Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 44, Number 3, 2015

ISSN: 1026597X, Vienna, Austria

Österreichische Zeitschrift für Statistik

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Austrian Journal of Statistics; Information and Instructions

GENERAL NOTES

The Austrian Journal of Statistics is an open-access journal with a long history and is published approximately quarterly by the Austrian Statistical Society. Its general objective is to promote and extend the use of statistical methods in all kind of theoretical and applied disciplines. Special emphasis is on methods and results in official statistics.

Original papers and review articles in English will be published in the Austrian Journal of Statistics if judged consistently with these general aims. All papers will be refereed. Special topics sections will appear from time to time. Each section will have as a theme a specialized area of statistical application, theory, or methodology. Technical notes or problems for considerations under Shorter Communications are also invited. A special section is reserved for book reviews.

All published manuscripts are available at

http://www.ajs.or.at

(old editions can be found at http://www.stat.tugraz.at/AJS/Editions.html)

Members of the Austrian Statistical Society receive a copy of the Journal free of charge. To apply for a membership, see the website of the Society. Articles will also be made available through the web.

PEER REVIEW PROCESS

All contributions will be anonymously refereed which is also for the authors in order to getting positive feedback and constructive suggestions from other qualified people. Editor and referees must trust that the contribution has not been submitted for publication at the same time at another place. It is fair that the submitting author notifies if an earlier version has already been submitted somewhere before. Manuscripts stay with the publisher and referees. The refereeing and publishing in the Austrian Journal of Statistics is free of charge. The publisher, the Austrian Statistical Society requires a grant of copyright from authors in order to effectively publish and distribute this journal worldwide.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge.

ONLINE SUBMISSIONS

Already have a Username/Password for Austrian Journal of Statistics?
Go to http://www.ajs.or.at/index.php/ajs/login

Need a Username/Password?
Go to http://www.ajs.or.at/index.php/ajs/user/register

Registration and login are required to submit items and to check the status of current submissions.

AUTHOR GUIDELINES

The original LaTeX-file guidelinesAJS.zip (available online) should be used as a template for the setting up of a text to be submitted in computer readable form. Other formats are only accepted rarely.

SUBMISSION PREPARATION CHECKLIST

- The submission has not been previously published, nor is it before another journal for consideration (or an explanation has been provided in Comments to the Editor).
- The submission file is preferable in LaTeXfile format provided by the journal.
- All illustrations, figures, and tables are placed within the text at the appropriate points, rather than at the end.
- The text adheres to the stylistic and bibliographic requirements outlined in the Author Guidelines, which is found in About the Journal.

COPYRIGHT NOTICE

Manuscripts should be unpublished and not be under consideration for publication elsewhere. By submitting an article, the author(s) certify that the article is their original work, that they have the right to submit the article for publication, and that they can grant the above license.

Austrian Journal of Statistics

Volume 44, Number 3, 2015

Editor-in-chief: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

Österreichische Zeitschrift für Statistik

Jahrgang 44, Heft 3, 2015

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Impressum

Editor:	Matthias Templ, Statistics Austria & Vienna University of Technology
Editorial Board:	Peter Filzmoser, Vienna University of Technology Herwig Friedl, TU Graz Bernd Genser, University of Konstanz Peter Hackl, Vienna University of Economics, Austria Wolfgang Huf, Medical University of Vienna, Center for Medical Physics and Biomedical Engineering Alexander Kowarik, Statistics Austria, Austria Johannes Ledolter, Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien & Management Sciences, University of Iowa Werner Mueller, Johannes Kepler University Linz, Austria Josef Richter, University of Innsbruck Milan Stehlik, Department of Applied Statistics, Johannes Kepler University, Linz, Austria Wolfgang Trutschnig, Department for Mathematics, University of Salzburg Regina Tüchler, Austrian Federal Economic Chamber, Austria Helga Wagner, Johannes Kepler University Walter Zwirner, University of Calgary, Canada
Book Reviews:	Ernst Stadlober, Graz University of Technology
Printed by	Statistics Austria, A-1110 Vienna

Published approximately quarterly by the Austrian Statistical Society, C/o Statistik Austria
Guglgasse 13, A–1110 Wien

Contents

Editorial

This volume include five scientific papers, one interview and one book review.

The first contribution analyses the impact of height on the occurence of Type II diabetes. The analysis of Type II diabetes prevalence under certain aspects is a highly newsworthy topic from an epidemiological point of view. The relation between Type II diabetes and height have not been yet accurately described.

In the second paper (*Transmuted Modified Inverse Rayleigh Distribution*) a new distribution for modelling reliability data is introduced.

The third contribution – *Covariance Structure of Compositional Tables* – deals with a new and interesting topic. Compositional tables are a continuous counterpart to contingency tables, and every probability table itself is of compositional nature. The paper gives contributions to the analysis of the covariance structure of compositional tables.

The fourth paper with the title *A New Class of Generalized Modified Weibull Distribution with Applications* defines a five parameter distribution that includes many well-known distributions. In addition, the authors also provide the code in R. The last paper analysis the incidence of stroke in Upper Austria with respect to diabetes. The authors found a strong relation between stroke risk and diabetes.

The Austrian Journal of Statistics started an interview series. Interviews with two elected former presidents of the Austrian Statistical Society were already published in former issues. Peter Hackl was full professor at the Vienna University of Economics, director general of Statistics Austria and also a former president of the Austrian Statistical Society. The new interview with him shows very interesting historical remarks and discusses new challanges in statistics.

A new book from Andreas Quatember about data quality in complex sample survey is discussed by Ernst Stadlober.

Finally, an obituary to a honorary member of the Austrian Statistical Society is given.

Matthias Templ
(Editor-in-Chief)

Statistics Austria & Vienna University of Technology
Wiedner Hauptstr. 8–10
A–1040 Vienna, Austria
E-mail: matthias.templ@gmail.com

Vienna/, July 2015

Austrian Journal of Statistics
October 2015, Volume 44, 3–15.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i3.17

On the Selection of Relevant Covariates and Correlation Structure in Longitudinal Binary Models: Analysing the Impact of the Height of Type II Diabetes

Md Erfan Hoque
University of Dhaka

Mahfuzur Rahman Khokan
University of Dhaka

Wasimul Bari
University of Dhaka

Abstract

To examine the impact of height on the occurrence of Type II diabetes, a longitudinal binary data set has been analyzed. The relevant covariates were selected by using quasi-likelihood based information criteria (QIC) and correlation information criteria (CIC) was used to select the correlation structure appropriate for the repeated binary responses. The consistent and efficient estimates of regression parameters were obtained from the generalized estimating equations (GEE). With the selected covariates height, education level, gender and unstructured correlation structure, it is found that there exists a statistically significant inverse relationship between height of an individual and the development of Type II diabetes. Risk Ratios for different covariates along with standard errors and confidence intervals are also given.

Keywords: correlation information criteria, generalized estimating equations, longitudinal binary data, quasi-likelihood based information criteria, risk ratio.

1. Introduction

The prevalence of diabetes particularly type 2 diabetes is increasing day by day and it becomes an emerging epidemic in the world. Among the regions, Southeast Asia region is affected markedly by this and according to WHO report approximately 79.5 million diabetic patients will live in this area,which is more than 26% of the world's total diabetic population (e.g. IDF, 1998). The prevalence rates in India, Pakistan and China are 12.1%, 11.1%, and 6.1% respectively; where in Bangladesh this rate is 8.1% in urban and 2.3% in rural. That is, Bangladesh as a developing country is facing a high prevalence of diabetes.

It has been well established that the increased prevalence of metabolic syndrome components and resultant increased risk of type 2 diabetes mellitus are associated with obesity (see, e.g. Janghorbani *et al*., 2010, WHO, 2000). Epidemiological studies have demonstrated that different anthropometric measures of obesity such as body mass index (BMI), waist circumference (WC), waist-height ratio (WHtR), waist-hip ratio (WHR) are strong and consistent predictors of type 2 diabetes (see, e.g. Janghorbani *et al*., 2010, Schulze *et al*, 2006). The relationship between increased BMI, WHtR and type 2 diabetes mellitus risk may be due to a direct or to inverse effect of height. It implies that height may play an important role for the incidence of diabetes. The association between height

of respondent and risk of type 2 diabetes mellitus has been investigated by several epidemiological studies but it is still unclear whether height affects the association. Also, the role of height as risk factor for type 2 diabetes mellitus remains uncertain. In most but not all studies, height appears to be inversely related with diabetes. There have been contravening reports about possible association of height and diabetes (see, e.g. Sicree et al, 2008, Snijder *et al*., 2003, Bozorgmanesh *et al*., 2011, Wang *et al*., 1997, Njolstand et al., 1998): a positive association was found in a studies (e.g. Wang *et al*.,1997), whereas no association (e.g. Lorenzo *et al*., 2009) or an inverse relation was reported in others (see, e.g. Snijder *et al*.,2003, Njolstand et al., 1998). Also, there was an association only in women (e.g. Bozorgmanesh *et al*., 2011) or men (e.g. Schulze *et al*., 2006). Hence, it would be interesting to find out a relationship between height and risk of type 2 diabetes mellitus. In this paper, we try to investigate this relationship in the context of Bangladesh using BIRDEM data.

The studies mentioned in the literature to explore the relationship between height and diabetes is based on the cross-sectional or follow-up designs. There exists no literature that deals with this relationship in the context of repeated observations obtained from an individual over a short period of time under a longitudinal study setup. Now-a-days, analysis of repeated observations has been extensively used in the biomedical studies. For example, the disease status of the patients may vary from time to time and covariates related with the disease behave differently with the changes in disease status. To analyze these types of data, observation at a single point provides misleading inferences about the disease status or the disease risk factor relationship. To overcome this problem longitudinal analysis plays an important role to draw valid inference. Note that repeated responses are likely to be correlated as these are collected from an individual. Therefore, it is necessary to take this correlation into account to estimate regression parameters consistently and efficiently. Using quasi-likelihood function, Liang and Zeger (1986) proposed the 'working' correlation based generalized estimating equations (GEE) for the purpose of estimation of regression parameters as well as the correlation parameters. In this paper, an attempt has been made to examine how height of an individual affects his/her diabetes status controlling relevant socioeconomic and demographic factors using longitudinal binary data. For the purpose of analysis, data have been obtained from Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic Disorders (BIRDEM).

One of the important features of any regression analysis is the model selection. In a longitudinal study, repeated responses along with a large number of covariates are collected from each individual of the study. Including all covariates in the regression analysis may result in a complex model and may lead to less precise estimates of parameters of interest. To overcome this problem, a subset of important covariates needs to be considered for the regression analysis so that model predictability and parsimony increase. There exist a number of subset selection criteria and procedures for linear regression models. Among them, likelihood function based Akaike's Information Criterion (AIC) (see, e.g. Akaike, 1973) is widely used. Since the construction of likelihood function is very much complicated in the longitudinal setup, Pan (2001a) proposed a modification of AIC based on the GEE, which is known as quasi-likelihood under the independent model information criterion (QIC). The other non-likelihood function bases criteria for model selection are: bootstrap smoothed cross-validation (BCV) [see, e.g. Pan (2001b)] that minimizes the expected predictive bias (EPB); bias-corrected bootstrap approaches to estimate the predictive mean squared error (PMSE) of a model and use the PMSE for model selection [see, e.g. Pan and Lee (2001)]; a generalized version of Mallows's C_p (GC_p) suitable for both parametric and non-parametric models [see, e.g. Cantoni et al. (2005)]; a cross-validation Markov Chain Monte Carlo (MCMC) procedure [see, e.g. Cantoni et al. (2008)].

Another issue that needs to address in the longitudinal setup is to select an appropriate correlation structure for the repeated responses. The QIC (e.g. Pan, 2001a) can also be used to select the appropriate 'working' correlation structure. Hin and Wang (2009) argued that the QIC measures are more sensitive to changes in the mean structure than changes in the covariance structure. As a remedy, Hin and Wang (2009) proposed correlation information criterion (CIC) for selecting the appropriate correlation structure.

Since the main focus of this paper is to measure the impact of height on the occurrence of diabetes, other covariates along with height are selected by using QIC (Pan, 2001a) and the correlation structure

for the repeated responses is selected by the CIC (e.g. Hin and Wang, 2009). Finally, longitudinal model is fitted by GEE (e.g. Liang and Zeger, 1986). In Section 2, a longitudinal binary model, GEE, QIC, CIC, and risk ratio estimation are described mathematically. A longitudinal binary model with selected covariates and correlation structure is illustrated to the data obtained from BIRDEM to determine the potential determinants of diabetes in Section 3. This paper concludes in Section 4 with a short discussion.

2. Methods

2.1. Longitudinal binary model

Suppose that y_{it} is the binary response obtained from individual i, $i = 1, \cdots\cdots, N$ at time point $t = 1, \cdots, \cdots, T$. Also, suppose that $\mathbf{x_{it}} = (x_{it1}, \cdots, x_{itj}, \cdots x_{itp})^{'}$ is the $p \times 1$ vector of covariates associated with the response y_{it}. Furthermore, suppose that the marginal probability distribution of y_{it} is a number of exponential family of distributions, i.e.,

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it})\}\varphi + b(y_{it}\varphi)], \tag{2.1}$$

(Liang and Zeger, 1986), where $a(\cdot)$ and $b(\cdot)$ are of known functional form. It can be shown that $\theta_{it} = \mathbf{x_{it}^{'}}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_j, \cdots, \beta_p)^{'}$ is the $p \times 1$ vector of regression coefficients. In equation (2.1), φ is the scale parameter and for binary response $\varphi = 1$. The marginal mean and variance of Y_{it} can be expressed as $\mu_{it} = E(Y_{it}) = a^{'}(\theta_{it})$ and $\sigma_{itt} = var(Y_{it}) = a^{''}(\theta_{it})$. For binary response, $\mu_{it} = [1 + exp(-\mathbf{x_{it}^{'}}\boldsymbol{\beta})]^{-1}$ and $\sigma_{itt} = \mu_{it}(1 - \mu_{it})$ The response vector for individual i is given by $\mathbf{Y_i} = (Y_{i1}, \cdots, Y_{it}, \cdots, Y_{iT})^{'}$ with mean $\boldsymbol{\mu_i} = (\mu_{i1}, \cdots, \mu_{it}, \cdots, \mu_{iT})^{'}$. Under a longitudinal set up, the repeated responses $Y_{i1}, \cdots Y_{it}, \cdots, Y_{iT}$ are likely to be correlated. Here, variance of $\mathbf{Y_i}$ can be expressed as

$$\Sigma_i = var\ (\mathbf{Y_i}) = A_i^{\frac{1}{2}} C(\rho)\ A_i^{\frac{1}{2}},$$

where $C(\rho)$ is the correlation matrix for response vector $\mathbf{Y_i}$ and $A_i = diag[\sigma_{i11}, \cdots, \sigma_{itt}, \cdots, \sigma_{iTT}]$. Note that the correlation matrix $C(\rho)$ is usually unknown. Here, the main parameter of interest is regression parameter $\boldsymbol{\beta}$ and the correlation parameter ρ is known as nuisance parameter. To obtain consistent as well as efficient estimates for $\boldsymbol{\beta}$, one needs to take the correlation parameter ρ into account. Since the probability distribution of $\mathbf{Y_i}$ is cumbersome, it would be difficult to obtain the maximum likelihood estimates of regression parameter $\boldsymbol{\beta}$ and correlation parameter ρ. As a remedy, Liang and Zegar (1986) proposed quasi-likelihood function based estimating equation for $\boldsymbol{\beta}$, which is well known as GEE. Note that GEE is constructed assuming 'working' correlation for response $\mathbf{Y_i}$. Liang and Zegar (1986) also proposed method of moments estimates for correlation parameters under different working correlation structures.

2.2. GEE for regression parameter

For known correlation parameter ρ, the GEE for regression parameter $\boldsymbol{\beta}$ is given by

$$\sum_{i=1}^{N} U_i(\boldsymbol{\beta}, \mathbf{y_i}, C) = 0, \tag{2.2}$$

[Liang and Zeger, (1986)] with $U_i(\boldsymbol{\beta}, \mathbf{y_i}, C) = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu_i}^{'} \Sigma_i^{-1}(\mathbf{y_i} - \boldsymbol{\mu_i})$, where $\boldsymbol{\mu_i}$ and Σ_i are defined in Section 2.1. The estimating equations given in (2.2) can be solved for $\boldsymbol{\beta}$ by using Newton-Raphson iterative procedure. The estimate, denoted by $\widehat{\boldsymbol{\beta}_C}$ under working correlation $C(\rho)$, obtained at the m^{th} $(m = 1, 2, 3, \cdots, \cdots, \cdots)$ iteration is given by

$$\widehat{\boldsymbol{\beta}_C}^{(m)} = \widehat{\boldsymbol{\beta}}^{(m-1)} + [A]^{-1}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(m-1)}} \left[\sum_{i=1}^{N} U_i(\boldsymbol{\beta}, \mathbf{y_i}, C)\right]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(m-1)}}$$

where $A = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu_i}^{'} \Sigma_i^{-1} \frac{\partial}{\partial \boldsymbol{\beta}^{'}} \boldsymbol{\mu_i}$. Note that $\widehat{\boldsymbol{\beta}_C}$ is asymptotically distributed as normal with mean $\boldsymbol{\beta}$ and the variance $V_{\widehat{\boldsymbol{\beta}_C}}$. The sandwich or robust estimate of $V_{\widehat{\boldsymbol{\beta}_C}}$ is given by

$$\widehat{V_{\widehat{\boldsymbol{\beta}_C}}} = A^{-1} \left\{ \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu_i}^{'} \Sigma_i^{-1} \left(\mathbf{y_i} - \boldsymbol{\mu_i}\right) \left(\mathbf{y_i} - \boldsymbol{\mu_i}\right)^{'} \Sigma_i^{-1} \frac{\partial}{\partial \boldsymbol{\beta}^{'}} \boldsymbol{\mu_i} \right\} A^{-1}, \tag{2.3}$$

with replacing $\boldsymbol{\beta}$ and ρ with their respective estimates. The estimation of ρ depends on the 'working' correlation structure. One may assume independence, exchangeable, first-order autoregressive (AR-1), or unstructured correlation structure for the repeated responses. For independence structure, $Corr\left(Y_{it}, Y_{it'}\right) = 0$; exchangeble, $Corr\left(Y_{it}, Y_{it'}\right) = \rho$; AR-1, $Corr\left(Y_{it}, Y_{it'}\right) = \rho^{\left|t-t^{'}\right|}$; and unstructured, $Corr(Y_{it}, Y_{it'})$
$= \rho_{itt'}$ with $t \neq t^{'}$. The estimator of ρ for different correlation structure is given by Liang and Zeger, (1986, section 3.3). The main purpose of this paper is to determine the effects of height on the occurrence of type-II diabetes along with other relevant covariates. For selecting relevant covariates from the available covariates, one may use QIC (e.g. Pan, 2001a). A short discussion on QIC is given below.

2.3. Quasi-likelihood based information criterion (QIC)

Pan (2001a) proposes QIC by modifying Akaike's information criterion (AIC) [e.g. Akaike, (1973)]. When the formulation of likelihood function is tractable, one may use AIC for the purpose of model selection. Akaike (1973) defined AIC as $AIC = -2 \ln L\left(\widehat{\boldsymbol{\beta}}\right) + 2p$, where $L\left(\widehat{\boldsymbol{\beta}}\right)$ is the likelihood function evaluated at $\widehat{\boldsymbol{\beta}}$ and p is number of regression parameters. In longitudinal setup, it may not be possible to construct the likelihood function. In this case, following AIC, Pan (2001a) proposed QIC, which is based on quasi-likelihood function under independent correlation structure. Mathematically, QIC may be defined as

$$QIC\left(C\right) = -2 \sum_{i=1}^{N} Q_i\left(\widehat{\boldsymbol{\beta}_C}, \mathbf{y_i}, I\right) + 2\, trace\left(\widehat{\Omega_I}\, \widehat{V_{\widehat{\boldsymbol{\beta}_C}}}\right), \tag{2.4}$$

where $Q_i\left(\widehat{\boldsymbol{\beta}_C}, \mathbf{y_i}, I\right)$ is the quasi-likelihood function under independence correlation structure, I, evaluated at estimated regression coefficient obtained under a 'working ' correlation structure C. Note that for binary repeated responses, one can express

$$\sum_{i=1}^{N} Q_i\left(\widehat{\boldsymbol{\beta}_C}, \mathbf{y_i}, I\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[y_{it} ln \frac{\mu_{it,}}{1-\mu_{it}} + \ln(1-\mu_{it}) \right]$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{T} \left[y_{it} \mathbf{x_{it}}^{'} \widehat{\boldsymbol{\beta}_C} + \ln(1 - e^{\mathbf{x_{it}}^{'} \widehat{\boldsymbol{\beta}_C}}) \right]$$

In (2.4), the expression for $\widehat{V_{\widehat{\boldsymbol{\beta}_C}}}$ is given in (2.3) and $\widehat{\Omega_I}$ is defined as

$$\widehat{\Omega_I} = -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{'}} \sum_{i=1}^{N} Q_i\left(\widehat{\boldsymbol{\beta}_C}, \mathbf{y_i}, I\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x_{it}}\, \mu_{it}\left(1-\mu_{it}\right) \mathbf{x_{it}}^{'}.$$

Like AIC, a model with minimum QIC is chosen to be the best model. Pan (2001a) also proposed to use QIC for selecting a correlation structure appropriate for the repeated responses. But, Hin and Wang (2009) argued that QIC cannot be used for correlation structure selection because of the following reasons. The first term of QIC depends neither on 'working' correlation nor on the correlation structure. In addition, the quasi-likelihood function is constructed assuming an independence correlation structure. Therefore, the first term has not contributed in selecting correlation structure. Though

second term of QIC reflects the 'working' correlation through sandwich estimator $\widehat{V_{\widehat{\beta}_C}}$, QIC is heavily influenced by the first term. Hence, QIC is not an appropriate tool to select the correlation structure. Hin and Wang (2009) proposed to use only the second term for selecting the correlation structure and this measure is known as correlation information criterion (CIC). That is,

$$CIC = tr(\widehat{\Omega_I}\,\widehat{V_{\widehat{\beta}_C}}) \quad . \tag{2.5}$$

The correlation structure for which the binary longitudinal model provides the minimum CIC will be chosen to analyze the data.

2.4. Risk ratio estimation

In this paper, the adjusted risk ratio (RR) is used to compare the rate of incidence of diabetes among the categories of a covariate. Mathematically, the RR for a covariate x_j having values x_j^1 and x_j^2 can be defined as

$$RR = \frac{R\left(\bar{x}_1, \cdots, x_j = x_j^1, \cdots, \bar{x}_p\right)}{R\left(\bar{x}_1, \cdots, x_j = x_j^2, \cdots, \bar{x}_p\right)}, \tag{2.6}$$

$$\text{with} \quad R\left(x_1, \cdots, x_j, \cdots, x_p\right) = \left[1 + exp\left(-\sum_{j=1}^{p} x_j\beta_j\right)\right]^{-1}$$

where all other covariates will be considered at respective mean values (e.g. Kleinbaum and Klein, 2005). The estimated RR can be computed from (2.6) by replacing β_j's with their corresponding estimates obtained from GEE. The $100\ (1-\alpha)\,\%$ confidence interval for RR is

$$\widehat{RR} \pm\ Z_{\alpha/2}\sqrt{var\left(\widehat{RR}\right)},$$

where $var(\widehat{RR}) = RR^2[(x_j^1)^2(1 - R(\bar{x}_1, \cdots, x_j = x_j^1, \cdots, \bar{x}_p))^2 - (x_j^2)^2(1 - R(\bar{x}_1, \cdots, x_j = x_j^2, \cdots, \bar{x}_p))^2\,]var(\widehat{\beta}_j)$.

3. Analysis of impact of height on type II diabetes

The main objective of this paper is to examine the impact of height of an individual on the occurrence of Type II diabetes controlling selected important factors by using repeated observations obtained from each individual considered in the study. For this purpose, the longitudinal data collected by BIRDEM has been used.

3.1. Data and variables

The data set consists of 2297 individuals each having 4 observations. An individual is defined whether diabetic or not by observing the glucose level after two hours of 75 gms glucose load at each visit. If the observed glucose level is less than 11.1 mmol/liter, then the patient is categorized as non-diabetic and otherwise diabetic (WHO, 2007; WHO/IDF, 2006). The main covariate of interest in this paper is height of an individual. It is found that mean height is 158.88 cm with standard deviation 8.8, and maximum and minimum heights are 193 cm and 109 cm, respectively. Besides height, age, heredity [HRD: Yes, No (ref)], education level [EDU: Yes, No (ref)], gender [Male, Female (ref)], physical exercise [PHEX: Yes, No (ref)], place of residence [AREA: Urban, Rural (ref)], and other complications [COM: Yes, No (ref)] are taken into consideration as these covariates are found to have significant impact on the occurrence of Type II diabetes in other studies (Njostad *et al*. 1998, Lorenzo *et al*. 2009, Schulze *et al*. 2006).

Among the individuals, the mean age is 53.64 years with standard deviation 11.85 and the maximum and minimum age are 106.4 and 13.3, respectively. Parents of 40.6% of individuals have diabetes. Most of the individuals (89.5%) have at least primary education. It is observed that 65.3% of individuals are male and 34.7% are female. Most of the individuals are not involved with physical exercise (96.6%). This data set is based on urban as 97.5% of individuals are from urban. Regarding complications other than diabetes, 96.5% of individuals have no other complications.

3.2. Selection of the best set of covariates

For the purpose of selection of covariates for the occurrence of Type II diabetes, we consider longitudinal binary models under different correlation structures. Since four responses from each individual are collected, they are likely to be correlated. Therefore, in this analysis, we do not consider independence as a 'working' correlation structure. The correlation structures considered are exchangeable, AR-1, and unstructured. Since height is the main covariate of interest, we consider this covariate in all possible models. The values of QIC are calculated using equation (2.4) for all possible models under different correlation structures. This result is shown in Table 1. It is clear from Table 1 that Model 20 produces minimum value under all three correlation structures. Hence, the selected covariates for the analysis are height, education level and gender.

3.3. Selection of the best correlation structure

To obtain estimates for the regression coefficients of the selected covariates, one may need a 'working' correlation that is appropriate for the repeated responses. To choose the appropriate correlation structure, one can compute the values of CIC for different correlation structures using equation (2.5) and then select the correlation that produces the minimum value of CIC. The values of CIC under exchangeable, AR-1, and unstructured correlations with the previously selected covariates are given in Table 2. It is clear from the table that unstructured correlation is appropriate for the longitudinal binary data obtained from BIRDEM.

3.4. Estimation of regression parameters using GEE

For the consistent and efficient estimates of the regression coefficients, one may solve the estimating equation given in equation (2.2) with the selected covariates and unstructured correlation structure by using Newton-Raphson iterative process. Estimates along with standard error, p-value, and 95% confidence interval are given in Table 3. From this table, it reveals that height is negatively associated with the occurrence of Type II diabetes and this effect is found to be statistically significant as p-value is 0.004. Education and gender have also negative significant effects on the diabetes with p-values 0.00 and 0.074, respectively.

The correlation parameters in a longitudinal setup are considered as the nuisance parameters and these parameters can be estimated by the method of moments (e.g. Liang and Zeger, 1986). The moment estimates of correlation parameters are given below. Note that the values in the parentheses are the standard errors of the estimators.

$$\widehat{C(\rho)} = \begin{bmatrix} 1 & 0.853\ (0.017) & 0.800\ (0.018) & 0.759\ (0.018) \\ & 1 & 0.889\ (0.016) & 0.835\ (0.017) \\ & & 1 & 0.909\ (0.016) \\ & & & 1 \end{bmatrix}$$

Since all the estimates of correlation parameters are more than 0.75, there exists a high correlation among the binary responses. Therefore, it is essential to take the correlation structure into account for the estimation of regression parameters.

Note that maximum and minimum heights are found to be 193 cm and 109 cm, whereas the mean height is 158.88 cm with standard deviation 8.8 cm. It indicates that the data contain outliers with respect to height. To examine the impact of height controlling other covariates, the GEE estimates

are also obtained after deleting outliers. The outliers were detected by the 'robust three sigma' rule (Maronna *et al.*, 2006). After deleting outliers, the GEE estimates under selected model using the unstructured correlation matrix for constant, height, education, and gender are 3.914, -0.0179, -0.574, and -0.213 with standard errors 1.02, 0.007, 0.144, and 0.121, respectively. The corresponding p-values are 0.00, 0.008, 0.00, and 0.057. It is observed that there is a little difference in the values of estimates before and after deleting outliers.

3.5. Estimation of risk ratio

To examine the association of a covariate with the occurrence of Type II diabetes controlling other covariates in the model, one may compute the adjusted risk ratio (RR) using equation (2.6). The adjusted RR and its standard error with 95% confidence interval for the selected covariates are given in Table 4. Since height is considered as continuous, we compute the quartile values first [e.g. first quartile (Q_1), second quartile (Q_2), and third quartile (Q_3)] and then RRs for Q_3 versus Q_1, Q_3 versus Q_2, and Q_2 versus Q_1. The values of first, second, and third quartiles are 152, 160, and 165 cm, respectively. From Table 4, it is found that the RR for Q_3 versus Q_1 is 0.78. It implies that an individual with height 165 cm is 22% less to have Type II diabetes compared to an individual with height 152 cm. The RR for Q_3 versus Q_2 is found to be 0.91, which implies that an individual with median height is 10% [((1/0.91)-1)$\times$100%] more like to develop Type II diabetes than an individual with height 165 cm. Finally, while comparing the second and first quartiles, an individual with second quartile height is 14% less likely to be a Type II diabetic patient compared to an individual with first quartile height. Note that all the RRs for height are statistically highly significant as p-values are 0.00. Therefore, an individual with shorter height is substantially at a higher risk of developing of Type II diabetes.

Education and gender are also found to have statistically significant impact on the occurrence of Type II diabetes with p-values 0.00 for both cases. Educated individuals are at 42% less risk for developing diabetes than their counterparts. On the other hand, male is 19% less likely to have diabetes than female.

4. Discussion

Generally, risk factors of Type II diabetes are modifiable and preventable. Therefore, early identification and preventive behavior for these risk factors can reduce the risk of developing Type II diabetes by 90% (see e.g. CDC, 2009). In this paper, an attempt has been made to identify the potential risk factors for Type II diabetes and to establish a relationship between height of an individual and the occurrence of diabetes by analyzing the longitudinal binary model obtained from BIRDEM. No study has been conducted in Bangladesh to identify the risk factors of diabetes by considering the longitudinal data. For the purpose of analysis, along with height, we first chose the important factors from the available covariates by using QIC (Pan, 2001a), which is appropriate for variable selection when the response is multivariate discrete variable and formulation of full likelihood function is mathematically involved. To obtain the efficient estimates for the regression parameters, one needs to consider a correlation structure appropriate for the repeated responses. After selecting the relevant covariates, we select the correlation structure using CIC (see e.g. Hin and Wang, 2009). Finally, estimates of regression parameters are obtained by solving the GEE (e.g. Liang and Zeger, 1986).

The selected covariates in this analysis are height, education level and gender and the appropriate correlation structure selected for the repeated responses is unstructured. It is found that education plays an important role for preventing the occurrence of Type II diabetes and male is at more risk of developing diabetes compared to female. One of main objectives of this paper is to examine the relationship between height and diabetes. This analysis reveals the fact that the probability of occurring diabetes deceases as the height of individual increases. That is, shorter height is associated with a higher occurrence of diabetes. One of the explanations of this inverse relation is that taller individuals have more muscle mass and muscle is the major tissue involved in uptake of glucose, against the fixed

glucose load of 75 grams (see e.g. Sicree *et al*., 2008). The dilution effect of total body water may contribute in establishing the results (e.g. Sicree *et al*.,2008).

In this study, a severe metabolic disturbance is identified in a shorter individual than a taller one regarding the occurrence of Type II diabetes. Therefore, developing diabetes may be reduced by controlling the factors that may influence the height. The height may be controlled by genetic and non-genetic (early-life and childhood) factors (e.g. Hirschhorn *et al*., 2001; Park *et al.*, 2004; Li *et al.*, 2006). Naturally, the next generation is likely to have shorter height, if most of the family members of the family are of short height. Note that genetic factors are totally beyond the control of human. The non-genetic factors that may affect the height are maternal smoking during pregnancy, birth weight, ill health during childhood and adolescence, and mental condition during childhood and adolescence. Non-genetic factors can be controlled to some extent by leading a healthy life style from childhood.

Acknowledgements We would like to thank Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic Disorders (BIRDEM), Bangladesh to make data available for analysis. We also thank anonymous reviewer and editor for their valuable comments and suggestions that led to significant improvements in the presentation.

References

1. International Diabetes Federation (1998). Diabetes around the World.

2. Janghorbani M., and Amini M. (2010). Comparison of Body Mass Index with Abdominal Obesity Indicators and Waist-to-stature Ratio for Prediction of Type 2 Diabetes: the Isfahan Diabetes Prevention Study. Obesity Research & Clinical Practice **4**: e25-e32.

3. WHO (2000). Obesity: Preventing and Managing the Global Epidemic. Report of a WHO consultation. World Health Organization Technical Report 2000; **894: i-xii**, 1-253.

4. Schulze M. B, Heidemann C, Schienkiewitz A. Bergmann M. M, Hoffmann K, and Boeing H (2006). Comparison of Anthropometric Characteristics in Predicting the Incidence of Type 2 Diabetes in the EPIC-Potsdam Study. Diabetes Care **29:** 1921-1923

5. Sicree, R. A., Zimmet, P. Z., Dunstan, D. W., Cameron, A. J., Wel-born, T. A., and Shaw, J. E. (2008). Differences in Height Explain Gender Differences in the Response to the Oral Glucose Tolerance Test the Aus Diab Study. Diabetic Medicine **25(3):**296-302

6. Snijder M. B., Dekker J. M, Visser, M, Bouter L. M, Stehouwer C. D. A, Kostense P. J, Yudkin J. S, Heine R. J, Nijpels G, and Seidell J. C (2003). Association of Hip and Thing Circumferences Independent of Waist Circumference with the Incidence of Type 2 Diabetes: the Hoorn Study. The American Journal of Clinical Nutrition **77:** 1192-1197

7. Bozorgmanesh M., Hadaegh F, Zabetian A, and Azizi F. (2011). Impact of Hip Circumference and Height on Incident Diabetes: Result from 6-year Follow-up in the Tehran Lipid and Glucose Study. Diabetic Medicine **28:** 1330-1336

8. Wang S. L., Pan W. H, Hwu C. M, Ho L. T, Lo C. H, Lin S. L, and Jong Y. S. (1997). Incidence of NIDDM and the Effects of Gender, Obesity, and Hyperinsulinaemia in Taiwan. Diabetologia **40:** 1431-1438

9. Njolstad I., Amesen E, and Lund-Larsen P. G (1998). Sex-differences in Risk Factors for Clinical Diabetes Mellitus in a General Population**:** a 12-years Follow-up of the Finnmark Study. American journal of Epidemiology **147:** 49-58.

10. Lorenzo C., Williams K, Stern M. P, and Haffner S. M. (2009). Height, Ethnicity and the Incidence of Diabetes: the San Antonio Heart Study. Metabolism **58:** 1530-1535.

11. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika **73:** 13–22.

12. Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). Robust Statistics: Theory and Methods. Wiley Series in Probability and Statistics.

13. Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. Akademiai Kiado, Budapest: 267-281.

14. Pan, W. (2001a). Akaike's Information Criterion in Generalized Estimating Equations. Biometrics **57:** 120-125.

15. Pan, W. (2001b). Model Selection in Estimating Equations. Biometrics **57:** 529-534.

16. Pan, W., and Lee, C. T. (2001). Bootstrap Model Selection in Generalized Linear Models. Journal of Agricultural, Biological & Environmental Statistics **6:** 49-61.

17. Cantoni, E., Flemming, J. M., and Ronchetti, E. (2005). Variable Selection for Marginal Longitudinal Generalized Linear Models. Biometrics **61:** 507-514.

18. Cantoni, E., Flemming, J. M., and Ronchetti, E. (2008). Longitudinal Variable Selection by Cross-validation in the Case of Many Covariates. Statistics in Medicine **26:** 919–930.

19. Hin, L. Y., and Wang, Y. G. (2009). Working-correlation-structure Identification in Generalized Estimating Equations. Statistics in Medicine **28(4):** 642-658.

20. Kleinbaum D. G., and Klein M. (2005). Survival Analysis: A Self-Learning Text, 2^{nd} edition. ISBN: Springer-Verlag New York, Inc; 105-127.

21. WHO. (2007). World Health Organization. "Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications: Report of a WHO Consultation. Part 1. Diagnosis and classification of diabetes mellitus".

22. WHO/IDF. (2006). Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia: Report of a WHO/IDF Consultation. Geneva: World Health Organization. p. 21. ISBN 978-92-4-159493-6.

23. Centers for Disease Control and Prevention (CDC) and National Center for Chronic Disease Prevention and Health Promotion. (2009). The Power of Prevention: Chronic Disease: The Public Health Challenge of the 21st Century. Atlanta, GA:CDC, http://www.cdc.gov/chronicdisease/pdf/2009-power-of-prevention.pdf

24. Hirschhorn J. N., Lindgren C. M., Daly M. J. *et al.* (2001). Genomewide Linkage Analysis of Stature in Multiple Populations Reveals Several Regions with Evidence of Linkage to Adult Height. American Journal of Human Genetics **69:** 106-116.

25. Park H. S., Yim K. S., and Cho S. I. (2004). Gender Differences in Familial Aggregation of Obesity-related Phenotypes and Dietary Intake Pattern in Korean Families. Annals of Epidemiology **14:** 486-491.

26. Li J. K., Ng M. C., So W. Y. *et al.* (2006). Phenotype and Genetic Clustering of Diabetes and Metabolic Syndrome in Chinese Families with Type 2 Diabetes Mellitus. Diabetes/Metabolism Research and Reviews **22:** 46-52

Table 1: Values of QIC for all possible models, keeping variable Height fixed under different correlation structures

Model	covariates								QIC values		
no.	height	age	hrd	edu	gender	phex	area	com	exchange	AR-1	unstructured
1	x	-	-	-	-	-	-	-	12260.9	12261.5	12261.1
2	x	x	-	-	-	-	-	-	12263.5	12279.4	12270.8
3	x	-	x	-	-	-	-	-	12261.9	12262.2	12261.8
4	x	-	-	x	-	-	-	-	12216.8	12217.6	12217.1
5	x	-	-	-	x	-	-	-	12256.3	12257.0	12256.5
6	x	-	-	-	-	x	-	-	12265.2	12265.8	12265.3
7	x	-	-	-	-	-	x	-	12265.4	12266.2	12265.5
8	x	-	-	-	-	-	-	x	12265.4	12264.7	12265.0
9	x	x	x	-	-	-	-	-	12264.5	12280.1	12271.5
10	x	x	-	x	-	-	-	-	12218.9	12234.5	12226.3
11	x	x	-	-	x	-	-	-	12259.8	12275.5	12267.0
12	x	x	-	-	-	x	-	-	12268.0	12283.5	12274.8
13	x	x	-	-	-	-	x	-	12268.0	12284.0	12275.3
14	x	x	-	-	-	-	-	x	12267.8	12282.3	12274.4
15	x	-	x	x	-	-	-	-	12218.0	12218.3	12217.8
16	x	-	x	-	x	-	-	-	12257.3	12257.7	12257.0
17	x	-	x	-	-	x	-	-	12266.2	12266.5	12266.0
18	x	-	x	-	-	-	x	-	12266.0	12267.0	12266.3
19	x	-	x	-	-	-	-	x	12266.4	12265.5	12265.8
20	**x**	-	-	**x**	**x**	-	-	-	**12211.9**	**12212.6**	**12212.0**
21	x	-	-	x	-	x	-	-	12221.9	12222.6	12222.0
22	x	-	-	x	-	-	x	-	12221.0	12222.0	12220.9
23	x	-	-	x	-	-	-	x	12221.3	12220.7	12221.0
24	x	-	-	-	x	x	-	-	12260.8	12261.4	12260.8
25	x	-	-	-	x	-	x	-	12260.7	12261.5	12260.8
26	x	-	-	-	x	-	-	x	12261.0	12260.0	12260.5
27	x	-	-	-	-	x	x	-	12269.7	12270.5	12269.7
28	x	-	-	-	-	x	-	x	12269.7	12269.0	12269.0
29	-	-	-	-	-	-	x	x	12269.9	12269.4	12269.5
30	x	x	x	x	-	-	-	-	12219.8	12235.0	12227.0
31	x	x	x	-	x	-	-	-	12260.8	12276.2	12268.0
32	x	x	x	-	-	x	-	-	12268.6	12284.2	12276.0
33	x	x	x	-	-	-	x	-	12268.9	12284.8	12276.0
34	x	x	x	-	-	-	-	x	12268.8	12283.0	12275.1
35	x	x	-	x	x	-	-	-	12214.9	12230.3	12222.0
36	x	x	-	x	-	x	-	-	12223.8	12239.5	12231.2
37	x	x	-	x	-	-	x	-	12222.7	12238.6	12230.2
38	x	x	-	x	-	-	-	x	12223.0	12237.4	12229.9
39	x	x	-	-	x	x	-	-	12264.1	12279.9	12271.2
40	x	x	-	-	x	-	x	-	12264.2	12280.1	12271.4
41	x	x	-	-	x	-	-	x	12264.2	12278.5	12270.7
42	x	x	-	-	-	x	x	-	12272.1	12288.2	12279.4
43	x	x	-	-	-	x	-	x	12272.0	12286.4	12278.5
44	x	x	-	-	-	-	x	x	12272.3	12287.0	12278.9
45	x	-	x	x	x	-	-	-	12212.8	12213.4	12212.8
46	x	-	x	x	-	x	-	-	12222.9	12223.4	12222.8
47	x	-	x	x	-	-	x	-	12221.7	12222.4	12221.7
48	x	-	x	x	-	-	-	x	12222.3	12221.5	12221.7
49	x	-	x	-	x	x	-	-	12261.8	12262.2	12261.6
50	x	-	x	-	x	-	x	-	12260.7	12261.5	12260.8
51	x	-	x	-	x	-	-	x	12261.0	12260.0	12260.5
52	x	-	x	-	-	x	x	-	12269.7	12270.5	12269.7
53	x	-	x	-	-	x	-	x	12269.7	12269.0	12269.0
54	x	-	x	-	-	-	x	x	12269.9	12269.4	12269.5
55	x	-	-	x	x	x	-	-	12217.1	12217.8	12217.2
56	x	-	-	x	x	-	x	-	12215.6	12216.5	12215.8
57	x	-	-	x	x	-	-	x	12216.4	12215.8	12216.1
58	x	-	-	x	-	x	x	-	12225.8	12226.7	12226.0
59	x	-	-	x	-	x	-	x	12226.3	12225.8	12226.0

60	x	-	-	x	-	-	x	x	12225.2	12224.8	12224.9
61	x	-	-	-	x	x	x	-	12265.1	12266.0	12265.2
62	x	-	-	-	x	x	-	x	12265.3	12264.7	12264.8
63	x	-	-	-	x	-	x	x	12265.3	12264.8	12264.9
64	x	-	-	-	-	x	x	x	12274.1	12273.7	12273.7
65	x	x	x	x	x	-	-	-	12215.9	12231.1	12222.9
66	x	x	x	x	-	x	-	-	12224.8	12240.2	12231.9
67	x	x	x	x	-	-	x	-	12223.7	12239.3	12230.9
68	x	x	x	x	-	-	-	x	12224.1	12238.1	12230.6
69	x	x	x	-	x	x	-	-	12265.0	12280.6	12271.9
70	x	x	x	-	x	-	x	-	12265.2	12280.8	12272.1
71	x	x	x	-	x	-	-	x	12265.2	12279.2	12271.4
72	x	x	x	-	-	x	x	-	12273.1	12288.9	12280.1
73	x	x	x	-	-	x	-	x	12272.9	12287.2	12279.2
74	x	x	x	-	-	-	x	x	12273.3	12287.7	12279.6
75	x	x	-	x	x	x	-	-	12220.0	12235.4	12227.2
76	x	x	-	x	x	-	x	-	12218.7	12234.3	12225.9
77	x	x	-	x	x	-	-	x	12219.2	12233.2	12225.9
78	x	x	-	x	-	x	x	-	12227.7	12243.6	12235.1
79	x	x	-	x	-	x	-	x	12228.1	12242.3	12234.8
80	x	x	-	x	-	-	x	x	12227.1	12241.4	12233.8
81	x	x	-	-	x	x	x	-	12268.5	12284.5	12275.6
82	x	x	-	-	x	x	-	x	12268.5	12282.8	12274.9
83	x	x	-	-	x	-	x	x	12268.6	12283.1	12275.1
84	x	x	-	-	-	x	x	x	12276.5	12291.1	12283.0
85	x	-	x	x	x	x	-	-	12218.0	12219.0	12217.9
86	x	-	x	x	x	-	x	-	12217.0	12217.3	12216.6
87	x	-	x	x	x	-	-	x	12217.4	12216.6	12216.8
88	x	-	x	x	-	x	x	-	12226.8	12227.5	12226.7
89	x	-	x	x	-	x	-	x	12227.0	12226.5	12226.7
90	x	-	x	x	-	-	x	x	12226.2	12225.6	12225.6
91	x	-	x	-	x	x	x	-	12266.1	12266.7	12266.0
92	x	-	x	-	x	x	-	x	12266.3	12265.4	12265.6
93	x	-	x	-	x	-	x	x	12266.2	12265.5	12265.6
94	x	-	x	-	-	x	x	x	12275.0	12274.4	12274.4
95	x	-	-	x	x	x	x	-	12220.8	12222.0	12221.0
96	x	-	-	x	x	x	-	x	12221.6	12221.0	12221.2
97	x	-	-	x	x	-	x	x	12220.2	12219.8	12219.9
98	x	-	-	x	-	x	x	x	12230.2	12229.9	12229.9
99	x	-	-	-	x	x	x	x	12270.0	12269.0	12269.2
100	x	x	x	x	x	x	-	-	12221.0	12236.2	12227.9
101	x	x	x	x	x	-	x	-	12219.6	12235.0	12226.7
102	x	x	x	x	x	-	-	x	12220.0	12234.0	12226.6
103	x	x	x	x	-	x	x	-	12228.7	12244.3	12235.8
104	x	x	x	x	-	x	-	x	12229.1	12243.1	12235.5
105	x	x	x	x	-	-	x	x	12228.0	12242.2	12234.5
106	x	x	x	-	x	x	x	-	12269.5	12285.2	12276.4
107	x	x	x	-	x	x	-	x	12269.5	12284.0	12275.6
108	x	x	x	-	x	-	x	x	12269.6	12284.0	12275.8
109	x	x	x	-	-	x	x	x	12277.4	12292.0	12283.7
110	x	x	-	x	x	x	x	-	12223.8	12239.4	12231.0
111	x	x	-	x	x	x	-	x	12224.3	12238.3	12230.9
112	x	x	-	x	x	-	x	x	12223.0	12237.2	12229.6
113	x	x	-	x	-	x	x	x	12232.0	12246.4	12238.7
114	x	x	-	-	x	x	x	x	12272.9	12287.4	12279.3
115	x	-	x	x	x	x	x	-	12221.8	12222.6	12221.7
116	x	-	x	x	x	x	-	x	12222.6	12221.8	12221.9
117	x	-	x	x	x	-	x	x	12221.1	12220.6	12220.6
118	x	-	x	x	-	x	x	x	12231.2	12230.6	12230.6
119	x	-	x	-	x	x	x	x	12270.7	12270.0	12269.9
120	x	-	-	x	x	x	x	x	12225.4	12225.0	12225.0
121	x	x	x	x	x	x	x	-	12224.8	12240.2	12231.7
122	x	x	x	x	x	x	-	x	12225.3	12239.1	12231.6
123	x	x	x	x	x	-	x	x	12224.0	12238.0	12230.0
124	x	x	x	x	-	x	x	x	12233.0	12247.2	12239.4

125	x	x	x	-	x	x	x	x	12273.9	12288.2	12280.0
126	x	x	-	x	x	x	x	x	12228.1	12242.3	12234.7
127	x	-	x	x	x	x	x	x	12226.3	12225.8	12225.7
128	x	x	x	x	x	x	x	x	12229.0	12243.1	12235.4

Table 2: Correlation Information Criterion (CIC)values under different correlation structures

Correlation Structures	Exchangeable	AR-1	Unstructured
Value of CIC	14.1	14.1	14.0

Table 3: GEE estimates of regression coefficients under selected model using unstructured correlation with standard errors, p-values and 95 % confidence intervals

Variables	Estimates	Standard Error	p-value	95 % Confidence Interval
Constant	4.1501	0.988	0.000	(2.21, 6.09)
Height	-0.020	0.007	0.004	(-0.03, -0.01)
Education	-0.569	0.144	0.000	(-0.85, -0.29)
Gender	-0.214	0.120	0.074	(-0.45, 0.02)

Table 4: Risk Ratios for covariates under selected model with standard errors, p-values and 95 % confidence intervals

Variables	Risk Ratio	Standard Error	p-value	95 % Confidence Interval
Height				
Q3 vs. Q1	0.78	0.021	0.000	(0.74, 0.82)
Q3 vs. Q2	0.91	0.015	0.004	(0.88, 0.94)
Q2 vs. Q1	0.86	0.018	0.000	(0.82, 0.89)
Education	0.58	0.081	0.000	(0.42, 0.73)
Gender	0.81	0.095	0.000	(0.62, 0.99)

Affiliation:

Md Erfanul Hoque
Department of Statistics, Biostatistics & Informatics,
University of Dhaka
Dhaka-1000,
Bangladesh
E-mail: imerfan49@yahoo.com

Mahfuzur Rahman Khokan
Department of Statistics, Biostatistics & Informatics,
University of Dhaka
Dhaka-1000,
Bangladesh
E-mail: mahfuz_sbi34@yahoo.com

Wasimul Bari
Department of Statistics, Biostatistics & Informatics,
University of Dhaka
Dhaka-1000,
Bangladesh
E-mail: w_bari@yahoo.com

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 44
October 2015

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2014-03-03
Accepted: 2014-09-04

Austrian Journal of Statistics
October 2015, Volume 44, 17–29.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i3.21

Transmuted Modified Inverse Rayleigh Distribution

Muhammad Shuaib Khan
University of Newcastle

Robert King
University of Newcastle

Abstract

We introduce the transmuted modified Inverse Rayleigh distribution by using quadratic rank transmutation map (QRTM), which extends the modified Inverse Rayleigh distribution. A comprehensive account of the mathematical properties of the transmuted modified Inverse Rayleigh distribution are discussed. We derive the quantile, moments, moment generating function, entropy, mean deviation, Bonferroni and Lorenz curves, order statistics and maximum likelihood estimation The usefulness of the new model is illustrated using real lifetime data.

Keywords: modified inverse Rayleigh distribution, moments, order statistics, maximum likelihood estimation.

1. Introduction

The inverse Rayleigh (IR) distribution is the special case of the inverse Weibull (IW) distribution for modeling lifetime data. Trayer (1964) introduced the (IR)distribution. Gharraph (1993), Mukarjee and Maitim (1996) discussed some properties of the (IR) distribution. Voda (1972) also discussed some properties of the maximum likelihood estimator for the IR distribution. Mohsin and Shahbaz (2005) studied the comparison of the negative moment estimator with maximum likelihood estimator of the IR distribution. Recently Khan (2014), studied the modified inverse Rayleigh (MIR) distribution and discussed its theoretical properties. The cumulative distribution function (cdf) of the MIR distribution is given by

$$G(x;\alpha,\beta) = \exp\left\{-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\}, \quad x>0, \tag{1.1}$$

where $\alpha>0$ and $\beta>0$ are the scale parameters. The density function corresponding to (1.1) is

$$g(x;\alpha,\beta) = \left(\alpha+\frac{2\beta}{x}\right)\left(\frac{1}{x}\right)^2\exp\left\{-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\}, \quad x>0. \tag{1.2}$$

The behavior of instantaneous failure rate of the modified inverse Rayleigh distribution has increasing and decreasing reliability patterns for engineering system or component failure rate for lifetime data. The two parameter modified inverse Rayleigh distribution is the extended model of the inverse Rayleigh distribution and has nice physical interpretation. The inverse Rayleigh (IR) distribution is the special case of the modified inverse Rayleigh (MIR) distribution when $\alpha=0$ and the MIR distribution

coincides with the inverse exponential distribution for $\beta = 0$.
Khan and King (2012), proposed the modified inverse Weibull distribution and presented comprehensive description of the mathematical properties along with its reliability behavior. Khan et al. (2008), studied the flexibility of the inverse Weibull distribution. Aryal et al. (2009) studied the transmuted extreme value distribution with application to climate data. Aryal et al. (2011), proposed the transmuted Weibull distribution and studied various structural properties of this model for analyzing reliability data. More recently Khan and King (2013), proposed the transmuted modified Weibull distribution and studied its mathematical properties. Khan and King (2013) also proposed the transmuted generalized inverse Weibull distribution with application to reliability data. Khan, King and Hudson (2013), studied the transmuted generalized exponential distribution and studied its various structural properties with an application to survival data. More recently Merovci (2013), studied the transmuted Rayleigh distribution. In this research article, we propose the three parameter transmuted modified inverse Rayleigh distribution denoted as the TMIR which is a new generalization of the modified inverse Rayleigh distribution and discuss its statistical properties and applications. The new extended distribution contains five submodels such as the TIR (transmuted inverse Rayleigh), TIE (transmuted inverse exponential), modified inverse Rayleigh, inverse Rayleigh and inverse exponential distributions.
A random variable *X* is said to have transmuted distribution if its distribution function is given by

$$F(x) \quad = \quad (1+\lambda)\, G(x) - \lambda\, G(x)^2, \tag{1.3}$$

where $G(x)$ is the CDF of the base distribution. It is important to note that at $\lambda = 0$ we have the distribution of the base random variable, Shaw et al.(2009).

The article is organized as follows, In Section 2, we present the analytical shapes of the probability density, distribution function, reliability function and hazard function of the subject model. A range of mathematical properties are considered in Section 3, we demonstrate the quantile functions, moment estimation, moment generating function. In Section 4, we derived the entropies, mean deviation, Bonferroni and Lorenz curves. In Section 5, we derive density functions of the pdf of rth order statistics and the rth moment of order statistics $X_{(r)}$. In Section 6, Maximum likelihood estimates (MLE_s) of the unknown parameters and the asymptotic confidence intervals of the TMIR distribution are discussed. In Section 7. we fit the TMIR distribution to illustrate its usefulness. Concluding remarks are addressed in Section 8.

2. Transmuted modified inverse Rayleigh distribution

A positive random variable *x* has the three parameters TMIR distribution with scale parameters $\alpha, \beta > 0$ and the transmuted parameter $|\lambda| \leq 1$ is given by

$$f(x;\alpha,\beta,\lambda) \quad = \quad \left(\alpha + \frac{2\beta}{x}\right)\left(\frac{1}{x}\right)^2 \exp\left\{-\frac{\alpha}{x} - \beta\left(\frac{1}{x}\right)^2\right\} u_2(x), \tag{2.1}$$

$$u_g(x) = \left\{1 + \lambda - g\lambda \exp\left\{-\frac{\alpha}{x} - \beta\left(\frac{1}{x}\right)^2\right\}\right\}, \quad g = 1, 2, \tag{2.2}$$

The cumulative distribution function CDF corresponding to (2.1) is given by

$$F(x;\alpha,\beta,\lambda) \quad = \quad \exp\left\{-\frac{\alpha}{x} - \beta\left(\frac{1}{x}\right)^2\right\} u_1(x). \tag{2.3}$$

Here α and β are the scale parameters and λ is the transmuting parameter representing the different patterns of the TMIR distribution. The probability density function given in (2.1) with their associated reliability function, hazard function and cumulative hazard function are given in (2.4-2.6) respectively

$$R(x;\alpha,\beta,\lambda) \quad = \quad 1 - \exp\left\{-\frac{\alpha}{x} - \beta\left(\frac{1}{x}\right)^2\right\} u_1(x), \tag{2.4}$$

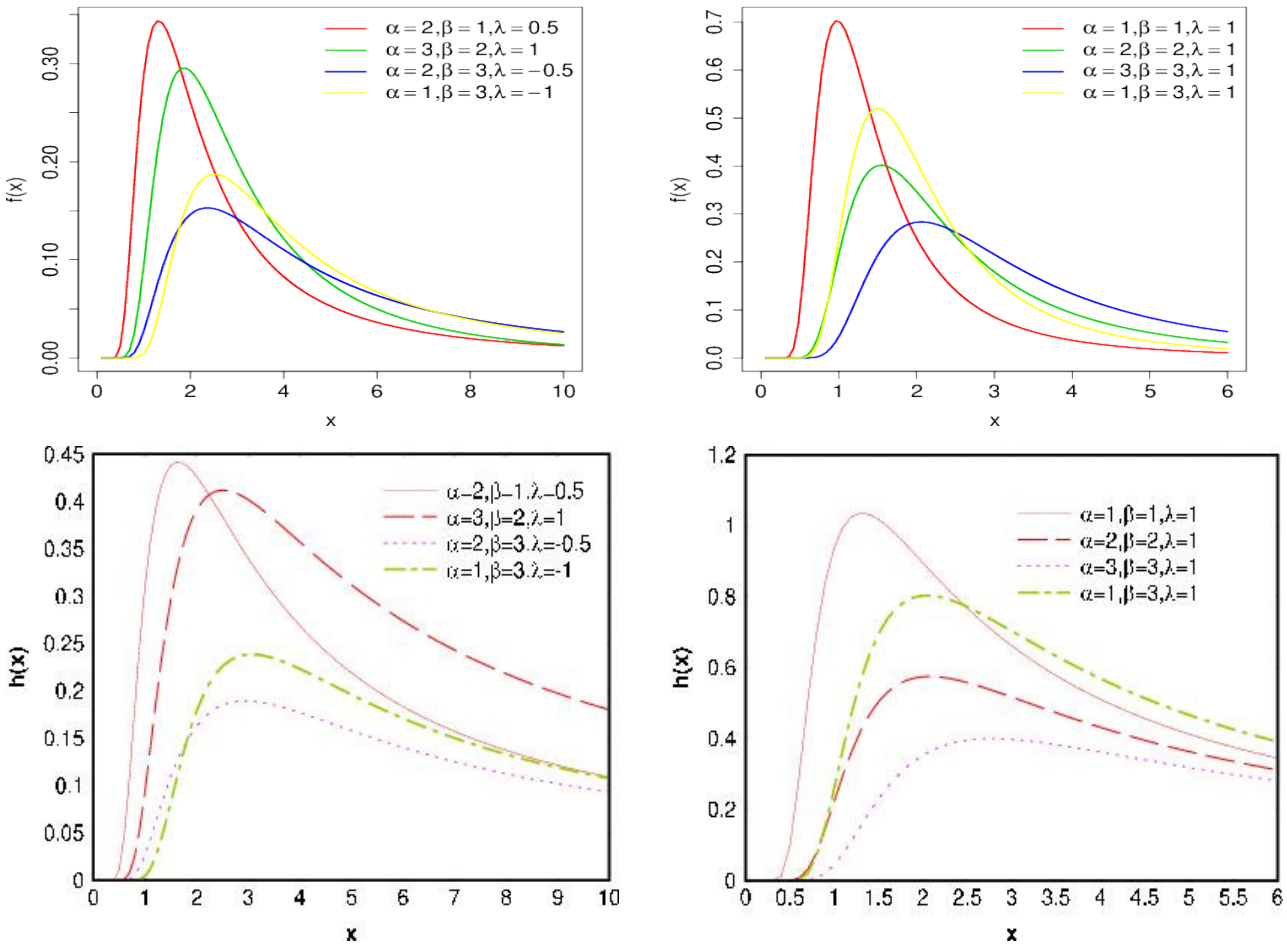

Figure 1: Plots of the TMIR pdf and hf for some parameter values.

$$h(x;\alpha,\beta,\lambda) = \frac{\left(\alpha+\frac{2\beta}{x}\right)\left(\frac{1}{x}\right)^2 \exp\left\{-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\} u_2(x)}{1-\exp\left\{-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\} u_1(x)}, \tag{2.5}$$

and

$$H(x;\alpha,\theta,\lambda) = -\ln\left[1-\exp\left\{-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\} u_1(x)\right]. \tag{2.6}$$

Fig. 1 shows the different patterns of the density function (pdf) and hazard function (hf) of the TMIR distribution. It illustrate that the behavior of instantaneous failure rate of the TMIR distribution has upside-down bathtub shape curves.

3. Moments and quantiles

In this section we obtain some statistical properties of the TMIR distribution.

3.1. Quantile and median

The quantile $F^{-1}(u)$ of the TMIR distribution is the real solution of the following equation

$$F^{-1}(u) = \frac{2\beta}{-\alpha+\sqrt{\alpha^2-4\beta\ln\left(\frac{(1+\lambda)-\sqrt{(1+\lambda)^2-4\lambda u}}{2\lambda}\right)}}, \tag{3.1}$$

where u has the uniform $U(0,1)$ distribution.

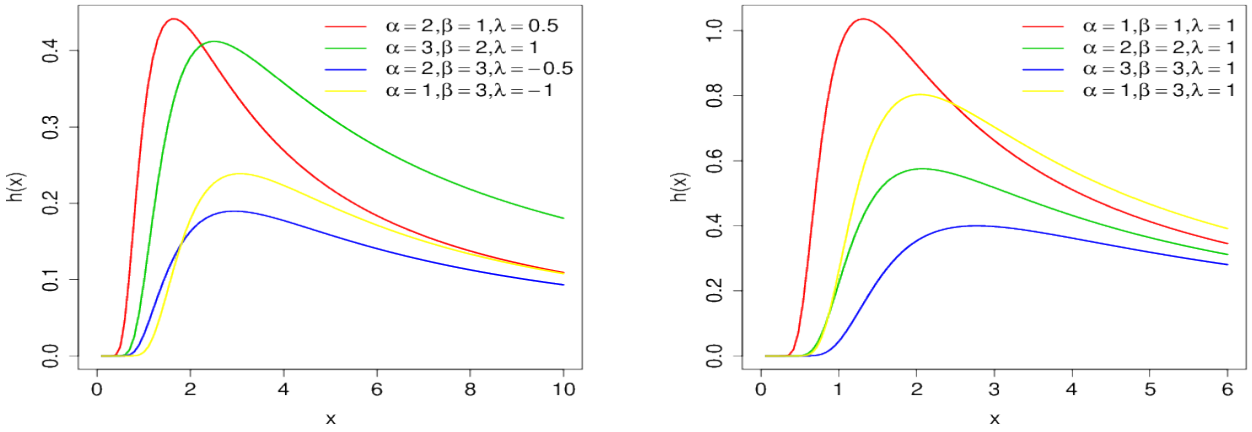

Figure 2: Median and coefficient of quantile deviation of the TMIRD.

The random number for the TMIR distribution is performed by generating uniform numbers and then applying the quantile function using equation (3.1). By substituting $u = 0.5$ in (3.1) we obtain the median of the TMIR distribution. Fig. 2 shows the median and quartile deviation life of the TMIR distribution when $\alpha = 2$ and $\beta = 3$. To illustrate the skewness and kurtosis we consider the measure based on quantiles. The skewness and kurtosis measures can now be calculated from quantiles using Bowley and Percentile coefficient of kurtosis. The Bowley Skewness and Percentile coeffficient of kurtosis when $\alpha = 2$, $\lambda = 0.5$ as a function of β are illustrated in Fig. 3 respectively. It is important to note that as the parameter β increases the behavior of the Bowley Skewness and Percentile coeffficient of kurtosis are decreases asymptotically.

3.2. Moments

Theorem 1. *If X has the* TMIR$(x;\alpha,\beta,\lambda)$ *with* $|\lambda| \leq 1$*, then the kth moment of X is given by*

$$\mu_k = (1+\lambda)\sum_{p=0}^{\infty}\frac{(-1)^p\beta^p\alpha^{k-2p}}{p!}z_2(k,p) - \lambda\sum_{p=0}^{\infty}\frac{(-1)^p(2\beta)^p(2\alpha)^{k-2p}}{p!}z_1(k,p),$$

$$z_g(k,p) = \left[\Gamma\left(2p-k+1\right) + \frac{g\beta}{\alpha^2}\Gamma\left(2p-k+2\right)\right], \quad g = 1,2.$$

Proof. By definition

$$\mu_k = \int_0^{\infty} x^{k-2}\left(\alpha+\frac{2\beta}{x}\right)\exp\left\{-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\}u_2(x)dx.$$

so that

$$\mu_k = (1+\lambda)\int_0^{\infty} x^{k-2}\left(\alpha+\frac{2\beta}{x}\right)\exp\left\{-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\}dx$$
$$-2\lambda\int_0^{\infty} x^{k-2}\left(\alpha+\frac{2\beta}{x}\right)\exp\left\{-\frac{2\alpha}{x}-2\beta\left(\frac{1}{x}\right)^2\right\}dx.$$

The above expression can be obtained by using

$$\exp\left\{-g\beta\left(\frac{1}{x}\right)^2\right\} = \sum_{p=0}^{\infty}\frac{(-1)^p(g\beta)^p x^{-2p}}{p!}, \quad g = 1,2.$$

the above integral yields the following kth moment,

$$\begin{aligned}\mu_k &= (1+\lambda)\sum_{p=0}^{\infty}\frac{(-1)^p\beta^p\alpha^{k-2p}}{p!}\left[\Gamma(2p-k+1)+\frac{2\beta}{\alpha^2}\Gamma(2p-k+2)\right] \\ &\quad -\lambda\sum_{p=0}^{\infty}\frac{(-1)^p(2\beta)^p(2\alpha)^{k-2p}}{p!}\left[\Gamma(2p-k+1)+\frac{\beta}{\alpha^2}\Gamma(2p-k+2)\right].\end{aligned} \tag{3.2}$$

□

Theorem 2. *If X is a random variable that has the* TMIR$(x;\alpha,\beta,\lambda)$ *with* $|\lambda|\leq 1$*, then the moment generating function* (mgf) *of X is given by*

$$\begin{aligned}M_x(t) &= (1+\lambda)\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\frac{(-1)^p\beta^pt^q}{\alpha^{2p-q}p!q!}J_2(p,q)-\lambda\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\frac{(-1)^p(2\beta)^pt^q}{(2\alpha)^{2p-q}p!q!}J_1(p,q), \\ J_h(p,q) &= \left[\Gamma(2p-q+1)+\frac{h\beta}{\alpha^2}\Gamma(2p-q+2)\right], \quad h=1,2.\end{aligned}$$

Proof. By definition

$$M_x(t) = \int_0^{\infty}\left(\alpha+\frac{2\beta}{x}\right)\left(\frac{1}{x}\right)^2\exp\left\{tx-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\}u_2(x)dx.$$

so that

$$\begin{aligned}M_x(t) &= (1+\lambda)\int_0^{\infty}\left(\alpha+\frac{2\beta}{x}\right)\left(\frac{1}{x}\right)^2\exp\left\{tx-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\}dx \\ &\quad -2\lambda\int_0^{\infty}\left(\alpha+\frac{2\beta}{x}\right)\left(\frac{1}{x}\right)^2\exp\left\{tx-\frac{2\alpha}{x}-2\beta\left(\frac{1}{x}\right)^2\right\}dx.\end{aligned}$$

Using the Taylor series expansions the above integral reduces to

$$\begin{aligned}M_x(t) &= (1+\lambda)\sum_{q=0}^{\infty}\frac{t^q}{q!}\int_0^{\infty}x^{q-2}\left(\alpha+\frac{2\beta}{x}\right)\exp\left\{-\frac{\alpha}{x}-\beta\left(\frac{1}{x}\right)^2\right\}dx \\ &\quad -2\lambda\sum_{q=0}^{\infty}\frac{t^q}{q!}\int_0^{\infty}x^{q-2}\left(\alpha+\frac{2\beta}{x}\right)\exp\left\{-\frac{2\alpha}{x}-2\beta\left(\frac{1}{x}\right)^2\right\}dx,\end{aligned}$$

the above integral yields the following moment generating function

$$\begin{aligned}M_x(t) &= (1+\lambda)\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\frac{(-1)^p\beta^pt^q}{\alpha^{2p-q}p!q!}\left[\Gamma(2p-q+1)+\frac{2\beta}{\alpha^2}\Gamma(2p-q+2)\right] \\ &\quad -\lambda\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\frac{(-1)^p(2\beta)^pt^q}{(2\alpha)^{2p-q}p!q!}\left[\Gamma(2p-q+1)+\frac{\beta}{\alpha^2}\Gamma(2p-q+2)\right].\end{aligned} \tag{3.3}$$

□

Based on Theorem 1, the coefficient of skewness and coefficient of kurtosis of the TMIR$(x;\alpha,\beta,\lambda)$ are obtained from the well known relations $\beta_1=\mu_3/\mu_2^{\frac{3}{2}}$ and $\beta_2=\mu_4/\mu_2^2$, respectively.

4. Entropy and mean deviation

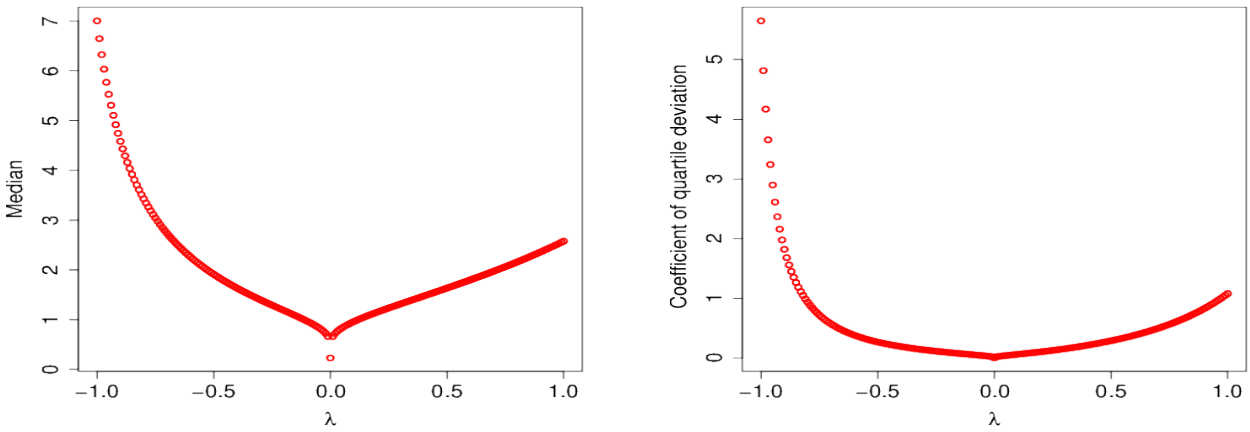

Figure 3: Bowley skewness and percentile kurtosis of the TMIRD.

The entropy of a random variable X with probability density $\text{TMIR}(x;\alpha,\beta,\lambda)$ is a measure of variation of the uncertainty. A large value of entropy indicates the greater uncertainty in the data. The Rényi entropy (1960), $I_R(\rho)$, for X is a measure of variation of uncertainty and is defined as

$$I_R(\rho) = \frac{1}{1-\rho}\log\left\{\int_0^\infty f(x)^\rho dx\right\}, \tag{4.1}$$

where $\rho > 0$ and $\rho \neq 1$. Suppose X has the $\text{TMIR}(x;\alpha,\beta,\lambda)$ then by substituting (2.1) and (2.2) in (4.1), we obtain

$$I_R(\rho) = \frac{1}{1-\rho}\log\left\{\int_0^\infty \left(\alpha+\frac{2\beta}{x}\right)^\rho \left(\frac{1}{x}\right)^{2\rho} \exp\left\{-\frac{\alpha\rho}{x}-\beta\rho\left(\frac{1}{x}\right)^2\right\} u_2(x)^\rho dx\right\},$$

the TMIR Rényi entropy reduces to

$$I_R(\rho) = \frac{1}{1-\rho}\log\left\{\sum_{i=0}^\infty\sum_{j=0}^\infty (-1)^j(1+\lambda)^\rho\alpha^\rho \binom{\rho}{i}\binom{\rho}{j}\left(\frac{2\beta}{\alpha}\right)^i\left(\frac{2\lambda}{1+\lambda}\right)^i \xi_{i,j}dx\right\},$$

where

$$\xi_{i,j} = \int_0^\infty \left(\frac{1}{x}\right)^{2\rho+j} \exp\left\{-(\rho+i)\left\{\frac{\alpha}{x}+\beta\left(\frac{1}{x}\right)^2\right\}\right\} dx.$$

The above integral can be calculated as

$$\xi_{i,j} = \sum_{k=0}^\infty \frac{(-1)^k\beta^k(\rho+i)^k}{k!}\left(\frac{\Gamma\left(j+2(k+\rho)-1\right)}{[\alpha(\rho+i)]^{j+2(k+\rho)+1}}\right).$$

and thus we obtain the TMIR Rényi entropy as

$$I_R(\rho) = \frac{\rho}{1-\rho}\log\alpha + \frac{\rho}{1-\rho}\log(1+\lambda) + \frac{1}{1-\rho} \log\left\{\sum_{i=0}^\infty\sum_{j,k=0}^\infty \frac{(-1)^{j+k}\beta^k(\rho+i)^k V_{i,j}}{k!}\left(\frac{\Gamma\left(j+2(k+\rho)-1\right)}{[\alpha(\rho+i)]^{j+2(k+\rho)+1}}\right)\right\}.$$

where

$$V_{i,j} = \binom{\rho}{i}\binom{\rho}{j}\left(\frac{2\beta}{\alpha}\right)^i\left(\frac{2\lambda}{1+\lambda}\right)^i$$

The β-or(q-entropy) was introduced by Havrda and Charvat (1967), and is defined as

$$I_H(q) = \frac{1}{q-1}\left\{1-\int_0^\infty f(x)^q dx\right\}, \tag{4.2}$$

where $q > 0$ and $q \neq 1$. Suppose X has the TMIR$(x;\alpha,\beta,\lambda)$ then by substituting (2.1) and (2.2) in (4.2), we obtain

$$I_H(q) = \frac{1}{q-1}\left\{1-\int_0^\infty \left(\alpha+\frac{2\beta}{x}\right)^q \left(\frac{1}{x}\right)^{2q} \exp\left\{-\frac{\alpha q}{x}-\beta q\left(\frac{1}{x}\right)^2\right\} u_2(x)^q dx\right\},$$

the above integral yields the TMIR q-entropy is

$$I_H(q) = \frac{1}{q-1}\left\{1-\sum_{i=0}^{\infty}\sum_{j,k=0}^{\infty}\frac{(-1)^{j+k}\beta^k(q+i)^k\xi_{i,j}}{k!}\left(\frac{\Gamma\left(j+2(k+q)-1\right)}{[\alpha(q+i)]^{j+2(k+q)+1}}\right)\right\},$$

where

$$\xi_{i,j} = \alpha^q(1+\lambda)^q\binom{q}{i}\binom{q}{j}\left(\frac{2\beta}{\alpha}\right)^i\left(\frac{2\lambda}{1+\lambda}\right)^i.$$

The degree of scatter in a population is widely measured by the totality of deviations from the mean and median. If X has the TMIR$(x;\alpha,\beta,\lambda)$, then we derive the mean deviation about the mean and about the median from the following equations of Gauss et al. (2013)

$$\delta_1 = \int_0^\infty |x-\mu| f(x)dx \qquad \text{and} \qquad \delta_2 = \int_0^\infty |x-M| f(x)dx.$$

The mean μ is given in equation (3.2) and the median M is obtained from equation (3.1). These measures are calculated using the relationships:

$$\delta_1 = 2[\mu F(\mu) - \psi(\mu)] \qquad \text{and} \qquad \delta_2 = \mu - 2\psi(M).$$

The quantity $\psi(q)$ used to determine the Bonferroni and Lorenz curves, which have applications in econometrics and finance, reliability and survival analysis, demography, insurance and biomedical sciences is given by

$$\begin{aligned}\psi(q) = {} & (1+\lambda)\sum_{h=0}^{\infty}\frac{(-1)^h\beta^h}{\alpha^{2h}h!}\left[\alpha\gamma\left(2h+1,\frac{\alpha}{q}\right)+\frac{2\beta}{\alpha}\gamma\left(2h+2,\frac{\alpha}{q}\right)\right] \\ & -\lambda\sum_{h=0}^{\infty}\frac{(-1)^h(2\beta)^h}{(2\alpha)^{2h}h!}\left[\alpha\gamma\left(2h+1,\frac{\alpha}{q}\right)+\frac{\beta}{\alpha}\gamma\left(2h+2,\frac{2\alpha}{q}\right)\right].\end{aligned} \tag{4.3}$$

By using (4.3), one obtains the Bonferroni and the Lorenz curve as

$$B(P) = \frac{\psi(q)}{P\mu}, \qquad \text{and} \qquad L(P) = \frac{\psi(q)}{\mu}.$$

5. Order statistics

The density of the rth order statistic $X_{(r)}$ of a random sample drawn from the TMIR$(x;\alpha,\beta,\lambda)$ distribution with $|\lambda| \leq 1$, follows from Arnold et al. (1), with the density function of $X_{(r)}$ is given by

$$f_{r:n}(x) = \frac{(F(x))^{r-1}(1-F(x))^{n-r}f(x)}{B(r,n-r+1)}, \quad x>0. \tag{5.1}$$

By setting $\delta = \exp\left\{-\frac{\alpha}{x} - \beta\left(\frac{1}{x}\right)^2\right\}$, substituting (2.1) and (2.3) into (5.1), we obtain

$$f_{r:n}(x) = n\binom{n-1}{r-1}\sum_{k=0}^{n-r}\binom{n-r}{k}(-1)^k\delta^{r+k}V_{r:k}(x),$$

where

$$V_{r:k}(x) = \left(\alpha + \frac{2\beta}{x}\right)\left(\frac{1}{x}\right)^2 u_1(x)^{r+k-1}u_2(x).$$

This leads to the combining terms of the order statistics of the TMIR distribution, given by

$$f_{r:n}(x) = n\binom{n-1}{r-1}\sum_{k=0}^{n-r}\sum_{m=0}^{\infty}\Im_{k,m}\left(\alpha + \frac{2\beta}{x}\right)\left(\frac{1}{x}\right)^2 z_{k,m}u_2(x), \tag{5.2}$$

where

$$\Im_{k,m} = n\binom{n-r}{k}\binom{r+k-1}{m}(-1)^{k+m}(1+\lambda)^{r+k-1}\left(\frac{\lambda}{1+\lambda}\right)^m,$$

and

$$z_{k,m} = \exp\left\{-(r+k+m)\left(\frac{\alpha}{x} + \beta\left(\frac{1}{x}\right)^2\right)\right\}.$$

Using (5.2), the sth moment of the rth order statistics $X_{(r)}$ is given by

$$\mu_s^{n:r} = n\binom{n-1}{r-1}\sum_{k=0}^{n-r}\sum_{m=0}^{\infty}\Im_{k,m}\left\{(1+\lambda)\sum_{i=0}^{\infty}\frac{(-1)^i\beta^i\tau_{i,s,0}}{C^{-i}i!} - 2\lambda\sum_{i=0}^{\infty}\frac{(-1)^i\beta^i\tau_{i,s,1}}{(C+1)^{-i}i!}\right\},$$

where $c = r+k+m$,

$$\tau_{i,s,g} = \alpha\left((c+g)\alpha\right)^{s-2i-1}\Gamma\left(2i-s+1\right) + 2\beta\left((c+g)\alpha\right)^{s-2i-2}\Gamma\left(2i-s+2\right).$$

6. Maximum likelihood estimation

Consider the random samples $x_1, x_2, ..., x_n$ consisting of n observations from the TMIR distribution and $\Theta = (\alpha, \beta, \lambda)^T$ be the parameter vector. The likelihood function of (2.1)is given by

$$L(\alpha,\beta,\lambda) = \prod_{i=1}^{n}\left(\alpha + \frac{2\beta}{x_i}\right)\left(\frac{1}{x_i}\right)^2\exp\left\{-\frac{\alpha}{x_i} - \beta\left(\frac{1}{x_i}\right)^2\right\} \times\left\{1+\lambda-2\lambda\exp\left\{-\frac{\alpha}{x_i} - \beta\left(\frac{1}{x_i}\right)^2\right\}\right\}. \tag{6.1}$$

By taking the logarithm of (6.1), we have the log-likelihood function

$$\log L = \sum_{i=1}^{n}\log\left(\alpha + \frac{2\beta}{x_i}\right) + \sum_{i=1}^{n}log\left(\frac{1}{x_i}\right)^2 - \sum_{i=1}^{n}\left(\frac{\alpha}{x_i}\right) -\beta\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)^2 + \sum_{i=1}^{n}\log\left\{1+\lambda-2\lambda\exp\left\{-\frac{\alpha}{x_i} - \beta\left(\frac{1}{x_i}\right)^2\right\}\right\}. \tag{6.2}$$

Differentiating (6.2) with respect to α, β and λ, then equating it to zero, we obtain the estimating equations are

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{n}\left\{\alpha+\frac{2\beta}{x_i}\right\}^{-1} - \sum_{i=1}^{n}\left(\frac{1}{x_i}\right) + 2\lambda\sum_{i=1}^{n}\frac{\exp\left\{-\frac{\alpha}{x_i}-\beta\left(\frac{1}{x_i}\right)^2\right\}\left(\frac{1}{x_i}\right)}{\left\{1+\lambda-2\lambda\exp\left\{-\frac{\alpha}{x_i}-\beta\left(\frac{1}{x_i}\right)^2\right\}\right\}}, \tag{6.3}$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n}\left\{\alpha+\frac{2\beta}{x_i}\right\}^{-1}\left(\frac{2}{x_i}\right) - \sum_{i=1}^{n}\left(\frac{1}{x_i}\right)^2 + 2\lambda\sum_{i=1}^{n}\frac{\exp\left\{-\frac{\alpha}{x_i}-\beta\left(\frac{1}{x_i}\right)^2\right\}\left(\frac{1}{x_i}\right)^2}{\left\{1+\lambda-2\lambda\exp\left\{-\frac{\alpha}{x_i}-\beta\left(\frac{1}{x_i}\right)^2\right\}\right\}}, \tag{6.4}$$

and

$$\frac{\partial \log L}{\partial \lambda} = \sum_{i=1}^{n}\frac{1-2\exp\left\{-\frac{\alpha}{x_i}-\beta\left(\frac{1}{x_i}\right)^2\right\}}{\left\{1+\lambda-2\lambda\exp\left\{-\frac{\alpha}{x_i}-\beta\left(\frac{1}{x_i}\right)^2\right\}\right\}}, \tag{6.5}$$

It is more convenient to use quasi Newton algorithm to numerically maximize the log-likelihood function given in equation (6.2) to yield the ML estimators $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\lambda}$ respectively. For finding the interval estimation and testing the hypothesis of the subject model, we required the observed information matrix is given by

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\\ \hat{\lambda}\end{pmatrix} \sim N\left[\begin{pmatrix}\alpha\\ \beta\\ \lambda\end{pmatrix}, \begin{pmatrix}\hat{V}_{11} & \hat{V}_{12} & \hat{V}_{13}\\ \hat{V}_{21} & \hat{V}_{22} & \hat{V}_{23}\\ \hat{V}_{31} & \hat{V}_{32} & \hat{V}_{33}\end{pmatrix}\right],$$

the expected information matrix is given by

$$V^{-1} = -E\begin{pmatrix}\frac{\partial^2 \log L}{\partial \alpha^2} & \frac{\partial^2 \log L}{\partial \alpha \partial \beta} & \frac{\partial^2 \log L}{\partial \alpha \partial \lambda}\\ \frac{\partial^2 \log L}{\partial \beta \partial \alpha} & \frac{\partial^2 \log L}{\partial \beta^2} & \frac{\partial^2 \log L}{\partial \beta \partial \lambda}\\ \frac{\partial^2 \log L}{\partial \alpha \partial \lambda} & \frac{\partial^2 \log L}{\partial \lambda \partial \beta} & \frac{\partial^2 \log L}{\partial \lambda^2}\end{pmatrix}. \tag{6.6}$$

Solving the inverse matrix for the observed information matrix (6.6), yields the asymptotic variance and co-variances of the ML estimators $\hat{\alpha}, \hat{\beta}$, and $\hat{\lambda}$. By using (6.6) approximate $100(1-\alpha)\%$ asymptotic confidence intervals for α, β and λ are

$$\hat{\alpha} \pm Z_{\frac{\alpha}{2}}\sqrt{\hat{V}_{11}}, \qquad \hat{\beta} \pm Z_{\frac{\alpha}{2}}\sqrt{\hat{V}_{22}}, \qquad \hat{\lambda} \pm Z_{\frac{\alpha}{2}}\sqrt{\hat{V}_{33}},$$

where $Z_{\frac{\alpha}{2}}$ is the upper αth percentile of the standard normal distribution.

7. Application

This section illustrates the usefulness of the TMIR distribution with real data. The data consist of thirty successive values of March precipitation (in inches) given by Hinkley (1977) and given below
0.77, 1.74, 0.81, 1.2, 1.95, 1.2, 0.47, 1.43, 3.37, 2.2, 3, 3.09, 1.51, 2.1, 0.52, 1.62, 1.31, 0.32, 0.59,

0.81, 2.81, 1.87, 1.18, 1.35, 4.75, 2.48, 0.96, 1.89, 0.9, 2.05.
Four distributions are fitted to the precipitation data using maximum likelihood estimation. The estimated parameters for the TMIR distribution with their corresponding 95% C.I are given in Table 2. The summary statistics of the fitted TMIR, TIR, MIR and IR distributions are given in Table 1.

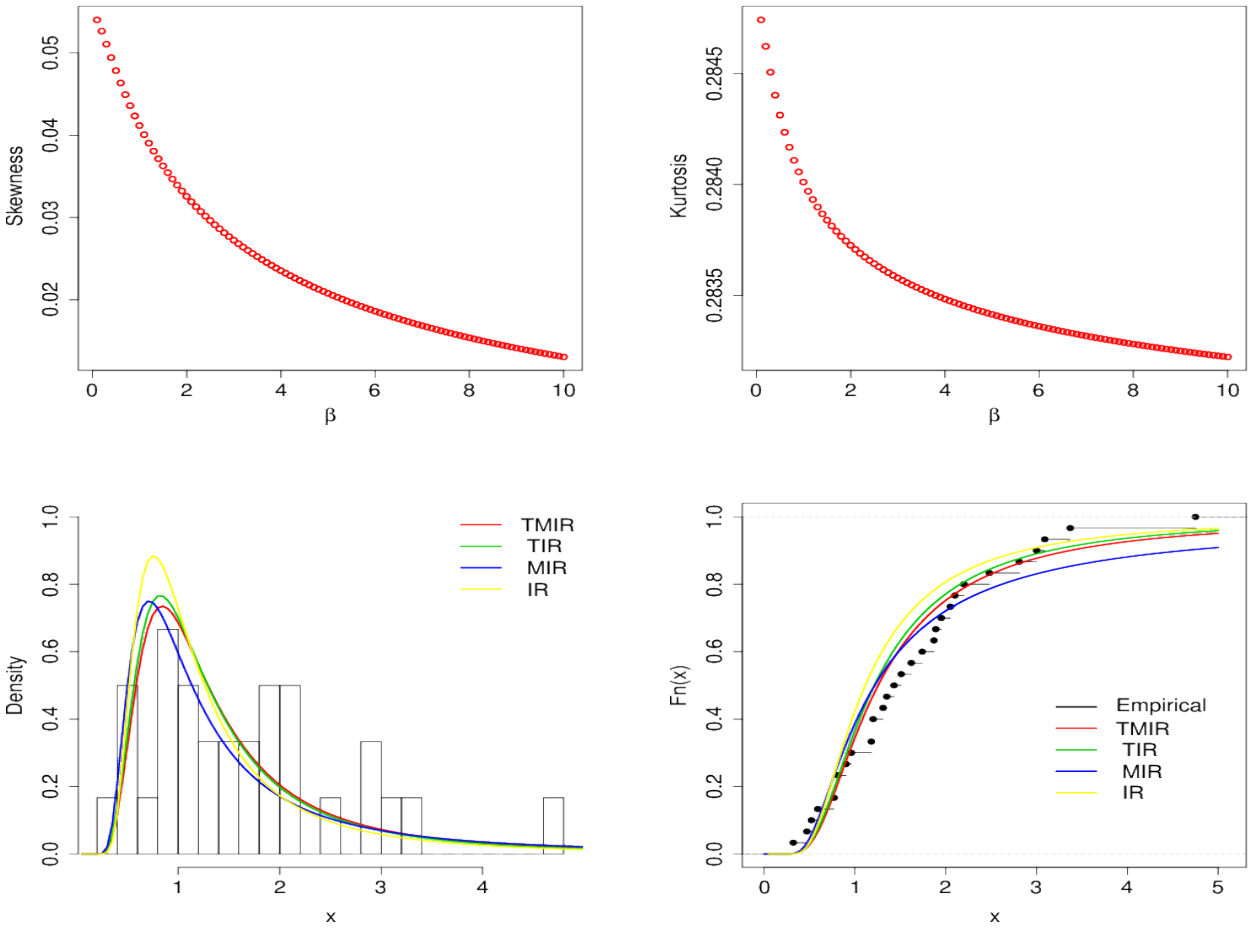

Figure 4: Estimated Reliability and Survival functions for four fitted models.

Table 1: *Summary Statistics for* TMIR, TIR, MIR *and* IR *distributions*

Distribution	Quartile deviation	Bowley Skewness	Percentile kurtosis
TMIR	0.1378	-0.0881	0.2866
TIR	0.1361	-0.0898	0.2864
MIR	0.4672	0.3928	0.1726
IR	0.3741	0.3069	0.2096

The MLE$_S$ of the parameters (with their standard errors) and their corresponding values of the Coefficient of determination, mean square error (MSE) and Kolmogorov-Smirnov (K-S) test values are displayed in Table 3. The likelihood ratio (LR) statistic for testing the hypothesis H_0 IR v.s H_A: TMIR is 4.0786 with their corresponding p-value 0.04343. Hence we reject the null hypothesis in favour of the TMIR distribution, because the p-value is small. Fig. 4 illustrates the four fitted models with empirical functions for the precipitation data. These graphs illustrate that the TMIR distribution fits well. The hazard plot of the estimated TMIR distribution has increasing and then decreasing instantaneous failure rate. As we can see from these numerical results in Table 3, the Coefficient of determination of the TMIR

Table 2: *Estimated Parameters of the* TMIR *distribution*

Parameter	ML Estimate	Standard Error	95% Confidence Interval Lower	Upper
α	0.0212	0.2388	-0.4469	0.4893
β	0.6472	0.2654	0.1268	1.16758
λ	-0.6703	0.2612	-1.1825^{*}	-0.1581

Table 3: *Estimates of the model parameters for precipitation data and the* K-S *test, Coefficient of determination and associated* MSE *values*

Distribution	TMIR	TIR	MIR	IR
α	0.0212	-	0.3598	-
	(0.2388)		(0.3745)	
β	0.6472	0.6285	0.5881	0.8588
	(0.2654)	(0.1583)	(0.2975)	(0.1568)
λ	-0.6703	-0.6701	-	-
	(0.2613)	(0.2661)		
K-S	0.1395	0.1626	0.1641	0.2206
R^2	0.9442	0.9199	0.9137	0.8382
MSE	0.0726	0.0862	0.0801	0.1212

distribution is higher than the other three sub-models and the values of mean square error (MSE) and Kolmogorov-Smirnov (K-S) test of the TMIR distribution are the smallest among those of the four fitted distribution. Therefore the TMIR distribution can be chosen as the best model for lifetime data analysis. Fig. 4 also illustrate that the TMIR distribution gives a better fit than the other three sub-models.

8. Conclusion

We proposed a new distribution, named the TMIR distribution, which is an extension of the MIR distribution. The TMIR distribution provides better results than the MIR, TIR and IR distributions. In this model the new parameter λ provides more flexibility in modeling reliability data. We derive the quantile function, moments, moment generating function, entropies, mean deviation, Bonferroni and Lorenz curves. We also derive the Sth moment of rth order statistics and the kth moment of rth median order statistics. We discuss the maximum likelihood estimation and obtain the fisher information matrix. The usefulness of the new model is illustrated in an application to real data using MLE. We hope that the proposed model may attract wider application in the analysis of reliability data.

9. Acknowledgements

The authors are grateful to the referee for helpful comments and suggestions.

References

[1] Arnold, B. C, Balakrishnan A. N, Nagaraja H. N. *A First Course in Order Statistics*. John Wiley & Sons, New York; 1992. MR 1178934 (94a:62076)

[2] Aryal, G.R, Tsokos C.P. Transmuted Weibull Distribution: A Generalization of the Weibull

Probability Distribution. *European Journal of Pure and Applied Mathematics*, Vol. 4, No. 2:89–102, 2011.

[3] Ammar M. Sarhan and Mazen Zaindin. Modified Weibull Distribution. *Applied Sciences*, Vol. 11: 123–136, 2009.

[4] Bonferroni C.E. Elmenti di Statistica Generale. *Libreria Seber*, Firenze, 1930.

[5] Gauss M. Cordeiro, Antonio Eduardo Gomes, Cibele Queiroz da-Silva, Edwin M. M. Ortega. The Beta Exponentiated Weibull Distribution. *Journal of Statistical Computation and Simulation*, Vol. 83, No. 1 : 114–138, 2013. http://dx.doi.org/10.1080/00949655.2011.615838

[6] Gharraph, M.K. Comparison of Estimators of Location Measures of an Inverse Rayleigh Distribution. *The Egyptian Statistical Journal*, 37:295–309, 1993.

[7] Hinkley, D. On Quick Choice of Power Transformations. *The American Statistician*, 26:67–69, 1977.

[8] Havrda, J. and Charvat, F. Quantification Method in Classification Processes: Concept of Structural α-entropy. Kybernetika, 3: 30–35, 1967.

[9] Khan, M.S. Modified Inverse Rayleigh Distribution. *International Journal of Computer Applications*, Vol. 87, No. 13:28–33, February 2014.

[10] Khan, M.S, King R. Transmuted Modified Weibull Distribution: A Generalization of the Modified Weibull Probability Distribution. *European Journal of Pure and Applied Mathematics*, Vol. 6, No. 1:66–88, 2013.

[11] Khan, M.S, King R. Transmuted Generalized Inverse Weibull Distribution. *Journal of Applied Statistical Sciences*, Vol. 20 (3):15–32 , 2013.

[12] Khan, M.S, King Robert, Hudson Irene. Transmuted Generalized Exponential Distribution. *57th Annual Meeting of the Australian Mathematical Society*, September 30-October 3, 2013 at the University of Sydney, Australia.

[13] Khan, M.S, King Robert. Modified Inverse Weibull Distribution. *J. Stat. Appl. Pro*, Vol. 1, No. 2, 115–132, 2012.

[14] Khan, M.S, Pasha, G.R and Pasha, A.H. Theoretical Analysis of Inverse Weibull Distribution. *WSEAS Transactions on Mathematics*, 7(2), 30–38, 2008.

[15] Lorenz, M. O. Methods of Measuring the Concentration of Wealth. *The American Statistical Association*, Vol. 9, No. 9 (70): 209–219, 1905.

[16] Mohsin and Shahbaz. Comparison of Negative Moment Estimator with Maximum Likelihood Estimator of Inverse Rayleigh Distribution, *PJSOR*, Vol.1: 45–48, 2005.

[17] Mukarjee, S.P. and Maitim, S.S. A Percentile Estimator of the Inverse Rayleigh Parameter. *IAPQR Transactions*, 21, 63–65, 1996.

[18] Merovci, F. Transmuted Rayleigh Distribution. *Austrian Journal of Statistics*, Volume 42, Number 1, 21-31, 2013.

[19] Renyi, Alfred. On Measures of Information and Entropy. *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 1960. 547?-561, 1961.

[20] R Development Core Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria, 2011.

[21] Trayer, V. N. Doklady Acad, Nauk, Belorus, U.S.S.R, 1964.

[22] Voda, V. Gh. On the Inverse Rayleigh Random Variable, Pep. *Statist. App. Res., JUSE*, 19, 13-21, 1972.

[23] Shaw, W. T., and Buckley, I. R. The Alchemy of Probability Distributions: Beyond Gram-Charlier Expansions, and a Skew-kurtotic Normal Distribution from a Rank Transmutation Map. arXiv preprint, arXiv:0901.0434, 2009.

Affiliation:

Muhammad Shuaib Khan
School of Mathematical and Physical Sciences,
University of Newcastle
Callaghan, NSW 2308,
Australia
E-mail: shuaib.stat@gmail.com

Robert King
School of Mathematical and Physical Sciences,
University of Newcastle
Callaghan, NSW 2308,
Australia
E-mail: robert.king@newcastle.edu.au

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 44 *Submitted:* 2014-03-17
October 2015 *Accepted:* 2014-07-29

Austrian Journal of Statistics
October 2015, Volume 44, 31–44.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i3.80

Covariance Structure of Compositional Tables

Kamila Fačevicová
Palacký University in Olomouc

Karel Hron
Palacký University in Olomouc

Abstract

Recent experience with interpretation of orthonormal coordinates in compositional data shows clearly a necessity of their better understanding in terms of logratios that form the primary source of information within the logratio methodology. This is even more crucial in the special case of compositional tables, where both balances and coordinates with odds ratio interpretation are involved. The aim of the paper is to provide a decomposition of covariance structure of orthonormal coordinates in compositional tables in terms of logratio variances, which could serve this purpose. For their better interpretability, the formulas are also accompanied with appropriate comments and graphical illustrations, and implications for the prominent case of 2×2 compositional tables are discussed.

Keywords: compositional tables, covariance structure, orthonormal coordinates.

1. Introduction

Although the logratio methodology seems to be nowadays a well-established approach to statistical analysis of compositional data, i.e. multivariate observations carrying relative information (Aitchison 1986; Pawlowsky-Glahn and Buccianti 2011), it is also suitable for more complex data structures, where not the absolute values but rather ratios are of primary interest. One of them are compositional tables (Egozcue, Díaz-Barrero, and Pawlowsky-Glahn 2008; Egozcue, Pawlowsky-Glahn, Templ, and Hron 2014; Fačevicová, Hron, Todorov, Guo, and Templ 2014a; Fačevicová, Hron, Todorov, and Templ 2014b), a continuous counterpart to well-known contingency tables (Agresti 2002). Besides the difference in the nature of data, (cells of the contingency table are discrete counts, while parts of the compositional tables are continuous values), the main difference between is that, on the one hand, a contingency table collects results from n independent observations, while, on the other hand, a compositional table itself represents one observation. The analysis of relationships between row and column factors is thus based on sample of n compositional tables. Furthermore, by applying the Aitchison geometry (Egozcue and Pawlowsky-Glahn 2006), and following the principles of compositional data analysis, it is possible to decompose the original table into its independent and interactive parts (the latter capturing relations between both factors), while assuming geometric marginals instead of the standard arithmetic ones. Moreover, in Fačevicová *et al.* (2014b) orthonormal coordinates were assigned to both independence and interaction tables that enable to perform statistical analysis using standard methods and focus only on coordi-

nates of the interaction table. The problem of analysis of relationship between factors from a sample of tables, which would need to be handled using three-dimensional contingency tables or log-linear models in the standard case, thus can directly transferred to standard statistical treatment (like hypotheses testing) in coordinates; for example, independence of factors corresponds to zero coordinates of the interaction table.

While coordinates of the independence table can be interpreted through balances (Egozcue and Pawlowsky-Glahn 2005), coordinate representation of the interaction one needs to be formulated in sense of odds ratios (Fačevicová *et al.* 2014b). Although motivation for the latter coordinates is quite intuitive as odds ratios become popular to represent also contingency tables (Agresti 2002), interpretation of the coordinate system as a whole may seem to be too complex for practical purposes. On the way to enhance the interpretability, one possibility is to analyze covariance structure of coordinates (Fišerová and Hron 2011) to see which ratios contribute (in positive or negative sense) to values of individual variances and covariances. Nevertheless, a specific structure of compositional tables and their respective coordinates requires a deeper insight as balances form just one part of the coordinate system.

The aim of the presented paper is to analyze covariance structure of orthonormal coordinates for compositional tables in terms of elements of the variation matrix (Aitchison 1986), i.e., as linear combinations of variances of single logratios, which seems to be a necessary step in further development of any reasonable coordinate representation of compositional tables. The paper is organized as follows. In the next section, basics of compositional tables and their decomposition into independent and interactive parts are recalled. Section 3 is devoted to the covariance structure of coordinates itself, where the corresponding formulas (that might seem to be rather complex) are always illustrated with a graphical scheme to allow their better understanding. In Section 4 some implications for the special case of 2×2 compositional tables are briefly mentioned. Section 5 presents an illustrative example and Section 6 concludes.

2. Orthonormal coordinates of I x J compositional tables

A $I \times J$ compositional table $\mathbf{x}$ is a special case of compositional data that are arranged into a form of table to indentify relation between two (row and column) factors. They are formed by parts $x_{ij} > 0$ for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$ which carry only relative information. Consequently, their sum κ is arbitrary (like $\kappa = 1$ for the case of proportions), reached formally using the closure operation

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa x_{ij}}{\sum_{k,l=1}^{I,J} x_{kl}} \right)_{i,j=1}^{I,J} .$$

The sample space of representations of $I \times J$ compositional tables is the simplex, $(IJ-1)$-dimensional subset of $\mathbf{R}^{IJ}$ defined as

$$\mathcal{S}^{IJ} = \left\{ \mathbf{x} = (x_{ij})_{i,j=1}^{I,J} \mid x_{ij} > 0,\ i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J;\ \sum_{i,j=1}^{I,J} x_{ij} = \kappa \right\} .$$

To follow specific features of compositional tables (as a special case of compositional data), the Aitchison geometry on the simplex is defined, see Egozcue and Pawlowsky-Glahn (2006) for details. This geometry has the same algebraic-geometrical structure as the standard Euclidean geometry in real space and is represented by operations of perturbation, power transformation, and the Aitchison inner product. According to Egozcue *et al.* (2008) these operations are defined for compositional tables $\mathbf{x}$ and $\mathbf{y}$ and $\alpha \in \mathbf{R}$ as

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left(x_{ij} y_{ij}\right)_{i,j=1}^{I,J}, \qquad \alpha \odot \mathbf{x} = \mathcal{C}\left(x_{ij}^{\alpha}\right)_{i,j=1}^{I,J}$$

and

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2IJ} \sum_{i,j} \sum_{k,l} \ln \frac{x_{ij}}{x_{kl}} \ln \frac{y_{ij}}{y_{kl}} \quad .$$

Compositional table $\mathbf{n} = \mathcal{C}\left(x_{ij} = 1\right)_{i,j=1}^{I,J}$ stands for the neutral element in the $(IJ-1)$-dimensional vector space $(\mathcal{S}^{IJ}, \oplus, \odot)$.

Each compositional table $\mathbf{x}$ can be expressed as

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{e}_1 \rangle_A \odot \mathbf{e}_1 \oplus \ldots \oplus \langle \mathbf{x}, \mathbf{e}_{IJ-1} \rangle_A \odot \mathbf{e}_{IJ-1},$$

where $(\mathbf{e}_1, \ldots, \mathbf{e}_{IJ-1})$ form an orthonormal basis in $(IJ-1)$-dimensional simplex (with respect to the Aitchison geometry), resulting in $(IJ-1)$-dimensional real vector of orthonormal coordinates

$$\mathbf{z} = h(\mathbf{x}) = (z_1, \ldots, z_{IJ-1}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_A, \ldots, \langle \mathbf{x}, \mathbf{e}_{IJ-1} \rangle_A).$$

Consequnetly, the following relations between the Aitchison and the Euclidean geometries can be derived,

$$h(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot h(\mathbf{x}) + \beta \cdot h(\mathbf{y}), \qquad \langle \mathbf{x}, \mathbf{y} \rangle_A = \langle h(\mathbf{x}), h(\mathbf{y}) \rangle_E,$$

i.e. h is an isometric mapping from $\mathcal{S}^{IJ}$ to $\mathbf{R}^{IJ-1}$ (we refer also to isometric logratio (ilr) transformation (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal 2003)).

Within the framework of the Aitchison geometry it is possible to decompose the original compositional table into its independent and interactive parts, $\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int}$, see Egozcue *et al.* (2008) for details. The independent part (independence table) is compositional table with elements

$$\mathbf{x}_{ind} = \left(x_{ij}^{ind} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} x_{kj} x_{il} \right)^{\frac{1}{IJ}} \right)_{i,j=1}^{I,J} \tag{1}$$

and the interactive part (interaction table)

$$\mathbf{x}_{int} = \left(x_{ij}^{int} = \left(\prod_{k=1}^{I} \prod_{l=1}^{J} \frac{x_{ij}}{x_{kj} x_{il}} \right)^{\frac{1}{IJ}} \right)_{i,j=1}^{I,J}. \tag{2}$$

According to Fačevicová *et al.* (2014b), the independence and the interaction tables can be expressed in $I + J - 2$ and $(I-1)(J-1)$ nonzero orthonormal coordinates, respectively, the remaining coordinates (up to the total number of $IJ-1$ variables) being zero. The coordinates of the independence table can be expressed as balances

$$z_i^r = \sqrt{\frac{(I-i)J}{I-i+1}} \ln \frac{(x_{i1} \ldots x_{iJ})^{1/J}}{(x_{i+1,1} \ldots x_{IJ})^{1/(IJ-iJ)}}, \quad i = 1, \ldots, I-1, \tag{3}$$

and

$$z_j^c = \sqrt{\frac{I(J-j)}{J-j+1}} \ln \frac{(x_{1j} \ldots x_{Ij})^{1/I}}{(x_{1,j+1} \ldots x_{IJ})^{1/(IJ-Ij)}}, \quad j = 1, \ldots, J-1, \tag{4}$$

representing the row and column information (logratios), respectively, conveyed by the independence table; coordinates of the interaction table can be chosen as

$$z_{rs}^{int} = \frac{1}{\sqrt{r \cdot s \cdot (r-1) \cdot (s-1)}} \ln \prod_{i=1}^{r-1} \prod_{j=1}^{s-1} \frac{x_{ij} x_{rs}}{x_{is} x_{rj}}, \tag{5}$$

with an odds ratio structure. These two sets of coordinates together form an orthonormal coordinate representation of the original compositional table $\mathbf{x}$. Covariance structure in terms of elements of the variation matrix (Aitchison 1986), especially for coordinates of the interaction table, will be studied in detail in the next section.

3. Covariance structure of coordinates of the compositional table

In the following, covariance structure of the above mentioned coordinate representation will be expressed as linear combinations of variances of logratios. At first, covariance structure of the interaction table is introduced, followed by the independence table structure, and finally also mutual relations between both tables (expressed through the corresponding covariances) are analyzed.

Variances of logratios form the elemental information on variability in compositional tables and are summarized in $IJ \times IJ$ variation matrix

$$\mathbf{T} = \begin{pmatrix} \operatorname{var}\left(\ln \frac{x_{11}}{x_{11}}\right) & \operatorname{var}\left(\ln \frac{x_{11}}{x_{12}}\right) & \cdots & \operatorname{var}\left(\ln \frac{x_{11}}{x_{IJ}}\right) \\ \operatorname{var}\left(\ln \frac{x_{12}}{x_{11}}\right) & \operatorname{var}\left(\ln \frac{x_{12}}{x_{12}}\right) & \cdots & \operatorname{var}\left(\ln \frac{x_{12}}{x_{IJ}}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{var}\left(\ln \frac{x_{IJ}}{x_{11}}\right) & \operatorname{var}\left(\ln \frac{x_{IJ}}{x_{12}}\right) & \cdots & \operatorname{var}\left(\ln \frac{x_{IJ}}{x_{IJ}}\right) \end{pmatrix}. \tag{6}$$

As it is usual within the logratio methodology, all coordinates are logcontrasts, i.e. they can be expressed in form

$$z = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} \ln x_{ij} = \mathbf{a}' \ln \mathbf{x}, \text{ where } \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} = 0.$$

Also the covariance structure can be derived accordingly (Aitchison 1986).

Proposition 3.1 *Variances and covariances for logcontrasts* $\mathbf{a}' \ln \mathbf{x}$ *and* $\mathbf{b}' \ln \mathbf{x}$ *of a* IJ*-part compositional table* $\mathbf{x}$ *are*

$$\operatorname{var}(\mathbf{a}' \ln \mathbf{x}) = -\frac{1}{2}\mathbf{a}'\mathbf{T}\mathbf{a}, \tag{7}$$

$$\operatorname{cov}(\mathbf{a}' \ln \mathbf{x}, \mathbf{b}' \ln \mathbf{x}) = -\frac{1}{2}\mathbf{a}'\mathbf{T}\mathbf{b}. \tag{8}$$

Since the possible logcontrast representation of coordinates (2), (3) and (4), Equations (7) and (8) are crucial to derive of their covariance structure. As the interaction table is usually of main interest for the analysis, we start with variances of its respective coordinates.

Theorem 3.2 *Consider an arbitrary coordinate* z_{rs}*, for* $r = 2, \ldots, I$ *and* $s = 2, \ldots, J$ *of the interaction table* $\mathbf{x}_{int}$ *from (5). Its variance is formed by three parts,*

$$\operatorname{var}(z_{rs}) = A_1 - B_1 - C_1. \tag{9}$$

The first part, increasing the variance, is

$$\begin{aligned} A_1 \;=\; & \frac{1}{rs(s-1)} \sum_{i=1}^{r-1}\sum_{j,j'=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{rj'}}\right) + \frac{1}{rs(r-1)} \sum_{i,i'=1}^{r-1}\sum_{j=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i's}}\right) + \\ & + \frac{r-1}{rs} \sum_{j=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{rj}}{x_{rs}}\right) + \frac{s-1}{rs} \sum_{i=1}^{r-1} \operatorname{var}\left(\ln \frac{x_{is}}{x_{rs}}\right). \end{aligned} \tag{10}$$

The variance of the coordinate is reduced by parts

$$B_1 \;=\; \frac{1}{2}\frac{1}{rs(r-1)(s-1)} \sum_{i,i'=1}^{r-1}\sum_{j,j'=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) + \frac{1}{rs} \sum_{i=1}^{r-1}\sum_{j=1}^{s-1} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{rs}}\right) \tag{11}$$

and

$$C_1 = \frac{1}{2}\frac{r-1}{rs(s-1)}\sum_{j,j'=1}^{s-1}\operatorname{var}\left(\ln\frac{x_{rj}}{x_{rj'}}\right)+\frac{1}{2}\frac{s-1}{rs(r-1)}\sum_{i,i'=1}^{r-1}\operatorname{var}\left(\ln\frac{x_{is}}{x_{i's}}\right)+ \\ +\frac{1}{rs}\sum_{i=1}^{r-1}\sum_{j'=1}^{s-1}\operatorname{var}\left(\ln\frac{x_{is}}{x_{rj'}}\right). \tag{12}$$

Proof: *When parts of the compositional table* $\mathbf{x}$ *are rearranged in form of composition* $\mathbf{x}_r = (x_{11}, x_{12}, \ldots, x_{1J}, x_{21}, \ldots, x_{IJ})$*, coordinate* z_{rs} *of the interaction table can be expressed as* $z_{rs} = \mathbf{a}' \ln \mathbf{x}_r$*, where for elements of the coefficient vector* $\mathbf{a} = (a_{11}, a_{12}, \ldots, a_{1J}, a_{21}, \ldots, a_{IJ})$ *the following relations hold,*

$a_{ij} = 1/\sqrt{rs(r-1)(s-1)}$	for	$i = 1, \ldots, r-1$	$j = 1, \ldots, s-1$
$a_{ij} = -(r-1)/\sqrt{rs(r-1)(s-1)}$	for	$i = r$	$j = 1, \ldots, s-1$
$a_{ij} = -(s-1)/\sqrt{rs(r-1)(s-1)}$	for	$i = 1, \ldots, r-1$	$j = s$
$a_{ij} = (r-1)(s-1)/\sqrt{rs(r-1)(s-1)}$	for	$i = r$	$j = s$
$a_{ij} = 0$		*otherwise.*	

Equation (9) is then consequence of Proposition 3.1.

◇

From Theorem 3.2 it is clear that variance of the coordinate z_{rs} is formed by nine groups of logratio variances. Four of them increase the overall variability and the other five reduce it. The first four groups are represented by A_1, which is formed by logratios of "inner" parts of the partial table or part x_{rs}, with its last row and column (i.e. r-th row and s-th column of the original table $\mathbf{x}$) except of the part x_{rs} itself:

- variances of logratios between an inner part of the partial table and a part from its last row (except of x_{rs}),
- variances of logratios between an inner part of the partial table and a part from its last column (except of x_{rs}),
- variances of logratios between a part from the last row (except of x_{rs}) and x_{rs} itself,
- variances of logratios between a part from the last column (except of x_{rs}) and x_{rs} itself.

The variance of z_{rs} is reduced by B_1 and C_1, formed by variances of logratios corresponding to remaining possible relations between parts of the above defined groups (inner tables, last row/column without x_{rs}, part x_{rs} itself). Concretely, B_1 consists of

- variances of logratios between inner parts of the partial table,
- variances of logratios between an inner part and x_{rs}.

Similarly, C_1 is formed by

- variances of logratios between parts from the last row (except of x_{rs}),
- variances of logratios between parts from the last column (except of x_{rs}),
- variances of logratios between parts from the last row and the last column (except of x_{rs}).

The above relations can be expressed also graphically, as shown in Figure 1.

Covariances between coordinates of the interaction table are derived in the next theorem.

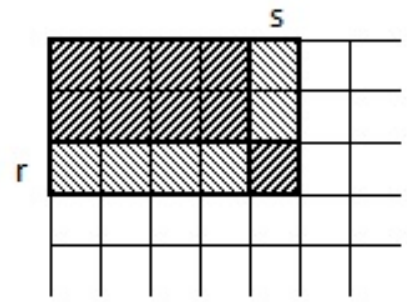

Figure 1: Variance of coordinate z_{rs} is increased by variances of logratios between a part from area highlighted by (/) and a part from the second area highlighted by (\) – A_1. The variance of z_{rs} is reduced by variances of logratios between two parts from area (/) – B_1 or two parts from (\) – C_1.

Theorem 3.3 *Consider two coordinates of the interaction table $z_{r_1s_1}, z_{r_2s_2}$, for $r_1, r_2 = 2, \ldots, I$ and $s_1, s_2 = 2, \ldots, J$. Then for their covariance the following holds,*

$$\mathrm{cov}(z_{r_1s_1}, z_{r_2s_2}) = K(A_2 + B_2 - C_2 - D_2), \tag{13}$$

where

$$\begin{aligned} A_2 &= (s_2-1)\sum_{i_1=1}^{r_1-1}\sum_{i_2=1}^{r_2-1}\sum_{j_1=1}^{s_1-1}\mathrm{var}\left(\ln\frac{x_{i_1j_1}}{x_{i_2s_2}}\right) + (r_2-1)\sum_{i_1=1}^{r_1-1}\sum_{j_1=1}^{s_1-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{i_1j_1}}{x_{r_2j_2}}\right) + \\ &+(r_1-1)(s_1-1)(s_2-1)\sum_{i_2=1}^{r_2-1}\mathrm{var}\left(\ln\frac{x_{r_1s_1}}{x_{i_2s_2}}\right) + (r_1-1)(r_2-1)(s_1-1)\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{r_1s_1}}{x_{r_2j_2}}\right), \end{aligned} \tag{14}$$

$$\begin{aligned} B_2 &= (s_1-1)\sum_{i_1=1}^{r_1-1}\sum_{i_2=1}^{r_2-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{i_1s_1}}{x_{i_2j_2}}\right) + (s_1-1)(s_2-1)(r_2-1)\sum_{i_1=1}^{r_1-1}\mathrm{var}\left(\ln\frac{x_{i_1s_1}}{x_{r_2s_2}}\right) + \\ &+(r_1-1)\sum_{j_1=1}^{s_1-1}\sum_{i_2=1}^{r_2-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{r_1j_1}}{x_{i_2j_2}}\right) + (r_1-1)(r_2-1)(s_2-1)\sum_{j_1=1}^{s_1-1}\mathrm{var}\left(\ln\frac{x_{r_1j_1}}{x_{r_2s_2}}\right), \end{aligned} \tag{15}$$

$$\begin{aligned} C_2 &= \sum_{i_1=1}^{r_1-1}\sum_{i_2=1}^{r_2-1}\sum_{j_1=1}^{s_1-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{i_1j_1}}{x_{i_2j_2}}\right) + (r_2-1)(s_2-1)\sum_{i_1=1}^{r_1-1}\sum_{j_1=1}^{s_1-1}\mathrm{var}\left(\ln\frac{x_{i_1j_1}}{x_{r_2s_2}}\right) + \\ &+(r_1-1)(s_1-1)\sum_{i_2=1}^{r_2-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{r_1s_1}}{x_{i_2j_2}}\right) + (r_1-1)(r_2-1)(s_1-1)(s_2-1)\mathrm{var}\left(\ln\frac{x_{r_1s_1}}{x_{r_2s_2}}\right), \end{aligned} \tag{16}$$

$$\begin{aligned} D_2 &= (s_1-1)(s_2-1)\sum_{i_1=1}^{r_1-1}\sum_{i_2=1}^{r_2-1}\mathrm{var}\left(\ln\frac{x_{i_1s_1}}{x_{i_2s_2}}\right) + (s_1-1)(r_2-1)\sum_{i_1=1}^{r_1-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{i_1s_1}}{x_{r_2j_2}}\right) + \\ &+(r_1-1)(s_2-1)\sum_{j_1=1}^{s_1-1}\sum_{i_2=1}^{r_2-1}\mathrm{var}\left(\ln\frac{x_{r_1j_1}}{x_{i_2s_2}}\right) + (r_1-1)(r_2-1)\sum_{j_1=1}^{s_1-1}\sum_{j_2=1}^{s_2-1}\mathrm{var}\left(\ln\frac{x_{r_1j_1}}{x_{r_2j_2}}\right) \end{aligned} \tag{17}$$

and $K = \frac{1}{2}\frac{1}{\sqrt{r_1r_2s_1s_2(r_1-1)(r_2-1)(s_1-1)(s_2-1)}}$.

Proof: *The covariances are obtained using the general formula (8), where the corresponding*

coefficient vectors $\mathbf{a}^1$ *and* $\mathbf{a}^2$ *have elements*

$$a_{ij}^k = \begin{cases} 1/\sqrt{r_k s_k (r_k-1)(s_k-1)} & \text{for} \quad i = 1,\ldots,r_k-1 \quad j = 1,\ldots,s_k-1 \\ -(r_k-1)/\sqrt{r_k s_k (r_k-1)(s_k-1)} & \text{for} \quad i = r_k \quad\quad\;\; j = 1,\ldots,s_k-1 \\ -(s_k-1)/\sqrt{r_k s_k (r_k-1)(s_k-1)} & \text{for} \quad i = 1,\ldots,r_k-1 \quad j = s_k \\ (r_k-1)(s_k-1)/\sqrt{r_k s_k (r_k-1)(s_k-1)} & \text{for} \quad i = r_k \quad\quad\;\; j = s_k \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

and $k = 1, 2$.

◇

Similarly as for the case of variances, there is a group of logratio variances that increases the overall covariance between coordinates (A_2 and B_2) and the remaining variances reduce it (C_2 and D_2). Specifically, for construction of logratios in A_2 the following parts are employed,

- an inner part of the first partial table and a part from the last column of the second partial table (except of x_{r_2,s_2}),
- an inner part of the first partial table and a part from the last row of the second partial table (except of x_{r_2,s_2}),
- the part x_{r_1,s_1} and a part from the last column of the second partial table (except of x_{r_2,s_2}),
- the part x_{r_1,s_1} and a part from the last row of the second partial table (except of x_{r_2,s_2}),

where we always deal with two "virtual" tables corresponding to the coordinates of interest. Similarly, B_2 is formed by variances of logratios of

- a part from the last column of the first partial table (except of x_{r_1,s_1}) and an inner part of the second partial table,
- a part from the last column of the first partial table (except of x_{r_1,s_1}) and the part x_{r_2,s_2},
- a part from the last row of the first partial table (except of x_{r_1,s_1}) and an inner part of the second partial table,
- a part from the last row of the first partial table (except of x_{r_1,s_1}) and the part x_{r_2,s_2}.

On the other hand, the covariance is reduced by C_2, involving logratios between

- an inner part of the first partial table and an inner part of the second partial table,
- an inner part of the first table and the part part x_{r_2,s_2},
- the part x_{r_1,s_1} and an inner part of the second partial table,
- parts x_{r_1,s_1} and x_{r_2,s_2},

and by D_2 consisting of logratios, formed by

- a part from the last column of the first partial table (except of x_{r_1,s_1}) and a part from the last column of the second partial table (except of x_{r_1,s_1}),
- a part from the last column of the first partial table (except of x_{r_1,s_1}) and a part from the last row of the second partial table (except of x_{r_2,s_2}),
- a part from the last row of the first partial table (except of x_{r_1,s_1}) and a part from the last column of the second partial table (except of x_{r_2,s_2}),

- a part from the last row of the first partial table (except of x_{r_1,s_1}) and a part from the last row of the second partial table (except of x_{r_2,s_2}).

Also covariance between two coordinates of the interaction table could supported by its graphical representation, see Figure 2.

Since coordinates of the independence table (3), (4) are balances obtained from sequential binary partitions, dividing rows and columns of the original table, respectively (Egozcue and Pawlowsky-Glahn 2005), their variances and covariances are obtained as direct consequence of Fišerová and Hron (2011).

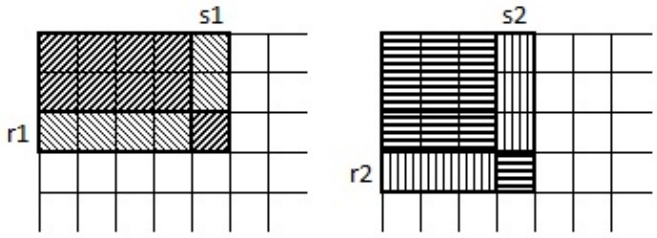

Figure 2: Covariance between coordinates z_{r_1,s_1} and z_{r_2,s_2} is increased by variances of logratios between a part of the first partial table from area highlighted by (/) and a part of the second partial table from area highlighted by (|) – A_2. The second group of variances increasing the covariance between coordinates are connected to logratios between parts from (\) and (−) areas – B_2. The covariance is reduced by variances of logratios between parts from (/) and (−) area – C_2 or two parts from (\) and (|) – D_2.

Theorem 3.4 *Consider coordinates of the independence table z_k^r for $k = 1, \ldots, I-1$ and z_l^c for $l = 1, \ldots, J-1$, then their variances are*

$$\begin{aligned} \operatorname{var}(z_k^r) \;=\; & K \sum_{i'=k+1}^{I} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{kj}}{x_{i'j'}}\right) - \frac{K}{2}(I-k) \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{kj}}{x_{kj'}}\right) - \\ & - \frac{K}{2(I-k)} \sum_{i,i'=k+1}^{I} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right), \end{aligned} \tag{19}$$

where $K = \frac{1}{J(I-k+1)}$, for balances between rows, and

$$\begin{aligned} \operatorname{var}(z_l^c) \;=\; & K \sum_{i,i'=I}^{I} \sum_{j'=l+1}^{J} \operatorname{var}\left(\ln \frac{x_{il}}{x_{i'j'}}\right) - \frac{K}{2}(J-l) \sum_{i,i'=1}^{I} \operatorname{var}\left(\ln \frac{x_{il}}{x_{i'l}}\right) - \\ & - \frac{K}{2(J-l)} \sum_{i,i'=1}^{I} \sum_{j,j'=l+1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right), \end{aligned} \tag{20}$$

where $K = \frac{1}{I(J-l+1)}$, for balances between columns.

The variances of these coordinates are enlarged by variances of logratios between a part from the k-th row/l-th column and any part from the subsequent rows/columns. On the other hand, the variances of z_k^r and z_l^c are reduced by variances of logratios between parts from the same row/column.

According to relation (8) there are three main options how to get covariance between coordinates of the independence table, depending on concrete balances of interest. All these possible covariances are summarized in the following theorem.

Theorem 3.5 *Consider three coordinates of the independence table $z_{k_1}^r$, $z_{k_2}^r$ and z_k^r, for $k_1, k_2, k = 1, \ldots, I-1$, $k_1 \neq k_2$, computed using expression (3), and three coordinates $z_{l_1}^c$, $z_{l_2}^c$*

and z_l^c, *for* $l_1, l_2, l = 1, \ldots, J-1$, $l_1 \neq l_2$, *computed from (4). Then*

$$\begin{aligned}\operatorname{cov}(z_{k_1}^r, z_{k_2}^r) &= \frac{K}{(I-k_2)} \sum_{i'=k_2+1}^{I} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{k_1 j}}{x_{i'j'}}\right) + \frac{K}{(I-k_1)} \sum_{i=k_1+1}^{I} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{k_2 j'}}\right) - \\ &\quad - K \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{k_1 j}}{x_{k_2 j'}}\right) - \frac{K}{(I-k_1)(I-k_2)} \sum_{i=k_1+1}^{I} \sum_{i'=k_2+1}^{I} \sum_{j,j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right),\end{aligned} \tag{21}$$

where $K = \frac{1}{2J}\sqrt{\frac{(I-k_1)(I-k_2)}{(I-k_1+1)(I-k_2+1)}}$, *for row balances,*

$$\begin{aligned}\operatorname{cov}(z_{l_1}^c, z_{l_2}^c) &= \frac{K}{(J-l_2)} \sum_{i,i'=1}^{I} \sum_{j'=l_2+1}^{J} \operatorname{var}\left(\ln \frac{x_{il_1}}{x_{i'j'}}\right) + \frac{K}{(J-l_1)} \sum_{i,i'=1}^{I} \sum_{j=l_1+1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'l_2}}\right) - \\ &\quad - K \sum_{i,i'=1}^{I} \operatorname{var}\left(\ln \frac{x_{il_1}}{x_{i'l_2}}\right) - \frac{K}{(J-l_1)(J-l_2)} \sum_{i,i'=1}^{I} \sum_{j=l_1+1}^{J} \sum_{j'=l_2+1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right),\end{aligned} \tag{22}$$

where $K = \frac{1}{2I}\sqrt{\frac{(J-l_1)(J-l_2)}{(J-l_1+1)(J-l_2+1)}}$, *for column balances, and*

$$\begin{aligned}\operatorname{cov}(z_k^r, z_l^c) &= \frac{K}{(J-l)} \sum_{i'=1}^{I} \sum_{j=1}^{J} \sum_{j'=l+1}^{J} \operatorname{var}\left(\ln \frac{x_{kj}}{x_{i'j'}}\right) + \frac{K}{(I-k)} \sum_{i=1}^{I} \sum_{i'=k+1}^{I} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{il}}{x_{i'j'}}\right) - \\ &\quad - K \sum_{i'=1}^{I} \sum_{j=1}^{J} \operatorname{var}\left(\ln \frac{x_{kj}}{x_{i'l}}\right) - \frac{K}{(I-k)(J-l)} \sum_{i=k+1}^{I} \sum_{i'=1}^{I} \sum_{j=1}^{J} \sum_{j'=l+1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right),\end{aligned} \tag{23}$$

where $K = \frac{1}{2}\sqrt{\frac{(I-k)(J-l)}{IJ(I-k+1)(J-l+1)}}$, *between row and column balances.*

To complete the covariance structure of coordinates of the compositional table $\mathbf{x}$, covariances between coordinates of the interaction and independence tables are necessary. They are provided in the last theorem.

Theorem 3.6 *Consider coordinate of the interaction table* z_{rs}, *for* $r = 2, \ldots, I$ *and* $s = 2, \ldots, J$, *and two coordinates of the independence table,* z_k^r, *for* $k = 1, \ldots, I-1$, *and* z_l^c, *for* $l = 1, \ldots, J-1$. *Then for covariances between coordinates of the interaction and independence tables the following hold,*

$$\operatorname{cov}(z_{rs}, z_k^r) = K \cdot (A_3 - B_3), \tag{24}$$

where

$$\begin{aligned}A_3 &= \frac{1}{J(I-k)} \sum_{i=1}^{r-1} \sum_{i'=k+1}^{I} \sum_{j=1}^{s-1} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{i'j'}}\right) + \frac{s-1}{J} \sum_{i=1}^{r-1} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{is}}{x_{kj'}}\right) + \\ &\quad + \frac{r-1}{J} \sum_{j=1}^{s-1} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{rj}}{x_{kj'}}\right) + \frac{(r-1)(s-1)}{J(I-k)} \sum_{i'=k+1}^{I} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{rs}}{x_{i'j'}}\right),\end{aligned} \tag{25}$$

$$\begin{aligned}B_3 &= \frac{1}{J} \sum_{i=1}^{r-1} \sum_{j=1}^{s-1} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{ij}}{x_{kj'}}\right) + \frac{s-1}{J(I-k)} \sum_{i=1}^{r-1} \sum_{i'=k+1}^{I} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{is}}{x_{i'j'}}\right) + \\ &\quad + \frac{r-1}{J(I-k)} \sum_{i'=k+1}^{I} \sum_{j=1}^{s-1} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{rj}}{x_{i'j'}}\right) + \frac{(r-1)(s-1)}{J} \sum_{j'=1}^{J} \operatorname{var}\left(\ln \frac{x_{rs}}{x_{kj'}}\right),\end{aligned} \tag{26}$$

for $K = \frac{1}{2}\frac{1}{\sqrt{rs(r-1)(s-1)}}\sqrt{\frac{J(I-k)}{I-k+1}}$, *and*

$$\mathrm{cov}(z_{rs}, z_l^c) = K \cdot (A_4 - B_4), \tag{27}$$

where

$$\begin{aligned} A_4 &= \frac{1}{I(J-l)}\sum_{i=1}^{r-1}\sum_{i'=1}^{I}\sum_{j=1}^{s-1}\sum_{j'=l+1}^{J}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'j'}}\right) + \frac{s-1}{I}\sum_{i=1}^{r-1}\sum_{i'=1}^{I}\mathrm{var}\left(\ln\frac{x_{is}}{x_{i'l}}\right) + \\ &+\frac{r-1}{I}\sum_{i'=1}^{I}\sum_{j=1}^{s-1}\mathrm{var}\left(\ln\frac{x_{rj}}{x_{i'l}}\right) + \frac{(r-1)(s-1)}{I(J-l)}\sum_{i'=1}^{I}\sum_{j'=l+1}^{J}\mathrm{var}\left(\ln\frac{x_{rs}}{x_{i'j'}}\right), \end{aligned} \tag{28}$$

$$\begin{aligned} B_4 &= \frac{1}{I}\sum_{i=1}^{r-1}\sum_{i'=1}^{I}\sum_{j=1}^{s-1}\mathrm{var}\left(\ln\frac{x_{ij}}{x_{i'l}}\right) + \frac{s-1}{I(J-l)}\sum_{i=1}^{r-1}\sum_{i'=1}^{I}\sum_{j'=l+1}^{J}\mathrm{var}\left(\ln\frac{x_{is}}{x_{i'j'}}\right) + \\ &+\frac{r-1}{I(J-l)}\sum_{i'=1}^{I}\sum_{j=1}^{s-1}\sum_{j'=l+1}^{J}\mathrm{var}\left(\ln\frac{x_{rj}}{x_{i'j'}}\right) + \frac{(r-1)(s-1)}{I}\sum_{i'=1}^{I}\mathrm{var}\left(\ln\frac{x_{rs}}{x_{i'l}}\right), \end{aligned} \tag{29}$$

for $K = \frac{1}{2}\frac{1}{\sqrt{rs(r-1)(s-1)}}\sqrt{\frac{I(J-l)}{J-l+1}}$.

Proof: *The assertion of the theorem is a direct consequence of Proposition 3.1 and Equations (3), (4) and (5).*

◇

Similarly as for the case of interaction table, also the above results could be interpeted graphically. Because Theorems 3.4 and 3.5 represent a special case of balances, that were in detail analyzed in (Fišerová and Hron 2011), in Figure 3 we focus just on covariances, resulting from Theorem 3.6.

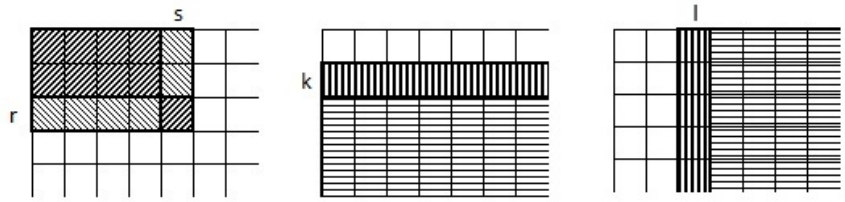

Figure 3: Covariance between a coordinate of the interaction table, z_{rs} (left), and coordinates of the independence table, z_k^r (middle) or z_l^c (right), is increased by variances of logratios between parts from areas (/) and (−), or (\) and (|), respectively, and reduced by variances of logratios between parts from areas (/) and (|), or (\) and (−), respectively.

4. Implications for 2 x 2 compositional tables

In practice, 2×2 compositional (and also contingency) tables represent a prominent special case that requires a special treatment (Fačevicová *et al.* 2014a; Agresti 2002). From Equations (3), (4) and (5) it is easy to see that for coordinate representation of the compositional table

$$\mathbf{x} = \mathcal{C}\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$$

it is sufficient to consider the following coordinates,

$$z_1^{ind} = \frac{1}{2}\ln\frac{x_{11}x_{12}}{x_{21}x_{22}}, \quad z_2^{ind} = \frac{1}{2}\ln\frac{x_{11}x_{21}}{x_{12}x_{22}} \quad \text{and} \quad z^{int} = \frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}}. \tag{30}$$

By applying the above theorems their covariance structure can be easily derived,

$$\begin{aligned}
\mathrm{var}(z^{int}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) + \mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) + \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right) + \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)\right.\\
&\quad \left. -\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right) - \mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)\right],\\
\mathrm{var}(z_1^{ind}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) + \mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right) + \mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right) + \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)\right.\\
&\quad \left. -\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) - \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right],\\
\mathrm{var}(z_2^{ind}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) + \mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right) + \mathrm{var}\left(\ln\frac{x_{21}}{x_{12}}\right) + \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right.\\
&\quad \left. -\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) - \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)\right],\\
\mathrm{cov}(z^{int}, z_1^{ind}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) - \mathrm{var}\left(\ln\frac{x_{22}}{x_{12}}\right)\right],\\
\mathrm{cov}(z^{int}, z_2^{ind}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) - \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right],\\
\mathrm{cov}(z_1^{ind}, z_2^{ind}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right) - \mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)\right].
\end{aligned}$$

Moreover, from the above covariance structure it is also interesting to see that coordinates (30) are uncorrelated (or even independent under the assumption of normality) if, and only if

$$\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) = \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right), \quad \mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) = \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right), \quad \mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right) = \mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right). \tag{31}$$

In other words, it means that zero covariances can be easily expressed in terms of logratio variances. Consequently, the above relations could be used, e.g., by designing simulation settings for 2×2 compositional tables using elements of the variation matrix as a source of elemental information in covariance structure of compositional tables.

Following Fačevicová *et al.* (2014a), it is possible to assign also another system of orthonormal coordinates to a 2×2 compositional table. Specificaly, we get

$$z_1^{ind} = \frac{1}{\sqrt{2}}\ln\frac{x_{12}}{x_{21}}, \quad z_2^{ind} = \frac{1}{\sqrt{2}}\ln\frac{x_{11}}{x_{22}}, \quad z^{int} = \frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}}, \tag{32}$$

for the interaction and independent tables, respectively, and the covariance structure changes as follows,

$$\begin{aligned}
\mathrm{var}(z^{int}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) + \mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) + \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right) + \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right.\\
&\quad \left. -\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right) - \mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)\right],\\
\mathrm{var}(z_1^{ind}) &= \frac{1}{2}\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right),\\
\mathrm{var}(z_2^{ind}) &= \frac{1}{2}\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right),\\
\mathrm{cov}(z^{int}, z_1^{ind}) &= \frac{1}{4\sqrt{2}}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) + \mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) - \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right) - \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right],\\
\mathrm{cov}(z^{int}, z_2^{ind}) &= \frac{1}{4\sqrt{2}}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) + \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right) - \mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) - \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right)\right],\\
\mathrm{cov}(z_1^{ind}, z_2^{ind}) &= \frac{1}{4}\left[\mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) + \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right) - \mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) - \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right)\right].
\end{aligned}$$

Now, although coordinates of the independent table are formed just by (scaled) logratios, the covariance structure becomes more complex than before. For example, coordinates (32) are *mutually* uncorrelated (independent) if, and only if

$$\mathrm{var}\left(\ln\frac{x_{11}}{x_{12}}\right) = \mathrm{var}\left(\ln\frac{x_{11}}{x_{21}}\right) = \mathrm{var}\left(\ln\frac{x_{12}}{x_{22}}\right) = \mathrm{var}\left(\ln\frac{x_{21}}{x_{22}}\right). \tag{33}$$

In other words, it means that $\mathrm{var}\left(\ln\frac{x_{12}}{x_{21}}\right)$ and $\mathrm{var}\left(\ln\frac{x_{11}}{x_{22}}\right)$ are influential just for variances of coordinates z_1^{ind}, z_2^{ind}, z^{int}, forming also natural constraints for their possible values.

5. Numerical example

To illustrate the presented theoretical outputs, let us consider the sample of eighteen 2×3 compositional tables, each reflecting population structure in European country according to age and BMI index ((weight in kg)/(height in m)^2), with values $25-44, 45-64, 65-84$ and under- or normal weight and overweight or obesity, respectively. The data set is an aggregated version of data from Fačevicová *et al.* (2014b). Table 5 shows an example of compositional table from the sample.

Table 1: Structure of population in Austria in 2008 according to age and BMI index (in proportions).

AUT	$25-44$	$45-64$	$65-84$
under or normal	0.249	0.144	0.074
over or obesity	0.171	0.221	0.140

Firstly, each table from the sample has been expressed in coordinates and, consequently, their descriptive statistics has been calculated. The sample mean is

$$\bar{\mathbf{z}} = (0.409, 0.294, -0.450, 0.578, 0.637),$$

but for our purposes the covariance structure of the sample is of primary interest. The variation matrix (6), as a source of elemental information in compositional tables, equals

$$\mathbf{T} = \begin{pmatrix} 0 & 0.037 & 0.083 & 0.024 & 0.030 & 0.069 \\ 0.037 & 0 & 0.030 & 0.077 & 0.050 & 0.051 \\ 0.083 & 0.030 & 0 & 0.127 & 0.098 & 0.065 \\ 0.024 & 0.077 & 0.127 & 0 & 0.019 & 0.078 \\ 0.030 & 0.050 & 0.098 & 0.019 & 0 & 0.040 \\ 0.069 & 0.051 & 0.065 & 0.078 & 0.040 & 0 \end{pmatrix}.$$

For example, using this matrix and equation (9), variance of the first coordinate of the interaction table, z_{22}, can be obtained as

$$\begin{aligned} \mathrm{var}(z_{rs}) &= \tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{11}}{x_{21}}\right) + \tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{11}}{x_{12}}\right) + \tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{21}}{x_{22}}\right) + \tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{12}}{x_{22}}\right) \\ &\quad -\tfrac{1}{8}\mathrm{var}\left(\ln\tfrac{x_{11}}{x_{11}}\right) - \tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{11}}{x_{22}}\right) - \tfrac{1}{8}\mathrm{var}\left(\ln\tfrac{x_{21}}{x_{21}}\right) - \tfrac{1}{8}\mathrm{var}\left(\ln\tfrac{x_{12}}{x_{12}}\right) \\ &\quad -\tfrac{1}{4}\mathrm{var}\left(\ln\tfrac{x_{12}}{x_{21}}\right) \\ &= \tfrac{1}{4}t_{14} + \tfrac{1}{4}t_{12} + \tfrac{1}{4}t_{45} + \tfrac{1}{4}t_{25} - \tfrac{1}{8}t_{11} - \tfrac{1}{4}t_{15} - \tfrac{1}{8}t_{44} - \tfrac{1}{8}t_{22} - \tfrac{1}{4}t_{24} \\ &= 0.0057. \end{aligned}$$

By comparing with the corresponding elements of the variation matrix we can conclude that none of logratios contributes exceptionally (in the positive sense) to variability of the coordinate. In the negative sense, the logratio ln(*underweight or normal weight in age 45-64/overweight or obesity in age 25-44*) shows a dominant effect. Similarly, also other variances and covariances can be derived (and further analysed for structural patterns), resulting in a covariance matrix

$$\mathrm{var}(\mathbf{z}) = \begin{pmatrix} 0.006 & 0.003 & -0.010 & 0.007 & 0.004 \\ 0.003 & 0.013 & -0.007 & -0.004 & -0.0005 \\ -0.010 & -0.007 & 0.051 & -0.021 & -0.012 \\ 0.007 & -0.004 & -0.021 & 0.051 & 0.026 \\ 0.004 & -0.0005 & -0.012 & 0.026 & 0.026 \end{pmatrix}.$$

Finally, note that by considering both markedly nonzero means of coordinates of the interaction table (first two elements of the vector $\overline{\mathbf{z}}$) and their corresponding small variances, we can conclude that, based on the considered sample, age and BMI index are not dependent.

6. Discussion

Recent experience with orthonormal coordinates for compositional data (Reimann, Filzmoser, Fabian, Hron, Birke, Demetriades, Dinelli, and Ladenberger 2012; Filzmoser and Walczak 2014) shows clearly the necessity of their better understanding in terms of logratios, which could be achieved also by decomposing the corresponding covariance structure. This is even more crucial for compositional tables, where both balances and coordinates with odds ratio interpretation are involved. Obviously, due to complex character of the above formulas for covariance structure in compositional tables, they will be rather rarely used for practical computations. Therefore, the formulas are also accompanied with comments and graphical illustrations to better understand their logical structure that is much more important for the aim of the paper. Consequently, similarly as for the case of balances (Fišerová and Hron 2011), we are convinced that decomposition of variances and covariances as linear combinations of logratio variances enhances interpretability of coordinates of compositional tables, using logratios as the primary source of information in compositional data.

Acknowledgments Authors gratefully acknowledge the support of the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic) and the grant PrF_2014_028 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

References

Agresti A (2002). *Categorial Data Analysis (2 ed.)*. J. Wiley & Sons, New York.

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Egozcue J, Díaz-Barrero J, Pawlowsky-Glahn V (2008). "Simplicial Geometry for Compositional Data." In *Daunis-i-Estadella J, Martín-Fernández JA (eds) Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop*. University of Girona, Spain.

Egozcue J, Pawlowsky-Glahn V (2005). "Groups of Parts and Their Balances in Compositional Data Analysis." *Mathematical Geology*, **37**, 795–828.

Egozcue J, Pawlowsky-Glahn V (2006). "Simplicial Geometry for Compositional Data." In *Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional data analysis in the geosciences: From theory to practice*. Geological Society, London.

Egozcue J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology*, **35**, 279–300.

Egozcue J, Pawlowsky-Glahn V, Templ M, Hron K (2014). "Independence in Contingency Tables using Simplicial Geometry." *Communications in Statistics - Theory and Methods, to appear.*

Fačevicová K, Hron K, Todorov V, Guo D, Templ M (2014a). "Logratio Approach to Statistical Analysis of 2×2 Compositional Tables." *Journal of Applied Statistics*, **41**, 944–958.

Fačevicová K, Hron K, Todorov V, Templ M (2014b). "Compositional Tables Analysis in Coordinates." *Submitted.*

Filzmoser P, Walczak B (2014). "What Can Go Wrong at the Data Normalization Step for Identification of Biomarkers?" *Journal of Chromatography A*, **1362**, 194–205.

Fišerová E, Hron K (2011). "On Interpretation of Orthonormal Coordinates for Compositional Data." *Math Geol*, **43**, 455–468.

Pawlowsky-Glahn V, Buccianti A (2011). *Compositional Data Analysis: Theory and Applications.* Wiley, Chichester.

Reimann C, Filzmoser P, Fabian K, Hron K, Birke M, Demetriades A, Dinelli E, Ladenberger A (2012). "The Concept of Compositional Data Analysis in Practice – Total Major Element Concentrations in Agricultural and Grazing Land Soils of Europe." *Science of the Total Enviroment*, **426**, 196–210.

Affiliation:

Kamila Fačevicová
Department of Mathematical Analysis and Applications of Mathematics, Department of Geoinformatics
Faculty of Science, Palack? University
Olomouc, Czech Republic
E-mail: kamila.facevicova@gmail.com

Austrian Journal of Statistics
October 2015, Volume 44, 45–68.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i3.36

A New Class of Generalized Modified Weibull Distribution with Applications

Broderick O. Oluyede
Georgia Southern University

Shujiao Huang
Georgia Southern University

Tiantian Yang
Georgia Southern University

Abstract

A new five parameter gamma-generalized modified Weibull (GGMW) distribution which includes exponential, Rayleigh, Weibull, modified Weibull, gamma-modified Weibull, gamma-modified Rayleigh, gamma-modified exponential, gamma-Weibull, gamma-Rayleigh, gamma-linear failure rate and gamma-exponential distributions as special cases is proposed and studied. Some mathematical properties of the new class of distributions including hazard function, quantile function, moments, distribution of the order statistics and Rényi entropy are presented. Maximum likelihood estimation technique is used to estimate the model parameters and applications to real datasets in order to illustrate the usefulness of the proposed class of models are presented.

Keywords: Gamma distribution, Modified Weibull distribution, Maximum likelihood estimation.

1. Introduction

Weibull distribution has been widely used for modeling data in a wide variety of areas including reliability, engineering, stochastic processes, survival analysis and renewal theory. In this paper, we present and study the mathematical properties of the gamma-generalized modified Weibull distribution. This class of distributions is flexible in accommodating all forms of hazard rate functions and contains several well known and new sub-models such as Weibull, Rayleigh, exponential, modified Weibull, gamma-modified Weibull, gamma-modified exponential, gamma-Weibull, gamma-Rayleigh, gamma-linear failure rate, gamma-extreme value, gamma-additive exponential and gamma-exponential distributions.

There are several extensions of the Weibull distribution and its sub-models including the exponentiated Weibull (Mudholkar, Srivastava, and Kollia 1996), which is a special case of the beta Weibull distribution proposed by (Lee, Famoye, and Olumolade 2007), generalized Rayleigh (Kundu and Rakab 2005), exponentiated exponential (Gupta and Kundu 1999), (Gupta and Kundu 2001), modified Weibull (Mudholkar, Srivastava, and Friemer 1995), exponentiated modified Weibull (Sarhan and Zaindin 2009), and a host of other distributions, some of which are presented in section 2 of this paper. Additional generalizations of Weibull distribution include (Famoye, Lee, and Olumolade 2005) where the authors discussed and presented results on the beta-Weibull distribution. (Nadarajah 2005) presented results on the modified Weibull

distribution. A host of researchers have also developed several parameter Weibull, modified Weibull and flexible Weibull distributions over the years. The two parameter Weibull extensions include (Bebbington, Lai, and Zitikis 2007), (Zhang and Xie 2011). The three parameter Weibull extensions include (Marshall and Olkin 1997), (Xie, Tang, and Goh 2002), (Nadarajah and Kotz 2005). Some of these extensions enable the accommodation of bathtub shape hazard rate function. (Carrasco, Ortega, and Cordeiro 2008) generalized the modified Weibull distribution of (Lai, Moore, and Xie 1998) to obtain the exponentiated modified Weibull distribution. The four parameter generalizations include the additive Weibull distribution of (Xie and Lai 1995), modified Weibull (Sarhan and Zaindin 2009), beta-Weibull proposed by (Famoye *et al.* 2005) and Kumaraswamy Weibull by (Cordeiro, Ortega, and Nadarajah 2010). The five parameter modified Weibull distribution include those introduced by (Phani 1987), beta modified Weibull by (Silva, Ortega, and Cordeiro 2010) and (Nadarajah, Cordeiro, and Ortega 2011). Additional results on the generalization of the Weibull distribution include work by (Singha, Jain, and Kumar 2012), as well as (Almalki and Yuan 2013) where results on a new modified Weibull distribution was presented. (Barlow and Campo 1975) discussed total time on test processes with application to failure data analysis. (Choudhury 2005) presented moments of the exponentiated Weibull distribution. The exponentiated Weibull distribution was also studied by (Nassar and Eissa 2003). (Haupt and Schabe 1992) presented a model for bathtub shaped failure rate function. (Hjorth 1980) studied a reliability function with increasing, decreasing and bathtub shaped failure rate functions, and (Rajarshi and Rajarshi 1988) gave a comprehensive review of bathtub shaped distributions.

For any continuous baseline cdf $F(x)$, and $x \in \mathbf{R}$, (Zografos and Balakrishnan 2009) defined the distribution (when $\psi = 1$ in equation (1)) with pdf $g(x)$ and cdf $G(x)$ (for $\delta > 0$) as follows:

$$g(x) = \frac{1}{\Gamma(\delta)\psi^{\delta}}[-\log(\overline{F}(x))]^{\delta-1}(1-F(x))^{1/\psi-1}f(x), \tag{1}$$

and

$$G(x) = \frac{1}{\Gamma(\delta)\psi^{\delta}}\int_0^{-\log(\overline{F}(x))} t^{\delta-1}e^{-t/\psi}dt = \frac{\gamma(\delta, -\psi^{-1}\log(\overline{F}(x)))}{\Gamma(\delta)}, \tag{2}$$

respectively, where $g(x) = dG(x)/dx$, $\Gamma(\delta) = \int_0^{\infty} t^{\delta-1}e^{-t}dt$ is the gamma function, and $\gamma(z,\delta) = \int_0^z t^{\delta-1}e^{-t}dt$ is the incomplete gamma function. The corresponding hazard rate function (hrf) is

$$h_G(x) = \frac{[-\log(1-F(x))]^{\delta-1}f(x)(1-F(x))^{1/\psi-1}}{\psi^{\delta}(\Gamma(\delta) - \gamma(-\psi^{-1}\log(1-F(x)),\delta))}. \tag{3}$$

When $\psi = 1$, this distribution is referred to as the ZB-G family of distributions. Also, (when $\psi = 1$), (Ristić and Balakrishnan 2011) proposed an alternative gamma-generator defined by the cdf and pdf

$$G_2(x) = 1 - \frac{1}{\Gamma(\delta)\psi^{\delta}}\int_0^{-\log F(x)} t^{\delta-1}e^{-t/\psi}dt, \quad x \in \mathbf{R},\ \delta > 0, \tag{4}$$

and

$$g_2(x) = \frac{1}{\Gamma(\delta)\psi^{\delta}}[-\log(F(x))]^{\delta-1}(F(x))^{1/\psi-1}f(x), \tag{5}$$

respectively. Note that if $\psi = 1$ and $\delta = n+1$, in equations (1) and (2), we obtain the cdf and pdf of the upper record values U given by

$$G_U(u) = \frac{1}{n!}\int_0^{-\log(1-F(u))} y^n e^{-y}dy, \tag{6}$$

and

$$g_U(u) = f(u)[-\log(1-F(u))]^n/n!. \tag{7}$$

Similarly, from equations (4) and (5), the pdf of the lower record values T is given by

$$g_{_L}(t) = f(t)[-\log(F(t))]^n/n!. \tag{8}$$

In this paper, we will consider and present a generalization of the generalized modified Weibull distribution via the family of distributions given in equation (5). (Zografos and Balakrishnan 2009) motivated the ZB-G model as follows. Let $X_{(1)}, X_{(2)}, ..., X_{(n)}$ be upper record values from a sequence of independent and identically distributed (i.i.d.) random variables from a population with pdf $f(x)$. Then, the pdf of the n^{th} upper record value is given by equation (1) when $\psi = 1$. A logarithmic transformation of the parent distribution F transforms the random variable X with density (1) to a gamma distribution. That is, if X has the density (1), then the random variable $Y = -\log[1-F(X)]$ has a gamma distribution $GAM(\delta;1)$ with density $k(y;\delta) = \frac{1}{\Gamma(\delta)}y^{\delta-1}e^{-y}$, $y > 0$. The opposite is also true, if Y has a gamma $GAM(\delta;1)$ distribution, then the random variable $X = G^{-1}(1 - e^{-Y})$ has a ZB-G distribution. In addition to the motivations provided by (Zografos and Balakrishnan 2009), we are interested in the generalization of the generalized modified Weibull distribution via the gamma-generator and establishing the relationship between weighted distributions and equations (1) and (5), respectively.

Weighted distributions applies to a variety of areas and provides an approach to dealing with model specification and data interpretation problems. It adjusts the probabilities of actual occurrence of events to arrive at a specification of the probabilities when those events are recorded. (Fisher 1934) introduced the concept of weighted distribution, in order to study the effect of ascertainment upon estimation of frequencies. (Patil and Rao 1978) used weighted distribution as stochastic models in the study of harvesting and predation. (Rao 1965) unified concept of weighted distribution and use it to identify various sampling situations. The usefulness and applications of weighted distribution to biased samples in various areas including medicine, ecology, reliability, and branching processes can also be seen in (Nanda and Jain 1999), (Gupta and Keating 1985), (Oluyede 1999) and in references therein. Let Y be a non-negative random variable with its natural pdf $f(y;\underline{\theta})$, where $\underline{\theta}$ is a vector of parameters, then the pdf of the weighted random variable Y^w is given by:

$$f^w(y;\underline{\theta},\underline{\beta}) = \frac{w(y,\underline{\beta})f(y;\underline{\theta})}{\omega}, \tag{9}$$

where the weight function $w(y,\underline{\beta})$ is a non-negative function, that may depend on the vector of parameters $\underline{\beta}$, and $0 < \omega = E(w(Y,\underline{\beta})) < \infty$ is a normalizing constant. In general, consider the weight function $w(y)$ defined as follows:

$$w(y) = y^k e^{ly} F^i(y)\overline{F}^j(y). \tag{10}$$

Setting $k = 0$; $k = j = i = 0$; $l = i = j = 0$; $k = l = 0$; $i \to i - 1$; $j = n - i$; $k = l = i = 0$ and $k = l = j = 0$ in this weight function, one at a time, implies probability weighted moments, moment-generating functions, moments, order statistics, proportional hazards and proportional reversed hazards, respectively, where $F(y) = P(Y \leq y)$ and $\overline{F}(y) = 1 - F(y)$. If $w(y) = y$, then $Y^* = Y^w$ is called the size-biased version of Y.

(Ristić and Balakrishnan 2011) provided motivations for the family of distributions given in equation (4) when $\psi = 1$, that is for $n \in \mathbf{N}$, equation (4) is the pdf of the n^{th} lower record value of a sequence of i.i.d. variables from a population with density $f(x)$. (Ristić and Balakrishnan 2011) used the exponentiated exponential (EE) distribution with cdf $F(x) = (1 - e^{-\beta x})^\alpha$, where $\alpha > 0$ and $\beta > 0$, to obtained and study the gamma-exponentiated exponential (GEE) model. See references therein for additional results on the GEE model. In this note, we obtain a natural extension of the generalized modified Weibull distribution, which we refer to as gamma-generalized modified Weibull (GGMW) distribution.

In section 2, some basic results, the gamma-generalized modified Weibull (GGMW) distribution, series expansion and its sub-models, quantile function, hazard and reverse hazard

functions are presented. Moments and moment generating function are given in section 3. Section 4 contains some additional useful results on the distribution of order statistics and Rényi entropy. In section 5, results on the estimation of the parameters of the GGMW distribution via the method of maximum likelihood are presented. Applications are given in section 6, and concluding remarks in section 7.

2. GGMW distribution, series expansion and sub-models

In this section, the GGMW distribution and some of its sub-models are presented. First consider the generalized modified Weibull (GMW) distribution (Sarhan and Zaindin 2009) given by

$$F_{GMW}(x,\alpha,\beta,\theta,\lambda) = 1-\exp(-\alpha x-\beta x^{\theta}e^{\lambda x}), \quad x\geq 0,\ \alpha,\ \beta,\ \theta,\ \lambda\geq 0. \tag{11}$$

We note that in (Sarhan and Zaindin 2009) paper, the parameter λ was taken to be zero. The parameters α and β control the scale of the distribution, θ controls the shape, whereas λ can be considered to be an accelerating factor in the imperfection time and a factor of fragility in the survival of the individual as time increases. By inserting the GMW distribution in equation (4), the survival function $\overline{G}_{GGMW}(x) = 1-G_{GGMW}(x)$ of the GGMW distribution is obtained as follows:

$$\begin{aligned}\overline{G}_{GGMW}(x) &= \frac{1}{\Gamma(\delta)\psi^{\delta}}\int_0^{-\log(1-e^{-\alpha x-\beta x^{\theta}e^{\lambda x}})} t^{\delta-1}e^{-t/\psi}dt\\ &= \frac{\gamma(-\psi^{-1}\log(1-e^{-\alpha x-\beta x^{\theta}e^{\lambda x}}),\delta)}{\Gamma(\delta)},\end{aligned} \tag{12}$$

where $x>0$, $\alpha,\ \beta,\ \theta,\ \lambda\geq 0$, $\delta>0$, $\psi>0$, and $\gamma(x,\delta)=\int_0^x t^{\delta-1}e^{-t}dt$ is the lower incomplete gamma function. The corresponding pdf is given by

$$\begin{aligned}g_{GGMW}(x) &= \frac{1}{\Gamma(\delta)\psi^{\delta}}[-\log(1-e^{-\alpha x-\beta x^{\theta}e^{\lambda x}})]^{\delta-1}\\ &\times\ (\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x])e^{-\alpha x-\beta x^{\theta}e^{\lambda x}}\\ &\times\ [1-e^{-\alpha x-\beta x^{\theta}e^{\lambda x}}]^{(1/\psi)-1}.\end{aligned} \tag{13}$$

If $F(x)=[1-e^{-\alpha x^{\eta}-\beta x^{\theta}e^{\lambda x}}]^{\phi}$, then the corresponding generalized gamma-generalized modified Weibull pdf is given by

$$\begin{aligned}g_{GGMW}(x) &= \frac{\phi}{\Gamma(\delta)\psi^{\delta}}[-\log(1-e^{-\alpha x^{\eta}-\beta x^{\theta}e^{\lambda x}})^{\phi}]^{\delta-1}\\ &\times\ (\alpha\eta x^{\eta-1}+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x])e^{-\alpha x^{\eta}-\beta x^{\theta}e^{\lambda x}}\\ &\times\ [1-e^{-\alpha x^{\eta}-\beta x^{\theta}e^{\lambda x}}]^{\phi+(1/\psi)-2}.\end{aligned} \tag{14}$$

In this note, we take $\phi=\eta=\psi=1$. The pdf in equation (14) is now given by

$$\begin{aligned}g_{GGMW}(x) &= \frac{1}{\Gamma(\delta)}[-\log(1-e^{-\alpha x-\beta x^{\theta}e^{\lambda x}})]^{\delta-1}\\ &\times\ (\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x])e^{-\alpha x-\beta x^{\theta}e^{\lambda x}}.\end{aligned} \tag{15}$$

If a random variable X has the GGMW density given in equation (15), we write $X\sim GGMW(\alpha,\beta,\theta,\lambda,\delta)$. The parameter δ is an extra shape parameter in the GGMW distribution. Let $y=e^{-\alpha x-\beta x^{\theta}e^{\lambda x}}$, $0<y<1$, $\alpha,\beta,\theta,\delta>0$, and $\lambda\geq 0$, then using the series representation $-\log(1-y)=\sum_{i=0}^{\infty}\frac{y^{i+1}}{i+1}$, we have

$$\Big[-\log(1-y)\Big]^{\delta-1} = y^{\delta-1}\Big[\sum_{m=0}^{\infty}\binom{\delta-1}{m}y^{m}\Big(\sum_{s=0}^{\infty}\frac{y^{s}}{s+2}\Big)^{m}\Big].$$

Applying the result on power series raised to a positive integer, with $a_s = (s+2)^{-1}$, that is,

$$\left(\sum_{s=0}^{\infty} a_s y^s\right)^m = \sum_{s=0}^{\infty} b_{s,m} y^s, \tag{16}$$

where $b_{s,m} = (sa_0)^{-1}\sum_{l=1}^{s}[m(l+1)-s]a_l b_{s-l,m}$, and $b_{0,m} = a_0^m$, (Gradshteyn and Ryzhik 2000), the GGMW pdf can be written as

$$\begin{aligned} g_{GGMW}(x) &= \frac{1}{\Gamma(\delta)}\sum_{m=0}^{\infty}\sum_{s=0}^{\infty}\binom{\delta-1}{m} b_{s,m} y^{m+s+\delta}(\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x]) \\ &= \frac{1}{\Gamma(\delta)}\sum_{m=0}^{\infty}\sum_{s=0}^{\infty}\binom{\delta-1}{m} b_{s,m} e^{-\alpha(m+s+\delta)x-\beta(m+s+\delta)x^{\theta}e^{\lambda x}} \\ &\times \frac{m+s+\delta}{m+s+\delta}(\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x]) \\ &= \sum_{m=0}^{\infty}\sum_{s=0}^{\infty}\binom{\delta-1}{m}\frac{b_{s,m}}{\Gamma(\delta)(m+s+\delta)} g_*(x;\alpha(m+s+\delta),\beta(m+s+\delta),\theta,\lambda), \end{aligned}$$

where $g_*(x;\alpha(m+s+\delta),\beta(m+s+\delta),\theta,\lambda)$ is the generalized modified Weibull pdf with parameters $\alpha(m+s+\delta)>0$, $\beta(m+s+\delta)>0$, $\theta>0$, and $\lambda \geq 0$. Let $C = \{(m,s) \in \mathbf{Z}_+^2\}$, then the weights in the GGMW pdf above are

$$w_\nu = \binom{\delta-1}{m}\frac{b_{s,m}}{(m+s+\delta)\Gamma(\delta)},$$

and

$$g_{GGMW}(x) = \sum_{\nu\in C} w_\nu g_*(x;\alpha(m+s+\delta),\beta(m+s+\delta),\theta,\lambda), \tag{17}$$

for $x>0$, $\delta>0, \alpha(m+s+\delta), \beta(m+s+\delta), \theta>0$, and $\lambda \geq 0$. It follows therefore that the GGMW density is linear combination of the generalized modified Weibull (GMW) densities. The statistical and mathematical properties of the GGMW distribution can be readily obtained from those of the generalized modified Weibull distribution.

For the convergence of equations (16) and (17), as well as elsewhere in this paper, note that for $\delta>0$,

$$[-\log(1-y)]^{\delta-1} = \left[y\left(1+y\sum_{s=0}^{\infty}\frac{y^s}{s+2}\right)\right]^{\delta-1}$$

so that

$$\left[1+y\sum_{k=0}^{\infty}\frac{y^k}{k+2}\right]^{\delta-1} = \sum_{k=0}^{\infty}\binom{\delta-1}{k} y^k\left(\sum_{s=0}^{\infty}\frac{y^s}{s+2}\right)^k$$

is convergent if and only if $0 < \left(y\sum_{k=0}^{\infty}\frac{y^k}{k+2}\right)^k < 1\ \forall y \in (0,1)$, since $0 < y = e^{-\alpha x-\beta x^{\theta}e^{\lambda x}} < 1$, for $x>0$, $\alpha,\beta,\theta>0$, and $\lambda \geq 0$. Now, $y\sum_{k=0}^{\infty}\frac{y^k}{k+2} = \frac{-\log(1-y)}{y} - 1$, so we must have $0 < \frac{-\log(1-y)}{y} - 1 < 1$. This leads to $1-y > \exp(-2y)$, and on the other hand $\exp(-y) = \sum_{k=0}^{\infty}\frac{(-1)^k y^k}{k!} > 1-y$. Thus, we have the system of inequalities $1-y > \exp(-2y)$ and $\exp(-y) > 1-y$, which is satisfied $\forall y \in (0, 0.7968)$. The implication here is that the inequality $0 < \left(y\sum_{k=0}^{\infty}\frac{y^k}{k+2}\right)^k < 1$ is not valid for all values of $0 < y = e^{-\alpha x-\beta x^{\theta}e^{\lambda x}} < 1$, and equations (16) and (17), and elsewhere in this paper are convergent only $\forall y \in (0, 0.7968)$. The series in equations (16) and (17), and elsewhere in this paper are not valid for all values of $0 <$

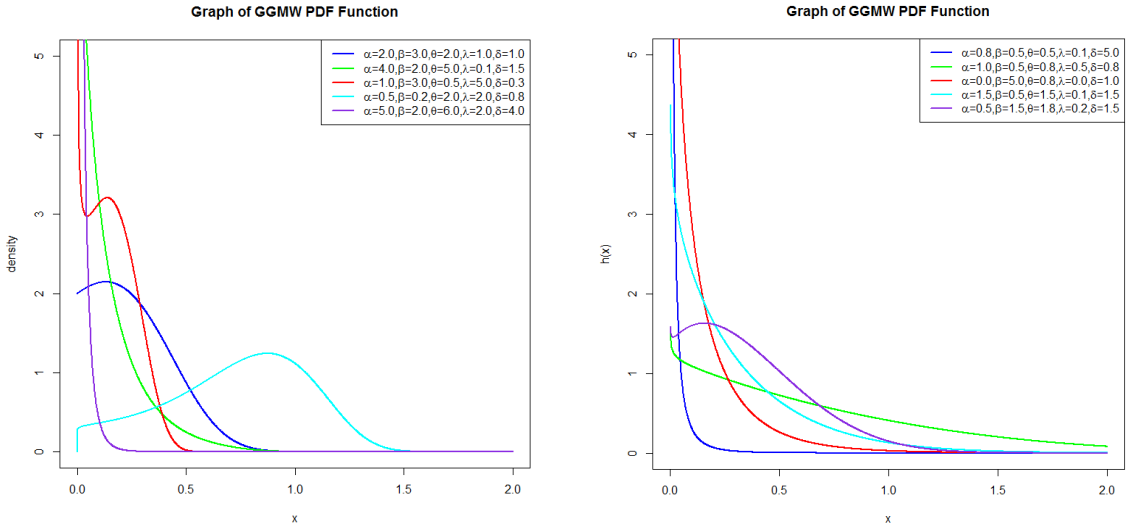

Figure 1: Graphs of GGMW pdf

$y = e^{-\alpha x - \beta x^{\theta} e^{\lambda x}} < 1$, but are convergent $\forall y \in (0, 0.7968)$, and not valid (convergent) for $y > 0.7986$.

Note that in general, $g_{GGMW}(x)$ is a weighted pdf with the weight function

$$w(x) = [-\log(1 - F(x))]^{\delta - 1}[1 - F(x)]^{\frac{1}{\psi} - 1}, \tag{18}$$

that is,

$$\begin{aligned} g_{GGMW}(x) &= \frac{[-\log(1 - F(x))]^{\delta - 1}[1 - F(x)]^{\frac{1}{\psi} - 1}}{\psi^{\delta}\Gamma(\delta)} f(x) \\ &= \frac{w(x)f(x)}{E_F(w(X))}, \end{aligned} \tag{19}$$

where $0 < E_F\{[-\log(1-F(x))]^{\delta-1}[1-F(x)]^{\frac{1}{\psi}-1}\} = \psi^{\delta}\Gamma(\delta) < \infty$, is the normalizing constant. Graphs of the pdf of GGMW distribution are given in the Figure 1 for selected values of the parameters. The plots show that the GGMW pdf can be decreasing or right skewed among several other possible shapes as seen in Figure 1. The distribution has positive asymmetry.

2.1. Quantile function

The quantile function of the GGMW distribution is given by the solution of the nonlinear equation

$$\frac{\gamma(-\log[1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}}], \delta)}{\Gamma(\delta)} = 1 - u. \tag{20}$$

That is, $-\log[1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}}] = \gamma^{-1}((1-u)\Gamma(\delta), \delta)$ and

$$\alpha x + \beta x^{\theta} e^{\lambda x} = -\log(1 - \exp(-\gamma^{-1}((1-u)\Gamma(\delta), \delta))). \tag{21}$$

We can simulate from the GGMW by solving the nonlinear equation

$$\alpha x + \beta x^{\theta} e^{\lambda x} + \log(1 - \exp(-\gamma^{-1}((1-u)\Gamma(\delta), \delta))) = 0, \tag{22}$$

where u is a uniformly distributed random variable on the interval $[0, 1]$. The inverse incomplete gamma function can be implemented by using numerical methods. Consequently, random numbers can be generated based the equation above. Table 1 lists the quantile for selected parameter values of the GGMW distribution.

Table 1: GGMW quantile for selected values

	$(\alpha, \beta, \theta, \lambda, \delta)$				
u	(1,1,1,1,1)	(2,1,2,1,1)	(6,4,3,6,1)	(5,3,3,5,6)	(0.1,0.3,0.4,0.2,0.3)
0.1	0.05132855	0.0512954	0.01755608	0.00001875	1.20674200
0.2	0.1056817	0.1053998	0.03714788	0.00007372	2.59472200
0.3	0.1637671	0.1627524	0.05924798	0.00018145	3.82692100
0.4	0.226598	0.2240198	0.08447059	0.00037058	4.95229000
0.5	0.2957024	0.2902609	0.11359275	0.00069065	6.01905300
0.6	0.3735554	0.3632644	0.14752769	0.00123419	7.07161000
0.7	0.4646056	0.4463389	0.18721167	0.00219605	8.16167800
0.8	0.5783069	0.5466338	0.23367423	0.00407477	9.37370000
0.9	0.7424909	0.6853097	0.29044828	0.00874208	10.92629700

2.2. Some sub-models of the GGMW distribution

The proposed model has several new and well known sub-models. Some of the sub-models of the GGMW distribution are listed in Table 2. They include the gamma-generalized modified Rayleigh (GGMR), gamma-generalized modified exponential (GGME), gamma-modified Weibull (GMW), gamma-modified exponential (GME), gamma-additive exponential (GAE), gamma-extreme value (GEV), gamma-Weibull (GW), modified Weibull (MW), Sardin and Zaindin modified Weibull (S-ZMW), modified Rayleigh (MR), modified exponential (ME), gamma-linear failure rate (GLFR), linear failure rate (LFR), extreme value (EV), Weibull (W) and exponential (E) distributions.

2.3. Hazard and reverse hazard functions

In this section, we present the hazard and reverse hazard functions, as well as graphs of the hazard function for selected values of the model parameters. Let X be a continuous random variable with distribution function G, and probability density function (pdf) g, then the hazard function, reverse hazard function and mean residual life functions are given by $h_G(x) = g(x)/\overline{G}(x)$, $\tau_G(x) = g(x)/G(x)$, and $\delta_G(x) = \int_x^\infty \overline{G}(u)du/\overline{G}(x)$, respectively. The functions $\lambda_G(x)$, $\delta_G(x)$, and $\overline{G}(x)$ are equivalent. (Shaked and Shanthikumar 1994). The hazard and reverse hazard functions are of the GGMW distribution are given by

$$h_{_G}(x) = \frac{\{-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})\}^{\delta-1} e^{-\alpha x-\beta x^\theta e^{\lambda x}}(\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x])}{\gamma(-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}}),\delta)}, \tag{23}$$

and

$$\tau_{_G}(x) = \frac{\{-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})\}^{\delta-1} e^{-\alpha x-\beta x^\theta e^{\lambda x}}(\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x])}{\Gamma(\delta)-\gamma(-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}}),\delta)}, \tag{24}$$

respectively. Plots of the hazard rate function for different combinations of the parameter values are given in Figure 2. The plot shows various shapes including monotonically increasing, monotonically increasing and bathtub shapes for five combinations of the values of the parameters. This flexibility makes the GGMW hazard rate function suitable for both monotonic and non-monotonic empirical hazard behaviors that are likely to be encountered in real life situations.

3. Moments and moment generating function

In this section, we obtain moments and moment generating function of the GGMW distribution. Let $X \sim GGMW(\alpha, \beta, \theta, \lambda, \delta)$, and $Y \sim GMW(\alpha, \beta, \theta, \lambda)$. Note that the r^{th} moment of the random variable Y is obtained as follows. By Taylor series expansion of the functions

Table 2: Sub-models of the gamma generalized modified Weibull distribution

Model	α	β	θ	λ	δ	$G(x)$	*Reference*
GGMR	-	-	2	-	-	$\frac{\gamma(-\log[1-e^{-\alpha x-\beta x^2 e^{\lambda x}}],\delta)}{\Gamma(\delta)}$	New
GGME	-	-	1	-	-	$\frac{\gamma(-\log[1-e^{-\alpha x-\beta x e^{\lambda x}}],\delta)}{\Gamma(\delta)}$	New
GMW	0	-	-	-	-	$\frac{\gamma(-\log[1-e^{-\beta x^{\theta} e^{\lambda x}}],\delta)}{\Gamma(\delta)}$	New
GME	0	-	1	-	-	$\frac{\gamma(-\log[1-e^{-\beta x e^{\lambda x}}],\delta)}{\Gamma(\delta)}$	New
GAE	-	-	1	0	-	$\frac{\gamma(-\log[1-e^{\alpha x-\beta x}],\delta)}{\Gamma(\delta)}$	New
GEV	0	1	0	-	-	$\frac{\gamma(-\log[1-e^{-e^{\lambda x}}],\delta)}{\Gamma(\delta)}$	New
GW	0	-	-	0	-	$\frac{\gamma(-\log[1-e^{-\beta x^{\theta}}],\delta)}{\Gamma(\delta)}$	Pinho, Cordeiro, and Nobre (2012)
MW	0	-	-	-	1	$1-e^{-\beta x^{\theta} e^{\lambda x}}$	Lai, Xie, and Murthy (2003)
S-ZMW	-	-	-	-	1	$1-e^{-\alpha x-\beta x^{\theta} e^{\lambda x}}$	Sarhan and Zaindin (2009)
LFR	-	-	2	0	1	$1-e^{-\alpha x-\beta x^2}$	Bain (1974)
EV	0	1	0	-	1	$1-e^{-e^{\lambda x}}$	Bain (1974)
Weibull	0	-	-	0	1	$1-e^{-\beta x^{\theta}}$	Weibull (1951)
Exponential	-	0	0	0	1	$1-e^{-\alpha x}$	Bain (1974)

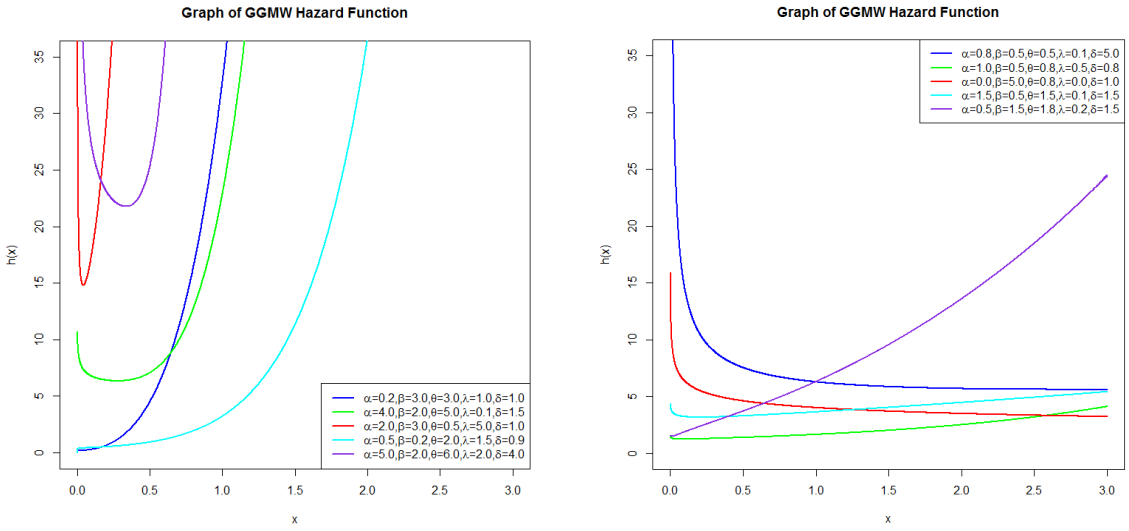

Figure 2: Graphs of GGMW hazard function

$e^{-\beta x^\theta e^{\lambda x}}$ and $e^{k\lambda x}$, we have:

$$\begin{aligned} E(Y^r) &= \int_0^\infty y^r d(1-e^{-\alpha y-\beta x^\theta e^{\lambda y}}) \\ &= \int_0^\infty r y^{r-1} e^{-\alpha y-\beta y^\theta e^{\lambda y}} dy \\ &= \sum_{k,n=0}^{\infty} \frac{r(-\beta)^n (n\lambda)^k}{k!n!} \int_0^\infty r y^{r+n\theta+k-1} e^{-\alpha y} dy \\ &= \sum_{k,n=0}^{\infty} \frac{r(-\beta)^n (n\lambda)^k}{k!n!} \alpha^{-(r+\theta n+k)} \Gamma(r+\theta n+k). \end{aligned} \tag{25}$$

Consequently, that the r^{th} raw moment of GGMW distribution is given by:

$$\mu_r' = E(X^r) = \sum_{\nu \in C} w_\nu E(Y^r),$$

where $Y \sim GMW(\alpha(m+s+\delta), \beta(m+s+\delta), \theta, \lambda)$. Note that, since $\sum_{r=0}^{\infty} \frac{t^r}{r!} x^r g_{GGMW}(x)$ converges and each term is integrable for all t close to zero, say (for $|t| < 1$), the moment generating function (MGF) of the GGMW distribution is given by:

$$\begin{aligned} M_X(t) &= \sum_{\nu \in C} \sum_{j=0}^{\infty} w_\nu \frac{t^j}{j!} E(Y^j) \\ &= \sum_{\nu \in C} \sum_{k,n,j=0}^{\infty} w_\nu \frac{t^j j(-\beta(k+s+\delta))^n (n\lambda)^k}{k!n!j!(\alpha(k+s+\delta))^{(j+\theta n+k)}} \Gamma(j+\theta n+k), \end{aligned} \tag{26}$$

where $\Gamma(a) = b^a \int_0^\infty t^{a-1} e^{-t} dt$ is the gamma function, and $r = 1, 2, ...$
Table 3 lists the first six moments for selected parameter values of GGMW distribution, where $Variance = E(Y^2) - E(Y)^2$, $Skewness = \frac{E(Y^3) - 3E(Y)\sigma^2 - E(Y)^3}{\sigma^3}$, and $Kurtosis = \frac{E(Y^4)}{\sigma^4} - 3$.

Theorem 3.1.

$$E\{[-\log(1-F(X))]^r [(1-F(X))^s]\} = \frac{\psi^{r+\delta} \Gamma(r+\delta)}{(s\psi+1)^\delta \psi^\delta \Gamma(\delta)}. \tag{27}$$

Proof:

$$\begin{aligned} E\{[-\log(1-F(X))]^r [(1-F(X))^s]\} &= \int_0^\infty \frac{f(x)}{\psi^\delta \Gamma(\delta)} [-\log(1-F(x))]^{r+\delta-1} \\ &\times [1-F(x)]^{s+(1/\psi)-1} dx \\ &= \frac{\psi^{r+\delta} \Gamma(r+\delta)}{(s\psi+1)^\delta \psi^\delta \Gamma(\delta)}. \end{aligned} \tag{28}$$

If $s = 0$ in equation (28), then we have

$$E[-\log(1-F(X))^r] = \frac{\psi^{r+\delta} \Gamma(r+\delta)}{\psi^\delta \Gamma(\delta)}. \tag{29}$$

Let $\psi^* = s + \frac{1}{\psi}$, then with $r = 0$ in equation (28), we obtain

$$\begin{aligned} E[(1-F(X))^s] &= \left(\frac{1}{\psi\psi^*}\right)^\delta \int_0^\infty \frac{(\psi^*)^\delta f(x)}{\Gamma(\delta)} [-\log(1-F(x))]^{\delta-1} \\ &\times [1-F(x)]^{\psi^*-1} dx \\ &= [s\psi+1]^{-\delta}. \end{aligned} \tag{30}$$

Table 3: GGMW moments for selected values

	$(\alpha,\beta,\theta,\lambda,\delta)$				
Moments	(1,2,0.5,0.5,1)	(1,2,0.5,1.5,2)	(1,4,2,1,6)	(1,1.5,2,1,2.5)	(2,0.9,1,1,3)
$E(Y)$	0.1798084	0.0360182	0.0130699	0.1442460	0.0502597
$E(Y^2)$	0.0883142	0.0050822	0.0007539	0.0394158	0.0071370
$E(Y^3)$	0.0649863	0.0011492	0.0000767	0.0142427	0.0016577
$E(Y^4)$	0.0608936	0.0003385	0.0000107	0.0060942	0.0005210
$E(Y^5)$	0.0674815	0.0001191	0.0000018	0.0029347	0.0002021
$E(Y^6)$	0.0848496	0.0000477	0.0000004	0.0015457	0.0000917
Variance	0.0559832	0.0037849	0.0005831	0.0186089	0.0046110
Skewness	2.1873821	2.9781690	3.6632310	1.2561066	2.6683440
Kurtosis	16.4292658	20.6287000	28.4321500	14.5986436	21.5064600

4. Order statistics and Rényi entropy

Order statistics play an important role in probability and statistics. The concept of entropy plays a vital role in information theory. The entropy of a random variable is defined in terms of its probability distribution and can be shown to be a good measure of randomness or uncertainty. In this section, we present Rényi entropy and the distribution of the order statistics for the GGMW distribution.

4.1. Rényi entropy

Rényi entropy is an extension of Shannon entropy. Rényi entropy is defined to be

$$I_R(v) = \frac{1}{1-v}\log\left(\int_0^\infty [g_{GGMW}(x;\alpha,\beta,\theta,\lambda,\delta)]^v dx\right), v \neq 1, v > 0. \tag{31}$$

Rényi entropy tends to Shannon entropy as $v \to 1$. Note that

$$\begin{aligned}\int_0^\infty g_{GGMW}^v(x)dx &= \left(\frac{1}{\Gamma(\delta)}\right)^v \int_0^\infty ((\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x])e^{-\alpha x-\beta x^\theta e^{\lambda x}})^v \\ &\times [-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})]^{v(\delta-1)}dx.\end{aligned} \tag{32}$$

Let $0 < y = e^{-\alpha x-\beta x^\theta e^{\lambda x}} < 0.7968$. Note that

$$\begin{aligned}((\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x]))^v &= \sum_{j=0}^{v}\binom{v}{j}\alpha^{v-j}\beta^j x^{j\theta-j}\sum_{n=0}^{\infty}\binom{j}{r}\frac{(j\lambda x)^n}{n!}\sum_{r=0}^{j}\theta^{j-r}(\lambda x)^r \\ &= \sum_{j=0}^{v}\sum_{r=0}^{j}\sum_{n=0}^{\infty}\binom{v}{j}\binom{j}{r}\alpha^{v-j}\beta^j\theta^{j-r}\lambda^r\frac{(j\lambda)^n}{n!}x^{n+r+j\theta-j}.\end{aligned}$$

Now, for $0 < e^{-v\beta x^\theta e^{\lambda x}} < 1$, $v > 0$, and applying Taylor series expansion, we have

$$e^{-v\beta x^\theta e^{\lambda x}} = \sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\frac{(-1)^l(v\beta)^l(l\lambda)^w}{l!w!}x^{l\theta+w}, \tag{33}$$

so that,

$$\begin{aligned} g^v(x) &= [\Gamma(\delta)]^{-v}\sum_{j=0}^{v}\sum_{r=0}^{j}\sum_{n,l,w,m,s=0}^{\infty}(-1)^l\binom{v}{j}\binom{j}{r}\binom{\delta(v-1)}{m} \\ &\times \alpha^{v-j}\beta^j\theta^{j-r}\lambda^r\frac{(j\lambda)^n}{n!}\frac{(v\beta)^l}{l!}\frac{(l\lambda)^w}{w!}b_{s,m} \\ &\times x^{n+r+j\theta-j+l\theta+w}e^{-(m+s+v\delta-v)\alpha x-(m+s+v\delta-v)\beta x^\theta e^{\lambda x}}e^{-v\alpha x} \\ &= [\Gamma(\delta)]^{-v}\sum_{j=0}^{v}\sum_{r=0}^{j}\sum_{n,l,w,m,s,k,i=0}^{\infty}(-1)^{l+k}\binom{v}{j}\binom{j}{r}\binom{\delta(v-1)}{m}b_{s,m} \\ &\times \alpha^{v-j}\beta^{j+l}\theta^{j-r}\lambda^{r+n+w}\frac{(j)^n(v)^l(l)^w}{n!l!w!} \\ &\times \frac{(m+s+v\delta-v)^k\beta^k(k\lambda)^i}{k!i!}x^{n+r+j\theta-j+l\theta+w+k\theta+i}e^{-(m+s+v\delta)\alpha x}. \end{aligned}$$

Using the fact that $\int_0^\infty t^{a-1}e^{-t}dt = \frac{\Gamma(a)}{b^a}$, we have

$$\begin{aligned} \int_0^\infty g_{GGMW}^v(x)dx &= [\Gamma(\delta)]^{-v}\sum_{j=0}^{v}\sum_{r=0}^{j}\sum_{n,l,w,m,s,k,i=0}^{\infty}(-1)^{l+k}\binom{v}{j}\binom{j}{r}\binom{\delta(v-1)}{m}b_{s,m} \\ &\times \alpha^{v-j}\beta^{j+l+}\theta^{j-r}\lambda^{r+n+w+i}\frac{(j)^n(v)^l(l)^w k^i(m+s+v\delta-v)^k}{n!l!w!k!i!} \\ &\times \frac{\Gamma(n+r+w+i+\theta(j+l+k)-j+1)}{(m+s+v\delta)^{n+r+w+i+\theta(j+l+k)-j+1}}, \end{aligned}$$

for $v > 0$, $v \neq 1$. Consequently, Rényi entropy for the GGMW distribution is given by

$$\begin{aligned} I_R(v) &= \frac{1}{1-v}\log\Bigg[[\Gamma(\delta)]^{-v}\sum_{j=0}^{v}\sum_{r=0}^{j}\sum_{n,l,w,m,s,k,i=0}^{\infty}(-1)^{l+k}\binom{v}{j}\binom{j}{r}\binom{\delta(v-1)}{m}b_{s,m} \\ &\times \alpha^{v-j}\beta^{j+l+}\theta^{j-r}\lambda^{r+n+w+i}\frac{(j)^n(v)^l(l)^w k^i(m+s+v\delta-v)^k}{n!l!w!k!i!} \\ &\times \frac{\Gamma(n+r+w+i+\theta(j+l+k)-j+1)}{(m+s+v\delta)^{n+r+w+i+\theta(j+l+k)-j+1}}\Bigg], \quad \text{for } v>0,\ v\neq 1. \end{aligned}$$

4.2. Order statistics

In this section, the pdf of the i^{th} order statistic and the corresponding moments are presented. Let $X_1, X_2, ..., X_n$ be independent and identically distributed GGMW random variables. The pdf of of the i^{th} order statistic for a random sample of size n for any gamma$-\overline{G}$ family with density (5) can be expressed as an infinite weighted sum of gamma$-\overline{G}$ densities. The pdf of the i^{th} order statistic from the GGMW pdf $g_{GGMW}(x)$ is given by

$$\begin{aligned} g_{i:n}(x) &= \frac{n!g(x)}{(i-1)!(n-i)!}[G(x)]^{i-1}[1-G(x)]^{n-i} \\ &= \frac{n!g(x)}{(i-1)!(n-i)!}\sum_{j=0}^{i-1}(-1)^j\binom{i-1}{j}[\overline{G}(x)]^{n-i+j} \\ &= \frac{n!g(x)}{(i-1)!(n-i)!}\sum_{j=0}^{i-1}(-1)^j\binom{i-1}{j}\left[\frac{\gamma(-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}}))}{\Gamma(\delta)}\right]^{n-i+j}. \end{aligned}$$

where $0 < y = e^{-\alpha x-\beta x^\theta e^{\lambda x}} < 0.7968$, $x > 0$, $\alpha, \beta, \theta, \delta > 0$, and $\lambda \geq 0$. Using the fact that $\gamma(x,\delta) = \sum_{m=0}^{\infty}\frac{(-1)^m x^{m+\delta}}{(m+\delta)m!}$, and setting $c_m = (-1)^m/((m+\delta)m!)$, we can write the pdf of the

i^{th} order statistic from the GGMW distribution as follows:

$$\begin{aligned} g_{i:n}(x) &= \frac{n!g(x)}{(i-1)!(n-i)!}\sum_{j=0}^{i-1}\binom{i-1}{j}\frac{(-1)^j}{[\Gamma(\delta)]^{n-i+j}}[-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})]^{\delta(n-i+j)} \\ &\times \left[\sum_{m=0}^{\infty}\frac{(-1)^m(\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}}))^m}{(m+\delta)m!}\right]^{n-i+j} \\ &= \frac{n!g(x)}{(i-1)!(n-i)!}\sum_{j=0}^{i-1}\binom{i-1}{j}\frac{(-1)^j}{[\Gamma(\delta)]^{n-i+j}}[-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})]^{\delta(n-i+j)} \\ &\times \sum_{m=0}^{\infty}d_{m,n-i+j}(-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}}))^m, \end{aligned}$$

where $d_0 = c_0^{(n-i+j)}$, $d_{m,n-i+j} = (mc_0)^{-1}\sum_{l=1}^{m}[(n-i+j)l-m+l]c_l d_{m-l,n-i+j}$. We note that

$$\begin{aligned} g_{i:n}(x) &= \frac{n!g(x)}{(i-1)!(n-i)!}\sum_{j=0}^{i-1}\sum_{m=0}^{\infty}\binom{n-i}{j}\frac{(-1)^j d_{m,i+j-1}}{[\Gamma(\delta)]^{n-i+j}}[-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})]^{\delta(n-i+j)+m} \\ &= \frac{n![-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})]^{\delta-1}f(x)}{(i-1)!(n-i)!\Gamma(\delta)}\sum_{j=0}^{i-1}\sum_{m=0}^{\infty}\binom{n-i}{j}\frac{(-1)^j d_{m,n-i+j}}{[\Gamma(\delta)]^{n-i+j}} \\ &\times [-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})]^{\delta(n-i+j)+m} \\ &= \frac{n!}{(i-1)!(n-i)!}\sum_{j=0}^{i-1}\sum_{m=0}^{\infty}\binom{i-1}{j}\frac{(-1)^j d_{m,n-i+j}}{[\Gamma(\delta)]^{n-i+j}} \\ &\times \frac{\Gamma(\delta(n-i+j)+m+\delta)}{\Gamma(\delta(n-i+j)+m+\delta)}\frac{[-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})]^{\delta(n-i+j)+m+\delta-1}}{\Gamma(\delta)} \\ &\times (\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x])e^{-\alpha x-\beta x^\theta e^{\lambda x}} \\ &= \frac{n!}{(i-1)!(n-i)!}\sum_{j=0}^{i-1}\sum_{m=0}^{\infty}\binom{i-1}{j} \\ &\times \frac{(-1)^j d_{m,n-i+j}\Gamma(\delta(n-i+j)+m+\delta)}{[\Gamma(\delta)]^{n-i+j+1}}f_{GGMW}(x), \end{aligned}$$

where

$$\begin{aligned} f_{GGMW}(x) &= \frac{1}{\Gamma(\delta(n-i+j)+m+\delta)}[-\log(1-e^{-\alpha x-\beta x^\theta e^{\lambda x}})]^{\delta(n-i+j)+m+\delta-1} \\ &\times (\alpha+\beta x^{\theta-1}e^{\lambda x}[\theta+\lambda x])e^{-\alpha x-\beta x^\theta e^{\lambda x}} \end{aligned} \tag{34}$$

is the GGMW pdf with parameters $\alpha,\beta,\theta>0,\lambda\geq 0$, and shape parameter $\delta^* = \delta(n-i+j)+m+\delta>0$. It follows therefore that the r^{th} moment is given by

$$E(X_{i:n}^j) = \sum_{\nu\in C}\sum_{j=0}^{i-1}\sum_{m,k,n=0}^{\infty}w_\nu \ell_{i,j,m}\frac{r(-\beta)^n(k+s+\delta^*)(n\lambda)^k}{k!n![\alpha(k+s+\delta^*)^{r+n\theta+k}]}\Gamma(r+n\theta+k),$$

where $\ell_{i,j,m} = \frac{n!}{(i-1)!(n-i)!}\frac{(-1)^j d_{m,n-i+j}\Gamma(\delta(n-i+j)+m+\delta)}{[\Gamma(\delta)]^{n-i+j+1}}$, and $\delta^* = \delta(n-i+j)+m+\delta>0$. We note that these moments are often used in several areas including reliability, survival analysis, biometry, engineering, insurance and quality control for the prediction of future failures times from a set of past or previous failures.

5. Maximum likelihood estimation

Let $X \sim GGMW(\alpha, \beta, \theta, \lambda, \delta)$ and $\mathbf{\Delta} = (\alpha, \beta, \theta, \lambda, \delta)^T$ be the parameter vector. The log-likelihood for a single observation x of X is given by

$$\begin{aligned} \ell = \ell(\mathbf{\Delta}) &= (\delta - 1)\log(-\log(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})) + \log(\alpha + \beta x^{\theta-1} e^{\lambda x}[\theta + \lambda x]) \\ &- \alpha x - \beta x^{\theta} e^{\lambda x} - \log(\Gamma(\delta)). \end{aligned} \tag{35}$$

The first derivative of the log-likelihood function with respect to the parameters $\mathbf{\Delta} = (\alpha, \beta, \theta, \lambda, \delta)^T$ are given by

$$\frac{\partial \ell}{\partial \alpha} = \frac{x(\delta - 1)e^{-\alpha x - \beta x^{\theta} e^{\lambda x}}}{(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})\log(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})} + \frac{1}{\alpha + \beta x^{\theta-1} e^{\lambda x}[\theta + \lambda x]} - x, \tag{36}$$

$$\frac{\partial \ell}{\partial \beta} = \frac{x^{\theta} e^{\lambda x}(\delta - 1)e^{-\alpha x - \beta x^{\theta} e^{\lambda x}}}{(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})\log(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})} + \frac{x^{\theta-1} e^{\lambda x}(\theta + \lambda x)}{\alpha + \beta x^{\theta-1} e^{\lambda x}[\theta + \lambda x]} - x^{\theta} e^{\lambda x}, \tag{37}$$

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= \frac{(\delta - 1)x^{\theta}\log(x)\beta e^{\lambda x} e^{-\alpha x - \beta x^{\theta} e^{\lambda x}}}{(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})\log(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})} + \frac{\beta x^{\theta-1} e^{\lambda x}[(\theta + \lambda x)\log(x) + 1]}{\alpha + \beta x^{\theta-1} e^{\lambda x}[\theta + \lambda x]} \\ &- \beta x^{\theta} e^{\lambda x}\log(x), \end{aligned} \tag{38}$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{(\delta - 1)x^{\theta+1}\beta e^{\lambda x} e^{-\alpha x - \beta x^{\theta} e^{\lambda x}}}{(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})\log(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})} + \frac{\beta x^{\theta} e^{\lambda x}(\theta + \lambda x + 1)}{\alpha + \beta x^{\theta-1} e^{\lambda x}[\theta + \lambda x]} - \beta x^{\theta+1} e^{\lambda x}, \tag{39}$$

and

$$\frac{\partial \ell}{\partial \delta} = \log(-\log(1 - e^{-\alpha x - \beta x^{\theta} e^{\lambda x}})) - \frac{\Gamma'(\delta)}{\Gamma(\delta)}. \tag{40}$$

The total log-likelihood function based on a random sample of n observations: $x_1, x_2, ..., x_n$ drawn from the GGMW distribution is given by $\ell_n = \ell(\mathbf{\Delta}) = \sum_{i=1}^{n} \ell_i(\mathbf{\Delta})$, where $\ell_i(\mathbf{\Delta})$, $i = 1, 2, ..., n$ is given by equation (35). The equations obtained by setting the above partial derivatives to zero are not in closed form and the values of the parameters $\alpha, \beta, \theta, \lambda, \delta$ must be found by using iterative methods. The maximum likelihood estimates of the parameters, denoted by $\hat{\mathbf{\Delta}}$ is obtained by solving the nonlinear equations $(\frac{\partial \ell}{\partial \alpha}, \frac{\partial \ell}{\partial \beta}, \frac{\partial \ell}{\partial \theta}, \frac{\partial \ell}{\partial \lambda}, \frac{\partial \ell}{\partial \delta})^T = \mathbf{0}$. It is convenient to apply or use nonlinear optimization algorithm such as quasi-Newton algorithm to numerically maximize the log-likelihood function.

We maximize the likelihood function using NLmixed in SAS as well as the function nlm in R (The R Development Core Team (2011)). These functions were applied and executed for wide range of initial values. This process often results or lead to more than one maximum, however, in these cases, we take the MLEs corresponding to the largest value of the maxima. In a few cases, no maximum was identified for the selected initial values. In these cases, a new initial value was tried in order to obtain a maximum.

The issues of existence and uniqueness of the MLEs are theoretical interest and has been studied by several authors for different distributions including Seregin (2010), Santos Silva and Tenreyro (2010), Zhou (2009), and Xia, Mi, and Zhou (2009). At this point we are not able to address the theoretical aspects (existence, uniqueness) of the MLE of the parameters of the GGMW distribution.

Note that for the five parameters of the GGMW distribution, all second order partial derivatives of the log-likelihood function exit, and are given in appendix A. The Fisher information matrix is given by $\mathbf{I}(\mathbf{\Delta}) = [\mathbf{I}_{\theta_i,\theta_j}]_{5X5} = E(-\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j})$, $i, j = 1, 2, 3, 4, 5$, can be numerically

obtained by MATHLAB, R or MAPLE software. The total Fisher information matrix $n\mathbf{I}(\boldsymbol{\Delta})$ can be approximated by

$$\mathbf{J}_n(\hat{\boldsymbol{\Delta}}) \approx \left[-\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\bigg|_{\boldsymbol{\Delta}=\hat{\boldsymbol{\Delta}}} \right]_{5X5}, \quad i,j = 1,2,3,4,5. \tag{41}$$

For a given set of observations, the matrix given in equation (41) is obtained after the convergence of the Newton-Raphson procedure in MATHLAB or R software. Elements of the observed information matrix are given in the appendix.

5.1. Asymptotic confidence intervals

In this section, we present the asymptotic confidence intervals for the parameters of the GGMW distribution. The expectations in the Fisher Information Matrix (FIM) can be obtained numerically. Let $\hat{\boldsymbol{\Delta}} = (\hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{\lambda}, \hat{\delta})$ be the maximum likelihood estimate of $\boldsymbol{\Delta} = (\alpha, \beta, \theta, \lambda, \delta)$. Under the usual regularity conditions and that the parameters are in the interior of the parameter space, but not on the boundary, (Ferguson 1996) we have: $\sqrt{n}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \xrightarrow{d} N_5(\underline{0}, I^{-1}(\boldsymbol{\Delta}))$, where $I(\boldsymbol{\Delta})$ is the expected Fisher information matrix. The asymptotic behavior is still valid if $I(\boldsymbol{\Delta})$ is replaced by the observed information matrix evaluated at $\hat{\boldsymbol{\Delta}}$, that is $J(\hat{\boldsymbol{\Delta}})$. The multivariate normal distribution $N_5(\underline{\mathbf{0}}, J(\hat{\boldsymbol{\Delta}})^{-1})$, where the mean vector $\underline{\mathbf{0}} = (0,0,0,0,0)^T$, can be used to construct confidence intervals and confidence regions for the individual model parameters and for the survival and hazard rate functions. That is, the approximate $100(1-\eta)\%$ two-sided confidence intervals for α, β, θ λ, and δ are given by:

$$\widehat{\alpha} \pm Z_{\frac{\eta}{2}}\sqrt{I^{-1}_{\alpha\alpha}(\hat{\boldsymbol{\Delta}})}, \quad \widehat{\beta} \pm Z_{\frac{\eta}{2}}\sqrt{I^{-1}_{\beta\beta}(\hat{\boldsymbol{\Delta}})}, \quad \widehat{\theta} \pm Z_{\frac{\eta}{2}}\sqrt{I^{-1}_{\theta\theta}(\hat{\boldsymbol{\Delta}})}, \quad \widehat{\lambda} \pm Z_{\frac{\eta}{2}}\sqrt{I^{-1}_{\lambda\lambda}(\hat{\boldsymbol{\Delta}})},$$

and $\widehat{\delta} \pm Z_{\frac{\eta}{2}}\sqrt{I^{-1}_{\delta\delta}(\hat{\boldsymbol{\Delta}})}$, respectively, where $I^{-1}_{\alpha\alpha}(\hat{\boldsymbol{\Delta}})$, $I^{-1}_{\beta\beta}(\hat{\boldsymbol{\Delta}})$, $I^{-1}_{\theta\theta}(\hat{\boldsymbol{\Delta}})$, $I^{-1}_{\lambda\lambda}(\hat{\boldsymbol{\Delta}})$ and $I^{-1}_{\delta\delta}(\hat{\boldsymbol{\Delta}})$ are the diagonal elements of $I^{-1}_n(\hat{\boldsymbol{\Delta}})$, and $Z_{\frac{\eta}{2}}$ is the upper $\frac{\eta}{2}^{th}$ percentile of a standard normal distribution.

The maximum likelihood estimates (MLEs) of the GGMW parameters α, β, θ, λ, and δ are computed by maximizing the objective function via the subroutine NLmixed in SAS and the function nlm in R. The estimated values of the parameters (standard error in parenthesis), -2log-likelihood statistic, Akaike Information Criterion, $AIC = 2p - 2\ln(L)$, Bayesian Information Criterion, $BIC = p\ln(n) - 2\ln(L)$, and Consistent Akaike Information Criterion, $AICC = AIC + 2\frac{p(p+1)}{n-p-1}$, where $L = L(\hat{\boldsymbol{\Delta}})$ is the value of the likelihood function evaluated at the parameter estimates, n is the number of observations, and p is the number of estimated parameters are presented. In order to compare the models, we use the criteria stated above. Note that for the value of the log-likelihood function at its maximum (ℓ_n), larger value is good and preferred, and for AIC, AICC and BIC, smaller values are preferred. GGMW distribution is fitted to the data sets and these fits are compared to the fits of the GGME, GGMR, GMW, GW, beta exponentiated Weibull (BEW) and beta Weibull (BW) distributions.

We can use the likelihood ratio (LR) test to compare the fit of the GGMW distribution with its sub-models for a given data set. For example, to test $\lambda = 0, \delta = 1$, the LR statistic is $\omega = 2[\ln(L(\hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{\lambda}, \hat{\delta})) - \ln(L(\tilde{\alpha}, \tilde{\beta}, \tilde{\theta}, 0, 1))]$, where $\hat{\alpha}$, $\hat{\beta}$, $\hat{\lambda}$, $\hat{\theta}$ and $\hat{\delta}$, are the unrestricted estimates, and $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\theta}$ are the restricted estimates. The LR test rejects the null hypothesis if $\omega > \chi^2_\epsilon$, where χ^2_ϵ denote the upper $100\epsilon\%$ point of the χ^2 distribution with 2 degrees of freedom.

6. Applications

In this section, we present examples to illustrate the flexibility and applicability of the GGMW distribution and its sub-models for data modeling. The GGMW distribution is also compared

Table 4: Estimation of GGMW model for waiting times data

	Estimates					Statistics				
Distribution	α	β	θ	λ	δ	-2LogLikelihood	AIC	AICC	BIC	SS
GGMW	0.3529	0.05611	1.6133	0.002399	0.1687	637.7	647.7	648.3	660.7	0.0574
	(0.01679)	(0.02513)	(0.1642)	(0.01078)	(0.01386)					
GGME	0.000223	0.4152	1	0.02139	0.1875	640.5	648.5	649	659	0.0929
	(0.292)	(0.2142)		(0.006622)	(0.02897)					
GMW	0	0.2887	1.3239	0.000107	0.1629	634.8	642.8	643.2	653.2	0.0271
		(0.07322)	(0.1062)	(0.005437)	(0.01887)					
GME	0	0.3892	1	0.01897	0.2058	640.9	646.9	647.2	654.7	0.1024
		(0.09978)		(0.006636)	(0.03738)					
GAE	0.2108	0.1808	1	0	0.2776	647.9	653.9	654.1	661.7	0.3585
	(0.09663)	(0.09657)			(0.141)					
GEV	0	1	0	0.09372	0.3111	727.1	731.1	731.2	736.3	1.1924
				(0.000906)	(0.02987)					
GW	0	0.2443	1.3435	0	0.1829	635.3	641.3	641.6	649.2	0.0329
		(0.07656)	(0.08978)		(0.0236)					
BW	k	λ	a	b						
	1.2455	2.1348	1.7298	0.1509		634.2	642.2	642.7	652.7	0.0239
	(0.1008)	(0.4812)	(0.5264)	(0.01871)						
BEW	k	λ	α	a	b					
	1.135	2.6752	1.6422	1.3894	0.2681	633.9	643.9	644.6	657	0.0159
	(0.2224)	(1.4164)	(1.3948)	(0.7837)	(0.2345)					

with the non-nested beta exponentiated Weibull (BEW), and beta Weibull (BW) distributions. The pdf of the BEW distribution (Cordeiro, Gomes, da Silva, and Ortega 2013) is given by

$$g(x) = \frac{\alpha k \lambda^k}{B(a,b)} x^{k-1} e^{(\lambda x)^k} (1 - e^{(\lambda x)^k})^{a\alpha - 1} [1 - (1 - e^{(\lambda x)^k})^{\alpha}]^{b-1}, \quad x > 0. \tag{42}$$

When $\alpha = 1$, we have the BW distribution.

The first data set is waiting times (in minutes) of 100 bank customers before service. See (Ghitany, Atieh, and Nadarajah 2008) for additional details. The second data set is failure times of a sample of $n = 30$ devices, see (Meeker and Escobar 1998). The third data set represent the survival times of 121 patients with breast cancer obtained from a large hospital in a period from 1929 to 1938, (Lee 1992).

Estimates of the parameters of GGMW distribution (standard error in parentheses), Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (AICC) and Bayesian Information Criterion (BIC) are given in Table 4 for the first data set, in Table 5 for the second data set and in Table 6 for the third data set.

The estimated covariance matrix for the GGMW distribution (Waiting Times Data) is given by

$$\begin{pmatrix} 0.00028 & -0.00110 & 0.00300 & 0.00020 & 0.00054 \\ -0.00110 & 0.00063 & -0.00294 & 0.00004 & -0.00001 \\ 0.00300 & -0.00294 & 0.02697 & -0.00141 & -0.00048 \\ 0.00020 & 0.00004 & -0.00141 & 0.00012 & 0.00003 \\ 0.00054 & -0.00001 & -0.00048 & 0.00003 & 0.00019 \end{pmatrix}$$

The 95% asymptotic confidence intervals for the GGMW model (Waiting Times Data) parameters are: $\alpha \in (0.3200, 0.3529)$, $\beta \in (0.0069, 0.1054)$, $\theta \in (1.2915, 1.9351)$, $\lambda \in (-0.0187, 0.0235)$, and $\delta \in (0.1415, 0.1959)$, respectively.

The estimated covariance matrix for the GGMW distribution (Meeker Data) is given by

$$\begin{pmatrix} 0.000015 & -0.000010 & 0.000399 & 0.000000 & -0.000001 \\ -0.000010 & 0.000008 & -0.000040 & -0.000005 & 0.000002 \\ 0.000399 & -0.000040 & 0.007974 & -0.000140 & -0.000110 \\ 0.000000 & -0.000005 & -0.000140 & 0.000007 & 0.000001 \\ -0.000001 & 0.000002 & -0.000110 & 0.000001 & 0.000145 \end{pmatrix}$$

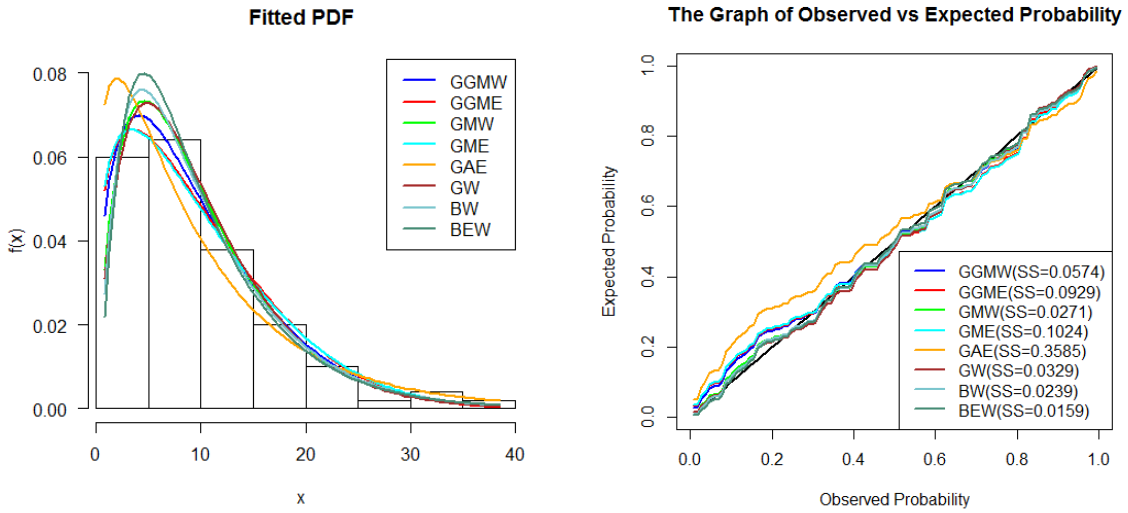

Figure 3: Graphs for waiting times data

Table 5: Estimation of GGMW model for meeker data

	Estimates					Statistics				
Distribution	α	β	θ	λ	δ	-2LogLikelihood	AIC	AICC	BIC	SS
GGMW	0.05354	0.004011	0.004549	0.02772	0.06625	345.2	355.2	357.7	362.2	0.1885
	(0.00392)	(0.002771)	(0.0893)	(0.002629)	(0.01203)					
GGME	0.000856	0.02468	1	0.005281	0.009626	418.7	426.7	428.3	432.3	4.7185
	(0.008775)	(0.002176)		(0.000289)	(0.001758)					
GMW	0	0.5256	0.3819	0.006687	0.06036	355.8	363.8	365.4	369.4	0.2396
		(0.06536)	(0.05393)	(0.000585)	(0.01095)					
GAE	0.006438	0.002438	1	0	0.677	370.2	376.2	377.1	380.4	0.3430
	(0.003377)	(0.003435)			(0.4661)					
GEV	0	1	0	0.01024	0.1007	366.3	370.3	370.7	373.1	0.2661
				(0.001629)	(0.0484)					
GW	0	0.000221	1.453	0	1.7066	368.3	374.3	375.2	378.5	0.3412
		(0.001258)	(0.672)		(2.1364)					
BW	k	λ	a	b						
	0.6935	1.8109	1.1725	0.04806		378.7	386.7	388.3	392.3	0.6369
	(0.03331)	(0.4279)	(1.1345)	(0.00938)						
BEW	k	λ	α	a	b					
	0.9895	8.0706	0.7834	0.9181	0.0426	371.2	381.2	383.7	388.2	0.2904
	(0.03881)	(1.1182)	(0.2485)	(0.5873)	(0.00799)					

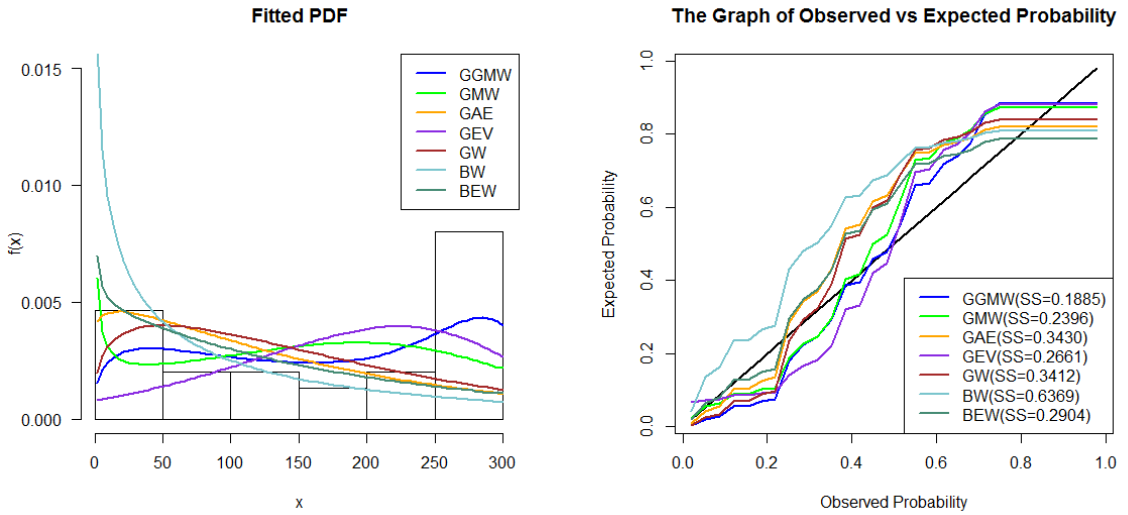

Figure 4: Graphs for meeker data

Table 6: Estimation of GGMW model for breast cancer data

	Estimates					Statistics				
Distribution	α	β	θ	λ	δ	-2LogLikelihood	AIC	AICC	BIC	SS
GGMW	0.1249	0.002047	0.1573	0.05334	0.1657	1155.3	1165.3	1165.9	1179.3	0.1153
	(0.06267)	(0.001812)	(0.8788)	(0.02297)	(0.06779)					
GGME	0.1293	0.002883	1	0.02348	0.1476	1155.8	1163.8	1164.2	1175.0	0.0927
	(0.01342)	(0.001374)		(0.002869)	(0.01646)					
GMW	0	0.07926	0.9986	0.003633	0.2248	1157.1	1165.1	1165.4	1176.3	0.0605
		(0.1335)	(0.1543)	(0.002969)	(0.3898)					
GME	0	0.07876	1	0.00361	0.2253	1157.1	1163.1	1163.3	1171.5	0.0604
		(0.1223)		(0.001698)	(0.3784)					
GAE	0.03852	0.03352	1	0	0.3249	1163.3	1169.3	1169.5	1177.7	0.1943
	(0.02057)	(0.02057)			(0.189)					
GEV	0	1	0	0.02336	0.2439	1243.4	1247.4	1247.5	1252.9	1.1443
				(0.000181)	(0.02174)					
GW	0	0.002764	1.3964	0	1.3417	1158.0	1164.0	1164.2	1172.4	0.0527
		(0.01402)	(0.6238)		(2.2681)					
BW	k	λ	a	b						
	0.7573	1.2899	0.2401	0.06145		1251.7	1259.7	1260.1	1270.9	3.5347
	(0.0495)	(0.3602)	(0.03752)	(0.006166)						
BEW	k	λ	α	a	b					
	0.7958	1.6244	1.2491	0.3575	0.06836	1223.6	1233.6	1234.1	1247.6	2.5138
	(0.007401)	(0.02646)	(0.2518)	(0.06094)	(0.006678)					

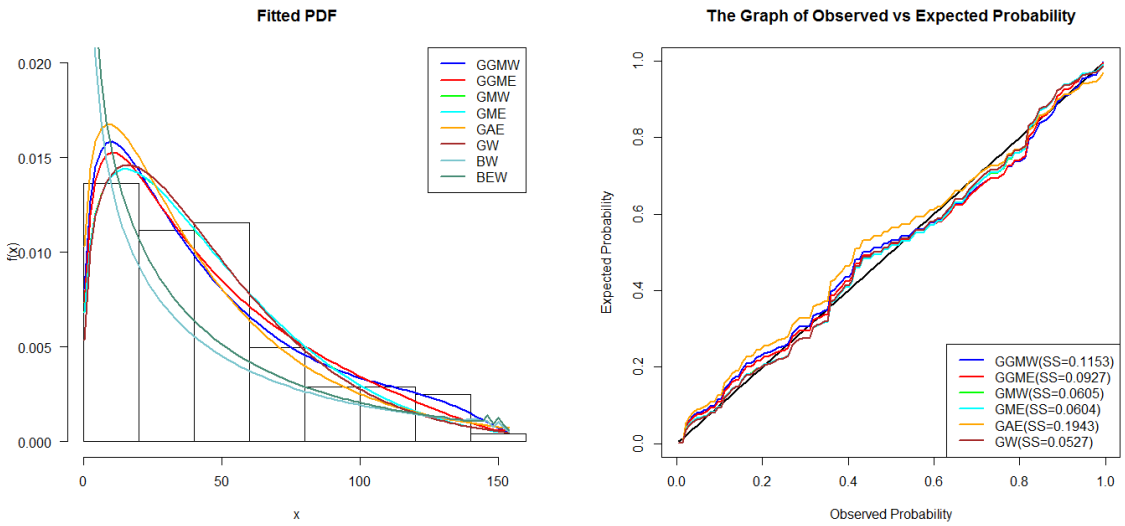

Figure 5: Graphs for breast cancer data

The estimated covariance matrix for the GGMW distribution (Breast Cancer Data) is given by

$$\begin{pmatrix} 0.003927 & 0.000059 & -0.04903 & 0.001194 & -0.00414 \\ 0.000059 & 3.28E-06 & -0.0009 & 0.000016 & -0.00006 \\ -0.04903 & -0.0009 & 0.7724 & -0.01968 & 0.05052 \\ 0.001194 & 0.000016 & -0.01968 & 0.000528 & -0.00122 \\ -0.00414 & -0.00006 & 0.05052 & -0.00122 & 0.004595 \end{pmatrix}$$

Plots of the fitted densities, the histogram of the data are given in Figure 3, Figure 4 and Figure 5. For the probability plot, we plotted $G_{GGMW}(x_{(j)}; \hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{\lambda}, \hat{\delta})$ against $\frac{j-0.375}{n+0.25}$, $j = 1, 2, \cdots, n$, where $x_{(j)}$ are the ordered values of the observed data. We also computed a measure of closeness of each plot to the diagonal line. This measure of closeness is given by the sum of squares

$$SS = \sum_{j=1}^{n} \left[G_{GGMW}(x_{(j)}; \hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{\lambda}, \hat{\delta}) - \left(\frac{j-0.375}{n+0.25} \right) \right]^2.$$

For waiting times data set, the LR test statistic of the hypothesis H_0: GGME against H_a: GGMW is $\omega = 2.8$. The p-value $= 0.094$. Therefore, there is no significant difference between GGMW and GGME distributions at the 5% level. However, there is a significant difference between GGME and GGMW distributions at the 10% level. The LR statistic of the hypothesis H_0: GEV against H_a: GGMW for waiting times data is $\omega = 89.4$ The p-value < 0.0001, we can conclude that there is a significance difference between GGMW and GEV distributions. There is no significant difference between the GGMW and GMW distributions. Also, there is no significant difference between the GW and GMW distributions. The values of the statistics AIC, AICC and BIC shows that the sub-model GW is a good fit for this data. Based on these statistics, the GW distribution could be chosen as the best model among these distributions. The values of the statistics are comparable to those of the non-nested BW distribution and those corresponding to the BEW distribution.

For Meeker data set, the LR test statistics of the hypothesis H_0: GGME against H_a: GGMW is $\omega = 73.5$. The p-value < 0.0001. Therefore, there is significant difference between GGMW and GGME distributions. The LR statistic of the hypothesis H_0: GMW against H_a: GGMW is $\omega = 10.6$. The p-value $= 0.0011$, we can conclude that there is a significance difference between GGMW and GMW distributions. The values of the statistics AIC, BIC, and AICC are smaller for the GGMW distribution. The values of these statistics points to the GGMW distribution as the "better" fit for Meeker data. Also, the values of AIC, BIC and AICC are better for the GMW and GGMW distributions when compared to the non-nested BW and BEW distributions.

For breast cancer data set, there is no significant difference between GGMW, GGME, GMW, GW and GME distributions based on the corresponding LR tests. The sub-models GME and GW seem to be the "best" fits for this data. The values of the statistics AIC, BIC and AICC are smaller for the GME distribution. The values of SS from the probability plots are 0.0604 and 0.0527 for the GME and GW distributions, respectively. The values of these statistics points to and supports the GW as well as the GME distributions as the better fits among the nested distributions. Also, the values of the statistics: AIC, BIC and AICC are far better for the GMW and GGMW distributions when compared to those of the non-nested BW and BEW distributions.

The conclusions based on the LR tests, fitted pdfs, the histograms of the data, and probability plots are in agreement with the statistics AIC, AICC and BIC for the selected models. The GW distribution provides a better fits for the waiting times data, while the GGMW distribution and GME as well as the GW distributions provides better fits for the Meeker and Escobar, and breast cancer data, respectively.

7. Concluding remarks

A new class of generalized modified Weibull distribution called the gamma-generalized modified Weibull (GGMW) distribution is proposed and studied. The GGMW distribution has several sub-models such as the GGMR, GGME, GAE, GLFR, LFR, GMW, GME, MW, MR, ME, Weibull, Raleigh and exponential distributions as special cases. The density of this new class of distributions can be expressed as a linear combination of GMW density functions. The GGMW distribution possesses hazard function with flexible behavior. We also obtain closed form expressions for the moments, distribution of order statistics and Renyi entropy. Maximum likelihood estimation technique was used to estimate the model parameters. Finally, the GGMW distribution and its sub-models was fitted to real data sets to illustrate the applicability and usefulness of this class of distributions.

Acknowledgements
The authors would like to thank the editor and the referee for carefully reading the paper and for their valuable comments, which greatly improved the presentation in this paper.

References

Almalki SJ, Yuan J (2013). "A New Modified Weibull Distribution." *Reliability Engineering and System Safety*, **111**, 164–170.

Bain L (1974). "Analysis for the Linear Failure Rate Life Testing Distribution." *Technometrics*, **16**(4), 551–559.

Barlow R, Campo R (1975). *Total Time on Test Processes and Applications to Failure Data Analysis*. Society for Industrial and Applied Mathematics.

Bebbington M, Lai C, Zitikis R (2007). "A Flexible Weibull Extension." *Reliability Engineering and System Safety*, **92**(6), 719–726.

Carrasco M, Ortega EM, Cordeiro G (2008). "A Generalized Weibull Distribution for Lifetime Modeling." *Computational Statistics and Data Analysis*, **53**(2), 450–462.

Choudhury A (2005). "A Simple Derivation of Moments of the Exponentiated Weibull Distribution." *Metrika*, **62**(1), 17–22.

Cordeiro G, Gomes A, da Silva C, Ortega M (2013). "The Beta Exponentiated Weibull Distribution." *Journal of Statistical Computation and Simulations*, **38**(1), 114–138.

Cordeiro G, Ortega E, Nadarajah S (2010). "The Kumaraswamy Weibull Distribution with Applications to Failure Data." *Journal of Franklin Institute*, **347**(8), 1399–1429.

Famoye F, Lee C, Olumolade O (2005). "The Beta-Weibull Distribution." *Journal of Statistical Theory and Applications*, pp. 121–138.

Ferguson T (1996). *A Course in Large Sample Theory*. Chapman and Hall.

Fisher R (1934). "The Effects of Methods of Ascertainment Upon the Estimation of Frequencies." *Annals of Human Genetics*, **6**(1), 439 – 444.

Ghitany M, Atieh B, Nadarajah S (2008). "Lindley Distribution and Its Applications." *Mathematics and Computers in Simulations*, **78**(4), 493–506.

Gradshteyn I, Ryzhik I (2000). *Tables of Integrals, Series and Products*. Academic Press.

Gupta R, Keating J (1985). "Relation for Reliability Measures under Length Biased Sampling." *Scandinavian Journal of Statistics*, **13**, 49–56.

Gupta R, Kundu D (1999). "Generalized Exponential Distributions." *Australian and New Zealand Journal of Statistics*, **43**, 117–130.

Gupta R, Kundu D (2001). "Exponentiated Exponential Distribution: An Alternative to Gamma and Weibull Distributions." *Biometrical Journal*, **43**, 117–130.

Haupt E, Schabe H (1992). "A New Model for A Lifetime Distribution with Bathtub Shaped Failure Rate." *Microelectronics and Reliability*, **32**, 633–639.

Hjorth U (1980). "A Reliability Distribution with Increasing, Decreasing, Constant and Bathtub Failure Rates." *Technometrics*, **22**, 99–107.

Kundu D, Rakab M (2005). "Generalized Rayleigh Distribution: Different Methods of Estimation." *Computational Statistics and Data Analysis*, **49**, 187–200.

Lai C, Moore T, Xie M (1998). "The Beta Integrated Model." *Proceedings International Workshop on Reliability Modeling and Analysis-From Theory to Practice*, pp. 153–159.

Lai C, Xie M, Murthy D (2003). "A Modified Weibull Distribution." *IEEE Transactions on Reliability*, **52**, 33–37.

Lee C, Famoye F, Olumolade O (2007). "Beta Weibull Distribution, Properties and Applications to Censored Data." *Journal of Mod. Appl. Statist. Meth*, **6**, 173–186.

Lee E (1992). *Statistical Methods for Survival Data Analysis*. John Wiley.

Marshall AW, Olkin I (1997). "A New Method for Adding a Parameter to a Family of Distributions with Applications to the Exponential and Weibull Families." *Biometrika*, **84**(3), 641–652.

Meeker W, Escobar L (1998). *Statistical Methods for Reliability Data*. John Wiley.

Mudholkar G, Srivastava D, Friemer M (1995). "The Exponentiated Weibull Family: A Reanalysis of the Bus-motor-failure Data." *Technometrics*, **37**, 436–445.

Mudholkar G, Srivastava D, Kollia G (1996). "A Generalization of the Weibull Distribution with Application to the Analysis of Survival Data." *Journal of the American Statistical Association*, **91**, 1575–1583.

Nadarajah S (2005). "On the Moments of the Modified Weibull Distribution." *Reliability Engineering and System Safety*, **90**, 114–117.

Nadarajah S, Cordeiro GM, Ortega EMM (2011). "General Results for the beta-Modified Weibull Distribution." *Journal of Statistical Computation and Simulation*, **81**(10), 1211–1232.

Nadarajah S, Kotz S (2005). "On Some Recent Modifications of Weibull Distribution." *IEEE Transactions Reliability*, **54**, 561–562.

Nanda K, Jain K (1999). "Some Weighted Distribution Results on Univariate and Bivariate Cases." *Journal of Statistical Planning and Inference*, **77**(2), 169 – 180.

Nassar M, Eissa F (2003). "On the Exponentiated Weibull Distribution." *Communications in Statistics - Theory and Methods*, **32**(7), 1317–1336.

Oluyede B (1999). "On Inequalities and Selection of Experiments for Length-Biased Distributions." *Probability in the Engineering and Informational Sciences*, **13**(2), 129–145.

Patil G, Rao C (1978). "Weighted Distributions and Size-Biased Sampling with Applications to Wildlife and Human Families." *Biometrics*, **34**(6), 179–189.

Phani KK (1987). "A New Modified Weibull Distribution Function." *Communications of the American Ceramic Society*, **70**(8), 182–184.

Pinho L, Cordeiro G, Nobre J (2012). "The Gamma-Exponentiated Weibull Distribution." *Journal of Statistical Theory and Applications*, **11**(4), 379–395.

Rajarshi S, Rajarshi M (1988). "Bathtub Distributions: A Review." *Communications in Statistics-Theory and Methods*, **17**, 2521–2597.

Rao C (1965). "On Discrete Distributions Arising out of Methods of Ascertainment." *The Indian Journal of Statistics*, **27**(2), 320 – 332.

Ristić M, Balakrishnan N (2011). "The Gamma-Exponentiated Exponential Distribution." *J. Statist. Comp. and Simulation*, **82**(8), 1191–1206.

Santos Silva JMC, Tenreyro S (2010). "On the Existence of Maximum Likelihood Estimates in Poisson Regression." *Economics Letters*, **107**, 310–312.

Sarhan AM, Zaindin M (2009). "Modified Weibull Distribution." *Applied Sciences*, **11**, 123–136.

Seregin A (2010). "Uniqueness of the Maximum Likelihood Estimator for K-monotone Densities." *Proceedings of the American Mathematical Society*, **138**(12), 4511–4515.

Shaked M, Shanthikumar J (1994). *Stochastic Orders and Their Applications*. Academic Press.

Silva G, Ortega E, Cordeiro G (2010). "The Beta Modified Weibull Distribution." *Lifetime Data Analysis*, **16**, 409–430.

Singha N, Jain K, Kumar SS (2012). "The Beta Generalized Weibull Distribution: Properties and Applications." *Reliability Engineering and System Safety*, **102**, 5–15.

The R Development Core Team (2011). "A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing*.

Weibull WA (1951). "Statistical Distribution Function of Wide Applicability." *Journal of Applied Mechanics*, **18**, 293–296.

Xia J, Mi J, Zhou YY (2009). "On the Existence and Uniqueness of the Maximum Likelihood Estimators of Normal and Log-normal Population Parameters with Grouped Data." *Journal of Probability and Statistics*.

Xie M, Lai C (1995). "Reliability Analysis Using an Additive Weibull Model with Bathtub-shaped Failure Rate Function." *Reliability Engineering and System Safety*, **52**, 87–93.

Xie M, Tang Y, Goh T (2002). "A Modified Weibull Extension with Bathtub Failure Rate Function." *Reliability Engineering and System Safety*, **76**, 279–285.

Zhang T, Xie M (2011). "On the Upper Truncated Weibull Distribution and Its Reliability Implications." *Reliability Engineering and System Safety*, **96**(1), 194–200.

Zhou C (2009). "Existence and Consistency of the Maximum Likelihood Estimator for the Extreme Index." *J. Multivariate Analysis*, **100**, 794–815.

Zografos K, Balakrishnan N (2009). "On Families of Beta- and Generalized Gamma-Generated Distribution and Associated Inference." *Stat. Method*, **6**, 344–362.

APPENDIX

Let $A(x_i;\alpha,\beta,\theta,\lambda) = (1-e^{-\alpha x_i-\beta x_i^\theta e^{\lambda x_i}})\log(1-e^{-\alpha x_i-\beta x_i^\theta e^{\lambda x_i}})$, $B(x_i;\alpha,\beta,\theta,\lambda) = e^{-\alpha x_i-\beta x_i^\theta e^{\lambda x_i}} + \log(1-e^{-\alpha x_i-\beta x_i^\theta e^{\lambda x_i}})$, and $C(x_i;\alpha,\beta,\theta,\lambda) = (1-e^{-\alpha x_i-\beta x_i^\theta e^{\lambda x_i}} - \beta x_i^\theta e^{\lambda x_i})\log(1-e^{-\alpha x_i-\beta x_i^\theta e^{\lambda x_i}}) - \beta x_i^\theta e^{-(\alpha-\lambda)x_i-\beta x_i^\theta e^{\lambda x_i}}$. Elements of the observed information matrix of the GGMW distribution are given by

$$\begin{aligned}\frac{\partial^2\ell}{\partial\alpha^2} &= \sum_{i=1}^{n}\frac{(1-\delta)x_i^2 e^{-\alpha x_i-\beta x_i^\theta e^{\lambda x_i}}B(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} \\ &- \sum_{i=1}^{n}\frac{1}{[\alpha+\beta x_i^{\theta-1}e^{\lambda x_i}(\theta+\lambda x_i)]^2}. \end{aligned} \tag{43}$$

$$\begin{aligned}\frac{\partial^2\ell}{\partial\alpha\partial\beta} &= \sum_{i=1}^{n}\frac{(1-\delta)x_i^{\theta+1} e^{-(\alpha-\lambda)x_i-\beta x_i^\theta e^{\lambda x_i}}B(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} \\ &- \sum_{i=1}^{n}\frac{x_i^{\theta-1}e^{\lambda x_i}(\theta+\lambda x_i)}{[\alpha+\beta x_i^{\theta-1}e^{\lambda x_i}(\theta+\lambda x_i)]^2}. \end{aligned} \tag{44}$$

$$\begin{aligned}\frac{\partial^2\ell}{\partial\alpha\partial\theta} &= \sum_{i=1}^{n}\frac{(1-\delta)\beta x_i^{\theta+1} e^{-(\alpha-\lambda)x_i-\beta x_i^\theta e^{\lambda x_i}}\log(x_i)B(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} \\ &- \sum_{i=1}^{n}\frac{\beta x_i^{\theta-1}e^{\lambda x_i}[(\theta+\lambda x_i)\log(x_i)+1]}{[\alpha+\beta x_i^{\theta-1}e^{\lambda x_i}(\theta+\lambda x_i)]^2}. \end{aligned} \tag{45}$$

$$\begin{aligned}\frac{\partial^2\ell}{\partial\alpha\partial\lambda} &= \sum_{i=1}^{n}\frac{(1-\delta)\beta x_i^{\theta+2} e^{-(\alpha-\lambda)x_i-\beta x_i^\theta e^{\lambda x_i}}B(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} \\ &- \sum_{i=1}^{n}\frac{\beta x_i^{\theta}e^{\lambda x_i}(\theta+\lambda x_i+1)}{[\alpha+\beta x_i^{\theta-1}e^{\lambda x_i}(\theta+\lambda x_i)]^2}. \end{aligned} \tag{46}$$

$$\frac{\partial^2\ell}{\partial\alpha\partial\delta} = \sum_{i=1}^{n}\frac{e^{-\alpha x_i-\beta x_i^\theta e^{\lambda x_i}}x_i}{A(x_i;\alpha,\beta,\theta,\lambda)}. \tag{47}$$

$$\begin{aligned}\frac{\partial^2\ell}{\partial\beta^2} &= \sum_{i=1}^{n}\frac{(1-\delta)x_i^{2\theta} e^{-(\alpha-2\lambda)x_i-\beta x_i^\theta e^{\lambda x_i}}B(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} \\ &- \sum_{i=1}^{n}\frac{x_i^{2\theta-2}e^{2\lambda x_i}(\theta+\lambda x_i)^2}{[\alpha+\beta x_i^{\theta-1}e^{\lambda x_i}(\theta+\lambda x_i)]^2}. \end{aligned} \tag{48}$$

$$\begin{aligned}\frac{\partial^2\ell}{\partial\beta\partial\theta} &= \sum_{i=1}^{n}\frac{(\delta-1)x_i^{\theta} e^{-(\alpha-\lambda)x_i-\beta x_i^\theta e^{\lambda x_i}}\log(x_i)C(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} \\ &+ \sum_{i=1}^{n}\frac{x_i^{\theta-1}e^{\lambda x_i}[(\theta+\lambda x_i)\log(x_i)+1]\alpha}{[\alpha+\beta x_i^{\theta-1}e^{\lambda x_i}(\theta+\lambda x_i)]^2} - \sum_{i=1}^{n}x_i^\theta e^{\lambda x_i}\log(x_i). \end{aligned} \tag{49}$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \lambda} = \sum_{i=1}^{n} \frac{(\delta-1)x_i^{\theta+1} e^{-(\alpha-\lambda)x_i - \beta x_i^{\theta} e^{\lambda x_i}} C(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} + \sum_{i=1}^{n} \frac{x_i^{\theta} e^{\lambda x_i}(\theta+\lambda x_i+1)\alpha}{[\alpha+\beta x_i^{\theta-1} e^{\lambda x_i}(\theta+\lambda x_i)]^2} - \sum_{i=1}^{n} x_i^{\theta+1} e^{\lambda x_i}. \quad (50)$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \delta} = \sum_{i=1}^{n} \frac{e^{-(\alpha-\lambda)x_i - \beta x_i^{\theta} e^{\lambda x_i}} x_i^{\theta}}{A(x_i;\alpha,\beta,\theta,\lambda)}. \quad (51)$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = \sum_{i=1}^{n} \frac{(\delta-1)\beta x_i^{\theta} e^{-(\alpha-\lambda)x_i - \beta x_i^{\theta} e^{\lambda x_i}} (\log(x_i))^2 C(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} + \sum_{i=1}^{n} \frac{\beta x_i^{\theta-1} e^{\lambda x_i}\left\{\left[(\theta+\lambda x_i)(\log(x_i))^2 + 2\log(x_i)\right]\alpha - \beta x_i^{\theta-1} e^{\lambda x_i}\right\}}{[\alpha+\beta x_i^{\theta-1} e^{\lambda x_i}(\theta+\lambda x_i)]^2} - \beta \sum_{i=1}^{n} x_i^{\theta} e^{\lambda x_i} (\log(x_i))^2. \quad (52)$$

$$\frac{\partial^2 \ell}{\partial \theta \partial \lambda} = \sum_{i=1}^{n} \frac{(\delta-1)\beta x_i^{\theta+1} e^{-(\alpha-\lambda)x_i - \beta x_i^{\theta} e^{\lambda x_i}} \log(x_i) C(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} + \sum_{i=1}^{n} \frac{\beta x_i^{\theta} e^{\lambda x_i}\left\{\left[(\theta+\lambda x_i+1)\log(x_i)+1\right]\alpha - \beta x_i^{\theta-1} e^{\lambda x_i}\right\}}{[\alpha+\beta x_i^{\theta-1} e^{\lambda x_i}(\theta+\lambda x_i)]^2} - \beta \sum_{i=1}^{n} x_i^{\theta+1} e^{\lambda x_i} \log(x_i). \quad (53)$$

$$\frac{\partial^2 \ell}{\partial \theta \partial \delta} = \sum_{i=1}^{n} \frac{\beta e^{-(\alpha-\lambda)x_i - \beta x_i^{\theta} e^{\lambda x_i}} x_i^{\theta} \log(x_i)}{A(x_i;\alpha,\beta,\theta,\lambda)}. \quad (54)$$

$$\frac{\partial^2 \ell}{\partial \lambda^2} = \sum_{i=1}^{n} \frac{(\delta-1)\beta x_i^{\theta+2} e^{-(\alpha-\lambda)x_i - \beta x_i^{\theta} e^{\lambda x_i}} C(x_i;\alpha,\beta,\theta,\lambda)}{A^2(x_i;\alpha,\beta,\theta,\lambda)} + \sum_{i=1}^{n} \frac{\beta x_i^{\theta+1} e^{\lambda x_i}[(\theta+\lambda x_i+2)\alpha - \beta x_i^{\theta-1} e^{\lambda x_i}]}{[\alpha+\beta x_i^{\theta-1} e^{\lambda x_i}(\theta+\lambda x_i)]^2} - \beta \sum_{i=1}^{n} x_i^{\theta+2} e^{\lambda x_i}. \quad (55)$$

$$\frac{\partial^2 \ell}{\partial \lambda \partial \delta} = \sum_{i=1}^{n} \frac{\beta e^{-(\alpha-\lambda)x_i - \beta x_i^{\theta} e^{\lambda x_i}} x_i^{\theta+1}}{A(x_i;\alpha,\beta,\theta,\lambda)}. \quad (56)$$

$$\frac{\partial^2 \ell}{\partial \delta^2} = -n\Psi'(\delta), \quad \text{where} \quad \Psi(\delta) = \frac{d\log(\Gamma(\delta))}{d\delta} = \frac{\Gamma'(\delta)}{\Gamma(\delta)}. \quad (57)$$

```
## define GGMW pdf
GGMW_pdf <- function(alpha, beta, theta, lambda, delta, x){
  (1/gamma(delta)) * ((-log(1-exp(-alpha * x - beta *
  (x^theta) * (exp(lambda * x)))))^(delta-1)) *
  (alpha + beta * (x^(theta - 1)) * (exp(lambda * x)) *
  (theta + lambda * x)) * (exp(-alpha * x - beta *
  (x^theta) * (exp(lambda * x))))
}
## define GGMW cdf
GGMW_cdf <- function(alpha, beta, theta, lambda, delta, x){
  1 - pgamma(-log(1 - exp(-alpha * x - beta * (x^theta) *
  exp(lambda * x))), delta)
}
## define GGMW Hazard
GGMW_hazard <- function(alpha, beta, theta, lambda, delta, x){
  GGMW_pdf(alpha, beta, theta, lambda, delta, x) /
  (1 - GGMW_cdf(alpha, beta, theta, lambda, delta, x))
}
## define GGMW moments
GGMW_moments <- function(alpha, beta, theta, lambda, delta, k){
  f <- function(alpha, beta, theta, lambda, delta, k, x){
         (x^k) * (GGMW_pdf(alpha, beta, theta, lambda, delta, x))
       }
  y <- integrate(f, lower = 0, upper = Inf, subdivisions = 10000,
                 alpha = alpha, beta = beta, theta = theta,
                 lambda = lambda, delta = delta, k = k)
  return(y)
}
## define GGMW quantile
GGMW_quantile <- function(alpha, beta, theta, lambda, delta, u){
  f <- function(x){alpha * x + beta * (x^theta) * (exp(lambda * x)) +
                     log(1 - exp(-qgamma(1 - u, delta)))
                   }
  rc <- uniroot(f, lower=0, upper=100, tol = 1e-9)
  result <- rc$root
  # check
  error <- GGMW_cdf(alpha, beta, theta, lambda, delta, result) - u
  return(list("result" = result, "error" = error))
}
```

Affiliation:

Broderick O. Oluyede
Department of Mathematical Sciences
Georgia Southern University
Statesboro, GA 30460
E-mail: boluyede@georgiasouthern.edu

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/

Volume 44 *Submitted:* 2014-05-12
October 2015 *Accepted:* 2015-01-27

Austrian Journal of Statistics
October 2015, Volume 44, 69–83.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i3.38

Incidence of stroke in the diabetic and non-diabetic population in Upper Austria (2008-2012) and related effect measures

Karl Schableger
Oberösterreichische Gebietskrankenkasse

Lisa Inreiter
Oberösterreichische Gebietskrankenkasse

Abstract

Background and Purpose: Although it is generally known that diabetes has a negative effect on the stroke incidence, only a limited number of long-term population-based studies focus on the comparison of incidence rates of stroke in diabetics and non-diabetics. Hence, the aim of this study was to estimate the risk of stroke in the diabetic and the non-diabetic population. **Methods:** For this study, data from the Upper Austrian stroke register and the statutory Upper Austrian health insurance (1.3 million members) was used to analyse all first strokes from 2008-2012. This was done by assessing stroke incidence for the total, the diabetic and the non-diabetic population. The analysis was mainly conducted on an age/sex-specific basis. Moreover, age/sex-standardized incidence rates were calculated as well. In addition, effect measures like the relative risk, the attributable risk among exposed and the population attributable risk were computed. **Results:** Out of the total cohort of 1,319,761 subjects, 17,663 had a first stroke (mean age (Sd.): 71.6 (14.3) years; 46.0 per cent male). Among these, 19.5 per cent were classified as diabetics. Concerning the stroke standardized incidence rates of the Upper Austrian population (per 100,000 person years), the following results were obtained for the diabetic and the non-diabetic population respectively: men: 571.9 (95%-confidence interval: 530.1-613.6), 319.3 (95%-confidence interval: 311.3-327.2); women: 600.9 (95%-confidence interval: 559.3-642.5), 343.5 (95%-confidence interval: 335.7-351.3). The age-standardized relative risk was found to be 1.79 (95%-confidence interval: 1.66-1.93) for men and 1.75 (95%-confidence interval: 1.63-1.88) for women. Attributable risks among exposed are as follows: men: 0.44 (95%-confidence interval: 0.40-0.48); women: 0.43 (95%-confidence interval: 0.39-0.47). For the population attributable risks 0.08 (95%-confidence interval: 0.04-0.11) was obtained for men and 0.07 (95%-confidence interval: 0.04-0.09) for women. **Conclusion:** This investigation showed that the stroke risk in the diabetic population is significantly higher compared to the non-diabetic population.

Keywords: stroke, diabetes, epidemiology, Health Service Research, effect measures.

1. Introduction

Over the past few years a number of studies have confirmed that diabetes is a risk factor for stroke events (Almdal et al. 2004; Grysiewiez et al. 2008; Icks et al. 2011). The Austrian Stroke Society published a position paper on the prevention of stroke among diabetics and on the treatment of diabetic stroke patients which stresses the importance of this issue (Österreichische Schlaganfallgesellschaft,

Positionspapier 2010). In 1989 the World Health Organization (WHO) and the International Diabetes Federation (IDF) announced the reduction of the incidence of stroke in diabetes as one of the prior objectives of the St. Vincent declaration in Europe. Their major goal was to bring the incidence of stroke in diabetics to the same level as in the non-diabetic population. Therefore, the knowledge of incidence rates is important and has been dealt with by many studies. However, it must be pointed out that a considerable number of these studies analyse only small cohorts, are meta-analyses or reviews (Lackland et al. 2014, Truelsen et al. 2006). In fact, the number of long term population based European studies are only limited. Icks et al. (2011) even argued that, studies conducted over a period of several years on a large population, investigating stroke incidence among diabetics and non-diabetics are lacking. Icks et al. (2011) for instance conducted such a large population based study for Germany.

For this reason, population based epidemiologic studies dealing with multiple years are important in order to determine incidence rates. Furthermore, studies that treat original data on a large proportion of a population do also generate a useful base of comparison for reviews, cohort and clinical studies.

Thus, the aim of this study was to estimate the risk of stroke for diabetics and non-diabetics using data of a stroke register and insurance data covering almost the complete Upper Austrian population over a period of five years. This was done by determining (age/sex-standardized) incidence rates, relative and attributable risks. Moreover, in order to offer a base of comparison to other studies, the results were not only standardized to the Upper Austrian but also to the Austrian and European population.[1]

Section 2 gives brief information about diabetes and stroke for readers with minor medical background. Data, material and used effect measures are described in Section 3. Results are presented in Section 4, while study findings, conclusions, study limitations and strength are discussed in Section 5.

2. Medical descriptions

2.1. Diabetes mellitus

Diabetes mellitus or simply diabetes is characterized by chronic hyperglycaemia (high blood sugar), which results from defects in insulin secretion and/or defects in insulin action. Suffering from diabetes mellitus may result in lasting damage, dysfunction or failure of various organs like heart, kidneys, eyes and blood vessels. The majority of cases of diabetes can be classified into two major categories that are referred to as type 1 and type 2 diabetes.

The main cause for type 1 diabetes is β-cell destruction. Due to the body's failure to produce insulin, most of the cases require insulin injections in order to be able to survive.

Type 2 diabetes is the most common form of diabetes. Patients with type 2 diabetes suffer from insulin resistance due to the fact that their cells fail to use insulin properly. Moreover, they usually display relative rather than absolute insulin deficiency. As the duration of this complaint advances, β-cell failure increases as well. However, people suffering from this kind of diabetes normally do not need insulin treatment. The risk of getting this kind of diabetes increases with obesity, physical inactivity and age. Moreover, individuals whose parents or siblings are suffering from the disease as well as people with hypertension or dyslipidaemia have an increased risk of developing type 2 diabetes. (Kahn and Sempson 1989; p. 45)

2.2. Stroke

A stroke or "brain attack" is defined as an interruption of blood flow to an area of the brain caused by either a blood clot blocking an artery or a breaking blood vessel. Due to this reason, brain cells begin to die and as a result brain damage occurs. Strokes can be classified into two main categories: ischaemic and haemorrhagic strokes. (National Stroke Association 2013)

[1] The interpretation of the standardized incidence rates for Austria and Europe has to be done carefully, due to the projection of regional data to other populations.

If the blood flow of an artery supplying the brain is interrupted due to the disturbance of blood vessels caused for instance by blood clots, then this may cause cerebral thrombosis or cerebral embolism. This is often followed by a cerebral infarction or a transient cerebral ischaemic attack. These are globally subsumed under ischaemic strokes. (Ross 2012; p. 13)

The result of a breaking blood vessel inside or around the brain is known as bleeding or haemorrhage. While subarachnoid haemorrhage affects arteries at the surface of the brain, intracerebral and intracranial haemorrhages affect vessels within the brain and the skull respectively. (Ross 2012; p. 13)

3. Materials and methods

3.1. Definition of the study population and statistical analysis

For this study data from the Upper Austrian stroke register (UASR) was used. This register is operated by the statutory Upper Austrian health insurance company called "Oberösterreichische Gebietskrankenkasse" (OÖGKK) in cooperation with the regional government of Upper Austria ("Landesregierung Oberösterreich") and is known as "Oberösterreichische Integrierte Versorgung Schlaganfall" (OÖIVS). Besides the 14 Upper Austrian hospitals, there are a number of other institutions that are reporting data to the UASR. These include a pension insurance company, three rehabilitation centres and an ambulance company.

As a regional health insurance company, the OÖGKK insures about 86 per cent of the Upper Austrian residents. The only population groups that are not covered are farmers, teachers, officials and employers.

From 2008 to 2012 1,386,000 people were protected by the OÖGKK. However, not every single individual lived through this whole period of 5 years, as there are some who were born (about 80,000) during these years and others who died (about 55,000). By looking at the standardized person years from 2008 to 2012, it is visible that about 1.3 million people were secured, which roughly represents the amount of individuals protected by the OÖGKK.

For all members of this insurance company not only demographic data exists, but also pharmaceutical prescriptions and hospitalizations are available. Furthermore, data from all subjects who were members of the OÖGKK and for which the OÖGKK bears the costs are included.

Starting in 2007, the UASR includes 97.6 per cent of all hospitalized strokes in Upper Austria (n = 35,195). Following the WHO definition, strokes were defined with the help of "International Classification of Diseases-10" (ICD-10) codes of hospital admissions (BMG, 2012). These included subarachnoid haemorrhage (I60), intracerebral haemorrhage (I61), intracranial haemorrhage (I62) and cerebral infarction (I63). Apart from transient global amnesia (G45.4), transient cerebral ischaemic attacks and related syndromes (G45) were included as well. Moreover, strokes that were not specified as haemorrhage or infarction (I64) were also considered.

To carry out this investigation, the register was accumulated with data from the statutory Upper Austrian health insurance company. In order to identify diabetics and to determine the number of person years, 275,600,810 individual performance information of the reference year 2007 and the observation period 2008 to 2012 were used. Therefore, some of the challenges were due to the big data set.

Like Icks et al., only individuals who did not have a brain attack in 2007 (a period free from stroke of at least one year) were taken into account. Among these people only first strokes between 2008 and 2012 were counted. Due to this, the inclusion of recurrent strokes was avoided. (Icks et al. 2011)

Diabetic subjects were identified by a procedure established by Köster et al.. This procedure leads to a diagnosis of diabetes given that the individuals showed at least one of the following three characteristics: (a) diabetes diagnoses (ICD E10-E14) in at least three of four consecutive quarters, (b) at least two prescriptions of anti-diabetic medication within twelve months, (c) at least one prescription of an anti-diabetic medication and one diabetes diagnosis or one measurement of blood glucose or glycated hemoglobin (HbA1c) within twelve months. (Köster et al. 2006)

However, Köster et al.'s algorithm, which only considers a single year, was slightly redefined. As this study evaluates diabetic subjects over a period of five years, the subjects were not monitored at an annual monthly rate but continuously over the whole period of interest. By doing this, diabetics could be defined more accurately.

Initially diabetics were determined during 2007. However, it must not be forgotten that a considerable number of people showed diabetic symptoms for the first time during the period of investigation. Table 1 shows for instance that almost 20,000 people changed from non-diabetics to diabetics during the five years of observation.

Instead of assigning the individuals for the whole period to diabetics and non-diabetics based only on their status in 2007, the months of each individual equal to the time they lived as diabetics or non-diabetics were counted. Therefore, all those who were diagnosed as diabetics between 2008 and 2012 were also taken into account. Due to the calculation of the person years (months) during this period, they could be weighted according to the duration of their disease. The number of person years where these individuals did not show diabetic symptoms was deleted in order to make sure that they did not have a stroke for at least one year.

Table 1: Number of people (per year) showing diabetic symptoms for the first time

Year	n
2007	41,315
2008	4,628
2009	4,504
2010	4,346
2011	4,029
2012*	1,511
Overall	60,330

* Due to the determination of diabetics similar to a smoothing average the number of diabetics is lower at the end of the observation period. Esteve et al. (1994) mentioned that it is essential to calculate the number of person years of an observation in an exact way. Therefore, the final date of the follow-up must be taken into account for each individual (Esteve et al. 1994; p.15). Moreover, following the procedure of Icks et al.'s study of 2011, the cohort of this work represents individuals that were protected by the OÖGKK in 2007. For this reason, those who were born during the period of observation (2008-2012) were eliminated as they were not part of the original standard population. However, people dying during this time were taken into account with an amount of person years equal to the time they lived through. Table 2 shows that in the case of this study the person years of most of the age strata are very close to five. The only exception is the sixth age strata (85+). However, this group does not hold a very high number of individuals, so it has only a limited effect. Nevertheless, the exact values were considered.

Moreover, similar to the study by Köster et al., a classification of diabetic subjects according to type 1 or type 2 was not possible due to the fact that there was no detailed diagnosis related information in the data of the statutory Upper Austrian health insurance company.

Finally, the cohort consisted of a total of 1,319,761 subjects. In the course of this work the five year stroke incidence for the period of 2008 to 2012 was analysed. This was mainly done on a sex-specific basis. Moreover, for the total, the diabetic and the non-diabetic population age/sex-specific as well as age/sex-standardized incidence rates of strokes were calculated. Therefore, the data was categorized into the following six age strata: 0-44, 45-54, 55-64, 65-74, 75-84, 85+.

As a further part of the analysis, relative risks (diabetic vs. non-diabetic population) were calculated from the (standardized Upper Austrian, Austrian and European) incidence rates. Attributable risks

Table 2: Mean of observed person years in the observation period 2008-2012

Age (years)	Diabetic	Non-diabetic
0-44	4.994	4.956
45-54	4.968	4.906
55-64	4.919	4.804
65-74	4.823	4.653
75-84	4.453	4.181
85+	3.716	3.400
Overall	4.935	4.628

among exposed and population attributable risks due to diabetes were determined. In addition to these measures, a number of confidence intervals were calculated.

The analyses were conducted with the Statistical Analysis System (SAS 9.3) and will be described in more detail below.

3.2. Incidence rates

Assuming a Poisson distribution the amount of brain attacks occurring in a particular age-time exposure cell is assumed to take the values $k = 0, 1, 2, \ldots$ with probabilities:

$$\Pr(K = k) = \mathrm{e}^{-\lambda m} \frac{(\lambda m)^k}{k!}, \tag{3.1}$$

where λ denotes the unknown rate and m is the amount of person years.

An estimate of the age specific incidence rate λ_i (for sex-specific analysis), that an individual might develop a stroke in age group i ($i = 1, \ldots, 6$) is:

$$\lambda_i \approx \frac{k_i}{m_i}, \tag{3.2}$$

where k_i denotes the number of observed strokes and m_i are the number of person years in the i^{th} age group for both sexes. (Breslow and Day, 1987, p. 53) Therefore, the calculation of $\hat{\lambda}_i$ requires an exact calculation of the person years of the observation, as described in section 3.1.

Let ω_i stand for the frequency of individuals in the i^{th} age group of the standard population, hence $\omega_i \cdot \hat{\lambda}_i$ represents the number of expected cases that might be observed in the i^{th} age group of the standard population if it were exposed to a level of risk defined by the rate $\hat{\lambda}_i$. The standardized incidence rate is then:

$$\lambda_s = \sum_{i=1}^{6} \omega_i \cdot \hat{\lambda}_i \tag{3.3}$$

with variance

$$\widehat{\mathrm{Var}}(\lambda_s) = \sum_{i=1}^{6} \frac{\omega_i^2}{m_i^2} \cdot k_i \tag{3.4}$$

(Breslow and Day 1987, p.59).

As Icks et al. (2011), a total of three age/sex-specific incidence rates were considered:

- All strokes in the total population $(\lambda_{\cdot i})$
- Strokes in individuals with diabetes in the population with diabetes (λ_{1i})
- Strokes in individuals without diabetes in the population without diabetes (λ_{0i})

Additionally, incidence rates for the standardized Upper Austrian, Austrian (both of 2007) and European population (of 2010)[2] were estimated.

3.3. Effect measures

For the calculation of the conventional effect measures two dichotomous variables were defined:

$$S = \begin{cases} 1 & \text{stroke} \\ 0 & \text{nostroke} \end{cases} \qquad D = \begin{cases} 1 & \text{diabetic} \\ 0 & \text{notdiabetic} \end{cases}$$

Let $\lambda_{1i} = \Pr[S = 1|D = 1]$ and $\lambda_{0i} = \Pr[S = 1|D = 0]$ be the stratum-specific incidence rates in the i^{th} age-stratum which also denote whether or not diabetes has been diagnosed.

The relative risk (RR) of stroke, defined as the ratio of the stratum-specific incidences is:

$$\text{RR}_i = \frac{\lambda_{1i}}{\lambda_{0i}} = \frac{\Pr[S = 1|D = 1]}{\Pr[S = 1|D = 0]} \tag{3.5}$$

(Breslow and Day 1980, p. 55).

The proportion

$$\text{ARE}_i = \frac{\lambda_{1i} - \lambda_{0i}}{\lambda_{1i}} = \frac{\text{RR}_i - 1}{\text{RR}_i} \tag{3.6}$$

is labelled as the attributable risk for exposed individuals (ARE) (Cole and MacMahon 1971, p. 242).

If δ_i denotes the proportion of diabetics in the strata population, then the total disease incidence is

$$\lambda_{\cdot i} = \delta_i \cdot \lambda_{1i} + (1 - \delta_i) \cdot \lambda_{0i}, \tag{3.7}$$

the population attributable risk (PAR) is calculated as follows:

$$\text{PAR}_i = \frac{\delta_i \cdot (\lambda_{1i} - \lambda_{0i})}{\delta_i \cdot \lambda_{1i} + (1 - \delta_i) \cdot \lambda_{0i}} = \frac{\delta_i \cdot (\text{RR}_i - 1)}{\delta_i \cdot \text{RR}_i + (1 - \delta_i)} \tag{3.8}$$

(Cole and MacMahon 1971, p. 242; Breslow and Day 1980, p. 74).

4. Results

4.1. Study population

During the period of observation 1,319,761 subjects were secured by the OÖGKK, 649,907 (49.2 %) men and 669,854 (50.8 %) women. These subjects, who account for 6,494,429 person years during 2008 and 2012 ($\sim$ 4.9 years per person), did not have a stroke for at least one year before study entry.

A total of 60,330 individuals were diagnosed as diabetic (30,516 men (50.6 %); 29,814 women (49.4 %)). With an average age (Sd.) of 74.0 (10.7) versus 71.0 (15.0) years diabetics were older than non-diabetics (men: 71.1 (10.2) vs. 67.9 (14.6); women: 76.7 (10.3) vs. 73.5 (14.6)).

During the period of 2008 to 2012 17,663 subjects had a first stroke. Their mean age was 71.6 (14.3) years and 46.0 per cent of them were male. Among the stroke patients 19.5 percent were classified as diabetics accounting for 20.7 percent male and 18.5 percent female. Table 4 gives a more detailed overview of the number of first strokes.

According to the distribution of strokes, the following results emerged: transient cerebral ischaemic attacks and related syndromes (G45), 30.2 %; (subarachnoid, intracerebral, intracranial) haemorrhage (I60-I62), 13.8 %; cerebral infarction (I63), 45.8 %; strokes that were not specified as haemorrhage or infarction (I64), 10.2 %.

[2] 2013 was a review of the European standard population of 1976. We use the quinquennial population 2010 from EU-28 which better reflect the characteristics of the EU-population.

Table 3: Description of the study population

	Total	Men	Women
Study population, n (%)	1,319,761 (100)	649,907 (49.2)	669,854 (50.8)
Diabetes, n (%)	60,330 (100)	30,516 (50.6)	29,814 (49.4)
Mean age, years (Sd.)	74.0 (10.7)	71.1 (10.2)	76.7 (10.3)
Stroke, n (%)	17,663 (100)	8,132 (46.0)	9,531 (54.0)
Mean age, years (Sd.)	71.6 (14.3)	68.5 (13.8)	74.1 (14.2)

Table 4: Description of the OÖGKK insurants with stroke, Upper Austria, 2008-2012

	Men		**Women**	
	diabetes (n=1,680)	No diabetes (n=6,452)	diabetes (n=1,763)	No diabetes (n=7,768)
Mean age, years (Sd.)	71.1 (10.2)	67.9 (14.6)	76.7 (10.3)	73.5 (14.9)
Age groups, n (%)				
0-44	29 (1.7)	609 (9.4)	21 (1.2)	525 (6.8)
45-54	131 (7.8)	879 (13.6)	68 (3.9)	638 (8.2)
55-64	421 (25.1)	1328 (20.6)	212 (12.0)	962 (12.4)
65-74	635 (37.8)	1758 (27.2)	533 (30.2)	1889 (24.3)
75-84	400 (23.8)	1510 (23.4)	702 (39.8)	2620 (33.7)
85+	64 (3.8)	368 (5.7)	227 (12.9)	1134 (14.6)
Age refers to age at time of first stroke.				

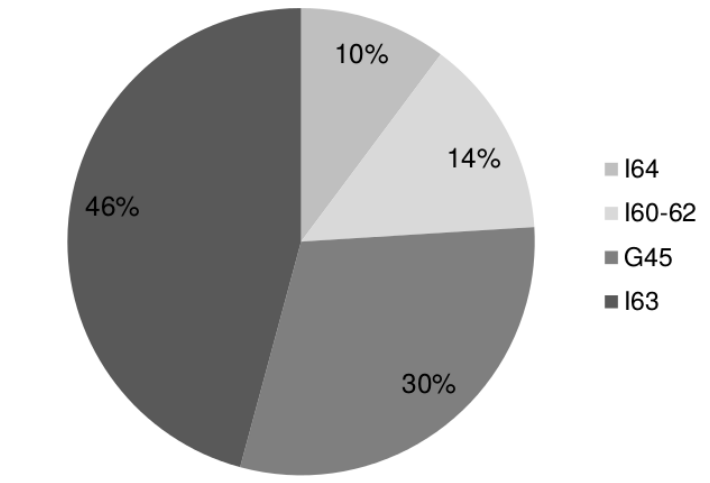

Figure 1: Distribution of strokes

4.2. Incidence of stroke

Table 5 presents details about the age/sex-specific incidence rates. Stroke incidence increased with age in both sexes. Consequently, the highest stroke incidence occurred for subjects being older than 85 years with 2437.6 strokes per 100,000 of the total study population (men: 2610.1; women: 2387.5), 3267.6 per 100,000 in diabetic subjects (men: 3318.6; women: 3253.4) and 2323.2 per 100,000 in non-diabetic subjects (men: 2516.7; women: 2266.7). The most drastic absolute increase of stroke incidence could be observed from the forth (65-74) to the fifth (75-84) age stratum (see also Figure 4 and 5). Regarding the female population, the total incidence rate for 75-84 year olds was more than two times as high as for the 65-74 year olds. A similar effect could be seen for the male population.

0-44 year old subjects suffering from diabetes showed a stroke incidence that was more than five times as high as for their non-diabetic complements. Diabetics of the age classes 45-54 and 55-64 showed an at least two-fold incidence rate compared to non-diabetics. The same effect could be observed by looking at men and women separately. The difference of stroke incidence between diabetics and non-diabetics declined a little for the latter three age strata, but the diabetic subjects still suffered from a stroke more often. (Details see 4.3)

Figure 2 visualizes the standardized incidence rates. Moreover, it stresses the difference between the incidences of stroke among diabetics relative to the one among non-diabetics.

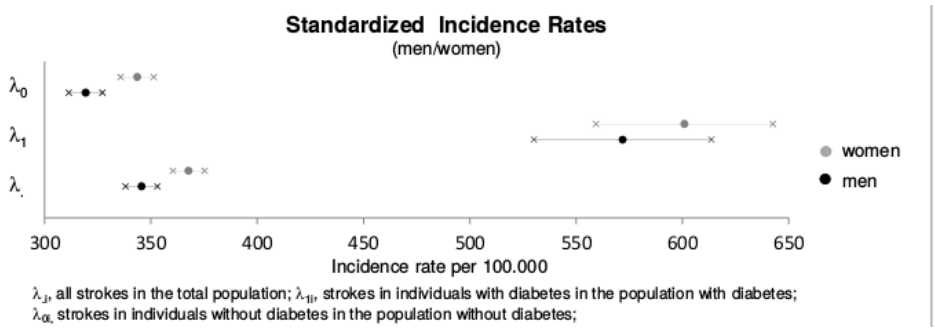

Figure 2: Standardized incidence rates for men and women, Upper Austria

Figure 2 illustrates that there is a significant difference between the stroke risk of women and men not suffering from diabetes. Yet no significant difference between the stroke risk of women and the stroke risk of men that are suffering from diabetes can be observed. This indicates that there must be a sex-specific effect that influences the natural stroke risk which disappears for diabetics (see 5.1).

For the standardized incidence rate of strokes (per 100,000 person years) of the Upper Austrian population, the result of the observation period (2008-2012) was 354.7 (95 % CI = 349.5-360.0) for the total population, 591.4 (95 % CI = 561.8-620.9) in diabetic subjects and 329.0 (95 % CI = 323.5-334.5) in non-diabetic subjects.

The standardized values of the Austrian and European population are quite similar.

4.3. Effect measures

The results of the effect measures that are given in Table 6 show that diabetes can be associated with the incidence of stroke.

The overall relative risk, the overall attributable risk for exposed and the overall population attributable risk are not represented as they are difficult to interpret. The main reason for this is that the first age group (0-44) is bigger than the other age groups. Hence, it has a strong impact on the overall result of the effect measures. However, this age stratum specific effect is taken into account in the overall relative and attributable risks that are standardized to the Upper Austrian, Austrian and European population.

For those that are 44 years old or younger (first age group: 0-44) diabetes has the worst influence on the incidence of stroke. The risk of getting a stroke is 5.44 (men: 5.55, women: 5.26) times higher for

Table 5: Standardized incidence rates for men and women, Upper Austria

	$\lambda_{\cdot i}$		λ_{1i}		λ_{0i}	
Men						
Age (years)						
0-44	29.1	(26.9-31.5)	156.0	(104.5-224.0)	28.1	(25.9-30.4)
45-54	231.1	(217.1-245.9)	434.9	(363.7-516.1)	216.1	(202.0-230.8)
55-64	629.9	(600.7-660.1)	1030.7	(934.6-1134.0)	560.8	(531.0-591.8)
65-74	1250.0	(1200.4-1301.1)	1798.0	(1660.8-1943.4)	1126.0	(1074.0-1179.9)
75-84	2118.8	(2024.9-2216.0)	2633.2	(2381.4-2904.3)	2014.6	(1914.2-2118.8)
85+	2610.1	(2369.8-2868.2)	3318.6	(2555.8-4237.8)	2516.7	(2266.1-2787.4)
All	254.0	(248.5-259.6)	1183.2	(1127.2-1241.1)	210.9	(205.8-216.1)
Standard: Upper Austrian population	345.6	(338.0-353.3)	571.9	(530.1-613.6)	319.3	(311.3-327.2)
Standard: Austrian Population	353.3	(345.5-361.1)	582.1	(540.3-623.9)	326.3	(318.1-334.4)
Standard: European Population	376.7	(368.2-385.1)	611.7	(569.4-653.9)	348.2	(339.4-357.0)
Women						
Age (years)						
0-44	26.5	(24.3-28.8)	135.2	(83.7-206.6)	25.7	(23.5-28.0)
45-54	161.8	(150.0-174.1)	376.1	(292.1-476.8)	152.5	(140.9-164.8)
55-64	383.1	(361.5-405.6)	707.6	(615.5-809.5)	347.9	(326.3-370.6)
65-74	926.8	(890.2-964.4)	1378.0	(1263.5-1500.1)	848.4	(810.5-887.5)
75-84	1926.5	(1861.5-1993.1)	2508.8	(2326.6-2701.4)	1813.7	(1744.9-1884.5)
85+	2387.5	(2262.3-2517.8)	3253.4	(2843.9-3705.3)	2266.7	(2136.7-2402.6)
All	289.5	(283.7-295.3)	1284.9	(1225.6-1346.3)	246.2	(240.7-251.7)
Standard: Upper Austrian population	367.7	(360.3-375.2)	600.9	(559.3-642.5)	343.5	(335.7-351.3)
Standard: Austrian Population	375.6	(368.0-383.3)	611.6	(569.9-653.2)	350.9	(342.9-358.8)
Standard: European Population	383.9	(376.1-391.7)	624.0	(582.5-665.5)	358.4	(350.3-366.5)

$\lambda_{\cdot i}$, all strokes in the total population; λ_{1i}, strokes in individuals with diabetes in the population with diabetes; λ_{0i}, strokes in individuals without diabetes in the population without diabetes

Table 6: Relative risk with confidence intervals (men/women)

	RR_i		ARE_i		PAR_i	
Men						
Age (years)						
0-44	5.55	(3.83-8.07)	0.82	(0.74-0.88)	0.04	(0.02-0.06)
45-54	2.01	(1.68-2.42)	0.50	(0.40-0.59)	0.07	(0.04-0.09)
55-64	1.84	(1.65-2.05)	0.46	(0.39-0.51)	0.11	(0.09-0.13)
65-74	1.60	(1.46-1.75)	0.37	(0.31-0.43)	0.10	(0.08-0.12)
75-84	1.31	(1.17-1.46)	0.23	(0.15-0.31)	0.05	(0.03-0.07)
85+	1.32	(1.01-1.72)	0.24	(0.01-0.42)	0.04	(0.00-0.08)
Standard: Upper Austrian population	1.79	(1.66-1.93)	0.44	(0.40-0.48)	0.08	(0.04-0.11)
Standard: Austrian population	1.78	(1.65-1.93)	0.44	(0.40-0.48)	0.08	(0.05-0.11)
Standard: European population	1.76	(1.63-1.89)	0.43	(0.39-0.47)	0.08	(0.04-0.11)
Women						
Age (years)						
0-44	5.26	(3.40-8.14)	0.81	(0.71-0.88)	0.03	(0.02-0.05)
45-54	2.47	(1.92-3.17)	0.59	(0.48-0.68)	0.06	(0.04-0.08)
55-64	2.03	(1.75-2.36)	0.51	(0.43-0.58)	0.09	(0.07-0.12)
65-74	1.62	(1.48-1.79)	0.38	(0.32-0.44)	0.08	(0.07-0.10)
75-84	1.38	(1.27-1.50)	0.28	(0.21-0.33)	0.06	(0.04-0.08)
85+	1.44	(1.24-1.66)	0.30	(0.20-0.40)	0.05	(0.03-0.07)
Standard: Upper Austrian population	1.75	(1.63-1.88)	0.43	(0.39-0.47)	0.07	(0.04-0.09)
Standard: Austrian population	1.74	(1.62-1.87)	0.43	(0.38-0.47)	0.07	(0.04-0.09)
Standard: European population	1.74	(1.62-1.87)	0.43	(0.39-0.47)	0.07	(0.04-0.09)

a diabetic person of this age group compared to a non-diabetic. The risk of a diabetic person of the second age group (45-54 years) is already considerably smaller, but still more than twice as high as the risk of a person not suffering from diabetes. The relative risk of getting a stroke declines for each age stratum but always stays significantly larger than 1. (see Figure 3).

Standardized individual attributable risk (ARE_s) values show that, 44 per cent of all strokes in the standardized male diabetic population are due to diabetes (women: 43 %). Looking at these results on an age/sex-specific basis, it can be seen that about 80 per cent of the strokes among diabetics in the first age stratum (0-44 years) are due to diabetes. For diabetic women of the second (45-54 years) and third (55-64 years) age class, more than half of the strokes could be avoided if diabetes did not exist any longer. Concerning the latter age groups, the impact of diabetes decreases a little but stays of considerable size. The same effect can be observed for male subjects.

Furthermore, the population attributable risk (PAR) tells us that 8 per cent of all strokes of the total Upper Austrian population are due to diabetes. The third (55-64 years) and fourth (65-74 years) age groups are exposed to the highest additional risk (men: 11 %, 10 % respectively; women: 9 %, 8 % respectively).

After the standardized incidence rates of Austria and Europe for diabetic and non-diabetic subjects are quite similar to those of Upper Austria (with equal proportion), we obtain same results for the standardized effect measures.

5. Discussion

5.1. Study findings

This study quantified incidence rates of brain attacks in order to compare the diabetic and non-diabetic population in Upper Austria between 2008 and 2012. Moreover, relative and attributable stroke risks due to diabetes were estimated. The results illustrated that there is a strong association between diabetes and stroke. By looking at the confounders age and sex, there are significant differences due to sex. Furthermore, the first age-group (0-44 years) differs significantly from the others. As the results of each stratum are higher for diabetics than for non-diabetics the intentions of the Austrian disease management program for patients suffering from diabetes or the diabetics-prevention campaign of the Austrian Diabetic Society (ÖDG) are important in order to decrease the number of diabetics and thus the occurrence of strokes.

Table 5 indicates that in each age strata the stroke risk was higher in men compared to women. However, looking at the sex-specific incidence rate of all age strata, it must be pointed out that more strokes occurred among female subjects. Uncritical use of these conflicting rates may lead to misinterpretations.[3]

By standardizing the incidence rate of strokes (per 100,000 person years) to the Upper Austrian, Austrian and European population the same effect can be observed. This is due to the fact that the amount of female subjects in the age group 85+ is a lot more elevated than the one of men. As the risk in this age group is very high, it has a strong impact on the overall incidence rate.

5.2. Comparison to other studies

As expected, the stroke incidence found in the course of this study is a bit higher compared to other studies (Truelsen et al. 2006; Icks et al. 2011). Truelsen et al. estimated stroke incidence rates in Europe with the help of available data on stroke from various studies. Among other countries, the age/sex-specific stroke incidence rates were also estimated for Austria. Although the estimated incidence rates are lower than the results obtained in this study, it can be seen that the overall trend is similar. In both cases, the incidence rate rises exponentially and men have a higher stroke risk than women. However, it must be kept in mind that the estimates of Truelsen et al. (2006) refer to the

[3] For this reason the overall risks have not been presented.

whole country, while the results of this study only relate to Upper Austria. The same effect can be observed on the results of Icks et al. (2011) who looked at the German population. The differences and similarities of the incidence of stroke of the three studies are visualized in Figure 4 and 5.

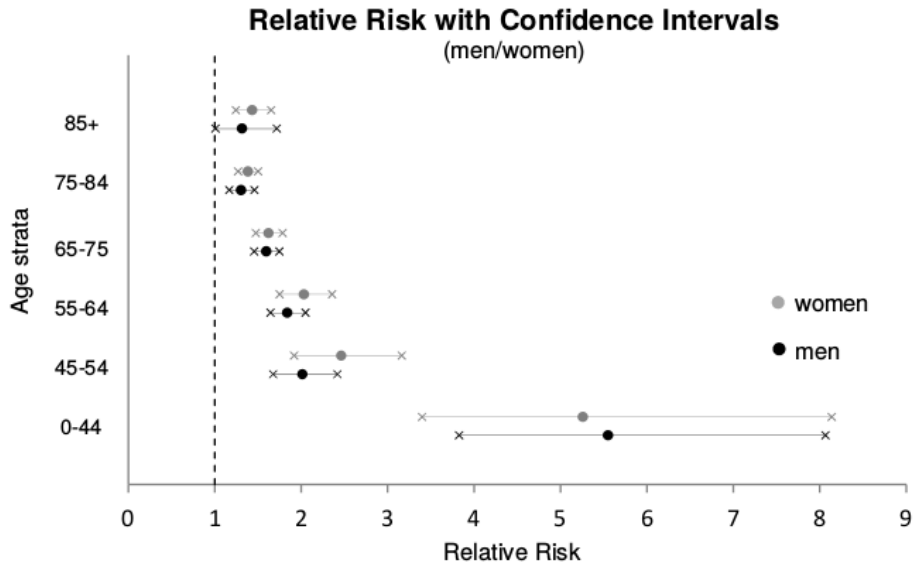

Figure 3: Incidence of stroke for men

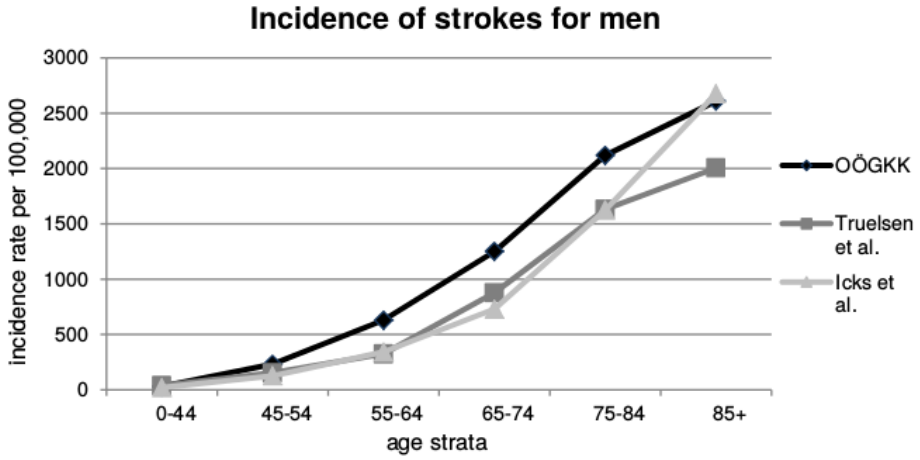

Figure 4: Incidence of stroke for women

Only a limited proportion of population-based studies concentrated on stroke rates in the diabetic and the non-diabetic population (Jeerakathil et al. 2007, Icks et al. 2011). Once again, the same overall trend can be observed when comparing the results of Icks et al. (2011) to this study. Nevertheless, Icks et al.'s results are a bit lower.

Regarding effect measures, various international studies (Icks et al. 2011; Chen et al. 2009; Lin et al. 2007; Kissela et al. 2005) obtained results that are in line with this investigation. Multiple works emphasized a strong association between diabetes and the outcome of stroke especially in lower age groups. Furthermore, it was frequently pointed out that there is no significant difference between the two sexes. Various studies referred to the fact that the relative stroke risk for diabetic women being older than 85 is considerably higher than the relative risk for diabetic men of the same age class which is probably due to the fact that women get older than men.

5.3. Limitations and strengths

Concerning this study, a number of limitations have to be taken into account. First of all, it has to

be said that due to the fact that health insurance data was used, it may have come to some misclassification because of coding mistakes. It is for instance possible that the incidence of stroke was overestimated as vertigo might have been diagnosed as very slight stroke and therefore recorded as cerebral ischaemic attack and related syndromes (G45). Especially during the first part of the observation period the occurrence of this coding mistake is very likely. But due to the existence of the UASR the record of strokes improved over time. For this reason, it would be interesting to replicate the study by only taking the second half of the five-year period into consideration or by ignoring G45 diagnoses. Moreover, the stroke incidence might have been underestimated because fatal strokes and strokes that were treated outside hospitals could not be considered as no data was available in these cases. However, according to Kolominsky-Rabas et al. (1998) about 95 % of strokes are hospitalized. Therefore, it can be assumed that the number of strokes not being hospitalized can be neglected. In addition, it has to be said that confounding variables were not analysed in this study. But Icks et al. (2011) showed that confounding variables like for instance hypertension, dyslipidemia and atrial fibrillation did not influence the stroke rate. A further limitation is that there might be some misclassification in the identification of diabetics, although Köster et al.'s (2007) algorithm to determine diabetics, which was already used by other studies (Icks et al. 2011), was considered. Another point of criticism is that the subjects included in this investigation might not be fully representative for the Upper Austrian population as farmers, teacher, officials and employers are not protected by the OÖGKK. Nevertheless, it can be assumed that this has no considerable impact on the results, as these population groups only represent a small proportion of the Upper Austrian population. This may probably not be neglected for Austrian and European population.

A strength of this study is that the investigations used original data. The database that was used was very large and therefore represents a considerable part (86 %) of the Upper Austrian population. So even if this study only covers a well-defined region it covers this region almost completely. Therefore, it can be used as comparison for studies that deal with large areas but only pick a limited number of cases from each region. Moreover, the time interval of five years that was investigated was large. In addition, by taking every lived through month into account the calculation of person years was very exact. Another strength is that the person years of subjects who were diagnosed as diabetics during the observation period of 2008 to 2012 were adjusted.

Acknowledgements This study was part of a summer research project and supported by the Upper Austrian Stroke Register and the statutory Upper Austrian health insurance. We would like to thank the 14 Upper Austrian hospitals and the number of other institutions that are reporting data to the UASR. For details see:

http://www.ooegkk.at/portal27/portal/ooegkkportal/channel_content/cmsWindow?p_tabid=5&p_menuid=62297&action=2

We thank Dr. Andreas Baierl for useful comments and the editor and referees for their valuable and constructive suggestions which helped us to improve the aim of this article.

References

Almdal, T; Scharling, H.;Jensen, J. S.;Vestergaard, H. (2004). The Independent Effect of Type 2 Diabetes Mellitus on Ischemic Heart Disease, Stroke and Death: A Population-based Study of 13,000 Men and Women with 20 Years of Follow-up. *Archives of Internal Medicin*, 164(13), 1422-1426.

Breslow, N. E.; Day, N. E. (1980). *Statistical Methods in Cancer Research - Volume 1 – The Analysis of Case-control Studies*. Lyon, International Agency for Research on Cancer.

Breslow, N. E.; Day, N. E. (1987). *Statistical Methods in Cancer Research - Volume 2 – The Design and Analysis of Cohort Studies*. Lyon, International Agency for Research on Cancer.

Bundesministerium für Gesundheit (Hrsg), (2012). *Internationale Statistische Klassifikation der Krankheiten und Verwandter Gesundheitsprobleme 10. Revision – BMG-Version 2013*, Wien. (http://bmg.gv.at/cms/home/attachments/1/1/2/CH1241/CMS1287572751172/icd-10_bmg_2013_-_systematisches_verzeichnis.pdf) [accessed on 26/07/13]

Chen, H. F.; Lee, S. P.; Li, C. Y. (2009). Sex Differences in the Incidence of Hemorrhagic and Ischemic Stroke among Diabetics in Taiwan. *Journal Womens Health*, 18 (5), 647-654.

Cole, P.; MacMahon, B., (1971). Attributable Risk Percent in Case-control Studies. *British Journal of Preventive & Social Medicine*, 25, 242-244

Diabetes Care and Research in Europe, (1989). *The Saint Vincent Declaration*. World Health Organization, ICP/CLR 034.

Esteve, J.; Benhamou, E.; Raymond, L. (1994). *Descriptive Epidemiology*. 4 ed.. Lyon, International Agency for Research on Cancer.

Fisz, M. (1989). *Wahrscheinlichkeitsrechnung und Mathematische Statistik*. Berlin, VEB Deutscher Verlag der Wissenschaften.

Grysiewiez, R.; Thomas, K.; Pandey, D. K. (2008). Epidemiology of Ischemic and Hemorrhagic Stroke: Incidence, Prevalence, Mortality and Risk Factors. *Neurologic clinics*, 26 (4), 871–895.

Icks, A.; Scheer, M.; Genz, J.; Giani, G.; Glaeske, G.; Hoffmann, F. (2011), Stroke in the Diabetic and Non-diabetic Population in Germany: Relative and Attributable Risks, 2005-2007. *Journal of Diabetes and its Complications*, 25, 90-96.

Jeerakathil, T.; Johnson, J. A.; Simpson, S. H.; Majumdar, S. R. (2007). Short-term Risk for Stroke is Doubled in Persons with Newly Treated Type 2 Diabetes Compared with Persons without Diabetes: A Population-based Cohort Study. *Stroke*, 38 (6), 1739-1743.

Kahn, H. A.; Sempson, C. T. (1989). *Statistical Methods in Epidemiology*. New York, Oxford University Press.

Kissela, B. M.; Khoury, J.; Kleindorfer, D.; Woo, D.; Schneider, A.; Alwell, K.; Miller, R.; Ewing, I.; Moomaw, C. J.; Szaflarski, J. P.; Gebel, J.; Shukla, R.; Broderick, J. P. (2005). Epidemiology of Ischemic Stroke in Patients with Diabetes – The Greater Cincinnati/Northern Kentucky Stroke Study. *Diabetes Care,* 28 (2), 355-359.

Kolominsky-Rabas, P. L.;Sarti, C.; Heuschmann, P. U.; Graf, C.; Siemonsen, S.; Neundoerfer, B.; Katalinic, A.; Lang, E.; Gassmann, K. G.; von Stockert, T. (1998). A Prospective Community-based Study of Stroke in Germany: The Erlangen Stroke Project (SEPro). Incidence and Case Fatality at 1, 3 and 12 Months. *Stroke*, 29 (12), 2501-2506.

Köster, I.; von Ferber, L.; Ihle, P.; Schubert, I.; Hauner, H. (2006). *The Cost Burden of Diabetes Mellitus: the Evidence from Germany-the CoDiM Study*. Köln, Springer-Verlag.

Lackland, D.; Roccella, E.; Deutsch, A. et al. (2014). Factors Influencing the Decline in Stroke Mortality: A Statement from the American Heart Association/American Stroke Association. *Stroke*, 45:315-353, 326-327

Lin, M.; Chen, Y.; Sigal, R. J. (2007). Stroke Associated with Diabetes among Canadians: Sex and Age Differences. *Neuroepidemiology,* 28 (1), 46-49.

National Stroke Association (2013). *What is a stroke?*

http://www.stroke.org/site/PageServer?pagename=stroke [accessed on 26/07/13].

Österreichische Diabetes Gesellschaft (ÖGD) (2013): Campaign 2013.

http://www.oedg.org/kampagnen.html [accessed on 15/12/2014]

Österreichische Schlaganfallgesellschaft (ÖGSF) (2010). *Schlaganfallprävention bei Patienten mit Diabetes mellitus Typ 2 – Positionspapier Dezember 2010*.

http://www.oegsf.at/aerzte/uploads/Positionspapiere/Positionspapier_Schlaganfallpraevention%20bei%20Patienten%20mit%20DM%20II_ Version%20Dez%202010.pdf [accessed on 27/08/2013]

Ross, J. (2012). *Nervous System*, 4 ed.. Elsevier Health Sciences.

Truelsen, T.; Piechowski-Jóźwiak, B.; Bonita, R.; Mathers, C.; Bogousslavsky, J.; Boysen, G. (2006). Stroke Incidence and Prevalence in Europe: a Review of Available Data. *European Journal of Neurology*, 13 (6), 581-598.

United Nations, Department od Economics and Social Affairs, 2012. World Population

Prospects. (http://esa.un.org/unpd/wpp/Excel-Data/population.htm)

Affiliation:

Karl Schableger
Oberösterreichische Gebietskrankenkasse
Gruberstrasse 77,
4020 Linz,
Austria
E-mail: karl.schableger@ooegkk.at

Lisa Inreiter
Oberösterreichische Gebietskrankenkasse
Gruberstrasse 77,
4020 Linz,
Austria

Austrian Journal of Statistics
October 2015, Volume 44, 85–95.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i3.116

Akademische und Offizielle Statistik vereint. Ein Interview mit Peter Hackl

Peter Hackl
Vormals Statistik Austria, WU Wien

Werner Müller
JKU Linz

Matthias Templ
TU Wien & Statistics Austria

Abstract

Das Interview mit Peter Hackl wurde von Werner Müller und Matthias Templ am 20.03.2014 gehalten. Es zeichnet ein Bild des beruflichen Werdeganges von Peter Hackl, von der Physik und des Welthandeles mit der Statistik in der richtigen Skala, long runners bei Marathons und Textbüchern, von seiner Zeit an der Wirtschaftsuniversität in Wien, Abstechern zur „Handelshögskolan“ , nach Abu Dhabi und Peer-Reviews für exotischen Institutionen wie Eurostat. Weiters werden seine zahlreichen Führungsrollen in der ÖSG und in der Statistik Austria beleuchtet. Im Blickfeld ist auch eine Diskussion über die Ausrichtung der Statistik.

Peter Hackl, ist am 18 August 1942 in Linz geboren. Er war unter anderem Fachstatistischer Generaldirektor der Statistik Austria, Präsident der Österreichischen Statistischen Gesellschaft und Universitätsprofessor an der Wirtschaftuniversität Wien. Er hat mehrere Bücher veröffentlicht, wie z.B. „*Einführung in die Ökonometrie*“ (Pearson Studium).

Keywords: Interview, Offizielle Statistik, Politikberatung.

Matthias Templ: *Wir leben eine gute Tradition. Zwei ehemaligen Präsidenten der Statistischen Gesellschaft wurden bereits für das Austrian Journal of Statistics interviewed, du bist schon der Dritte in der Interviewreihe. Vielen Dank für deine Zusage.*

Peter Hackl: Ich danke auch euch. Ihr habt mir ein sehr ehrenvolles Angebot gemacht, auf das es nicht ganz einfach war, ja oder nein zu sagen.

Werner Müller: *Das Erste, was in deiner Biografie auffällt: du bist ein gebürtiger OÖer, Linzer, und hast in Wien studiert.*

Peter Hackl: Es war eine Folge der Kriegsumstände, dass ich in Linz zur Welt gekommen bin. Meine Großeltern väterlicherseits waren aus dem Hausruck, und dort bin ich die ersten drei Jahre meines Lebens aufgewachsen. Einen Teil dieser Zeit lebte ich auch in

Innsbruck. Ich erzähle ganz gern, dass ich Linzer bin. Tatsache ist aber, dass ich fast mein ganzes Leben in Wien verbrachte.

Werner Müller: *Was natürlich noch auffällt ist, dass du ein Studium der techn. Physik an der TU Wien gemacht und nach dem Studium das Fach gewechselt hast.*

Peter Hackl:

Das stimmt so nicht ganz. Beide Fächer, sowohl die Statistik als auch die Physik, kann man als Disziplin der angewandten Mathematik sehen. Ich habe mir in Mathematik immer sehr leicht getan, in der Schule schon und auch im Studium. Daher hat das Studium sehr gut gepasst.

In der ersten Vorlesung der Experimentalphysik, zentrales Fach für die jungen Studierenden der Physik, ist der honorige Professor Regler in den Hörsaal gekommen, hat in den übervollen Saal geschaut - es gab ca. 200 Physikanfänger - und hat gesagt: „Ich kann Ihnen sagen, es werden nicht mehr von Ihnen als die Zahl der Finger einer Hand einen Job als Physiker bekommen“ . Es war dann nicht ganz so radikal, wie er das gesagt hat, aber es sind sehr viele meiner Studienkollegen nicht in der Physik geblieben; zum Teil sind sie auch im Ausland gelandet. Viele Studienkollegen haben sich der damals aufkommenden Informatik zugewendet. Statistik war damals kein populäres Fach. Mein Semesterkollege Peter Bauer hat sich im Rahmen seiner Dissertation mit einer statistischen Fragestellung befasst und ist so zur Biometrie gekommen. Ich habe während meiner Tätigkeit als Dissertant Paukerkurse in Statistik für WU-Studierende gehalten, damals noch an der Hochschule für Welthandel. Als ich mit meiner Doktorarbeit fertig wurde, ist eine Assistentenstelle am damals ziemlich neuen Statistikinstitut frei geworden. Ich bin in die Sprechstunde zu Professor Roppert gegangen und hab ihm meinen Werdegang geschildert einschließlich meiner Tätigkeit als Pauker.

Werner Müller: *Und er hat dich ein Integral lösen lassen?*

Peter Hackl: Nein (lacht). Professor Roppert hat mich nach zwei Tagen angerufen und ich habe dann einen Job gehabt, Assistent an der Welthandel. Das war im Sommer 1970.

Werner Müller: *Das waren die relativen Anfänge des Instituts.*

Peter Hackl: vorher ist es gegründet worden. Es war damals die Zeit, wo die Orientierung des Instituts in Richtung Statistik noch nicht sehr ausgeprägt war. Das Institut vermittelte den Studierenden eine Grundausbildung in Mathematik und in Statistik; daneben führte das Institut das Rechenzentrum der Welthandel. Mitarbeiter von Roppert waren Heinz Skala, der nach Berlin ging, und dessen Posten ich bekam, und Wolfgang Janko, inzwischen Emeritus des *Department of Information Systems and Operations* der Wirtschaftsuniversität.

Werner Müller: *Und Professor Derflinger ist ja erst später gekommen.*

Peter Hackl: Ja, er dürfte 1975 oder 1976 von der Uni Linz nach Wien gekommen sein und hat eine Abteilung des Instituts mit Schwerpunkt angewandte Statistik und Datenanalyse aufgebaut. Derflinger war gut in der Faktorenanalyse ausgewiesen. Die Abteilung Roppert befasste sich mit Themen in den Bereichen statistische Methoden, Wirtschaftsstatistik, operations research und Informatik.

Werner Müller: *War ein bisschen verpönt mit Daten...*

Peter Hackl: Im Statistik-Institut sind sehr unterschiedliche Themen behandelt worden; als Assistent hat man viele Freiheiten gehabt in diesem Institut. Professor Roppert hat es großartig verstanden, seine Mitarbeiter zu motivieren, und es wurden international herzeigbare Arbeiten produziert. Roppert konnte sehr gut das Gefühl vermitteln, dass gute Forschungsleistungen Voraussetzung einer erfolgreichen Karriere an der Uni sind, und es haben sich alle daran gehalten. Roppert hat seine Mitarbeiter auch sehr gefördert.

Werner Müller: *Die große Herausforderung waren ja die großen Studentenmengen. Die zu bewältigen und dafür eine sinnvolle Didaktik zu entwickeln, das war alles etwas Neuland; die Statistik- und Mathematikausbildung war ja komplett in eurer Hand.*

Peter Hackl: Ja, die war in unserer Hand. Ich kann mich an mein erstes Jahr an der WU erinnern: In den Proseminaren sind Studierende, die keinen Sitzplatz mehr fanden, dicht gedrängt auf den Stiegen des Audi Max gesessen. In meiner Zeit als Student und auch später gab es in der heutigen Zeit unvorstellbare Zustände. Zu den Vorlesungen gab es kaum begleitende Literatur: Passende und studierfreundliche Lehrbücher waren kaum auf dem Markt; es gab keine Skripten; auf beharrliche Nachfrage wurden oft gänzlich ungeeignete Bücher empfohlen. Weil ich aus eigener Erfahrung gewusst habe, wie schwierig es die Studierenden haben, habe ich dann angefangen, den Studierenden zur Vorbereitung auf die Proseminare hektographierte Beispielsammlungen auszuteilen, aus denen dann bei der ÖH verlegte Skripten wurden.

Werner Müller: *So sind dann die legendären Textbücher Hackl, Katzenbeisser, Panny entstanden.*

Peter Hackl: Ja, so hat es angefangen.

Werner Müller: *In wievielen Auflagen sind diese Textbücher erschienen?*

Peter Hackl: Unzählige, das war Policy des Oldenbourg-Verlags. Die Statistik hat es auf elf, die Mathematik auf acht Auflagen gebracht.

Matthias Templ: *Du hast erwähnt, dass die Ausrichtung des "Roppert'schen" Statistikinstituts relativ breit war. Das charakterisiert deine Forschungsausrichtung ebenfalls. Du bist sehr vielseitig und breit aufgestellt. Unter anderem hast dich beispielsweise auch früh mit Partial Least Squares Methoden beschäftigt.*

Peter Hackl: Ich habe viel über lineare Modelle gearbeitet. Ich habe das Glück gehabt hab, dass ich ziemlich am Anfang meiner Assistententätigkeit zu einem Gastaufenthalt an die *University of Edinburgh* gekommen bin, wo Peter Fisk am Statistikinstitut gearbeitet hat. Er war einer der ersten, der über Mehrgleichungssysteme in der Ökonometrie gearbeitet und darüber auch ein Buch geschrieben hat. Ich war zwei Sommer für jeweils zwei Monate in Edinburgh. Peter Fisk hat mich auf ein interessantes Thema aufmerksam gemacht: Jim Durbin hat gemeinsam mit zwei Autoren vom britischen Statistikamt eine Untersuchung über die Identifikation von Brüchen in Regressionsmodellen gemacht, eine methodische Arbeit mit Anwendungen auf bekannte Datensätze. Es gab einen Vortrag über diese Arbeit vor der *Royal Statistical Society* und einen diskutierten Beitrag in der Serie B des *Journal of the Royal Statistical Society*, der 1975 erschienen ist. Peter Fisk war Diskutant dieses Beitrages und er hat mir einen Vorabdruck des Artikels gegeben. Von den Verfahren von Durbin & Co ausgehend, habe ich eigene Ideen zur Diagnose von Strukturbrüchen entwickelt. Eines der Verfahren, die Durbin & Co vorgeschlagen haben, basiert auf kumulativen Summen von sogenannten rekursiven Residuen, das sind im Wesentlichen *one step ahead forecast errors*. Technisch unkompliziertere Indikatoren basieren auf gleitenden Summen (*moving sums*) von rekursiven Residuen, deren Verteilungseigenschaften einfachere Tests erlauben. Entsprechende Verfahren waren der Kern meiner Habilschrift.

Werner Müller: *Es sind zwei Papers entstanden, mit Peter Bauer.*

Peter Hackl: Tests auf Strukturbrüche sind wichtige Instrumente für die Diagnostik von Regressionsmodellen, etwa zur Anwendung in der Ökonometrie, wo bei der Analyse von ökonomischen Zeitreihen immer die Frage zu klären ist, ob das Modell richtig spezifiziert ist, insbesondere, ob die Modellparameter über den gesamten Zeitraum als konstant angesehen werden können. Eine naheliegende Anwendungen dieser Idee in der Prozesskontrolle erlaubt es zu entscheiden, ob die Prozessvariable einen gewünschten Niveauwert

einhält oder ob Abweichungen es notwendig machen, in den Prozess einzugreifen. Auch in dieser Situation können *moving sums* verwendet werden, wie in zwei methodischen Arbeiten gezeigt wird, die 1978 und 1980 in der Zeitschrift Technometrics erschienen sind. Für die *Encyclopedia of Statistical Sciences* von Kotz und Johnson durfte ich einen Eintrag „Moving Sums (MOSUM)"verfassen.

Werner Müller: *Es gibt immer noch Zitate davon, das sind so long runner.*

Peter Hackl: Diagnostik von Regressionsmodellen in der Anwendung von ökonomischen Fragestellungen, Tests auf Strukturbruch, Kontroll-Karten in der Prozesskontrolle etc. sind auch Themen, die von enormer Relevanz für die Anwendung sind.

Werner Müller : *Hast du dann schon begonnen die Arbeiten mit deinen schwedischen Freunden?*

Peter Hackl: In Schweden hat *partial least squares* (PLS) Regression eine lange Tradition. Herman Wold hat die zugrunde liegende Idee zum Modellieren von Relationen zwischen nicht-beobachtbaren Variablen entwickelt. Diese sogenannten *structural equation models* werden typischerweise in den Sozialwissenschaften verwendet. Mein Befassen mit PLS Regression ist nicht so weit entfernt von meinem Interesse für die Diagnostik von Regressionsmodellen. Mit Anders Westlund, einem Schüler von Herman Wold, und anderen Kollegen vom Institut für Wirtschaftsstatistik der Handelshögskolan in Stockholm sind einige Arbeiten entstanden, in denen Modelle entwickelt werden, die beispielsweise zum Schätzen von *customer satisfaction* verwendet werden können. Entsprechende Modelle werden in Schweden und einigen anderen Ländern verwendet, um jährliche nationale *customer satisfaction* Indikatoren zu schätzen.

Werner Müller: *Du hast mit dieser Gruppe eine enge Beziehung, du warst dann oft an der Handelshögskolan.*

Peter Hackl: Ja, ich habe einige Male das Stockholmer Statistikinstitut besucht. In einem gemeinsamen Projekt mit dem Institut für Marketing der WU haben wir auch für Österreich ein Kundenzufriedenheits-Barometer entwickelt und in einer empirischen Studie über die großen Einzelhandelsketten Spar, Billa und Co ausprobiert.

Werner Müller: *Und die Kollaborationen waren so erfolgreich, dass du mit dem Ehrendoktor gewürdigt wurdest.*

Peter Hackl: Die Zusammenarbeit mit dem Stockholmer Statistikinstitut hat in den 70er Jahren begonnen, und wir haben eine Reihe von gemeinsamen Forschungs- und Publikations-Projekte durchgeführt. In der zweiten Hälfte der 1980er Jahre war ich Koordinator einer internationalen Arbeitsgruppe *Statistical Analysis and Forecasting of Economic Structural Change* der IIASA in Laxenburg. In diesem Zusammenhang hat die Handelshögskolan eine internationale Konferenz veranstaltet; ein Ergebnis der Konferenz ist der 1991 bei Springer erschienene Band „Economic Structural Change: Analysis and Forecasting"mit Anders Westlund und mir als Herausgeber. Wir waren auch Gast-Herausgeber einer *Special Section* des *International Journal of Forecasting* zum Thema „*Forecasting in the Manufactoring Industry*", erschienen 1996, und eines *Special Issue* der Zeitschrift *Total Quality Management* zum Thema „*Customer Satisfaction: Theory and Measurement*", das 2000 erschien. Das Ehrendoktorat der Handelshögskolan wurde mir 1996 verliehen.

Werner Müller: *Gut, das waren die 80-iger und dann ist doch die Situation entstanden, wo ich auch persnlich betroffen war, dass dir die Möglichkeit eingeräumt wurde, eine eigene Abteilung zu gründen. Es war, glaube ich, nicht ganz friktionsfrei, und dann aber doch erfolgreich. Ich durfte dein Assistent werden, Anfang der 90-iger Jahre. Es war ja eine kleine Abteilung, die erst später größer geworden ist, aber am Anfang waren wir nur zu*

zweit plus dem Sekretariat und es ist uns ganz gut gelungen die Ökonometrieausbildung an der Wirtschaftsuniversität wieder auf neue Beine zu stellen. Was für mich recht überraschend war, ich bin ja vom IHS gekommen, dass das praktisch nicht vorhanden war.

Peter Hackl: Ja, das war wirklich merkwürdig. Es war in dieser Zeit im Curriculum der Volkswirte gar nicht vorgesehen, dass sich die Studierenden mit quantitativen Methoden vertieft auseinander setzen. Als Vizestudiendekan für Evaluierung, eine Funktion, die ich an der WU zwischen 1995 und 2000 ausübte, haben wir Verfahren entwickelt und systematisch dafür eingesetzt, nicht nur die didaktischen Leistungen der Lehrenden, sondern auch die Lehrprogramme zu bewerten und Verbesserungspotentiale aufzuzeigen. In diesem Rahmen haben wir auch das Curriculum des Volkswirtschaftsstudiums bewertet. Dabei haben wir mit Volkswirten gesprochen, die schon Jahre in der Berufspraxis stehen, beispielsweise in der ÖNB. Dabei wurde uns gesagt, das größte Handicap der Absolventen von der WU sei, dass sie keine Ausbildung in Ökonometrie bekommen. Da gab es wirklich eine Marktlücke. Unsere Abteilung für Wirtschaftsstatistik hat eine Einführung in die Ökonometrie seit den 80er Jahren angeboten; die von den Studierenden ganz gut angenommen worden ist. Ich habe den Studierenden auch ein Skriptum zur Verfügung gestellt, das die Basis meines Buches „Einführung in die Ökonometrie“war, das bei Pearson Studium verlegt und 2012 in zweiter Auflage erschienen ist.

Werner Müller: *Dann in den 90-iger Jahren, da war ich auch am Rande beteiligt, ist dann deine Kollaboration mit Valerii Fedorov und dein Interesse an der Versuchsplanung entstanden. Wir haben ja auch ein paar papers gemeinsam verfasst, die vielleicht nicht sonderlich erfolgreich waren, wenn die Zahl der Zitierungen oder so ansieht, aber immerhin das Buch, das am Rande dieser Aktivität entstanden ist, wird sehr gut angenommen und hat eine hohe Zitationsrate.*

Peter Hackl: Das geringe Interesse an den Aufsätzen hat wohl damit zu tun, dass die Anwendungssituationen, etwa aus dem Bereich der Biometrie, für die uns Daten zur Verfügung standen, eher exotisch waren. Unser Buch „*Model-Oriented Design of Experiments*“, das du angesprochen hast, ist 1997 erschienen und wird bis heute häufig zitiert. Ausgangspunkt für das Buch waren Unterlagen, die Valerii in seiner Vorlesung als Gastprofessor unseres Instituts den Studierenden zur Verfügung gestellt hat.

Mit Valerii Fedorov in Moskau 1989.

Werner Müller: *Jetzt ist es total in. Mit Versuchsplänen für klinische Studien beschäftigen sich heute ganze Institute damit, aber das Thema würde nun zu weit führen.*

Kommen wir in die Mitte der 90-iger Jahre, deine Tätigkeit als Präsident der Österreichischen Statistischen Gesellschaft. Du warst ja mehrere Perioden Präsident und hast die Präsidentschaft in einer schwierigen Phase übernommen, die nicht ganz friktionsfrei war. Vielleicht kannst du schildern, wie die Ausgangslage war.

Peter Hackl: Das war wirklich eine ungute Geschichte. Das Statut der ÖSG hat vorgesehen, dass die Statistische Gesellschaft von zwei Vorsitzenden geführt wird, von denen einer der Präsident des Statistischen Zentralamtes ist, der andere ein Vertreter der akademischen Statistik. An sich eine exzellente Idee, die seit Bestehen der ÖSG so konzipiert war, deren gutes Funktionieren aber von den handelnden Personen abhängt. Ab dem Jahr 1995 war ich gemeinsam mit Präsident Bader einer der Vorsitzenden der ÖSG. Schon in den Jahren davor ist die Zusammenarbeit im Vorstand der ÖSG schwierig geworden,

so dass eine Umstrukturierung angedacht wurde, die in das derzeit bestehende Statut mündete. Die damals eingeführte Arbeitsteilung zwischen Amtlicher, Akademischer und Angewandter Statistik und die Rotation der Präsidentschaft zwischen Vertretern dieser Bereiche haben sich inzwischen sehr gut bewährt. Sie hat zu einer stärkeren Einbindung der Angewandten Statistik und einem gewachsenen Interesse der Akademischen Statistik an Fragen der Angewandten und Amtlichen Statistik geführt. Das war und ist an der seit damals gewachsenen Zahl der Arbeitskreise und Veranstaltungen und an der sehr erfolgreichen Einbindung der Studierenden in die Aktivitäten der ÖSG ablesbar. Ich möchte in diesem Zusammenhang zwei Namen von Personen nennen, die für den Übergang und neuen Aufschwung der ÖSG Entscheidendes geleistet haben: Wilfried Grossmann ist das Funktionieren der ÖSG in der schwierigsten Phase zu verdanken, und Michaela Denk hat den Arbeitskreis Junge Statistik sehr erfolgreich aufgebaut.

Werner Müller: *Viele Dinge von denen wir jetzt noch zehren in unserer Struktur und Aktivitäten ist ja damals eingeführt wurde, wie Statistiktage, Förderpreise und all diese Dinge.*

Peter Hackl:

Es war sicher an der Zeit, neue Ideen einzubringen. Was wollen die ÖSG-Mitglieder, wie kann man die Mitglieder mehr einbeziehen, wie kann man mehr Arbeitsgruppen haben, die gezielt ihre Interessen verfolgen. Die Statistischen Kolloquien, den methodischen Arbeitskreis, haben Peter Bauer und ich Anfang der 70-iger Jahre vorgeschlagen, und in den späten 70-iger Jahren sind weitere Arbeitskreise entstanden.

Nach der Umstrukturierung der ÖSG haben wir die Zahl der Arbeitskreise massiv ausgeweitet. Jeden entsprechenden Vorschlag haben wir gefördert und unterstützt; vieles ist auch wieder eingeschlafen. Aber generell hat sich die Idee gut bewährt, und es ist damals viel entstanden. Auch die Zeitschrift der ÖSG, die Österreichische Zeitschrift für Statistik, haben wir auf neue Beine gestellt.

Matthias Templ: *Was hat schließlich bewegt zur Statistik Austria zu wechseln und dort die Führung als Fachstatistischer Generaldirektor zu übernehmen?*

Peter Hackl: Es war nicht wirklich angestrebt von mir. In meiner Zeit als Präsident der Statistischen Gesellschaft wurden die Beitrittsverhandlungen mit der EU geführt. Da spielte die Amtliche Statistik eine zentrale Rolle: Die Art und Weise, wie ein Land gesehen wird, wird maßgeblich durch Statistiken geprägt, vor allem die Wirtschaftsstatistiken und die makroökonomischen Kennzahlen des Landes. Die Harmonisierung der österreichischen Statistiken mit dem Acquis Communautaire, dem Kompendium der von den Mitgliedstaaten zu liefernden Statistiken, war für das Statistische Zentralamt keine einfache Aufgabe. Komplikationen und vor allem Verzögerungen nahmen Ausmaße an, die schließlich sogar von den Medien breit berichtet und sehr negativ kommentiert wurden. Schließlich hat die Politik dann beschlossen, das Statistikgesetz zu ändern und damit auch die ganze organisatorische Basis der Amtlichen Statistik. Die Statistische Gesellschaft hat sich in die entsprechenden Überlegungen und Bemühungen eingebracht. Im schließlich beschlossenen Statistikgesetz war vorgesehen, dass es zur Statistik Austria, wie das Statistikamt ab 2000 genannt wurde, auch einen Beirat geben soll, den Statistikrat. Er ist, verglichen mit den Beiräten anderer Statistikämter, durch seinen jährlichen Bericht an Parlament und Regierung ein ziemlich mächtiges Gremium. Unter anderem ist vorgesehen, dass ein Vertreter der Wissenschaft, der im Fach Statistik habilitiert ist, im Statistikrat einen Sitz hat. Allzu viele Vertreter der akademischen Statistik mit Affinität zur Amtlichen Statistik gab es nicht, und so hatte ich die Ehre, für den Zeitraum 2000 bis 2004 in den Statistikrat bestellt zu werden. In dieser Zeit musste ich mich naturgemäßsehr massiv mit der Amtlichen Statistik befassen. Die Kompetenzen des Statistikrats sind sehr breit angelegt: Kernangelegenheiten betreffen das Jahresprogramm, den Jahresbericht und den jährlichen Bericht an Parlament und

Regierung; daneben hat sich das Gremium mit verschiedensten anderen Fragen befasst, zentral darunter um Fragen zur Qualität der statistischen Produkte. Als die Amtszeit meines Vorgängers, Ewald Kutzenberger, zu Ende ging, wurde ich vom Bundeskanzleramt eingeladen, eine Bewerbung abzugeben. Nach einer längeren Nachdenkphase habe mich schließlich beworben. Das Ergebnis des Auswahlverfahrens ist euch ja bekannt.

Matthias Templ: *Für die methodische Statistik war es für mich persönlich ein großer Glücksfall, weil von dir der Kontakt zur Methodik intensiv gesucht wurde. Im Speziellen warst du auch interessierst modernere Methoden in der Statistik anzuwenden und zu entwickeln. Aufgrund der Freiheiten die wir von dir bekommen haben, konnten wir in der Zeit einige neue Entwicklungen vorantreiben von denen wir heute noch profitieren. Ich stelle es mir trotzdem schwierig vor neue Wege zu beschreiten, weil man immer die Reibungsstelle ist. Sobald man neue Sachen umsetzen will stößt man oft auf Widerstand und Unverständnis. Aber du bist immer sehr pragmatisch vorgegangen und bist jedem immer in respektvoller Weise begegnet.*

Peter Hackl: Ja, Respekt gegenüber anderen Menschen habe ich schon zu Hause gelernt. Auf der anderen Seite ist der Spielraum für neue Entwicklungen in dieser Funktion beschränkt. Ein Thema bei dem, im Nachhinein gesehen, wahrscheinlich mehr hätte passieren sollen, das aber ganz schwierig war, war der IT-Bereich. Die Statistik Austria hat damals Groß-IT verwendet; das war sehr teuer und sehr unbeweglich mit dem Vorteil einer hohen Sicherheit der Daten. Andererseits konnten viele Chancen der Technologie-Entwicklung nicht genutzt werden. In diesem Bereich habe ich sicherlich nicht alle Erwartungen erfüllen können. Beispielsweise hat sich Josef Richter im Statistikrat massiv für Metadaten-gesteuerte statistische Prozesse eingesetzt hat, ein Thema, das ganz richtig liegt, das man aber nicht so leicht hinbekommt.

Matthias Templ: *Das ist nach wie vor ein Dauerthema.*

Peter Hackl: Jetzt kommt es von der anderen Seite, weil von Eurostat das *generic statistical business process model* (GSBPM) forciert wird, ein Konzept für den statistischen Prozess, das für alle statistischen Produkte anwendbar sein soll. Es wird in vielen europäischen Statistikämtern schrittweise implementiert. Wir haben mit einem Entwurf für ein Metadaten-Repositorium angefangen. Die Umsetzung wäre eine riesige Investition geworden, und es war nicht abzusehen, was herauskommt. Der nächste Schritt des Projektes, die Steuerung der statistischen Prozesse, wäre ein Unternehmen gewesen, an dem einige andere Statistikämter gescheitert sind. Zu radikale Änderungen waren für die Statistik Austria nicht ohne weiteres möglich; man muss auch die Beschränkungen akzeptieren. Andererseits wurde im Rahmen der Strategie 2006-2010 eine große Zahl von innovativen Projekten durchgeführt. Vor allem ist hier der Umstieg von traditionellen Volks-, Gebäude- und Betriebszählungen auf eine registergestützte Zählung zu nennen, ein Vorhaben, dessen Konzeption und Realisierung vor allem dem leider so früh verstorbenen Peter Findl zu verdanken ist. Eine Auswahl weiterer innovativer Projekte: die Neukonzeption der Website, die Entwicklung des regionalstatistischen Online-Atlas, der Aufbau einer Arbeitsmarktstatistik-Datenbank, die Einrichtung des Forschungsschwerpunktes Plausibilitätsprüfung und Imputation, die Einrichtung eines Web-Portals für Studierende, die Kooperation mit Universitäten und anderen Forschungseinrichtungen, die Beteiligung an internationalen Forschungsprojekten wie dem FP7 Projekt Advanced Methods for Laeken Indicators (AMELI) und dem CENEX-Projekt ISAD, die Etablierung der vertieften Kooperation mit den Nationalen Statistischen Ämtern der Nachbarländer Tschechien, Slowakei, Ungarn und Slowenien. Insgesamt ist die Statistik Austria in dieser Zeit, vor allem nach der Umstrukturierung in der Zeit meines Vorgängers Ewald Kutzenberger, unter den europäischen Statistikämtern sehr gut dagestanden.

Matthias Templ: *Ich kann mich an den Peer Review Bericht aus dem Jahr 2008 erinnern, welcher sehr positiv ausgefallen ist.*

Peter Hackl: Aus den Bewertungen der *peer review* Berichte aus 2007 und 2008 kann man ein Ranking der Statistikämter ableiten: Die Statistik Austria hatte darin gemeinsam mit Finnland den ersten Platz. Ich glaube, auch jetzt steht die Statistik Austria recht gut da. In der Sozialstatistik hat Peter Findl tüchtigen, jungen Leuten die Chance gegeben, wirklich schöne Sachen zu machen und aufzubauen. Auch andere Bereiche haben sich sehr gut entwickelt.

Matthias Templ *Du hast vom GSBPM gesprochen, das ist eines von den Modellen, welches momentan sehr stark herumgeistert. Findest du es wichtig?*

Peter Hackl: Ich kann es nicht wirklich beurteilen, dazu müsste man mehr mit den Details vertraut sein. Aber ich habe mir ein paar Websites von Statistikämtern angeschaut und die Ämter von Deutschland, Bulgarien, Zypern und Ungarn als *peer* kennen gelernt. Alle reden vom GSBPM, das ist ja klar, schließlich gibt es eine Vorgabe von Eurostat. Alle denken darüber nach, wie das GSBPM implementiert werden kann. Ein paar sind schon weiter; die Deutschen haben schon einige Module, wie zur Erhebung und zum Validieren und Editieren der Daten. Ich glaube, dass es nirgends sehr rasch gehen wird, weil das Implementieren erstens einen großen Aufwand erfordert und zweitens das GSBPM sehr komplexe Teile enthält. Der Output kann auch von sehr unterschiedlicher Qualität sein. Ob das GSBPM wirklich das Gelbe vom Ei ist, traue ich mich nicht zu beurteilen. Das Konzept sieht vernünftig aus. Es umfasst die Aktivitäten des Statistikers vom Design des statistischen Produkts bis zur Publikation und dem Archivieren. In den meisten statistischen Produktionen braucht man nur Teile davon. Ob man das alles so umsetzen kann, ob es Interfaces gibt zwischen den einzelnen Modulen, die es erlauben, das System wie aus einem Guss zu verwenden, wird man sehen. Meines Wissens gibt es noch nirgends ein komplettes System von der Art des GSBPM. Aber ich glaube, dass es eine gute Sache ist, dass darüber nachgedacht wird, wie man den Statistikprozess systematischer machen kann. Es ist keine Frage, dass sehr viel Potenzial an Vereinfachung fr den laufenden Betrieb vorhanden ist, und eine bessere Kosteneffektivität erreicht werden kann und dringend notwendig ist.

Matthias Templ: *Du hast schon erwähnt, dass einem Peer-review als Generaldirektor der Statistik Austria unterzogen warst. Jetzt bist du selbst Peer-Reviewer.*

Peter Hackl: Ja, ich war eingeladen, in den genannten Ländern am *peer review* mitzuwirken. Es gibt eine neue Runde der *peer reviews*, die 2014/15 läuft. Man möchte wissen, ob die Empfehlungen umgesetzt wurden, die in der ersten Runde ausgesprochen wurden, wie weit die Ämter in der Umsetzung des *Code of Practice* sind, und, ein neues Thema, wie gut die Koordinierung aller mit der Amtlichen Statistik eines Landes befassten Institutionen funktioniert. Wird z.B. die Migrationsstatistik in einem Ministerium oder die Statistiken aus dem Umweltbereich von einer eigenen Anstalt produziert, so sollten, das ist die Vorstellung der EU, die verantwortlichen Stellen ebenfalls die Prinzipien des *Code of Practice* anwenden. Die neuen *peer review* Berichte werden auch dazu Empfehlungen für die Länder machen.

Matthias Templ: *Wäre das eine Empfehlung, dass die Statistik Austria auch Umweltdaten analysieren sollte, weil es Auswirkungen auf die Gesellschaft hat.*

Peter Hackl: In Österreich gibt es schon eine gute Zusammenarbeit mit dem Umweltbundesamt. In meiner Zeit ist da sehr viel gemeinsam besprochen worden, und so viel ich weiß, werden manche Statistiken aus dem Bereich Umwelt von der Statistik Austria produziert. Beim *peer review* Bericht geht es vor allem um die Koordinierungsfunktion, es geht um die professionelle Unabhängigkeit, die sicherstellt, dass alle mit Amtlicher Statistik befassten Institutionen Weisungsfreiheit genießen, und um weitere Punkte wie methodische Zusammenarbeit und Hilfestellung, Schulung, etc. Der *Code of Practice* hat sich in seiner Umsetzung weiterentwickelt, und daher hat man befunden, dass eine

neue Runde *peer reviews* gemacht werden soll. Die Arbeit der *peers* ist naturgemäßsehr interessant.

Werner Müller: *Auch beim Unterrichten bist ja noch aktiv. Vor wenigen Jahren ist eine neue Ausgabe von deinem Ökonometrielehrbuch erschienen, du unterrichtest Ökonometrie noch.*

Peter Hackl: Ja, ich halte für die Studierenden des PhD-Programms in Volkswirtschaft an der Masaryk Universität in Brno eine Pflichtlehrveranstaltung ab. Für einige wenige Studierende, die es genauer wissen wollen, gibt es jeweils im Sommersemester auch einen vertiefenden zweiten Teil. Ich halte diese Vorlesung recht gerne. Mein Ökonometriebuch kann ich leider nicht verwenden, da ein englisches Textbuch zu verwenden ist. Die Studierenden können ganz gut Englisch, Deutsch aber nur vereinzelt.

Matthias Templ: *Ich lese in deinem Lebenslauf noch von einigen interessanten Orten, welche du als statistischer Experte besucht hast, wie in Abu Dhabi, in ... klingt abenteuerlich.*

Peter Hackl: Meine Tätigkeiten waren - leider oder Gott sei Dank - zumindest bisher nicht abenteuerlich. Nach meiner Tätigkeit für die Statistik Austria wurde ich zu verschiedenen Beratungsarbeiten eingeladen. Auf meiner Liste finden sich - neben Abu Dhabi - die Ukraine, Bulgarien, Albanien, Georgien, Azerbaijan, Palestina, und andere. Meist geht es um Themen der Amtlichen Statistik, oft um das Bewerten des Statistikamtes. Das *Statistical Centre of Abu Dhabi* (SCAD) ist ziemlich jung, aber interessant. Das Management ist sehr ehrgeizig, aber die Voraussetzungen sind schwierig. In Abu Dhabi gibt es kein Statistikstudium. Das SCAD hat in wichtigen Bereichen ausländische Experten, natürlich begleitet von einem Chef, der ein Emirati ist. Abu Dhabi hat exzellente Manager, vor allem in der Ölindustrie, die Betriebsführung verstehen sie sehr gut, und auch im SCAD spielen Strategie und Managementtechniken eine wichtige Rolle. Ich hatte im Herbst 2012 die Ehre, eine Strategie, den sogenannten *Masterplan*, zu schreiben, der die wichtigen Entwicklungspotentiale aufzeigt und einen Aktionsplan für die kommenden Jahre vorschlägt. Seither gehöre ich auch dem *International Advisory Comittee* des Generaldirektors des SCAD an, fünf pensionierte Amtsleiter, die halbjährlich zu ein- oder zweitägigen Beratungen zusammen kommen. Noch ein Wort zum Ehrgeiz: Das SCAD wird die 2016-Konferenz der *International Association of Official Statistics* veranstalten. Meine Beratungsarbeiten sind sehr interessante Aufgaben. Man sieht, wo ein Statistikamt seine Probleme hat und wie es sich entwickeln kann. Wenn man da mithelfen kann, ist das sehr interessant.

Matthias Templ: *Ist es nicht zu heiß, einen Marathon in Abu Dhabi zu laufen?*

Peter Hackl: Das habe ich nicht vor, und ich bin gar nicht sicher, ob in Abu Dhabi einer veranstaltet wird. Ich bin viermal den Marathon gelaufen, habe aber dann Probleme mit beiden Knien gehabt. Seither bin ich nur mehr Halbmarathons gelaufen. Ich laufe aber regelmäßig mehrmals die Woche, vor allem aus Gesundheitsgründen, und glaube, dass sich das auch bezahlt macht.

Matthias Templ: *Die Statistik ist derzeit mit großen Herausforderungen konfrontiert. Was sind die Kernkompetenzen der Statistik und wo gibt es eine Abgrenzung zu anderen Fächern. Stichwort Big Data. Hast du eine Vision wie sich die Statistik in Zukunft entwickeln soll?*

Peter Hackl: Ich finde, dass sich das Ansehen und die Akzeptanz der Statistik in der Öffentlichkeit in den letzten Jahren sehr gut entwickelt hat, dass es aber nach wie vor riesige Defizite gibt, was die Rolle der Statistik im öffentlichen Leben spielt. Ich möchte das an einem Beispiel ausführen. Kürzlich habe ich einen Bericht im ORF Radio über die zunehmende Anzahl von resistenten Keimen gehört, die in Krankenhäusern immer mehr zum Problem werden. Eine ganz probate Möglichkeit, mit resistenten Keimen

umzugehen, wären Statistiken über die Zahl der Problemfälle, gegliedert nach Krankenhäusern, die auch noch entsprechend detailliert werden können. Solche Statistiken wären wichtige Hinweise für die Gesundheitspolitik, aber auch für Patienten; sie wären darüber hinaus Basis für das Verstehen der Ausbreitung dieser Keime und für die Planung von notwendigen Maßnahmen. Die Basisdaten gibt es zumindest teilweise. Warum nicht die Statistiken? Eine Linzer Expertin für Hygiene im Krankenhaus erklärte in dieser ORF-Sendung, es wäre gefährlich, der Öffentlichkeit diese Information zu geben: Journalisten würden Horrorgeschichten daraus machen, und die Menschen würden durch diese Fakten verunsichert und verängstigt. Man versteht: Es gibt keinen Druck, diese Statistiken zu produzieren, und der Grund dafür ist vermutlich, dass die Wichtigkeit eines faktenbezogener Umgangs mit der Realität von den Verantwortlichen gar nicht verstanden wird. Ein Beispiel dafür, dass in Österreich das Bereitstellen von Statistiken bisweilen geradezu böswillig verhindert wird, ist das Verweigern detaillierter Ergebnisse der österreichischen PISA-Daten. Ergebnisse nach Regionen und nach Schultypen würden wesentlich zur Versachlichung der laufenden Schuldiskussion beitragen. Ein nicht faktenbezogener Umgang mit der Realität führt zu schlechten und teuren Entscheidungen. In Zeiten der *Big Data* ist das Anliegen noch viel dringender. Das Streichen des Amtsgeheimnisses für die öffentliche Verwaltung wäre ein wichtiger Schritt, um den Umgang mit Informationen in der Öffentlichkeit zu verbessern. Die Statistik Austria hat in letzten Jahren sehr viel dazu beigetragen. Stichwörter aus meiner Zeit sind hier: neue Website, regionalstatistischer Online-Atlas, erleichterter Zugriff auf Daten für die Forschung, flächendeckende Metadaten.

Big Data ist vor allem ein riesen Hype. Symptomatisch ist, dass es noch keine tragfähige Definition für *big data* gibt. Die Verfügbarkeit von ungeheuer großen Datenmengen ist zweifelsohne eine Herausforderung für Statistiker.

Data Science ist ein Bereich, in dem die Statistiker aufpassen müssen, dass sie das Feld nicht zu sehr den Informatikern überlassen. Problematisch ist dabei auch, dass Informatiker inhaltlichen Beschränkungen oft nicht die notwendige Bedeutung beimessen; ein weiterer Grund dafür, dass sich Statistiker in diesen Bereich einbringen.

Mit Werner Müller, Raul Martín-Martín und Jesus Lopéz-Fidalgo (von links) in Neusiedl/See 2005.

Werner Müller: *Dies ist ja mittlerweile widerlegt, das das so gut funktioniert.*

Peter Hackl: Für die Amtliche Statistik können alternative Datenquellen wie das Internet eine sehr brauchbare Ergänzung zu Daten aus Erhebungen und administrativen Quellen sein. Inzwischen gibt es im Bereich der Amtlichen Statistik einige konkrete Anwendungen der Nutzung von solchen Daten; dabei wird auch gerne von *Big Data* gesprochen. Beispiele sind die Verwendung von Ticketpreisen von Fluglinien aus dem Internet für den Verbraucherpreisindex, die Verwendung von Daten von Telefonprovidern in der Tourismusstatistik, oder von *remote sensing* Daten in der Agrarstatistik. Diese Aufzählung zeigt auch, wie heterogen die sogenannten *big data* sind. Die Nachhaltigkeit dieses Begriffs ist jedenfalls zu bezweifeln.

Matthias Templ, Werner Müller: *Danke für das Gespräch.*

Die Interviewer bedanken sich herzlich bei Gabriele Mack-Niederleitner (Johannes Kepler Universität Linz) für die Transkription, und bei Klára Hrůzová für die Konvertierung in das AJS Format.

Affiliation:

Peter Hackl
Wohnhaft im Burgenland

Werner Müller
Johannes Kepler University Linz
Department of Applied Statistics
Altenberger Straße 69
A-4040 Linz, Austria
E-mail: werner.mueller@jku.at

Matthias Templ
Vienna University of Technology &
Statistik Austria
A-1040 Vienna, Austria
E-mail: matthias.templ@gmail.com

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 44 *Submitted:* 2015-03-20
October 2015 *Accepted:* 2015-04-20

Austrian Journal of Statistics
October 2015, Volume 44, 97–98.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i3.103

Reviewer: Ernst Stadlober
TU Graz

Datenqualität in Stichprobenerhebungen. Eine verständnisorientierte Einführung in Stichprobenverfahren und verwandte Themen.

Andreas Quatember
Springer Spektrum, Berlin, Germany, 2014.
ISBN 978-3-642-39605-2. 97–98 pp. EUR 14.99.
http://www.springer.com/us/book/9783642396052

Das Buch spiegelt die langjährige Erfahrung des Autors als Forscher und Lehrer auf dem Gebiet des Stichprobenziehens aus endlichen Grundgesamtheiten wider. Mit großem didaktischen Geschick wird die Leserschaft behutsam Schritt für Schritt in das Fachgebiet eingeführt. Kapitel 1 startet mit dem klassischen Schätzer für die Merkmalssumme bei einer Zufallsstichprobe, dem Horvitz-Thompson-Schätzer, zeigt dessen Unverzerrtheit, berechnet die theoretische Varianz und den unverzerrten Schätzer dieser Varianz. Für die uneingeschränkte oder einfache Zufallsauswahl werden in Kapitel 2 die Schätzer für Merkmalssummen, Mittelwerte, Anzahlen und Anteile angegeben, deren Genauigkeit diskutiert und Formeln für den erforderlichen Stichprobenumfang bei vorgegebener Genauigkeit hergeleitet. In Kapitel 3 wird gezeigt wie man die Genauigkeit von Schätzungen mittels Zusatzinformationen erhöhen kann. Dies wird durch Verhältnis- und Regressionsschätzer erreicht. Darüber hinaus kann die Populationsgrße durch das sehr populäre *capture-recapture* Verfahren ermittelt werden. Die Populationsverteilung und deren Quantile werden hingegen über gewichtete Stichprobenelemente geschätzt. Hier kommt die in der Statistik vielseitig einsetzbare Bootstrapmethode ins Spiel, welche Schätzungen für die komplexe Varianz der Schätzer liefert. Das Problem der Nicht-Antworten und Falsch-Antworten wird hier ebenfalls erörtert.

Durch die in der Praxis sehr wichtige geschichtete Zufallsauswahl kann ein Genauigkeitsgewinn dadurch erzielt werden, dass die Gesamtstichprobe proportional auf die Schichten aufgeteilt wird. Kapitel 4 beinhaltet die entsprechenden Schätzer und analysiert deren Eigenschaften. In den Kapiteln 5 bis 8 werden anspruchsvollere Designs wie die uneingeschränkte Klumpenauswahl, die zweistufige uneingeschränkte Zufallsauswahl, die größenproportionale Zufallsauswahl sowie nichtzufällige Auswahlverfahren untersucht.

Das Werk bietet eine sehr gelungene Einführung in die Problematik der Stichprobenverfahren, welche die theoretischen Ergebnisse mit mathemetischen Beweisen belegt und das Verständnis für den Stoff durch eingängige Beispiele erleichtert. Jedes Kapitel schließt mit einer Zusammenfassung des Inhalts, der neu eingeührten Notation und einem Literaturverzeichnis ab. Diese konsistente Struktur eröglicht es, dass die Leser gezielt nur die für die eigene Fragestellung relevanten Kapitel unabhängig von den anderen Teilen studieren können.

Für jene Leserinnen und Leser, welche die beschriebenen Stichprobenverfahren direkt in die Praxis umsetzen wollen, kann zusätzlich das Buch *G. Kauermann und H. Küchenhoff (2010): Stichproben. Eine praktische Umsetzung mit R. Springer, Berlin,* empfohlen werden.

Reviewer:

Ernst Stadlober
Institut für Statistik
Technische Universität Graz
E-mail: e.stadlober@tugraz.at

Nachruf Dr. Josef Schmidl (1922 – 2014)

von Dr. Kurt KLEIN

Im vergangenen Jahr verstarb ein Ehrenmitglied der ÖSG, das an der jüngeren Entwicklung der amtlichen Statistik großen Anteil hatte. Dr. Schmidl stammte aus Kärnten. Er kam im obersten Mölltal zur Welt und maturierte am Stiftsgymnasium in St. Paul im Lavanttal. Zurück aus dem Krieg begann er neben dem juristischen Studium in Wien seine Tätigkeit in der Statistik. Von den großen Zählungen der Nachkriegszeit über den Aufbau der Wirtschaftsstatistik führt sein Weg in leitende Aufgaben des damaligen Statistischen Zentralamtes, zuletzt als dessen Präsident 1981 – 1987. Wichtige Leistungen dieser Jahre: eine dauerhafte Regelung für die Zusammenarbeit von Bund und Ländern in der österreichischen Statistik; Vernetzung auf europäischer Ebene als Vorstufe für den späteren EU-Beitritt; systemorientierter Ausbau der Wirtschaftsstatistik.

Dr. Schmidl war 1951 Gründungsmitglied der ÖSG, 1982 – 1988 ihr Vorsitzender (gemeinsam mit Dr. Gerhard Bruckmann, danach mit Dr. Reinhard Viertl). 1996 wurde er durch die Ehrenmitgliedschaft ausgezeichnet.

In liebevoller Erinnerung

an Herrn

Dr. jur. Josef Schmidl

9.3.1922 - 28.9.2014

Präsident des Österr. Statistischen Zentralamtes i. R.
Großes Goldenes Ehrenzeichen für Verdienste
um die Republik Österreich
Großes Goldenes Ehrenzeichen des Landes Kärnten
Ehrenzeichen der Stadt St. Pölten

Unsere Gedanken
sind bei Dir.

Wer im Gedächtnis seiner Lieben lebt,
der ist nicht tot, der ist nur fern;
tot ist nur, wer vergessen wird.

ISSN 1026–597X

Contents