2

*Proof of Lemma 2.1.* Since \( \theta \in \mathbb{R}^p \), we may write it as \( \theta = (\theta_1, \ldots, \theta_p)' \). Let \( F(\theta) \) be the \( n \times p \) matrix with \( ij \)th element \( f_i(\theta_j) \). Let \( \Xi = \{ F(\theta) : \theta \in \Theta \} \) and \( \Xi_1 = \{ F(\theta) : \theta \in \Theta \} \). Any two vectors, \( v \in \Xi \) and \( v_c \in \Xi_1 \), are related as \( v = a1 + v_c \) where \( a \in \mathbb{R} \). We have

\[
\|y - v\|_{L_1} = \|y - a1 - v_c\|_{L_1} \geq \|y - \text{med}\{y - v_c\}1 - v_c\|_{L_1},
\]

with the last inequality due to the fact that the sample median minimizes the \( L_1 \) distance between a vector and the space spanned by \( 1 \). This implies, for any \( v \in \Xi \),

\[
\|y - v\|_{L_1} \geq \|y - \text{med}\{y - v_c\}1 - v_c\|_{L_1} = \|y - v_c\|_S, \tag{19}
\]

since \( \text{sgn}(y_i - \text{med}\{y - v_c\} - v_c) = \text{sgn}(R(y_i - v_c) - (n + 1)/2) \) and the sign scores sum to 0. Using the same argument we can show that

\[
\|y - \text{med}\{y - f(\bar{\theta}_s)\}1 - f(\bar{\theta}_s)\|_{L_1} = \|y - f(\bar{\theta}_s)\|_S. \tag{20}
\]

Putting (19) and (20) together completes the proof. \( \square \)
Proof of Lemma 2.2. The statement of the lemma can be written as,

\[
\lim_{n \to \infty} P\left( \left| 2n(n+1)^{-1} \sum_{i<j}|W_{ij}(\theta, \theta_0) - E(W_{ij}(\theta, \theta_0))| \right| > \delta \right) = 0 ,
\]

for all \( \delta > 0 \). Now applying Markov’s inequality followed by Minkowski’s, triangular and Jensen’s inequalities (see Petrov (1995)) we get,

\[
P\left( \left| 2n(n+1)^{-1} \sum_{i<j}|W_{ij}(\theta, \theta_0) - E(W_{ij}(\theta, \theta_0))| \right| > \delta \right)
\leq \frac{2\delta(n+1)}{[2\delta n(n+1)]^{-1}E\left[ \sum_{i<j}|W_{ij}(\theta, \theta_0) - E(W_{ij}(\theta, \theta_0))| \right]}
\leq \frac{2\delta(n+1)}{2E|W_{ij}(\theta, \theta_0) - E(W_{ij}(\theta, \theta_0))|}
\leq \frac{2\delta(n+1)}{[2\delta n(n+1)]^{-1}E\left[ |W_{ij}(\theta, \theta_0)| + |E(W_{ij}(\theta, \theta_0))| \right]}
\leq \frac{2\delta(n+1)}{[\delta n(n+1)]^{-1}E|W_{ij}(\theta, \theta_0)|}.
\]

But,

\[
|W_{ij}(\theta, \theta_0)| = ||\varepsilon_i(\theta) - \varepsilon_j(\theta)| - |\varepsilon_i - \varepsilon_j||
\leq |h_{ij}(\theta, \theta_0)|
\leq |h^*_i(\theta, \theta_0)| + |h^*_j(\theta, \theta_0)|.
\]

This implies that,

\[
[\delta n(n+1)]^{-1} E|W_{ij}(\theta, \theta_0)| \leq \frac{n-1}{\delta n(n+1)} \sum_{i=1}^{n} |h_i^*(\theta, \theta_0)|
\leq \frac{n-1}{\delta(n+1)} \left\{ n^{-1} \Delta_n(\theta, \theta_0) \right\}^{1/2} ,
\]

which goes to zero as \( n \to \infty \) by A3. This combined with (21) completes the proof. \( \square \)

Proof of Lemma 2.3. Note that,

\[
D_n(\theta) - D_n(\theta_0) = [2n(n+1)]^{-1} \sum_{i<j} \left\{ \|\varepsilon_i(\theta) + \varepsilon_j(\theta, \theta_0)\| - |\varepsilon_i - \varepsilon_j| \right\} .
\]

It is easy to show that if \( T \) is a random variable with distribution function \( F_T \) and \( F_T(0) = 1/2 \), then for any constant \( k \),

\[
E(|T + k| - |T|) = 2I(k \leq 0) \int_{0}^{-k} \{ |k| - x \} dF_T(x) + 2I(k > 0) \int_{0}^{k} \{ |k| + x \} dF_T(x) .
\]

(22)

Applying this we obtain,

\[
[2n(n+1)]^{-1} \sum_{i<j} E\left\{ \|\varepsilon_i(\theta) + \varepsilon_j(\theta, \theta_0)\| - |\varepsilon_i - \varepsilon_j| \right\}
= [n(n+1)]^{-1} \sum_{(i,j) \in A} \int_{0}^{-h_{ij}(\theta, \theta_0)} \{ |h_{ij}(\theta, \theta_0)| - x \} dG(x)
+ [n(n+1)]^{-1} \sum_{(i,j) \in B} \int_{0}^{h_{ij}(\theta, \theta_0)} \{ |h_{ij}(\theta, \theta_0)| + x \} dG(x) ,
\]

353
The result follows from Lemma 4.2. For any \( \theta \) it is uniformly continuous in \( B_\theta \). Thus by Lemma 4.1 we have that
\[
[2n(n+1)]^{-1} \sum_{i<j} E\{ |\varepsilon_i - \varepsilon_j| + h_{ij}(\theta, \theta_0)| - |\varepsilon_i - \varepsilon_j| \}
\geq [2n(n+1)]^{-1} \sum_{(i,j) \in A} |h_{ij}(\theta, \theta_0)| \{ G(-h_{ij}(\theta, \theta_0)/2) - 1/2 \}
+ [2n(n+1)]^{-1} \sum_{(i,j) \in B} |h_{ij}(\theta, \theta_0)| \{ 1/2 - G(-h_{ij}(\theta, \theta_0)/2) \}
\geq [2n(n+1)]^{-1} \sum_{i<j} |h_{ij}(\theta, \theta_0)| \times \min \{ G(|h_{ij}(\theta, \theta_0)/2|) - 1/2, 1/2 - G(-|h_{ij}(\theta, \theta_0)/2|) \}.
\]

Proof of Theorem 2.2. Let
\[ \Gamma_n(\theta) = D_n(\theta) - D_n(\theta_0), \]
and
\[ \mu_n(\theta) = E \{ \Gamma_n(\theta, \theta_0) \}. \]

For any \( \theta^* \in \Theta^* \subset \Theta \), where \( \Theta^* \) is a closed set not containing \( \theta_0 \), under A5 and Lemma 2.3, there exist a \( \beta = \beta(\Theta^*) > 0 \) and a \( n_0 = n_0(\Theta^*) \), such that for all \( n \geq n_0 \),
\[ \inf_{\Theta^*} \mu_n(\theta) \geq \beta. \] (23)

Because
\[ \Gamma_n(\theta) \leq \frac{n-1}{n+1} \{ n^{-1} \Delta_n(\theta, \theta_0) \}^{1/2}, \]
it is uniformly continuous in \( \theta \) on \( \Theta \), uniformly in \( n \). Moreover, by Lemma 2.2, we have \( \Gamma_n(\theta) - \mu_n(\theta) \to 0 \)
in probability for all \( \theta \in \Theta \). Thus by Lemma 4.1 we have that
\[ \lim_{n \to \infty} P \{ \inf_{\Theta^*} \Gamma_n(\theta) > 0 \} = 1. \]
The result follows from Lemma 4.2.

Proof of Lemma 2.4. Assume that there exists \( \theta \in \Theta \) distinct from \( \theta_0 \) such that \( f(\theta) = f(\theta_0) \) \( \lambda \)-a.e. For such \( \theta \), \( |h_{ij}(\theta, \theta_0)| = 0 \) \( \lambda \)-a.e. implying that A5 fails to hold.

Proof of Lemma 2.5. Since \( G(0) = 1/2 \), applying a first order Taylor series expansion of \( G \) about \( 0 \), we have
\[ n^{-2} \sum_{i<j} |h_{ij}(\theta, \theta_0)| \times \min \left\{ G\left( \frac{|h_{ij}(\theta, \theta_0)|}{2} \right) - 1/2, \frac{1}{2} - G\left( -|h_{ij}(\theta, \theta_0)|/2 \right) \right\} \]
\[ = n^{-2} \sum_{i<j} |h_{ij}(\theta, \theta_0)|^2 g(t)/2, \]
where \( t = -|h_{ij}(\theta, \theta_0)|/2, |h_{ij}(\theta, \theta_0)|/2 \). Since \( g \) is continuous there is an interval \( (-\delta, \delta) \) over which \( g > 0 \) for some \( \delta > 0 \). Moreover, since \( h_{ij} \) is continuous and \( \Theta^* \) is arbitrary, the interval \( (-|h_{ij}(\theta, \theta_0)|/2, |h_{ij}(\theta, \theta_0)|/2) \)
\[ \text{can be made a subset of } (-\delta, \delta). \] The result follows from A5.2.
Proof of Lemma 2.6. Let $\rho$ be any metric on $\Theta$. For $\delta > 0$, define $M_n(\delta) = \{ \theta \in \Theta: \rho(\hat{\theta}_{W,n}, \theta) \leq \delta/\sqrt{n} \}$. By Lemma 4.3, sufficient conditions for

$$\sqrt{n}(\hat{\theta}_{W,n} - \theta_n) \overset{P}{\to} 0,$$

are

$$\sup_{\theta \in M_n(\delta)} |D_n(\theta) - D_n^*(\theta)| \overset{P}{\to} 0, \quad \text{and}$$

$$\inf_{\theta \in \partial M_n(\delta)} \{ D_n(\theta) - D_n(\hat{\theta}_{W,n}) \} \geq \beta > 0,$$

for all $\delta > 0$ and sufficiently large $n$.

To verify (24) notice that

$$|D_n(\theta) - D_n^*(\theta)| \leq [2n(n + 1)]^{-1} \sum_{i \neq j} |e_i(\theta) - e_j(\theta) - e_i^*(\theta) + e_j^*(\theta)|$$

$$\leq \frac{n - 1}{2n(n + 1)} \sum_{i=1}^n |e_i(\theta) - e_i^*(\theta)|.$$ 

But, for $1 \leq i \leq n$,

$$|e_i(\theta) - e_i^*(\theta)| \leq |f_i(\theta) - f_i(\hat{\theta}_{W,n})| + |\{\nabla f_i(\theta)\}^T(\theta - \theta_0)|$$

$$\leq |f_i(\theta) - f_i(\hat{\theta}_{W,n})| + |f_i(\theta) - f_i(\hat{\theta}_{W,n})| + |\{\nabla f_i(\theta)\}^T(\hat{\theta}_{W,n} - \theta_0)|.$$ 

Moreover, since $f_i$ are uniformly continuous on $\Theta$ and $\hat{\theta}_{W,n}$ is weakly consistent for $\theta_0$,

$$|f_i(\theta) - f_i(\hat{\theta}_{W,n})| + \sup_{\theta \in M_n(\delta)} |f_i(\theta) - f_i(\hat{\theta}_{W,n})| \overset{P}{\to} 0. \quad (26)$$

We also have $\|\nabla f_i(\theta)\| < \infty$ by N1 and the compactness of $\Theta$. This gives us,

$$\sup_{\theta \in M_n(\delta)} |\{\nabla f_i(\theta)\}^T(\theta - \hat{\theta}_{W,n})| + \sup_{\theta \in M_n(\delta)} |\{\nabla f_i(\theta)\}^T(\hat{\theta}_{W,n} - \theta_0)| \overset{P}{\to} 0. \quad (27)$$

The expressions in (26) and (27) establish (24).

Proceeding to show (25) notice that by the definition of $\hat{\theta}_{W,n}$ and continuity of $D_n(\theta)$ in $\theta$ we have,

$$\inf_{\theta \in \partial M_n(\delta)} \{ D_n(\theta) - D_n(\hat{\theta}_{W,n}) \} \geq 0 \quad \text{a.s.} \quad (28)$$

We also have,

$$\inf_{\theta \in \partial M_n(\delta)} \{ D_n(\theta) - D_n(\theta_0) \} \geq \inf_{\theta \in M_n(\delta)} E\{ D_n(\theta) - D_n(\theta_0) \}$$

$$+ \inf_{\theta \in \partial M_n(\delta)} \left[ D_n(\theta) - D_n(\theta_0) - E\{ D_n(\theta) - D_n(\theta_0) \} \right].$$

But by Lemma 2.3 and assumption A5, there exist $\eta > 0$ and $n_0$ such that for all $n \geq n_0$,

$$\inf_{\theta \in \partial M_n(\delta)} E\{ D_n(\theta) - D_n(\theta_0) \} \geq \eta.$$ 

Also by Lemma 2.2,

$$\inf_{\theta \in \partial M_n(\delta)} \left[ D_n(\theta) - D_n(\theta_0) - E\{ D_n(\theta) - D_n(\theta_0) \} \right] \overset{P}{\to} 0.$$ 

Thus for sufficiently large $n$ we have,

$$\inf_{\theta \in \partial M_n(\delta)} \{ D_n(\theta) - D_n(\theta_0) \} \geq \eta. \quad (29)$$

The expressions (28) and (29) give (25). The proof is complete. \qed
6 Conclusion

In this article, we have proposed a robust estimator for nonlinear models. Its geometry is similar to that of the least squares (LS) estimator in that another norm (Wilcoxon) is substituted for the LS (Euclidean) norm. It generalizes the Wilcoxon R estimator for linear models. We then showed that it was a consistent estimator of the true parameter and obtained its asymptotic distribution. Its efficacy is the same as that of the Wilcoxon estimator for simple location and linear models. In particular, its asymptotic relative efficiency (ARE) relative to the the nonlinear LS estimator is the same as in these linear models. Under normal errors this ARE is the well known ratio 0.955. Hence, it is a highly efficient estimator. As with its linear model counterpart, it is more efficient than the LS estimator for heavier than normal tailed distributions. In a simulation study, these AREs were verified for a family of contaminated normal distributions.

As with the highly efficient Wilcoxon estimator for linear models, the Wilcoxon nonlinear estimator has bounded influence in the $y$-space but not in the model space. In a concurrent study, Abebe and McKean (2006) extended the Wilcoxon estimator to a class of weighted Wilcoxon estimators which have bounded influence in both the $y$-space and the model-space. These Wilcoxon nonlinear estimators are easy to compute. We discussed computational algorithms in Section 2.6.1.

REFERENCES


On tail probabilities of the t-Distribution for samples from the Uniform Distribution

R. Ayesha Ali\textsuperscript{1} and M. M. Ali\textsuperscript{2}
\textsuperscript{1}Dept. of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada, aali@uoguelph.ca
\textsuperscript{2}Dept. of Statistics & Actuarial Sciences, University of Western Ontario, London, ON, Canada

Abstract
This paper deals with the tail probabilities of the Student's $t$-distribution under the assumption that the sample arises from a uniformly distributed population. It is shown that the tail probability of the Student's $t$-distribution, $Pr(T \geq t)$, for sufficiently large values of $t$, under the assumption that the sample arises from a uniform population is greater than the corresponding tail probability of the usual Student's $t$-distribution under the normality assumption.

Keywords: tail probabilities, spherically symmetric distributions, $t$-distribution.

1 Introduction

In testing of hypotheses based on a particular statistic, only the tail probabilities of the test statistic are relevant. Hotelling (1961) raised this problem in connection with robustness of tests and since then, Hoq et al. (1978) also studied tail end probabilities. In particular, the $t$-distribution is often used for small sample problems when the underlying population distribution is non-normal. Hence, knowing how the exact tail probabilities of the $t$-distribution under non-normality assumption provides insight into the robustness of $t$-tests under violations of the usual normality assumption. However, the problem of comparing tail ends of $t$-distributions under various underlying assumptions regarding the population distribution, to the usual $t$-distribution under normality assumption is in itself a challenging mathematical problem.

The object of this paper is to compare the tail end probabilities of the $t$-distribution for a sample from a uniform population to the usual $t$-distribution. In particular we show that the probability content of the tail ends of the $t$-distribution is greater when the population distribution is uniform.

Since this problem is rather specialized, we feel that a review of robustness of tests is not necessary. However, the reports by Posten and Hatch (1966) and Govindarajulu and Leslie (1970) provide excellent surveys of the subject matter. We instead provide in Section 2 a brief review of key results needed to prove our main result, which is then stated and proved in Section 3.

2 Background

Consider a sample $X_1, X_2, \ldots, X_n$ of size $n$ from a population symmetric around location parameter $\mu$ and with scale parameter $\sigma$ so that

$$g(x; \mu, \sigma) = \frac{1}{\sigma} f \left( \frac{x - \mu}{\sigma} \right),$$

Let

$$T_n = \frac{\sqrt{n}(\bar{X} - \mu)}{s_X},$$

358
where $s_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$, denote the t-statistic.

We are interested in proving that, for all sufficiently large values of $t$, the $Pr\{T_n \geq t\}$ when the population has a uniform distribution over the interval $(-\frac{1}{2}, \frac{1}{2})$, is greater than the corresponding probability of the usual t-distribution under normality assumption.

In what follows, we will resort to qualitative arguments of a geometrical nature to prove the main result. Although the geometrical interpretation of the t-statistic is very well-known, we will sketch this interpretation in order to set up our notation and to shed some light on the nature of the problem at hand.

In addition we will require another result that is not so well-known, which states that the distribution of the Student’s t-statistic is the same as that of the usual t-distribution under normality assumption, for any member of the class of spherically symmetric distributions having center at $\mu$ and arbitrary scale parameter $\sigma$. It is to be noted that the normal distribution is a member of this class and the component variables of the joint density of a spherically symmetric distribution need not be independent, although they are uncorrelated. An extensive treatment of spherically symmetric distributions with a comprehensive bibliography is given by Fang and Anderson (1990).

2.1 Geometry of the t-distribution

Let $X_1, X_2, \ldots, X_n$ be $n$ random variables from a population symmetric around location parameter $\mu$ and with scale parameter $\sigma$ and let, as before,

$$T_n = \frac{\sqrt{n}(\bar{X} - \mu)}{s_X}.$$  \hfill (1)

When $\mu$ is assumed to have a known value, say $\mu_0$, then let the standardized random variable $Z_i$ be given by

$$Z_i = (X_i - \mu_0)/\sigma, \quad i = 1, \ldots, n.$$  

Obviously, $T_n$ can be written in terms of $Z_1, Z_2, \ldots, Z_n$ as follows

$$T_n = \frac{\sqrt{n}(\bar{Z})}{s_Z},$$  \hfill (2)

where $s_Z^2 = \frac{\sum_{i=1}^{n} (Z_i - \bar{Z})^2}{n-1}$. Note that the distribution of $T_n$ is free from the scale parameter $\sigma$. Hence, without loss of generality, the problem reduces to that of considering the distribution of the t-statistic based on $Z_1, Z_2, \ldots, Z_n$ with each $Z_i$ having mean 0 and $\sigma = 1$, for example.

Consider an $n$-dimensional Euclidean space and set up the Cartesian coordinates $Z_1, \ldots, Z_n$.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure1.png}
\caption{$n$-dimensional Euclidean space with coordinates $Z_1, \ldots, Z_n$. Let the angle $POQ$ be $\theta$, and angle $PQO$ is a right angle.}
\end{figure}

Let $O$ denote the origin and let $OM$ be the line equiangular to the axes so that the direction cosine of $OM$ with each axis is $1/\sqrt{n}$. Let $P$ be a general point having co-ordinates $(z_1, z_2, \ldots, z_n)$ and $Q$ be the point $(\bar{z}, \bar{z}, \ldots, \bar{z})$ on the equiangular line $OM$. Since $P$ and $Q$ lie on the hyperplane
\[ Z_1 + Z_2 + \cdots + Z_n = n \bar{Z} \]

and \( OQ \) is perpendicular to the above hyperplane, it follows that angle \( PQO \) is a right angle (see Figure 1). Letting \( \theta \) be angle \( POQ \), then

\[
\sqrt{n - 1} \cot(\theta) = \sqrt{n - 1} \frac{OQ}{PQ} = \frac{\sqrt{n - 1} \sqrt{n \bar{Z}^2}}{\sqrt{\sum_{i=1}^{n} (Z_i - \bar{Z})^2}} = \frac{\sqrt{n - 1} \sqrt{n \bar{Z}}}{\sqrt{\sum_{i=1}^{n} (Z_i - \bar{Z})^2}} = T_n.
\]

Hence

\[
\Pr\{ (z_1, z_2, \ldots, z_n) \mid T_n \geq t \} = \Pr \left\{ (z_1, z_2, \ldots, z_n) \mid \cot(\theta) \geq \frac{t}{\sqrt{n - 1}} \right\} = \Pr \left\{ (z_1, z_2, \ldots, z_n) \mid \theta \leq \cot^{-1} \left( \frac{t}{\sqrt{n - 1}} \right) \right\};
\]

this set represents the points inside a cone with axis along \( OM \), apex at \( O \), and with generator line \( OP \) subtending an angle \( \theta = \cot^{-1} \left( \frac{t}{\sqrt{n - 1}} \right) \) (See Figure 2).

![Figure 2](image)

**Figure 2**: \( \Pr\{ (z_1, z_2, \ldots, z_n) \mid T \geq t \} \) is the probability content of the cone shown with axis along \( OM \), and with subtending angle \( \theta \).

Hence with \( T_n \) as defined in (1) or (2),

\[
\Pr\{ T_n \geq t \} = \Pr \left\{ (z_1, z_2, \ldots, z_n) \mid \theta \leq \cot^{-1} \left( \frac{t}{\sqrt{n - 1}} \right) \right\}
\]

which is the probability content of the cone having axis \( OQ \), generator along the line \( OP \), and vertical angle \( POQ = \theta = \cot^{-1} \left( \frac{t}{\sqrt{n - 1}} \right) \).

### 2.2 Geometry of spherically symmetric distributions

Consider an \( n \)-dimensional random variable \( X = (X_1, X_2, \ldots, X_n)' \) having probability density of the form

\[
f(x_1, \ldots, x_n) = g \left( \sum_{i=1}^{n} x_i^2 \right) = g(X'X).
\]

Any orthogonal transformation \( Y = AX \) implies
\((Y'Y) = X'(A'A)X = (X'X),\)

so under any rotation of axes around the origin, \(g(X'X) = g(Y'Y)\). Further, on any hypersphere with origin as the centre we have constant density. In other words, equiprobable surfaces are concentric spheres with centre at the origin \(O\). It is clear that the radial projection of the probability mass along any radial line from \(O\) onto the surface of the unit sphere is constant. That is to say that the probability distribution of \((X_1, X_2, \ldots, X_n \mid \sum_{i=1}^{n} x_i^2 = 1)\) is uniformly distributed on the surface of the unit sphere.

\[
\begin{align*}
\text{Figure 3: Unit mass is uniformly distributed on surface of unit sphere. Hence Pr} & \{ (x_1, x_2, \ldots, x_n) \mid T_n \geq t \} \\
& \text{is mass contained in intersection of the cone and the unit sphere.}
\end{align*}
\]

Now consider the set of points
\[
\left\{ (x_1, x_2, \ldots, x_n) \mid \theta \leq \cot^{-1} \left( \frac{t}{\sqrt{n-1}} \right) \right\},
\]
describing a cone. The corresponding mass on the intersection of the cone and the unit sphere will be constant for any member of the class of spherically symmetric distributions (see Figure 3). Since the normal distribution belongs to this class, it follows that the distribution of the \(t\)-statistic for any member of the class of spherically symmetric distributions is identical to the usual \(t\)-distribution under normality assumption.

In particular, if \(X_1, X_2, \ldots, X_n\) have uniform distribution inside a sphere of any prescribed radius \(c\), then this distribution belongs to the class of spherically symmetric distributions. Hence the distribution of the \(t\)-statistic is the usual \(t\)-distribution under normality assumption.

In other words, if \(X_1, X_2, \ldots, X_n\) is uniformly distributed inside a sphere of radius \(c\), then
\[
f(x_1, \ldots, x_n) = \begin{cases} 
\frac{1}{K} & \text{inside } \sum_{i=1}^{n} x_i^2 \leq c^2 \\
0 & \text{elsewhere}
\end{cases}
\]

where \(K\), the normalizing constant, is given by
\[
K = \int_{\sum x_i^2 \leq c^2} dx_1 \cdots dx_n.
\]

Then \(Pr(T_n \geq t)\) equals the volume of the intersection of the cone defined in (3) and the sphere of radius \(c\).
3 The Problem

In order to avoid complications of a trivial nature, in what follows we will restrict ourselves to population probability distributions which are symmetric around \( \mu \) and hence \( \mu \) will be taken to be zero. We now formally state the problem:

**Theorem 3.1** Consider a random sample \( X_1, X_2, \ldots, X_n \) of size \( n \) from the Uniform distribution over the interval \( (-\frac{1}{2}, \frac{1}{2}) \). Also consider a random sample \( Z_1, Z_2, \ldots, Z_n \) of size \( n \) from the standard \( N(0,1) \) distribution. The tail probabilities of \( T_n \) based on \( X_1, X_2, \ldots, X_n \) are greater than the corresponding tail probabilities of \( T_n \) based on \( Z_1, Z_2, \ldots, Z_n \).

**Proof:** Let \( C \) denote the \( n \)-cube defined by \( -\frac{1}{2} \leq X_i \leq \frac{1}{2}, \ i = 1, \ldots, n \) and having volume one. We now seek a hypersphere of radius \( r_0 \), centred at the origin, and having unit volume, for which we can define a spherically symmetric distribution with constant density one at each point inside the hypersphere. Clearly, the volume of a hypersphere with radius \( \frac{1}{2} \), centred at \((0,0,\ldots,0)\), and which is fully inscribed inside \( n \)-cube \( C \), is less than 1; while the volume of a hypersphere with radius \( \sqrt{\pi}/2 \) and circumscribing the unit \( n \)-cube will be greater than 1. Hence, by continuity, there exists an \( r = r_0 \), where \( \frac{1}{2} < r_0 < \frac{\sqrt{\pi}}{2} \), and for which the volume of the hypersphere is 1 (see Figure 4). For any given \( n \), \( r_0 \) can be easily computed. Let \( S_{r_0} \) denote this hypersphere of radius \( r_0 \).

![Figure 4: (a) The sphere has mass less than the unit \( n \)-cube. (b) The hypersphere has mass greater than the unit \( n \)-cube. (c) The probability content of the intersection of the cone and the unit \( n \)-cube is greater than the probability content of the intersection of the cone and the hypersphere.](image)

Now consider \( Y_1, Y_2, \ldots, Y_n \) uniformly distributed inside hypersphere \( S_{r_0} \). Since the uniform distribution belongs to the class of spherically symmetric distributions, the distribution of \( T_n \) based on \( Y_1, \ldots, Y_n \) is identical to the distribution under normality assumption, i.e. to the distribution of \( T_n \) based on \( Z_1, \ldots, Z_n \).

Note that since \( \frac{1}{2} < r_0 < \frac{\sqrt{\pi}}{2} \), hypersphere \( S_{r_0} \) cuts the corners of \( n \)-cube \( C \) (see Figure 4(c)).

Let \( K \) denote the set of points contained in the cone described in (3), with vertical angle \( \theta \). For sufficiently small \( \theta \), the volume of the intersection of cone \( K \) and hypersphere \( S_{r_0} \) is contained in the intersection of cone \( K \) and the unit cube \( C \). Hence

\[
\Pr\{T \geq t \mid \text{sample arises from Uniform } (-\frac{1}{2}, \frac{1}{2}) \text{ population}\} = \frac{\text{vol}(K \cap C)}{\text{vol}(K \cap S_{r_0})} \geq \frac{\text{vol}(K \cap S_{r_0})}{\text{vol}(K \cap C)} = \Pr\{T \geq t \mid \text{the sample arises from a Normal population}\}.
\]

Hence the proof is complete.

**References**


