# Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

**Volume 43, Number 2, 2014**

# 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

The author(s) retain any copyright on the submitted material. The contributors grant the journal the right to publish, distribute, index, archive and publicly display the article (and the abstract) in printed, electronic or any other form.

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 43, Number 2, 2014

Editor: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

## Österreichische Zeitschrift für Statistik

### Jahrgang 43, Heft 2, 2014

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

## Impressum


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

**Page**

*Herwig FRIEDL and Matthias TEMPL:* Editorial ..... 91

*Andreas QUATEMBER:* The Finite Population Bootstrap – from the Maximum Likelihood to the Horvitz-Thompson Approach ..... 93

*Helga WAGNER and Regina TÜCHLER:* A Comparison of Bayesian Mixed Data Models for Austrian SILC Data ..... 103

*Faton MEROVCI, Ibrahim EELBATAL and Alaa AHMED:* The Transmuted Generalized Inverse Weibull Distribution ..... 119

*Manisha PAL and Montip TIENSUWAN:* The Beta Transmuted Weibull Distribution ..... 133

*Wilfried GROSSMANN, Werner MÜLLER and Matthias TEMPL:* Ein Interview mit Wilfried Grossmann ..... 151

News and Announcements ..... 163

# Editorial

This issue is published by a joint editorship. It contains some very interesting new methods from distribution fitting to Bayesian modelling, concepts like new views on the bootstrap, and applications to EU-SILC data, reliability data and failure times data as well as insights to the Austrian historical development of statistics in Vienna.

The first contribution highlights some insights to the bootstrap method by artificial populations. This approach is especially useful for teaching the concepts of the bootstrap and extends the original work on the bootstrap from Efron.
The second contribution present an application to the Austrian SILC data to analyse material deprivation and household income. The authors present three different MCMC methods for Bayesian regression models on these variables.
In the third contribution, the authors propose a maximum likelihood method for estimating the model parameters of the transmuted generalized inverse Weibull distribution. They applied it on failure times data.
The fourth contribution take investigations in the beta transmuted Weibull distribution and give an example on reliability data.

Finally, as last contribution, an interview with Wilfried Grossmann provides background information on data science and statistics at the University of Vienna as well as details on the relationship of official statistics and statistics at universities. Insights from the development from punching cards to the software environment R and *Big Data* is given.

Please note that all papers in this issue are also available online at

http://www.ajs.or.at

Herwig Friedl
Institute of Statistics
Graz University of Technology
Kopernikusgasse 24/III
A-8010 Graz, Austria
HFriedl@TUGraz.at
http://stat.tugraz.at/friedl.html

Matthias Templ
Statistics Austria &
Vienna University of Technology
Wiedner Hauptstr. 8–10
A–1040 Vienna, Austria
E-mail: matthias.templ@gmail.com
http://www.statistik.tuwien.ac.at/public/templ

Austrian Journal of Statistics
*June 2014, Volume 43, 93–102.*
*http://www.ajs.or.at/*

# The Finite Population Bootstrap - from the Maximum Likelihood to the Horvitz-Thompson Approach

**Andreas Quatember**
Johannes Kepler University

**Abstract**

The finite population bootstrap method is a computer-intensive alternative to estimate the sampling distribution of a sample statistic. In one possible approach, generation of an artificial population, the so-called "bootstrap population", becomes a necessary step between the original sample drawn and the resamples needed to mimic this distribution. For this, the main problem is how to create a bootstrap population that is adequately close-to-real for resampling. For this process, the bootstrap population need not be generated in reality. After an overview of different methods of the finite population bootstrap, this paper presents an approach, based on the idea behind the Horvitz-Thompson estimator, which allows not only whole units in the bootstrap population but also parts of whole units. In a simulation study, this method is compared with a more heuristic technique, taken from the bootstrap literature.

*Keywords*: survey methodology, sampling theory, bootstrap, variance estimation.

## 1. Introduction

The bootstrap technique was originally published by Efron (1979) for the problem of estimating the sampling distribution of a statistic $\widehat{\theta}$, depending on a random sample and an unknown probability distribution of a variable $y$ under study, on the basis of the observed sample. This procedure can be described in the following way (cf. Efron 1979, p.3):

1. An i.i.d. sample of size $n$ is drawn to observe an empirical distribution of the study variable.
2. From this empirical distribution, a bootstrap i.i.d. resample of size $n$ is considered.
3. The sampling distribution of the statistic of interest is approximated by the theoretical bootstrap distribution of it.

This bootstrap distribution equals the sampling distribution of the statistic if the empirical distribution of the variable equals its probability distribution. Efron (1979) considers as "the difficult part of the bootstrap procedure ... the actual calculation of the bootstrap distribution" (p.4). Three methods are possible: The direct theoretical calculation, an approximation

by Taylor series expansion, and a Monte Carlo approximation. The latter has turned out to be most common. In this case, $B$ resamples of the same size as that of the original sample are drawn with replacement from the empirical distribution of $y$, which can be seen as the Maximum-Likelihood (ML) estimator of the underlying probability distribution (cf. Chao and Lo 1994, pp.391). Within each of the $B$ bootstrap samples $s_1, ..., s_B$, the estimator $\widehat{\theta}_b$ of parameter $\theta$ is calculated in the same way as the statistic $\widehat{\theta}$ for the sample $s$ $(b = 1, ..., B)$. For large $B$, the distribution of $\widehat{\theta}_b$ is interpreted as an estimate of the sample distribution of $\widehat{\theta}$. Hence, the theoretical variance $V(\widehat{\theta})$ of $\widehat{\theta}$ is estimated by the Monte Carlo (MC) variance given by

$$V_{MC}(\widehat{\theta}) = \frac{1}{B-1} \cdot \sum_{b=1}^{B} (\widehat{\theta}_b - \overline{\theta})^2 \tag{1}$$

with

$$\overline{\theta} = \frac{1}{B} \cdot \sum_{b=1}^{B} \widehat{\theta}_b. \tag{2}$$

The statistic $\overline{\theta}$ is the mean value of the estimators $\widehat{\theta}_b$ in the $B$ bootstrap samples. For approximately normal sampling distributions, this variance estimator can be used for the calculation of an approximate confidence interval. For large $B$ and non-normal sampling distributions, a confidence interval can be calculated by applying the percentile method (cf. Efron 1981, pp.317ff).

With increasing memory space of computers, an application of this method to finite population sampling became desirable. Such an effort has to consider complex sampling designs consisting of complex estimators and sampling schemes drawing the sample units without replacement, and arbitrary sample inclusion probabilities at various stages of the sampling process (cf., as an example, the discussion of the sampling design of the Austrian PISA survey in Quatember and Bauer 2012). For this purpose, different approaches are available in the relevant literature (cf., for instance, Wolter 2007, p.200ff). One of them rescales the observations in the resamples drawn with replacement from the original sample in a way that the bootstrap variance (1) approximates the actual variance under a given sampling design (cf. Rao and Wu 1988). Another approach is to use the with-replacement bootstrap technique and adjust its bootstrap variance estimator to the parameter by an adequate choice of the size of the resamples (cf. McCarthy and Snowden 1985). Sitter (1992a) presented the "Mirror-Match Method", in which subsamples of the original sample are drawn repeatedly according to the original sampling plan with a subsample size chosen to appropriately match the original variance of the estimator. Antal and Tillé (2011) discuss another approach, in which different with- and without-replacement resampling procedures are mixed in a way that the bootstrap variance estimator calculated from resamples of the same size as that of the original sample under this mixture of resampling schemes, equals the interesting variance.

Furthermore, the finite population bootstrap approach, which is considered a natural extension of the technique by Efron (1979) to finite population sampling, generates an artificial population, the "bootstrap population", from the observed sample data. For this problem, the finite population $U$ of $N$ elements takes over the role of the unknown probability distribution in the i.i.d. bootstrap. The population elements are characterized by their values $y_k$ of $y$ and $x_k$ of a possible auxiliary variable $x$ $(k = 1, ..., N)$. Gross (1980) was the first to adapt the original bootstrap method to the specific case of a simple random sample without replacement (SI), but only with the restriction of integer design weights $\frac{N}{n} \in \mathbb{N}$. For this purpose, from an SI sample $s$, a set-valued finite population estimator $U_G^*$ of size $N_G^* = N$ of the true population $U$ of size $N$ is generated by replicating each sample value $y_k$ exactly $\frac{N}{n}$ times (cf. p.184) providing a variable" $y^*$ denoting these "clones" of the sample values. Hence, the bootstrap population $U_G^*$ can be interpreted as the finite population with the ML regarding the sample drawn (cf. Chao and Lo 1994, p.396). For $\frac{N}{n} = \frac{2{,}400}{400} = 6$, for example, the bootstrap population $U_G^*$ comprises six units of each sample value $y_k$ resulting in a population of total size $N_G^* = 2,400$ $(k = 1, ..., n)$.

This entire process can be seen as the application of the idea behind the unbiased Horvitz-Thompson (H-T) estimator of the total $t$ of $y$, that is

$$t_{H\text{-}T} = \sum_{k=1}^{n} y_k \cdot \frac{1}{\pi_k} \tag{3}$$

with first-order sample inclusion probabilities $\pi_k$, to the problem of generating an adequate bootstrap population. For SI sampling, Eq.(3) results in

$$t_{H\text{-}T} = \sum_{k=1}^{n} y_k \cdot \frac{N}{n}. \tag{4}$$

Obviously, this estimation strategy can be described by the generation of a so-called "pseudo-population" $U^*_{H\text{-}T}$, for which each sample value $y_k$ is replicated exactly $\frac{N}{n}$ times (cf. Quatember 2014, pp.20ff). Compared to the bootstrap population $U^*_G$ from above with $\frac{N}{n} \in \mathbb{N}$, the pseudo-population $U^*_{H\text{-}T}$ of the H-T process allows not only to contain $\lfloor \frac{N}{n} \rfloor$ whole units ($\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$) with the same value $y_k$ of variable $y$ but also an $(\frac{N}{n} - \lfloor \frac{N}{n} \rfloor)$-piece of a unit with that value when $\frac{N}{n}$ is not an integer ($\forall\ k \in s$). For $\frac{N}{n} = \frac{2{,}600}{400} = 6.5$, for example, the H-T pseudo-population $U_{H\text{-}T}$ comprises six whole and one half unit of each sample value $y_k$ $(k = 1, ..., n)$.

After $U^*_G$ is generated, $B$ resamples of size $n$ are drawn from $U^*_G$ following the original sampling method. In other words, the resamples are no i.i.d. samples of size $n$ from the original sample $s$. Instead, the resampling process from $U^*_G$ follows a multivariate hypergeometric distribution with parameters $N$, $n$, and $N$ times 1 (cf., for instance Ranalli and Mecatti 2012). Hence, each of the $n$ sample values $y_1, ..., y_n$ has the same probability $\frac{1}{n}$ of being chosen as the first value in the resample of same size $n$. After the first draw, the same value already drawn at the first step has a probability of $\frac{\frac{N}{n}-1}{N-1} = \frac{N-n}{n(N-1)}$ for being chosen also as the second element of the resample. The other $n-1$ values of $y$ in $s$, not selected as the first resample element, have a probability of $\frac{\frac{N}{n}}{N-1} = \frac{N}{n(N-1)}$ and so on. Generally, a value $y_k$ observed in $s$ has a probability

$$\frac{N - n \cdot h_{k,j-1}}{n \cdot (N - j + 1)} \tag{5}$$

of being selected at the $j$-th step of a resample selection from $U^*_G$ $(j = 1, ..., n)$. In (5), $h_{k,j-1}$ denotes the number of times the value $y_k$ was already selected in the first $j-1$ steps of the process to generate a resample ($h_{k,0} = 0\ \forall\ k \in s$).

This shows that the bootstrap population $U^*_G$ does not have to be generated in reality. The resample process from $U^*_G$ might as well be carried out by applying the probability mechanism described above directly to the sample $s$. This was also discussed by Ranalli and Mecatti (2012) as a resource and time saving alternative to the physical generation of the bootstrap population. These resamples form the basis for the estimation of the sampling distribution of the estimator $\widehat{\theta}$ (for example, the H-T estimator $t_{H\text{-}T}$) for parameter $\theta$ (for instance, the total $t$) in SI sampling. For this purpose, in each of the $B$ resamples $s_b$, the estimator $\widehat{\theta}_b$ has to be calculated in the same way as $\widehat{\theta}$ was calculated in the original sample $s$ $(b = 1, ..., B)$.

Considering, for instance, the estimation of parameter $t$ of variable $y$ by Formula (4), this means that within each SI resample $s_b$ of size $n$, an estimate $t_{H\text{-}T_b}$ is calculated using the replication variable $y^*$ in $U^*_G$:

$$t_{H\text{-}T_b} = \sum_{k=1}^{n} y^*_k \cdot \frac{N}{n}.$$

The MC variance (1) of these $B$ estimates serves as an estimator of the variance $V(t_{H\text{-}T})$ of $t_{H\text{-}T}$ under the SI sampling scheme. This variance estimator is approximately unbiased in large samples (cf., for instance Sitter 1992b, p.139).

Obviously, for general applicability in the practice of survey sampling, this idea has to be extended to

i) non-integer design weights, and

ii) general probability sampling, including stratification and clustering, ensuring a bootstrap population with the same structure as the original population (cf. Chao and Lo 1994, pp.398ff).

The techniques to generate a bootstrap population $U_\cdot^*$ proposed in the relevant literature for different sampling schemes deviate more or less from the ML principle applied by Efron (1979) and Gross (1980) in the generation of the bootstrap population (cf., for instance Bickel and Freedman 1984; Kuk 1989; Sitter 1992b; Chao and Lo 1994; Booth, Butler, and Hall 1994).

General probability proportional to size random sampling without replacement ($\pi$PS), where the selection process is carried out, for instance, by systematic selection of elements ordered randomly in a list (cf. Särndal, Swensson, and Wretman 1992, p.96f), is commonly used in large-scale surveys such as the PISA survey (cf. OECD 2012, ch.4). This sampling method has a positive impact on the efficiency of the HT estimator of $t$, when the study variable $y$ and the size variable $x$ are approximately proportionally related. But, the estimation of the variance of the HT estimator may be hard. In particular, the calculation of the second-order inclusion probabilities, needed for the HT variance estimator, can be cumbersome or even impossible. Holmberg (1998) proposed a bootstrap approach to estimate the variance for general $\pi$PS sampling. The total of size variable $x$ in $U$ is denoted as $t_x$. Under the restriction $x_k \cdot n \leq t_x \; \forall \; k \in U$, the design weight $\frac{1}{\pi_k} = \frac{t_x}{x_k n}$ of survey unit $k$ is decomposed into an integer part $\lfloor \frac{t_x}{x_k n} \rfloor$ and the "rest" $\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor$. To generate the pseudo-population $U_H^*$, the values $y_k$ and $x_k$ of each unit $k$ are jointly replicated $\lfloor \frac{t_x}{x_k n} \rfloor$ times and, independently from each other, randomly once more with probability $\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor$. This process creates a bootstrap population $U_H^*$ of size $N_H^*$ with an expected value of $E(N_H^*) = N$. After $U_H^*$ is generated, the sample inclusion probabilities $\pi_k$ have to be recalculated according to the variable $x^*$ consisting of the replicated sample values of $x$, before the resampling process can start. Then, a number of $B$ resamples of size $n$ are drawn from $U_H^*$ according to $\pi$PS sampling. The estimation of the parameter under study is done in each of these bootstrap samples in the same way as it was done in the original one. For large $N$ and $n$, the bootstrap variance estimator (1), for example, achieves approximate unbiasedness with respect to the variance of the H-T estimator of $t$ (cf. Holmberg 1998, p.381).

Barbiero and Mecatti (2010) aimed to simplify the procedure presented for $\pi$PS sampling by Holmberg (1998) and, at the same time, improve its efficiency with respect to the estimation of the variance of the H-T estimator of $t$. They propose to make "a more complete use of the auxiliary information" (Barbiero and Mecatti 2010, p.62) available for size variable $x$, in particular of its total $t_x$. According to these authors, the following understandable properties should apply to a bootstrap algorithm with respect to the estimation of a total $t$ of variable $y$ (cf. Barbiero and Mecatti 2010, pp.60ff):

1. Given the sample $s$, in a bootstrap population $U_\cdot^*$, the total $t_{x^*}$ of variable $x^*$ should be equal to the total $t_x$ of $x$ in $U$.

2. The total $t_{y^*}$ of variable $y^*$ in $U_\cdot^*$ should be equal to the H-T estimator $t_{H\text{-}T}$ of $t$ calculated in the original sample $s$.

3. For given $s$, over all $B$ resamples $s_b$, the H-T estimator of the total $t_{y^*}$ of $y^*$ in $U_\cdot^*$ should have an expectation equal to the H-T estimator of $t$ in the original sample $s$.

Obviously, these properties are desirable for an efficient estimation of $V(t_{HT})$ by $V_{MC}(t_{HT})$. For different bootstrap methods in the literature dealing with the generation of bootstrap populations, these three properties hold only for $\frac{1}{\pi_k} \in \mathbb{N} \; \forall \; k \in s$. Hence, Barbiero and Mecatti

(2010) at least proposed an "$x$-balanced $\pi$PS-bootstrap", where after replicating each sample unit $k$ a number of $\lfloor \frac{t_x}{x_k n} \rfloor$ times, further units are iteratively added to the bootstrap population $U^*_{BM}$ from a list where these units are sorted in decreasing order of their $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$-value or their ratio $\frac{1}{\pi_k} \cdot \frac{1}{\lfloor \frac{t_x}{x_k n} \rfloor + 1}$ until the minimum difference of $t_{x^*}$ and $t_x$ is achieved. Considering the ratios, for the same $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$-values, elements with a higher integer part $\lfloor \frac{t_x}{x_k n} \rfloor$ of their design weight $\frac{t_x}{x_k n}$ have a higher probability of again being added to $U^*_{BM}$, as compared with elements with a lower integer part. After $U^*_{BM}$ is generated, the probabilities $\pi_k$ have also to be recalculated before the resampling process can start.

But, these proposals for non-integer design weights also result in bootstrap populations not guaranteeing a size $N^*_{BM} = N$ for SI sampling, when $\frac{1}{\pi_k} \notin \mathbb{N}$ (cf. Ranalli and Mecatti 2012, p.4095). For $\lfloor \frac{t_x}{x_k n} \rfloor \in \mathbb{N}$ and SI sampling, the methods of Holmberg (1998) and Barbiero and Mecatti (2010) reduce to the original concept proposed by Gross (1980).

## 2. The proposed bootstrap method

All the methods described in the introductory section are more or less heuristic when it comes to the generation of a bootstrap population in the presence of non-integer design weights (cf. Rao and Wu 1988, pp.237). They all try to establish a bootstrap population to start the resampling process from it, which includes solely integer numbers of replications of the original sample values and, as a consequence, also of the total number of units in the bootstrap population.

In the following, a procedure is proposed, which is a direct application of the idea behind the H-T estimator of a total as it was described below Eq.(4) to the bootstrap population problem. It complements the proposals of Holmberg (1998) and Barbiero and Mecatti (2010) for the problem of non-integer design weights. This Horvitz-Thompson based bootstrap approach (HTB) also allows non-integer numbers of replications of the sample values of $y$ and $x$ to generate the bootstrap population $U^*_{HTB}$. Let each unit $k$ be replicated exactly $\frac{1}{\pi_k} = \frac{t_x}{x_k n}$ times. In this way, a bootstrap population $U^*_{HTB}$ is generated which contains not only $\lfloor \frac{t_x}{x_k n} \rfloor$ whole units with values $y_k$ and $x_k$ but also an additional $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$-piece of a unit with these values when $\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor > 0$ applies $(k \in s)$. In this way, $U^*_{HTB}$ has an expected size $N^*_{HTB}$ of $E(N^*_{HTB}) = \sum_s \frac{1}{\pi_k} = N$. For SI sampling with $\pi_k = \frac{n}{N}$, this means that a bootstrap population with size $N^*_{HTB} = N$ is guaranteed. In the resampling process based on the bootstrap population $U^*_{HTB}$, a whole unit $k$ belonging to this population has a resample inclusion probability proportional to its original $x$-value. But, for a $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$-piece of a unit, this probability is proportional to $(\frac{t_x}{x_k n} - \lfloor \frac{t_x}{x_k n} \rfloor)$ times $x$. Hence, after the generation of $U^*_{HTB}$ as a set-valued estimator of $U$, the design weights of the elements will not have to be recalculated.

From the point of view of the underlying probability mechanism, the value $y_k$ of the original sample $s$ $(k = 1, ..., n)$ has a probability of

$$\frac{t_x - n \cdot h_{k,j-1} \cdot x_k}{n \cdot (t_x - \sum_{s_{b_{j-1}}} x_i)} \tag{6}$$

for being selected into the $b$-th resample at the $j$-th step of the process of drawing $n$ resampling units $(j = 1, ..., n)$ when $\frac{t_x}{x_k n} - h_{k,j-1} \cdot x_k > 0$ applies. Otherwise, for the $j$-th draw, its inclusion probability is set to zero. In Eq. (6), $h_{k,j-1}$ denotes the number of times $y_k$ was already chosen within the first $j-1$ steps of the selection of $n$ units for resample $s_b$. Furthermore, $s_{b_{j-1}}$ denotes the subset of the resample $s_b$ after the $(j-1)$-th draw. Applying this probability mechanism in the resampling process can replace the resource consuming physical generation of the bootstrap population $U^*_{HTB}$. For $x_k = 1 \; \forall \; k \in U$ and $\frac{N}{n} \in \mathbb{N}$, the method reduces to the strategy of the SI technique of Gross (1980) as discussed under Section 1.

In each of the resamples drawn, the original estimator $\widehat{\theta}$ of parameter $\theta$ under study is calculated. In the case of the estimation of total $t$, for instance, the estimator can be the ordinary $H$-$T$ estimator, ratio or regression estimator, or other calibrated estimators based on poststratification or iterative proportional fitting (cf., for instance Alfons and Templ 2013, p.20f).

For the proposed HTB technique, regarding the three desired properties for efficient variance estimation, as mentioned in Section 1 (cf. Barbiero and Mecatti 2010, pp.60ff), the following applies:

1. The total $t_{x^*}$ of size variable $x^*$ in $U^*_{HTB}$ is given by: $t_{x^*} = \sum_{k=1}^{n} x_k \cdot \frac{1}{\pi_k} = t_x$.
2. For the total $t_{y^*}$ of variable $y^*$ in $U^*_{HTB}$, $t_{y^*} = \sum_{k=1}^{n} y_k \cdot \frac{1}{\pi_k} = t_{H\text{-}T}$ applies.
3. The expected value of the H-T estimator of the total $t_{y^*}$ of $y^*$ in $U^*_{HTB}$ yields $E^*(\sum_{k=1}^{n} y^*_k \cdot \frac{1}{\pi_k}) = t_{y^*} = t_{H\text{-}T}$ with $E^*$ denoting the expectation over all resamples, given $s$ and the sampling design.

Clearly, for usual design weights, the proposed HTB method it is not expected to perform much better than, for instance, the technique of Holmberg (1998). Nevertheless, it is shown that the three desirable properties regarding estimation quality mentioned by Barbiero and Mecatti (2010) always hold for this method. The proposed method to generate the bootstrap population might as well seem more understandable in terms of educational reasons than the heuristic methods from the literature, because it follows the same idea as the one behind the widely used H-T estimator when it comes to the composition of the bootstrap population. Moreover, it is not necessary that the bootstrap population needed for the resampling process is physically generated, which often may be cumbersome. The resampling can be done directly from the original sample applying the probability mechanism behind the sampling scheme. Additionally, the HTB bootstrap can still be used in situations, where other methods fail because of first-order sample inclusion probabilities $\pi_k$ of the population units which are close to one. For a $\pi$PS sample, this might provide resample inclusion probabilities that are outside the acceptable range (see Section 3).

## 3. Simulation study

A simulation study was undertaken to compare exemplarily the performance of the proposed H-T based technique (HTB) with the method (H) presented by Holmberg (1998). For this purpose, the Swedish MU 281 population as described in Appendix B of Särndal *et al.* (1992), formed the basis. This specific population was also used by Holmberg (1998) as basis for a simulation study. It consists of all but the largest three municipalities Stockholm, Gothenburg and Malmö.

Different sets of study variables $y$ and auxiliary size variables $x$ were used. For all the variables, in the simulation results, almost the same pattern appeared. Hence, as a typical example, the simulation results for study variable SS82 ($= y$), that is the number of social-democratic seats in municipal council, are presented. The total $t$ of $y$ in the population is 6,193. 10,000 simulations were conducted according to a $\pi$PS sampling scheme with size variable

i) P75 ($\equiv x_1$, the number of inhabitants in the municipality in 1975),

ii) $x_2 = 1 + \frac{x_1}{100}$, and

iii) $x_3 = 1 \; \forall \; k \in U$.

As the values of $x_1$ widely differ, so do the first-order sample inclusion probabilities $\pi_k = \frac{x_k n}{t_x}$. For size variable $x_2$, these probabilities are much closer. With size variable $x_3$, the $\pi$PS method reduces to an SI scheme.

The parameter to be estimated is the variance of the H-T estimator $t_{H\text{-}T}$ of $t$. For each simulation, the chosen sample sizes were $n_1 = 40$ and, if possible, $n_2 = 100$. The chosen number $B$ of bootstrap resamples was $B = 300$.

The simulation results are reported in Tables 1 and 2 for the two simulated sample sizes, as also for all three $\pi$PS sampling schemes where possible. Furthermore, the results are shown exemplarily for one of these setups in the form of boxplots in Figures 1 and 2. For $n_2 = 100$, with $y$ and $x_1$, no $\pi$PS design could be carried out because $\frac{x_{1k}n}{t_{x_1}}$ was greater than one for at least one $k \in U$. Moreover, for each setup, the relative simulation bias (in percent)

$$rb_{sim}(V_{\cdot}) = \frac{E_{sim}(V_{\cdot}) - V_{sim}(t_{H\text{-}T})}{V_{sim}(t_{H\text{-}T})} \cdot 100 \tag{7}$$

was computed as an indicator of its performance. The terms $E_{sim}(V_{\cdot})$ in Eq.(7) and $sd_{sim}(V_{\cdot})$ in Tables 1 and 2 denote the simulation mean values and standard deviations, respectively, of the MC bootstrap variance estimates $V_{\cdot}$ according to Eq.(1) with respect to the bootstrap method HTB ($V_{HTB}$) or H ($V_H$) within the 10,000 simulations. Furthermore, the term $V_{sim}(t_{H\text{-}T})$ in Eq.(7) denotes the variance of the H-T estimates within the 10,000 simulations as the reference value because for $\pi$PS sampling, the true variance of the H-T estimator cannot be calculated exactly. This reference term is substituted by the known variance $V(t_{SI})$ in the simulations of the SI method. In the tables that follow, the simulated standard deviations $sd_{sim}(N^*_{\cdot})$ of the unbiased 10,000 respective bootstrap population sizes $N^*_{\cdot}$ are also presented for both bootstrap methods. Eventually, the percentage coverage rates of approximate confidence intervals in the simulations using the variance estimates can be found in the tables for both approaches.

Table 1: Simulation results for three different sampling designs (sample size $n_1 = 40$)

| Sampling scheme | $\pi$PS $(y, x_1)$ | | $\pi$PS $(y, x_2)$ | | SI $(y)$ | |
|---|---|---|---|---|---|---|
| . | HTB | H | HTB | H | HTB | H |
| $rb_{sim}(V_{\cdot})$ | +1.09 | -1.18 | +0.71 | -2.40 | -1.90 | -2.18 |
| $sd_{sim}(V_{\cdot})$ | 182,044 | 180,328 | 9,662 | 9,688 | 20,867 | 21,055 |
| $sd_{sim}(N^*_{\cdot})$ | 36.34 | 36.28 | 6.53 | 7.06 | 0 | 1.00 |
| coverage $_{\cdot}$ (in %) | 92.61 | 92.34 | 94.23 | 94.23 | 93.72 | 93.64 |

Table 2: Simulation results for two different sampling designs (sample size $n_2 = 100$)

| Sampling scheme | $\pi$PS $(y, x_2)$ | | SI $(y)$ | |
|---|---|---|---|---|
| . | HTB | H | HTB | H |
| $rb_{sim}(V_{\cdot})$ | -2.80 | -0.81 | 2.14 | -0.68 |
| $sd_{sim}(V_{\cdot})$ | 1,936 | 1,975 | 4,054 | 4,084 |
| $sd_{sim}(N^*_{\cdot})$ | 3.28 | 4.99 | 0 | 3.90 |
| coverage $_{\cdot}$ (in %) | 94.98 | 94.70 | 94.54 | 94.30 |

As expected, no major improvement is found with respect to the performance of the variance estimator in comparison to the performance of the one presented by (Holmberg 1998). Nevertheless, in all the simulations, the HTB method strongly tends to perform slightly better with respect to relative bias and standard deviation. Figure 1 shows, as an example, the boxplots regarding the $\pi$PS sampling design with auxiliary variable $x_2$ and sample size $n = 100$.

A $\pi$PS sampling with auxiliary variable $x_1$ is defined only for $n \leq 49$ because for all elements $k$ of the population $U$, $\frac{x_{1k}n}{t_{x_1}} \leq 1$ has to apply. Whereas the HTB method is applicable for all sample sizes $n \leq 49$, method H does not work for sample sizes close to the upper limit of 49, because the inclusion probabilities have to be recalculated for a given bootstrap population $U^*_H$. Depending on the drawn sample $s$, this process may yield resample inclusion probabilities

outside the admissible range. In the HTB method, for the resamples drawn from the bootstrap population, no recalculation of the original first-order inclusion probabilities $\pi_k$ is required. Hence, it is sufficient that the original probabilities are within the admissible range.

Allowing not only integer numbers of clones of the original sample values in the bootstrap population has also an impact on the size $N^*_{HTB}$ of the bootstrap population. While for both the methods, size $N^*_{\cdot}$ is unbiased for $N$, the standard deviation of $N^*_{\cdot}$ is smaller for the HTB method in almost all simulation results. The difference between these standard deviations increases with less differing original first-order inclusion probabilities. This is shown in both the tables as well as in Figure 2.

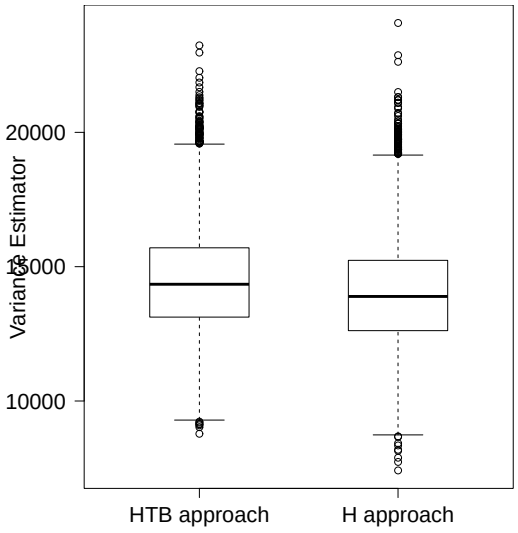


Figure 1: Boxplots of simulated variance estimates calculated according to the HTB and the H approaches for $\pi$PS sampling with size variable $x_2$ ($n_2 = 100$)

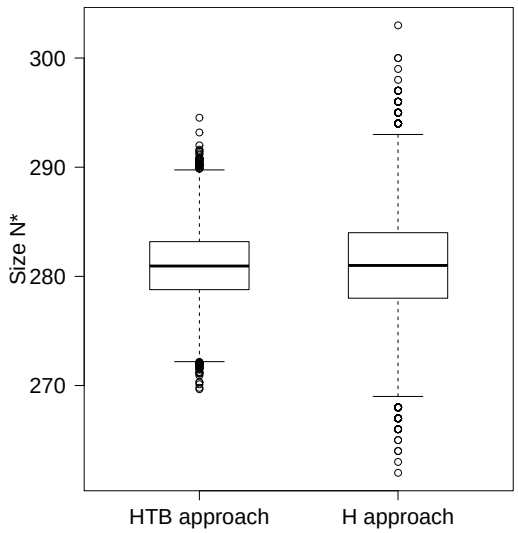


Figure 2: Boxplots of simulated bootstrap population sizes $N^*_{\cdot}$ calculated according to the HTB and H approaches for $\pi$PS sampling with size variable $x_2$ ($n_2 = 100$)

Eventually, in all the simulated cases, the coverage rates of the usual approximate confidence intervals, calculated by using the variance estimates of the HTB method are closer to the desired 95% level than the intervals calculated by using the variance estimates of the H method.

## 4. Conclusion

The H-T approach to the generation of the bootstrap population presented herein applies the idea behind the H-T estimator to the generation of a bootstrap population for finite populations. Overall, as expected, the simulation results indicate that the bootstrap estimator of the variance of a total based on this bootstrap population $U^*_{H\text{-}T}$ is slightly more efficient than the one proposed by Holmberg (1998). For the proposed method, the three properties, considered desirable by Barbiero and Mecatti (2010) for efficient variance estimation, hold. This approach has an effect on the precision of the variance estimates. Applying this approach, the size of the bootstrap population, a variable unbiased for the true size of the original population, has a smaller standard deviation as compared to that of the approach by Holmberg. Furthermore, this method, unlike other methods in the literature, does not require the recalculation of the inclusion probabilities in general $\pi$PS sampling. This also means that the method proposed here can be applied even in situations where other methods fail.

In practice, the generation of the bootstrap population will not have to be processed physically. The whole resampling procedure can be carried out using the probability mechanism behind the process.

However, further studies including other populations than the one used here, and topics such as the optimum number $B$ of resamples, or the estimation of other parameters than totals, are necessary to examine the suitability of this method in greater detail.

**Acknowledgements**

The author is grateful to the Associate Editor and two learned referees for their valuable comments and suggestions that led to improvement of the paper.

## References

Alfons A, Templ M (2013). "Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken." *Journal of Statistical Software*, **54**.

Antal E, Tillé Y (2011). "A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population." *Journal of the American Statistical Association*, **106**.

Barbiero A, Mecatti F (2010). "Bootstrap algorithms for variance estimation in $\pi$PS sampling." In P Mantovan, P Secchi (eds.), *Complex Data Modeling and Computationally Intensive Statistical Methods*, pp. 57–69. Springer, Milan.

Bickel PJ, Freedman DA (1984). "Asymptotic Normality and the Bootstrap in Stratified Sampling." *The Annals of Statistics*, **12**.

Booth JG, Butler RW, Hall P (1994). "Bootstrap Methods for Finite Populations." *Journal of the American Statistical Association*, **89**.

Chao MT, Lo SH (1994). "Maximum Likelihood Summary and the Bootstrap Method in Structured Finite Populations." *Statistica Sinica*, **4**, 389–406.

Efron B (1979). "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics*, **7**, 1–26.

Efron B (1981). "Censored Data and the Bootstrap." *Journal of the American Statistical Association*, **76**.

Gross S (1980). "Median Estimation in Sample Surveys." *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 181–184.

Holmberg A (1998). "A Bootstrap Approach to Probability Proportional-to-Size Sampling." *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 378–383.

Kuk AYC (1989). "Double Bootstrap Estimation of Variance under Systematic Sampling with Probability Proportional to Size." *Journal of Statistical Computation and Simulation*, **31**, 73–82.

McCarthy PJ, Snowden CB (1985). "The Bootstrap and Finite Population Sampling." *National Center of Health Statistics: Data Evaluation and Methods Research*, **2**.

OECD (ed.) (2012). *PISA 2009 Technical Report*. PISA, OECD Publishing. Available on [June 12, 2014]: http://dx.doi.org/10.1787/9789264167872-en.

Quatember A (2014). *Datenqualit?t in Stichprobenerhebungen*. Springer Spektrum, Berlin.

Quatember A, Bauer A (2012). "Genauigkeitsanalysen zu den ?sterreich-Ergebnissen der PISA-Studie 2009." In Eder, F (ed.), *PISA 2009 - Nationale Zusatzanalysen*, pp. 534–550. M?nster, Waxmann Verlag.

Ranalli MG, Mecatti F (2012). "Comparing Recent Approaches for Bootstrapping Sample Survey Data: A First Step Towards A Unified Approach." *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 4088–4099.

Rao JNK, Wu CFJ (1988). "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association*, **83**.

Särndal CE, Swensson B, Wretman J (1992). *Model Assisted Survey Sampling*. Springer, New York.

Sitter RR (1992a). "A Resampling Procedure for Complex Survey Data." *Journal of the American Statistical Association*, **87**.

Sitter RR (1992b). "Comparing Three Bootstrap Methods for Survey Data." *The Canadian Journal of Statistics*, **20**.

Wolter KM (2007). *Introduction to Variance Estimation*. Springer, New York.

## Affiliation:

Andreas Quatember
Institute of Applied Statistics
Johannes Kepler University
A-4040 Linz, Austria
E-mail: andreas.quatember@jku.at
URL: http://www.jku.at/ifas


http://www.ajs.or.at/
http://www.osg.or.at/
*Submitted:* 2013-12-19
*Accepted:* 2014-03-24

Austrian Journal of Statistics
*June 2014, Volume 43, 103–117.*
*http://www.ajs.or.at/*

# A Comparison of Bayesian Mixed Data Models for Austrian SILC Data

**Helga Wagner**
Johannes Kepler University

**Regina Tüchler**
Austrian Federal Economic Chamber

**Abstract**

In many applications multidimensional outcome variables measured on different scales are of interest. In this paper we consider regression modelling of a bivariate response with a normal and a binary component. We use three different approaches to model dependence: a joint logit-normal model for the two responses, a factorization model with linear dependence and a factorization model with flexible non-linear dependence. We apply these approaches to Austrian SILC data to analyse material deprivation and household income.

*Keywords*: joint modelling, logit-normal, factorization model, data augmentation, material deprivation, living conditions.

## 1. Introduction

In the past years the topic of well-being of societies became increasingly important in European politics. It is commonly agreed that the GDP does not sufficiently measure this concept and that complementary indicators are necessary to get a more comprehensive picture of living conditions. Initiatives that deal with this subject are the "GDP and beyond" initiative (http://www.beyond-gdp.eu/), the Stiglitz-Sen-Fitoussi Commission (Stiglitz, Sen, and Fitoussi 2009) and the Sponsorship Group on Measuring Progress, Well-being and Sustainable Development (Eurostat 2011). Currently scoreboards of indicators are being developed in European statistics. These scoreboards consist e.g. of economic indicators, social indicators, or environmental indicators. Since such scoreboards include many different measures we face an increasing need for analyses of dependencies between them and of driving factors. From the methodological point of view the analysis of such dependencies requires models which are able to deal with multidimensional data of mixed type. Our paper meets these needs. We develop mixed data models which incorporate continuous and binary data in regression type models. In our application we focus on social measures coming from the European survey on income and living conditions (EU-SILC).

In this paper we investigate the dependence between a monetary and a material aspect of living conditions in Austria. The monetary aspect is captured via the household income - the money each household has to make its living from, whereas the material aspect is represented by the so-called material deprivation indicator. According to the definition a household faces material deprivation if the members are not capable to meet certain predefined needs like e.g. TV, phone, holiday away from home.

We combine the continuous outcome variable household income and the binary outcome variable material deprivation in mixed data models and analyse their dependence on socio-demographic factors like e.g. the age or activity status of the main-income earner, the household type and migration status. We derive the importance of these explanatory variables by introducing variable selection.

We consider different modelling approaches to deal with multidimensional data of mixed type. We define the models for a continuous and a binary outcome, but they may easily be extended to other data types. In all joint models we use a linear regression model for the continuous and a logistic regression model for the binary outcome. Based on a representation of the binary outcome through a latent continuous utility, the joint bivariate distribution of the error terms is specified as a mixture of bivariate normal distributions in the first modelling approach, whereas the other models use a factorization of the joint bivariate error distribution.

Our paper is organized as follows: Section 2 introduces the different models and corresponding priors for the Bayesian analysis. MCMC estimation is explained in Section 3. In Section 4 we analyse Austrian SILC data and compare results from the three different models. We summarize the findings of the paper in Section 5.

# 2. Model specification

## 2.1. Regression model

Let $\mathbf{y}_i = (y_i^b, y_i^n)'$ denote a bivariate response observed for subjects $i = 1, \ldots, n$, where $y_i^b$ is a binary and $y_i^n$ a normal component. $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{id})$ denotes a $1 \times (d+1)$ vector of covariates. To specify a joint regression model for $\mathbf{y}_i$, we assume an underlying latent continuous variable $u_i$, which determines the value of the binary response: $y_i^b = 1$ if $u_i > 0$ and $y_i^b = 0$ otherwise, and model the bivariate response $(u_i, y_i^n)'$. In all models considered in this paper we use a regression specification for the mean of the bivariate response and model dependence via the error terms.

The joint regression model of $(u_i, y_i^n)$ thus is a SUR (seemingly unrelated regressions) model, given as

$$\begin{pmatrix} u_i \\ y_i^n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i \boldsymbol{\beta}^b \\ \mathbf{x}_i \boldsymbol{\beta}^n \end{pmatrix} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_i^b \\ \varepsilon_i^n \end{pmatrix}, \tag{1}$$

where $\boldsymbol{\beta}^n = (\beta_0^n, \beta_1^n, \ldots, \beta_d^n)'$ and $\boldsymbol{\beta}^b = (\beta_0^b, \beta_1^b, \ldots, \beta_d^b)'$ denote the regression coefficients including the intercept for the two responses. Further we assume that $\varepsilon_i^n \sim \mathcal{N}\left(0, \sigma^2\right)$ in all models, i.e. we specify a standard normal regression model for $y_i^n$,

$$y_i^n \sim \mathcal{N}\left(\mathbf{x}_i \boldsymbol{\beta}^n, \sigma^2\right) \tag{2}$$

and specify either the marginal distribution of $\varepsilon_i^b$ or the conditional distribution of $\varepsilon_i^b | \varepsilon_i^n$ as an (approximate) standard logistic distribution. In all models we use the representation of the standard logistic distribution $\mathcal{L}og(0,1)$ as a scale mixture of six normal components derived by Mohanan and Stefanski (1992). Table 1 gives the fixed variances $s_r^2$ and weights $w_r$ of this approximation, which was shown to be very accurate in Frühwirth-Schnatter and Frühwirth (2010).

Table 1: Variances and weights of the normal components in the finite mixture approximation of the standard logistic distribution.

| $r$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $s_r^2$ | 0.68159 | 1.2419 | 2.2388 | 4.0724 | 7.4371 | 13.772 |
| $100\, w_r$ | 1.8446 | 17.268 | 37.393 | 31.697 | 10.89 | 0.90745 |

Using this approximation the logistic regression model with linear predictor $\eta_i^b$ can be represented as

$$u_i = \eta_i^b + \epsilon_i, \qquad p(\epsilon_i) = \sum_{r=1}^{6} w_r \varphi(\epsilon_i; 0, s_r^2), \tag{3}$$

$$y_i^b = I_{(0,\infty)}(u_i), \tag{4}$$

where $\varphi(y_i; \mu, \sigma^2)$ denotes the pdf of the $\mathcal{N}\left(\mu, \sigma^2\right)$-distribution.

In our first model we specify the bivariate distribution of the error term $\boldsymbol{\varepsilon}_i$ in equation (1) as a finite scale mixture of bivariate normal distributions,

$$p(\boldsymbol{\varepsilon}) = \sum_{r=1}^{6} w_r \varphi_2(\boldsymbol{\varepsilon};\, \mathbf{0}, \boldsymbol{\Sigma}_r), \quad \boldsymbol{\Sigma}_r = \begin{pmatrix} s_r^2 & s_r\sigma\rho \\ s_r\sigma\rho & \sigma^2 \end{pmatrix}. \tag{5}$$

Here $\varphi_2(\mathbf{y};\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the pdf of the bivariate Normal distribution with moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The mixture components share the same correlation $\rho$, but the variances $s_r^2$ are component specific. We call this model the *logit-normal model* (LN), as the marginal distribution of $u_i$ is essentially logistic with mean $\eta_i^b = \mathbf{x}_i\boldsymbol{\beta}^b$, yielding a logit regression model for the binary response $y_i^b$.

As a further modelling approach we consider factorization models with a conditional logit model for $y_i^b$. In the *conditional linear model* (CL) we specify the predictor $\eta_i^b$ as a linear function of the covariates $\mathbf{x}_i$ and the standardized error of the normal model, i.e.

$$\eta_i^b = \mathbf{x}_i\boldsymbol{\beta}^b + \psi\, \frac{y_i^n - \mathbf{x}_i\boldsymbol{\beta}^n}{\sigma}.$$

The composite error of the latent utility is given as

$$\varepsilon_i^b = u_i - \mathbf{x}_i\boldsymbol{\beta}^b = \psi\frac{\varepsilon_i^n}{\sigma} + \epsilon_i,$$

where the error $\epsilon_i$ has a standard logistic distribution and hence the conditional distribution of the composite error is $\varepsilon_i^b|\varepsilon_i^n \sim \mathcal{L}og(\psi\,\varepsilon_i^n/\sigma, 1)$. Note, that marginally $\varepsilon_i^b$ has not a standard logistic distribution but a location mixture of logistic components with normal mixing distribution,

$$p(\varepsilon_i^b) = \int_{\mathcal{R}} p(\varepsilon_i^b|\mathcal{L}og(\psi\,\varepsilon_i^n/\sigma, 1))\; \varphi(\varepsilon_i^n; 0, 1)\; d\varepsilon_i^n,$$

and the conditional linear model therefore is not equivalent to the logit-normal model.

To highlight the difference between both models we make use of the finite mixture approximation of the standard logistic distribution: the joint error distribution in the conditional linear model is

$$p(\varepsilon^b, \varepsilon^n) = \sum_r w_r \varphi(\varepsilon^b; \psi\,\varepsilon^n/\sigma, s_r^2)\varphi(\varepsilon^n/\sigma) = \sum_r w_r \varphi_2((\varepsilon^b, \varepsilon^n/\sigma)'; \mathbf{0}, \boldsymbol{\Sigma}_r^*),$$

where $\boldsymbol{\Sigma}_r^* = \begin{pmatrix} s_r^2 + \psi^2 & \psi\sigma \\ \psi\sigma & \sigma^2 \end{pmatrix}$. This is a mixture of bivariate normal distributions, where correlations $\rho_r = \psi/\sqrt{s_r^2 + \psi^2}$ and variances $s_r^2 + \psi^2$ are component specific. An implication of this model is that dependence will be smaller for mixture components with higher variance. Note, that in contrast the correlation $\rho$ is constant for all mixture components in the logit-normal model.

The third model we consider is an extension of the conditional linear model, where the linear term $\mathbf{x}_i\boldsymbol{\beta}^b$ is combined with a smooth function of the standardized residuals $\varepsilon_i^n/\sigma$ to

$$\eta_i^b = \mathbf{x}_i\boldsymbol{\beta}^b + f(\frac{y_i^n - \mathbf{x}_i\boldsymbol{\beta}^n}{\sigma}). \tag{6}$$

The smooth function $f(z)$ is represented as a linear combination of B-spline basis functions $B_j$, see Lang and Brezger (2004),

$$f(z) = \sum_{j=1}^{J} \gamma_j B_j(z),$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ denotes the coefficients of the B-spline basis functions. As this model allows to capture non-linear dependence between the two error terms $\varepsilon^n$ and $\varepsilon^b$ it is more flexible than the conditional linear model, and we will call it the *conditional flexible model* (CF). We note here that also in the flexible model the marginal error distribution of $\varepsilon^b$ is no longer logistic.

### 2.2. Prior distributions

Bayesian model specification is completed by assigning prior distributions to the model parameters. We consider a prior of the structure $p(\boldsymbol{\beta}^n, \boldsymbol{\beta}^b, \sigma^2, \boldsymbol{\vartheta}) = p(\boldsymbol{\beta}^n)p(\boldsymbol{\beta}^b)p(\sigma^2)p(\boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta}$ denotes the model specific parameters.

Priors for the regression coefficients $\boldsymbol{\beta}^c$, $c = n, b$ could be specified as multivariate normal distributions. As we intend to perform variable selection we will however use spike and slab prior distributions for all regression effects. Spike and slab priors are mixtures of a spike component at zero, which allows to shrink small effects to zero and a rather flat slab component, see e.g. George and McCulloch (1997); Ishwaran and Rao (2005); Malsiner-Walli and Wagner (2011) for different variants of spike and slab priors. We introduce a vector of binary indicators $\boldsymbol{\delta}^c = (\delta_1^c, \ldots, \delta_d^c)'$ for $c = n, b$, with elements taking the value 1 if the corresponding coefficient is unrestricted and 0 otherwise. The prior for $\boldsymbol{\beta}^c$ can be specified hierarchically as

$$p(\boldsymbol{\beta}^c|\boldsymbol{\delta}^c) = p_{\text{slab}}(\beta_0^c) \prod_{j:\delta_j^c=1} p_{\text{slab}}(\beta_j^c) \prod_{j:\delta_j^c=0} p_{\text{spike}}(\beta_j^c), \tag{7}$$

$$p(\boldsymbol{\delta}^c) = \prod_{j=1}^{d} (\omega^c)^{\delta_j^c} (1-\omega^c)^{1-\delta_j^c}, \qquad \omega^c \sim \mathcal{B}\left(a_{0,c}, b_{0,c}\right), \tag{8}$$

where $\mathcal{B}()$ denotes the Beta distribution. Note, that for both components the intercept is not subject to selection in this specification, and hence is assigned a slab prior.

We will use independent normal slabs $p_{\text{slab}}(\beta_j^c) = \varphi(\beta_j^c; 0, B_0^c)$ and either Dirac spikes, i.e. a point mass at zero or continuous spikes specified as $\mathcal{N}\left(0, \alpha B_0^c\right)$ with $\alpha << 1$. Our choice of the spike component is dictated by convenience of MCMC sampling: Dirac spikes require computation of the marginal likelihood, which can be determined analytically only for normal and conditionally normal regression models. With the normal mixture approximation of the standard logistic distribution, marginal likelihoods are analytically available for all parameters in the joint logit-normal model and the conditional linear model and we use a Dirac spike for $\boldsymbol{\beta}^n$ and $\boldsymbol{\beta}^b$ in these models. In the conditional flexible model a Dirac spike is assigned to $\boldsymbol{\beta}^b$ but we use a continuous spike for the regression effects $\boldsymbol{\beta}^n$ of the normal response.

For the remaining model parameters we assign priors which are standard for Bayesian analysis. In the logit-normal model we assume prior independence of $\sigma$ and $\rho$ with $\theta = \ln \sigma \sim \mathcal{N}\left(d_0, D_0\right)$ and a standard normal prior truncated to $[-1, 1]$ for $\rho$. In both factorization models a $\mathcal{G}^{-1}\left(s_0, S_0\right)$-prior is specified for the error variance $\sigma^2$. We chose a normal prior, $\mathcal{N}\left(0, P_0\right)$ for $\psi$ in the conditional linear model. As an alternative a spike and slab prior distribution could be specified to determine whether there is linear dependence between the error terms in both models. Finally, in the conditional flexible model we assume the standard second order random walk prior for the spline coefficients $\boldsymbol{\gamma}$ with variance $\tau^2 \sim \mathcal{G}^{-1}\left(a, b\right)$, see Lang and Brezger (2004).

## 3. Bayesian inference

Bayesian inference for the models specified in Section 2 is feasible by sampling from the posterior distribution using data augmentation and MCMC methods. As we use the latent utility representation for the binary response and the finite mixture approximation of the standard logistic distribution, the latent utilities $\mathbf{u} = (u_1, \ldots, u_n)'$ as well as component indicators $\mathbf{r} = (r_1, \ldots, r_n)'$ have to be sampled additionally to the model parameters. We will use $\boldsymbol{\Theta}$ to denote the collection of all model parameters and the indices LN, CL, CF to address a specific model. The different modelling approaches

lend themselves to different MCMC schemes, which are convenient for posterior estimation. These are detailed below.

### 3.1. MCMC for the logit-normal model

Conditional on the auxiliary variables $\mathbf{u}$ and $\mathbf{r}$ the logit-normal model is a bivariate linear Gaussian regression model with regression coefficients $\boldsymbol{\beta} = \left((\boldsymbol{\beta}^n)', (\boldsymbol{\beta}^b)'\right)'$, where joint sampling of the binary indicators $\boldsymbol{\delta} = \left((\boldsymbol{\delta}^n)', (\boldsymbol{\delta}^b)'\right)'$ and the regression coefficients $\boldsymbol{\beta}$ is feasible. Posterior inference for all model parameters can be accomplished by the following sampling scheme:

(I) Sample the component indicators $\mathbf{r}$ from $p(\mathbf{r}|\boldsymbol{\beta}, \boldsymbol{\delta}, \rho, \theta, \mathbf{u}, \mathbf{y}) \propto \prod_{i=1}^n p(u_i|\boldsymbol{\beta}, \boldsymbol{\delta}, \rho, \theta, r_i, y_i^n)p(r_i)$.

(II) Sample $(\rho, \theta)$ and the latent utilities $\mathbf{u}$:

(IIa) Sample $\rho$ and $\theta$ together from the posterior $p(\rho, \theta|\boldsymbol{\beta}, \mathbf{r}, \mathbf{y}) = \prod_{i=1}^n p(\mathbf{y}_i|\boldsymbol{\beta}, \rho, \theta, r_i)p(\theta)p(\rho)$ using an MH-step.

(IIb) For $i = 1, \ldots, n$ sample the latent utilities $\mathbf{u}$ from the posterior $p(u_i|\boldsymbol{\beta}, \rho, \theta, r_i, \mathbf{y}_i)$.

(III) Sample the indicator variables and the regression coefficients $(\boldsymbol{\delta}, \boldsymbol{\beta})$ from the full conditional posterior $p(\boldsymbol{\delta}, \boldsymbol{\beta}|\rho, \theta, \omega^n, \omega^b, \mathbf{u}, \mathbf{r}, \mathbf{y})$.

(IV) For $c = n, b$ sample $\omega^c$ from the Beta posterior, $\mathcal{B}\left(\sum \delta_j^c + a_{0,c}, d - \sum \delta_j^c + b_{0,c}\right)$.

All sampling steps with the exception of step (IIa) are simple Gibbs steps. The component indicators in step (I) are sampled independently from the discrete distributions

$$P(r_i = r) \propto \phi\left(\frac{u_i - m_{i,r}}{s_{i,r}}\right)\pi_r, \quad r = 1, \ldots, 6,$$

where $m_{i,r}$ and $s_{i,r}$ are the parameters of the conditional normal distribution $u_i|y_i^n \sim \mathcal{N}\left(m_{i,r}, s_{i,r}^2\right)$, given as

$$m_{i,r} = \mathbf{x}_i\boldsymbol{\beta}^b + s_r\rho\,\frac{\varepsilon_i^n}{\sigma}, \tag{9}$$

$$s_{i,r} = s_r\sqrt{1-\rho^2}. \tag{10}$$

As proposal for the MH-step (IIa) we use a bivariate Student t-distribution with 10 degrees of freedom, where the mean is the ML estimate of the likelihood $p(\mathbf{y}|\boldsymbol{\beta}, \theta, \rho, \mathbf{r})$ after a few maximising iterations and the variance-covariance parameter is the inverse Hessian at this point. The full conditionals for the latent utilities $u_i$ in step (IIb) are the normal distributions $\mathcal{N}\left(m_{i,r}, s_{i,r}^2\right)$ truncated to $(-\infty, 0)$ if $y_i^b = 0$ and to $(0, \infty)$ if $y_i^b = 1$.

Details on sampling step (III), which is a standard step for variable selection in Gaussian regression models, are given in Appendix A.

### 3.2. MCMC for the conditional linear model

In the conditional linear model the joint likelihood of normal observations and latent utilities, conditional on the component indicators is given as

$$\prod_{i=1}^n p(u_i, y_i^n|r_i, \boldsymbol{\Theta}_{\text{CL}}) = \prod_{i=1}^n p(y_i^n|\boldsymbol{\beta}^n, \sigma^2)p(u_i|\boldsymbol{\Theta}_{\text{CL}}, s_{r_i}, y_i^n), \tag{11}$$

where

$$u_i|\boldsymbol{\Theta}_{\text{CL}}, s_{r_i}, y_i^n = \mathbf{x}_i\boldsymbol{\beta}^b + \psi\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}^n}{\sigma} + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim \mathcal{N}\left(0, s_{r_i}^2\right). \tag{12}$$

This suggests sampling $(\boldsymbol{\delta}^b, \boldsymbol{\beta}^b, \psi)$ and $(\boldsymbol{\delta}^n, \boldsymbol{\beta}^n)$ separately, as the full conditional posterior distribution of $(\boldsymbol{\delta}^b, \boldsymbol{\beta}^b, \psi)$ involves only the linear regression model (12).

Hence we use the following sampling scheme for the conditional linear model:

(I) Sample the indicators and regression coefficients $(\boldsymbol{\delta}^n, \boldsymbol{\beta}^n)$ jointly from $p(\boldsymbol{\delta}^n, \boldsymbol{\beta}^n|\sigma^2, \boldsymbol{\beta}^b, \psi, \mathbf{u}, \mathbf{r}, \mathbf{y})$.

(II) Sample the error variance $\sigma^2$ from its conditional posterior $\mathcal{G}^{-1}\left(s_0 + n/2, S_0 + \sum(y_i^n - \mathbf{x}_i\boldsymbol{\beta}^n)^2\right)$.

(III) Sample the auxiliary variables $\mathbf{u}$ and $\mathbf{r}$ from the full conditional $p(\mathbf{u}, \mathbf{r}|\sigma^2, \boldsymbol{\beta}^n, \boldsymbol{\beta}^b, \psi, \mathbf{y})$.

(IV) Sample the indicators and regression coefficients in the conditional logit model $(\boldsymbol{\delta}^b, \boldsymbol{\beta}^b, \psi)$ from the full conditional

$$p(\boldsymbol{\delta}^b, \boldsymbol{\beta}^b, \psi|\boldsymbol{\beta}^n, \sigma^2, \mathbf{u}, \mathbf{r}, \mathbf{y}).$$

(V) Sample $\omega^c, c = n, b$ from the $\mathcal{B}\left(\sum \delta_j^c + a_{0,c}, d - \sum \delta_j^c + b_{0,c}\right)$.

Sampling steps (III), (IV) and sampling of $\omega^b$ in step (V) are standard steps for Bayesian variable selection in a logit regression model, see e.g. Wagner and Duller (2012) for full details. If $\psi = 0$ the joint model decomposes into a linear and a logit regression model and the remaining steps (I), (II) and sampling of $\omega^n$ in step (V) perform Bayesian estimation with variable selection in the normal regression model. For $\psi \neq 0$, $\boldsymbol{\beta}^n$ is a regression parameter the heterogeneous linear regression model

$$\begin{aligned} y_i^n &= \mathbf{x}_i\boldsymbol{\beta}^n + \varepsilon_i, & \varepsilon_i &\sim \mathcal{N}\left(0, \sigma^2\right), \\ w_i &= \frac{\psi}{\sigma}\mathbf{x}_i\boldsymbol{\beta}^n + \tilde{\epsilon}_i, & \tilde{\epsilon}_i &\sim \mathcal{N}\left(0, s_{r_i}^2\right), \end{aligned}$$

with working observations defined as $w_i = \psi\frac{y_i^n}{\sigma} - (u_i - \mathbf{x}_i\boldsymbol{\beta}^b)$. For this model, the indicators $\boldsymbol{\delta}^n$ and the regression coefficients $\boldsymbol{\beta}^n$ can be sampled jointly in one Gibbs step, see in Appendix B for details.

### 3.3. MCMC for the conditional flexible model

In the flexible specification of the conditional model given as

$$u_i|y_i^n = \mathbf{x}_i\boldsymbol{\beta}^b + \sum_{j=1}^{J} B_j(\frac{y_i^n - \mathbf{x}_i\boldsymbol{\beta}^n}{\sigma})\gamma_j + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim \mathcal{N}\left(0, s_{r_i}^2\right) \tag{13}$$

the errors of the normal regression model enter nonlinearly. Hence the posterior of $\boldsymbol{\delta}^n$, marginalised over the regression effects $\boldsymbol{\beta}^n$ is not available in closed form. This is the reason why we choose a spike and slab prior with continuous spike for the regression effect $\boldsymbol{\beta}^n$ which allows to sample the indicators $\boldsymbol{\delta}^n$ conditional on $\boldsymbol{\beta}^n$ (see Malsiner-Walli and Wagner 2011, for more details). As the full conditional posteriors of $\boldsymbol{\beta}^n$ and $\sigma^2$ are not of closed form we sample these parameters using an MH-step.

The sampling scheme for posterior inference in the conditional flexible model consists of the following steps:

(I) Sample $\boldsymbol{\beta}^n$ and $\boldsymbol{\delta}^n$.

(Ia) Sample the regression coefficients $\boldsymbol{\beta}^n$ of the normal model from the full conditional $p(\boldsymbol{\beta}^n|\boldsymbol{\delta}^n, \boldsymbol{\beta}^b, \boldsymbol{\gamma}, \sigma^2, \mathbf{u}, \mathbf{r}, \mathbf{y})$ using an MH-step.

(Ib) Sample the indicators $\boldsymbol{\delta}^n$ from the full conditional $p(\boldsymbol{\delta}^n|\boldsymbol{\beta}^n, \omega^n) = \prod_{j=1}^{d} p(\delta_j^n = 1|\beta_j^n, \omega^n)$, where

$$p(\delta_j^n = 1) = \frac{1}{1 + \frac{1-\omega^n}{\omega^n}L_j}, \qquad L_j = \frac{p_{\text{spike}}(\beta_j^n)}{p_{\text{slab}}(\beta_j^n)}.$$

(II) Sample the error variance $\sigma^2$ from its conditional posterior $p(\sigma^2|\boldsymbol{\beta}^n, \boldsymbol{\beta}^b, \boldsymbol{\gamma}, \mathbf{u}, \mathbf{r}, \mathbf{y})$ using an MH-step.

(III) Sample the auxiliary variables $\mathbf{u}$ and $\mathbf{r}$ from the full conditional $p(\mathbf{u}, \mathbf{r}|\sigma^2, \boldsymbol{\beta}^n, \boldsymbol{\beta}^b, \boldsymbol{\gamma}, \mathbf{y})$.

(IV) Sample the indicators and regression coefficients in the conditional logit model $(\boldsymbol{\delta}^b, \boldsymbol{\beta}^b)$ from the full conditional

$$p(\boldsymbol{\delta}^b, \boldsymbol{\beta}^b|\boldsymbol{\beta}^n, \boldsymbol{\gamma}, \sigma^2, \mathbf{y}, \mathbf{u}, \mathbf{r}).$$

(V) Sample the spline coefficients $\boldsymbol{\gamma}$ from the model

$$u_i - \mathbf{x}_i\boldsymbol{\beta}^b = \sum_{j=1}^{J} B_j(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}^n}{\sigma})\gamma_j + \tilde{\epsilon}_i$$

and the hyper-parameter $\tau^2$ from $p(\tau^2|\boldsymbol{\gamma})$.

Sampling steps (III) and (IV) are described in Wagner and Duller (2012) and details on step (V) are given in Lang and Brezger (2004). For sampling $\boldsymbol{\beta}^n$ and $\sigma^2$ in steps (Ia) and (II) respectively, we use the posterior distributions resulting from the marginal regression model (2) as proposals. As these proposals ignore only the information on $\boldsymbol{\beta}^n$ and $\sigma^2$ contained in the binary observations, this strategy works well.

# 4. Analysis of household income and material deprivation

## 4.1. Data

Our data come from the European household survey EU-SILC, which focuses on income and living conditions but also includes questions about socio-demographic attributes. We combine the logarithm of household income and the material deprivation indicator to the bivariate mixed response, which is analysed applying the different models. According to European guidelines a person is hit by material deprivation if at least four out of the following nine criteria are fulfilled: (1) arrears on mortgage or rent payments, utility bills, hire purchase instalments or other loan payments; (2) household cannot afford paying for one week's annual holiday away from home; (3) household cannot afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day; (4) household cannot bear unexpected financial expenses of an amount which varies for different countries and is about 900 Euros for Austria; (5) household cannot afford a telephone (including mobile phone); (6) household cannot afford a colour TV; (7) household cannot afford a washing machine; (8) household cannot afford a car and (9) household is not able to pay for keeping its home adequately warm.

Our data set contains 3694 households from the EU-SILC 2009 survey in Austria BMASK (2011). Following Fusco, Guio, and Marlier (2010) we consider only those data sets where the main-income-earner of the household, i.e. the person with the highest income, is not retired and at least one adult person is less than 60 years old.

We include several covariates which may have an influence on the responses material deprivation and household income, respectively. Some of these covariates are associated with the main-income earner whereas other covariates are household variables. The variables of the main-income-earner are gender, age, activity status (with categories full-time work, part-time work, unemployed and out-of-labour-force), education (with categories lower education, medium education, higher education and university) and migration background. A person has migration background if he or she either now has or once had a non-EU/EFTA citizenship. To allow for a deviation from a pure linear relationship between the two responses and age we add the logarithm of age as predictor. The household variables are the type of household (with six categories: single, two adults/no children, single-parent household, two adults/one or two children, two adults/more than two children and other household), the type of building (categorized in single-family house, house with two families, multi-family house with three to nine households, multifamily house with more than nine households and other) and the population density (with categories high, medium and low).

## 4.2. Variable Selection

As a first step of data analysis we performed Bayesian variable selection to identify important regressors for both response variables. We use a uniform prior for the inclusion probabilities $\omega^c \sim \mathcal{B}(1,1)$, normal slabs with variance $B_0 = 5$ and set $\alpha = 0.005^2$ for the continuous spike. We use a standard

normal prior for $\theta$ in the logit-normal model and an improper $\mathcal{G}^{-1}(0,0)$-prior for $\sigma^2$ in both factorization models. Finally $\psi \sim \mathcal{N}(0,5)$ in the conditional linear model and in the conditional flexible model we use cubic P-splines with 41 inner knots on the interval $[-8,8]$ and $\tau^2 \sim \mathcal{G}^{-1}(0.001, 0.001)$.

To estimate posterior inclusion probabilities MCMC was run for 100 000 iterations after a burn-in of 10 000 draws where the first 5 000 draws are from the unrestricted model. Convergence was checked by running several chains. Integrated autocorrelation times are highest for the conditional flexible model, where they range from 12-82 for the regression coefficients of the binary and from 5-54 for the coefficients of the normal response. Posterior means are estimated by the means of all draws after burn-in.

The variable selection results were similar for the normal response log income in all three models, which differ only with respect to the specification for the binary response. Table 2 gives detailed results on estimated inclusion probabilities. Based on posterior inclusion probabilities larger 0.5 the following variables were selected: linear age effect, dummy variables for all categories of activity status and education, migration background, dummy variables for all household categories except for households with 2 adults and 1-2 children, the category "multifamily house with more than nine households" for the type of building variable and the category "low population density". For material deprivation there are slight differences in the three specifications. In all three models dummy variables for all categories of activity status and education, migration background, as well as two dummies for the type of building variable (3-9 families, multifamily house with more than nine households) were selected. The posterior inclusion probability for a logarithmic effect of age and the dummy for households with two adults and no children are close to 0.5 in all three models. The dummy variables for single parent households and for low population density have a posterior inclusion probability larger than 0.5 in the conditional linear model but slightly below 0.5 in the two other models.

To check for sensitivity with respect to the number of P-spline knots, the analysis was repeated with only 21 inner knots, yielding essentially the same results.

### 4.3. Model Selection

We compare model adequacy of different specifications of the three models by the DIC (Spiegelhalter, Best, Carlin, and Van der Linde 2002), defined as

$$DIC = \overline{D(\boldsymbol{\Theta})} + p_D(\mathbf{y}, \bar{\boldsymbol{\Theta}}(\mathbf{y})).$$

Here $\overline{D(\boldsymbol{\Theta})}$ is the posterior mean of the deviance

$$D(\boldsymbol{\Theta}) = -2 \log p(\mathbf{y}|\boldsymbol{\Theta}) + 2 \log f(\mathbf{y}),$$

where $f(\mathbf{y})$ is some fully specified standardizing term. $p_D(\mathbf{y}, \bar{\boldsymbol{\Theta}}(\mathbf{y}))$ is a measure of complexity,

$$p_D(\mathbf{y}, \bar{\boldsymbol{\Theta}}(\mathbf{y})) = E_{\boldsymbol{\Theta}|\mathbf{y}}(-2 \log p(\mathbf{y}|\boldsymbol{\Theta})) + 2 \log p(\mathbf{y}|\bar{\boldsymbol{\Theta}}(\mathbf{y})),$$

where $\bar{\boldsymbol{\Theta}}(\mathbf{y})$ is the posterior mean of $\boldsymbol{\Theta}$. The model with the smallest DIC is to be preferred. DIC is very popular for Bayesian model comparison, as it can readily be computed from the MCMC output. We follow here Celeux, Forbes, Robert, and Titterington (2006) and set $f(\mathbf{y}) = 1$.

For all models we compute the likelihood based on the factorization

$$\log p(\mathbf{y}|\boldsymbol{\Theta}) = \sum_{i=1}^{n} \left( \log p(y_i^n|\boldsymbol{\Theta}) + \log p(y_i^b|\boldsymbol{\Theta}, y_i^n) \right).$$

For the logit-normal model the conditional distribution of $y_i^b$ is given as

$$p(y_i^b|\boldsymbol{\Theta}, y_i^n) = \begin{cases} 1 - \sum \Phi(\frac{m_{i,r}}{s_{i,r}})\pi_r & \text{if } y_i^b = 0 \\ \sum \Phi(\frac{m_{i,r}}{s_{i,r}})\pi_r & \text{if } y_i^b = 1, \end{cases} \tag{14}$$

where conditional mean $m_{i,r}$ and standard deviation $s_{i,r}$ are given in equations (9) and (10).

Table 2: Posterior inclusion probabilities for the regression effects.

| | household income | | | material deprivation | | |
|---|---|---|---|---|---|---|
| inclusion probability | NL | CL | CF | NL | CL | CF |
| gender (base: male) | 0.20 | 0.23 | 0.15 | 0.30 | 0.36 | 0.23 |
| age (centered at 15) | | | | | | |
| linear (in 10 years) | **1.00** | **1.00** | **0.99** | 0.23 | 0.26 | 0.23 |
| log | 0.06 | 0.07 | 0.07 | 0.49 | **0.50** | **0.51** |
| activity status (base: full time) | | | | | | |
| part-time | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| unemployed | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| out-of-labour | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| education (base: lower) | | | | | | |
| medium | **1.00** | **1.00** | **1.00** | **0.65** | **0.76** | **0.64** |
| higher | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| university | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| migration (base: no migration) | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| type of household (base: single) | | | | | | |
| 2 adults/no children | **1.00** | **1.00** | **1.00** | 0.49 | **0.56** | **0.51** |
| single-parent | **0.84** | **0.89** | **0.74** | 0.41 | **0.54** | 0.42 |
| 2 adults/1-2 children | 0.02 | 0.03 | 0.03 | 0.18 | 0.22 | 0.20 |
| 2 adults/3+ children | **1.00** | **1.00** | **1.00** | 0.19 | 0.26 | 0.20 |
| other | **1.00** | **1.00** | **0.99** | 0.21 | 0.27 | 0.22 |
| type of building (base: single-family) | | | | | | |
| 2 families | 0.02 | 0.03 | 0.02 | 0.23 | 0.27 | 0.22 |
| 3-9 families | 0.45 | 0.47 | 0.22 | **0.95** | **0.92** | **0.92** |
| 10+ families | **1.00** | **1.00** | **0.98** | **1.00** | **0.99** | **0.99** |
| other | 0.15 | 0.18 | 0.13 | 0.40 | 0.45 | 0.40 |
| population density (base: high) | | | | | | |
| medium | 0.04 | 0.06 | 0.03 | 0.17 | 0.22 | 0.17 |
| low | **0.88** | **0.86** | **0.51** | 0.47 | **0.62** | 0.49 |

As the marginal models for the continuous outcomes are essentially the same in all three specifications, we further focus on a comparison of the conditional models for the binary outcome based on the posterior predictive distribution.

For model comparison we use the Brier score Brier (1950) for the binary outcome $y_i^b$,

$$S = \sum_{i=1}^{n} (p(Y_i^b = 1) - y_i^b)^2,$$

which takes the value 0 for a perfect forecast and the maximum value 1 for the worst forecast. An estimate of $S$ is obtained by replacing $p(Y_i^b = 1)$ by its posterior mean, given as

$$\hat{p}(Y_i^b = 1|\boldsymbol{\Theta}, y_i^n) = \frac{1}{m} \sum_{i=1}^{m} p(Y_i^b = 1|\boldsymbol{\Theta}^{(m)}, y_i^n).$$

We compare four specifications of the three models, which differ with regard to the included covariates. The full model includes all covariates for both responses. Models M1-M3 include only those regressors with inclusion probability larger than 0.5 for the normal response and differ with respect to the covariates included in the logit model. The sparsest model is M3, which includes only those regressors selected in all specifications (i.e. all dummies for activity status, education, migration, dummies for type of building 3-9 families and 10 and more families). M2 includes the regressors selected in at least two models (regressors in M3 plus log(age) and dummy for households with 2 adults/no children) and finally M1 includes all regressors selected in at least one specification (regressors in M2 and dummies for single-parent households and low population density). For each set of covariates the DIC is lowest for the conditional flexible model with the lowest overall value obtained for model M1 and and also the Brier Score is lowest for this model.

Table 3: Model comparison: DIC and Brier Score.

| | DIC | | | Brier Score | | |
|---|---|---|---|---|---|---|
| model | NL | CL | CF | NL | CL | CF |
| full | 5233.5 | 5238.3 | **5222.6** | 0.0414 | 0.0415 | **0.0408** |
| M1 | 5222.9 | 5228.1 | **5212.2** | 0.0416 | 0.0416 | **0.0408** |
| M2 | 5225.4 | 5231.6 | **5213.8** | 0.0418 | 0.0418 | **0.0410** |
| M3 | 5229.6 | 5235.4 | **5217.7** | 0.0420 | 0.0420 | **0.0412** |

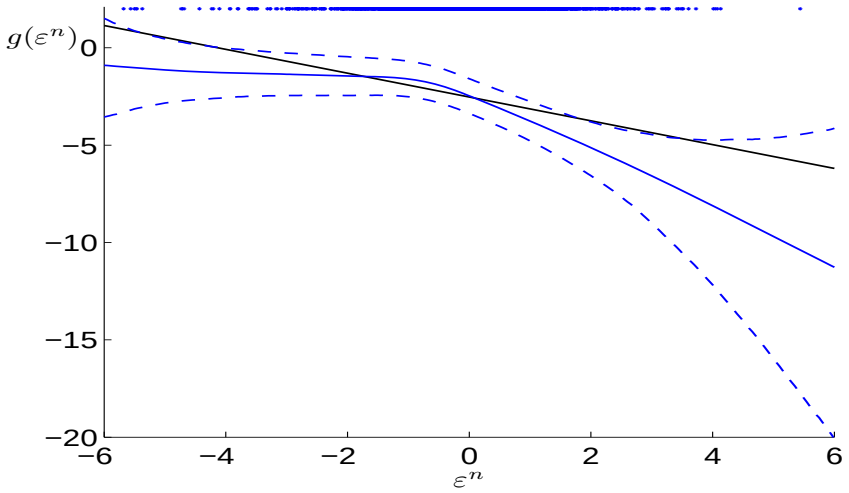


Figure 1: Flexible (full curved line; pointwise 90% HPD-intervals dashed) and linear dependence (full straight line).

Figure 1 compares conditional linear and flexible dependence in model M1. The plot shows the posterior mean of $g(\varepsilon^n) = \beta_0^b + f(\varepsilon^n/\sigma)$, where $\beta_0^b$ is the intercept in the conditional logit model, together with pointwise 90%-credible intervals. The estimated smooth function has a kink with almost zero slope before and a negative slope after the breakpoint. Hence the risk of material deprivation changes only little before the breakpoint but decreases quickly afterwards. Also shown (in black) is the estimated linear function $\beta_0^b + \psi(\varepsilon^n/\sigma)$ in the conditional linear model. The dots at the top of the figure indicate the residuals from the normal model $\hat{\varepsilon}_i = y_i^n - \mathbf{x}_i\hat{\boldsymbol{\beta}}^n$.

### 4.4. Results

In Section 4.3, based on DIC and Brier score, the conditional flexible model was selected and we report estimates for the regression effects on `log(income)` and on material deprivation in Table 4.

From Table 4 we see that age has a positive effect on the household-income. The activity status plays an important role. Naturally, full-time jobs yield the highest household-income and the smallest risk of material deprivation, whereas households with the main-income-earner working only on a part-time basis have less income and a higher risk of material deprivation. These effects are even stronger for households with a main-income-earner who is unemployed or out-of-labour-force. It is well-known that education has an important influence on the living conditions of households. This is also confirmed by our study. The higher the level of education the bigger is the estimated effect on the income and the smaller on the material deprivation response. The importance of the migration status for the economic situation of households is also revealed in our study as households with a main-income-earner who currently has or once had a non-EU/EFTA citizenship have a smaller income and a higher risk of material deprivation. The type of the household has an influence on the income variable. The income of households with two adults and no children is higher than the income of the baseline category single-household, whereas the income of single-parent households and of households with more than two children is smaller. Households living in a building with many flats have less income and are more likely in a situation of material deprivation. Households living in an area with low

Table 4: Posterior mean estimates (std.) of the regression effects.

| variable | log(income) | material deprivation |
|---|---|---|
| Intercept | 9.659 (0.033) | -2.468 (0.542) |
| gender (base: male) | . | . |
| age (centered at 15) | | |
| linear (in 10 years) | 0.071 (0.007) | . |
| log | . | -0.326 (0.151) |
| activity status (base: full-time) | | |
| part-time | -0.243 (0.025) | 1.066 (0.257) |
| unemployed | -0.396 (0.031) | 2.396 (0.231) |
| out-of-labour | -0.550 (0.037) | 1.917 (0.322) |
| education (base: lower) | | |
| medium | 0.128 (0.025) | -0.481 (0.215) |
| higher | 0.276 (0.028) | -1.746 (0.302) |
| university | 0.419 (0.029) | -1.831 (0.335) |
| migration (base: no migration) | -0.285 (0.033) | 1.535 (0.250) |
| type of household (base: single) | | |
| 2 adults/no children | 0.181 (0.018) | -0.405 (0.251) |
| single-parent | -0.110 (0.028) | 0.444 (0.241) |
| 2 adults/1-2 children | . | . |
| 2 adults/3+ children | -0.206 (0.029) | . |
| other | 0.109 (0.021) | . |
| type of building (base: single-family) | | |
| 2 families | . | . |
| 3-9 families | . | 0.643 ( 0.254) |
| 10+ families | -0.091 (0.017) | 0.958 (0.237) |
| other | . | . |
| population density (base: high) | | |
| medium | . | . |
| low | -0.054 (0.015) | -0.395 (0.231) |

population density have less income than households in areas with high or medium density.

The proportion of materially deprived estimated in this model is 0.0531 (std. dev. 0.0038) and corresponds exactly to the proportion of materially deprived in the sample.

In this application interest is also on the probability of material deprivation as a function of income. Though not of primary interest in joint regression modelling (where dependence is modelled in the error distribution) the probability of the binary response taking the value 1 as a function of the continuous response is available for given covariate and parameter values in the joint models considered here. For the conditional flexible model it can be estimated by the posterior mean of $p(Y^b = 1|y^n, \mathbf{x})$, i.e.

$$\hat{p}(Y^b = 1|y^n, \mathbf{x}) = \int p(Y^b = 1|y^n, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma) p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma|\mathbf{y}) d\boldsymbol{\beta} d\boldsymbol{\gamma} d\sigma.$$

From the MCMC draws we compute the estimate

$$\bar{p}(Y^b = 1|y^n) = \sum_{m=1}^{M} \frac{\exp(\eta^{(m)})}{1 + \exp(\eta^{(m)})},$$

where

$$\eta^{(m)} = \mathbf{x}\boldsymbol{\beta}^{b,(m)} + \sum_{j=1}^{J} B_j \left( \frac{y^n - \mathbf{x}\boldsymbol{\beta}^{n,(m)}}{\sigma^{(m)}} \right) \gamma_j^{(m)}.$$

Figure 2 shows the estimated probability of material deprivation for different households. In all plots the reference household with a main income earner of median age 42 and baseline values in all other covariates is compared to a household differing in only one covariate value. The estimated probability

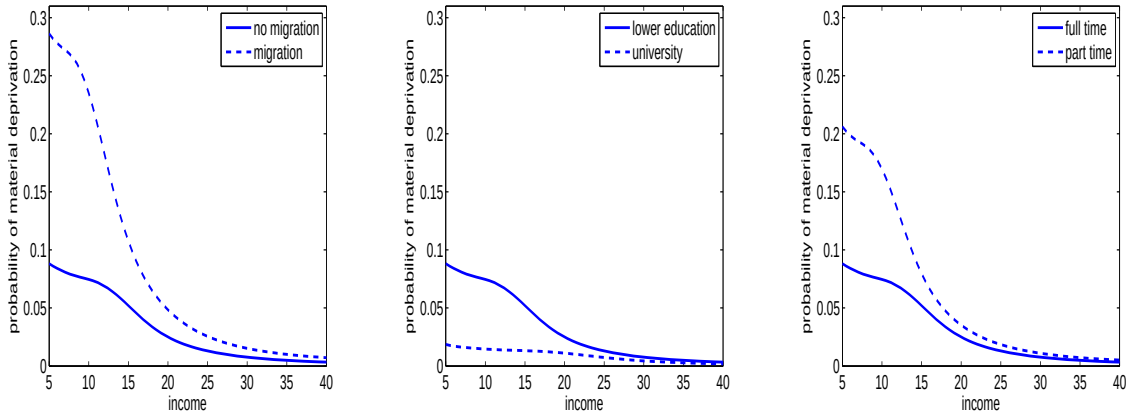


Figure 2: Probability of material deprivation for different households conditional on covariates and income (in 1 000 Euro). Main income earner of age 42 years, baseline values for all other covariates.

of material deprivation is shown for households with/without migration background of the main-income earner in the left panel, for households with lower education/university degree of the main-income-earner in the middle and for households with full-time/part-time employed main-income-earner in the right panel.

For a household income of 21 219 Euro, which is the median income in our sample, the probability of material deprivation is estimated as 2.1% for the reference household but almost doubles to 4.09% when the main-income-earner has migration background. The risk of material deprivation for an otherwise reference household, is estimated as 1.02% if the main-income-earner has a university degree and as 2.97% if the main-income-earner works only part-time.

Due to the non-linear dependence between the error terms each of the curves in Figure 2 has a kink with risk of material deprivation decreasing more pronouncedly afterwards. The income corresponding to this kink is around 10 000 Euro but differs with covariate values.

## 5. Conclusion

We presented three different approaches for joint regression modelling of a bivariate response with a normal and a binary component. In these models we use the latent utility specification for the binary response. The first two models, the normal-logit and the conditional linear model assume linear dependence between the latent utility and the normal response and differ with respect to the marginal error distribution of the latent utility whereas the third model, the conditional flexible model, allows for non-linear dependence between the error terms of the normal response and the latent utility. For each model Bayesian estimation and variable selection is feasible by straightforward MCMC sampling.

In joint regression modelling of log household income and material deprivation, based on the DIC and the Brier score, the conditional flexible model turned out to be preferred to the other models for each of the mean specifications we considered.

Extensions of the flexible model, where the normal distribution of the continuous response is replaced by a scale mixture of normal distributions with fatter tails, e.g. a t-distribution or a normal-gamma distribution Griffin and Brown (2010) is straightforward and requires only an additional sampling step to draw the scale parameters. Also, the logit model could be easily replaced by a probit or a robit model where the latent utility follows a normal or a t-distribution.

Finally we emphasize that our focus was on joint regression modelling of a binary and a normal outcome. All models considered here share the property that the error term and not the observed normal response enters as a regressor in the conditional model for the binary component. This implies that the effect of the continuous on the binary outcome is heterogeneous, i.e. it depends on

the regressors included in the linear predictor of the continuous response, as shown in Figure 2. A non-differential, smooth nonlinear effect of the endogenous continuous variable could be estimated in a model, where the continuous variable enters as endogenous regressor with smooth nonlinear effect in the logit model. Bayesian estimation of models allowing for a smooth effect of an endogenous regressor is considered in Chib, Greenberg, and Jeliazkov (2009) for continuous response with normal and in Wiesenfarth, Hisgen, Kneib, and Cadarso-Suarez (2012) for a flexible error distribution. As noted in Chib *et al.* (2009) extension to a binary response requires a further data augmentation step, in which the unobserved latent utility is sampled. A comparison of our analysis to this latter modelling approach would be an interesting task for future research.

# Appendix

## A Sampling indicator variables and regression coefficients in the logit-normal model

Let $\tilde{\mathbf{y}}_i$ denote the bivariate variables $\tilde{\mathbf{y}}_i = (u_i, y_i^n)'$ and $\tilde{\mathbf{y}}$ the stacked vector of all $\tilde{\mathbf{y}}_i$. By $\tilde{\mathbf{X}}$ we denote the corresponding regressor matrix in the joint regression model

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

and $\boldsymbol{\Sigma}$ is the block diagonal matrix with blocks $\boldsymbol{\Sigma}_{r_i}$ given in equation (5).

We sample the indicator variables one at a time from the posterior marginalised over $\boldsymbol{\beta}$, which is given as

$$p(\delta_j = 1 | \boldsymbol{\delta}_{\setminus j}, \rho, \theta, \tilde{\mathbf{y}}) = \frac{1}{1 + \frac{p(0, \boldsymbol{\delta}_{\setminus j})}{p(1, \boldsymbol{\delta}_{\setminus j})} R_j}, \quad R_j = \frac{p(\tilde{\mathbf{y}} | 0, \boldsymbol{\delta}_{\setminus j}, \rho, \theta)}{p(\tilde{\mathbf{y}} | 1, \boldsymbol{\delta}_{\setminus j}, \rho, \theta)},$$

where $\boldsymbol{\delta}_{\setminus j}$ includes all indicators but $\delta_j$. The posterior for $\delta_j$ involves the conditional marginal likelihoods of two heteroscedastic linear regression models with design matrices differing only by inclusion/exclusion of the j-th column of the matrix $\tilde{\mathbf{X}}$. The conditional marginal likelihood of a linear regression model is available in closed form as

$$p(\tilde{\mathbf{y}} | \boldsymbol{\delta}, \rho, \theta) \propto \frac{|\mathbf{B}_{\boldsymbol{\delta}}|^{1/2}}{|\mathbf{B}_{0,\boldsymbol{\delta}}|^{1/2}} \exp\big( -\frac{1}{2} (\tilde{\mathbf{y}}' \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{y}} - \mathbf{b}_{\boldsymbol{\delta}}' \mathbf{B}_{\boldsymbol{\delta}}^{-1} \mathbf{b}_{\boldsymbol{\delta}} + \mathbf{b}_{0,\boldsymbol{\delta}}' \mathbf{B}_{0,\boldsymbol{\delta}}^{-1} \mathbf{b}_{0,\boldsymbol{\delta}}) \big),$$

where $\mathbf{B}_{\boldsymbol{\delta}}$ and $\mathbf{b}_{\boldsymbol{\delta}}$ are the moments of the normal posterior

$$\begin{aligned} \mathbf{B}_{\boldsymbol{\delta}} &= (\tilde{\mathbf{X}}_{\boldsymbol{\delta}}' \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{X}}_{\boldsymbol{\delta}} + \mathbf{B}_0^{-1})^{-1}, \\ \mathbf{b}_{\boldsymbol{\delta}} &= \mathbf{B}_{\boldsymbol{\delta}} \tilde{\mathbf{X}}_{\boldsymbol{\delta}}' \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{y}}, \end{aligned}$$

and $\mathbf{X}_{\boldsymbol{\delta}}$ is the appropriate design matrix, including those regressors, for which the corresponding indicator variable takes the value 1.

Regression coefficients $\beta_j$ for which the corresponding indicator $\delta_j = 0$ are set to zero and the remaining elements $\boldsymbol{\beta}_{\boldsymbol{\delta}}$ are sampled from the normal posterior $\mathcal{N}(\mathbf{b}_{\boldsymbol{\delta}}, \mathbf{B}_{\boldsymbol{\delta}})$.

## B Details on posterior sampling the conditional linear model

We give details on sampling the indicators and regression coefficients for the normal response in the conditional linear model. As noted in Section 3.2 we deal with the heterogeneous linear regression model

$$\mathbf{y}^n = \mathbf{X}\boldsymbol{\beta}^n + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right), \tag{15}$$

$$\mathbf{w} = \mathbf{V}\boldsymbol{\beta}^n + \tilde{\boldsymbol{\epsilon}}, \qquad \tilde{\boldsymbol{\epsilon}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{S}\right), \tag{16}$$

where $\mathbf{y}^n = (y_1^n, \ldots, y_N^n)'$, and $\mathbf{w} = (w_1, \ldots, w_N)'$ is the column vectors of the (working) responses. $\mathbf{X}$ is the regressor matrix for the normal response, $\mathbf{V}$ is the matrix with rows $\mathbf{v}_i = \frac{\psi}{\sigma}\mathbf{x}_i$ and $\mathbf{S} =$

$\mathrm{diag}(s_{r_i}^2)$ is the diagonal matrix of the error variances. The conditional posterior inclusion probability of regression coefficient $\beta_k^n$ in this model is given as

$$p(\delta_k^n = 1|\boldsymbol{\delta}_{\setminus k}^n, \omega^n, \sigma^2, \mathbf{S}, \mathbf{y}^n, \mathbf{w}) = \frac{1}{1 + \frac{1-\omega^n}{\omega^n} \frac{p(\mathbf{y}^n, \mathbf{w}|\delta_k^n=0, \boldsymbol{\delta}_{\setminus k}^n, \sigma^2, \mathbf{S})}{p(\mathbf{y}^n, \mathbf{w}|\delta_k^n=1, \boldsymbol{\delta}_{\setminus k}^n, \sigma^2, \mathbf{S})}}.$$

Denoting by $\mathbf{X}_{\boldsymbol{\delta}}$ and $\mathbf{V}_{\boldsymbol{\delta}}$ the submatrices of $\mathbf{X}$ and $\mathbf{V}$ consisting of those columns for which the corresponding elements of the indicator vector $\boldsymbol{\delta}$ is equal to 1, and by $\mathbf{B}_{0,\boldsymbol{\delta}}$ the prior variance matrix for the corresponding elements of $\boldsymbol{\beta}^n$, the marginal likelihood $p(\mathbf{y}^n, \mathbf{w}|\boldsymbol{\delta}, \sigma^2, \mathbf{S})$ is explicitly available as

$$-2\log p(\mathbf{y}^n, \mathbf{w}|\sigma^2, \boldsymbol{\delta}) = (\log|\mathbf{B}_{N,\boldsymbol{\delta}}| - \log|\mathbf{B}_{0,\boldsymbol{\delta}}|) + \Big(\sum_{i=1}^{n}\big(\frac{(y_i^n)^2}{\sigma^2} + \frac{w_i^2}{s_{r_i}^2}\big) - \mathbf{b}'_{N,\boldsymbol{\delta}}(\mathbf{B}_{N,\boldsymbol{\delta}})^{-1}\mathbf{b}_{N,\boldsymbol{\delta}}\Big) + N\log(2\pi\sigma^2),$$

where

$$\mathbf{B}_{N,\boldsymbol{\delta}}^{-1} = \frac{1}{\sigma^2}\mathbf{X}'_{\boldsymbol{\delta}}\mathbf{X}_{\boldsymbol{\delta}} + \mathbf{V}'_{\boldsymbol{\delta}}\mathbf{S}^{-1}\mathbf{V}_{\boldsymbol{\delta}} + \mathbf{B}_{0,\boldsymbol{\delta}}^{-1},$$
$$\mathbf{b}_{N,\boldsymbol{\delta}} = \mathbf{B}_{N,\boldsymbol{\delta}}(\mathbf{X}'_{\boldsymbol{\delta}}\frac{\mathbf{y}^n}{\sigma^2} + \mathbf{V}'_{\boldsymbol{\delta}}\mathbf{S}^{-1}\mathbf{w}).$$

# References


BMASK (2011). "Armutsgefährdung und Lebensbedingungen in Österreich, Ergebnisse aus EU-SILC 2009. Studie der Statistik Austria im Auftrag des BMASK." *Sozialpolitische Studienreihe*, **5**.

Brier GW (1950). "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, **78**, 1–3.

Celeux G, Forbes F, Robert C, Titterington DM (2006). "Deviance Information Criteria for Missing Data Models." *Bayesian Analysis*, **4**, 651–674.

Chib S, Greenberg E, Jeliazkov I (2009). "Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection." *Journal of Computational and Graphical Statistics*, **18**, 321–348.

Eurostat (2011). "Sponsorship Group on Measuring Progress, Well-being and Sustainable Development, Final report." *EEA ESSC 2011/11/05/EN*.

Frühwirth-Schnatter S, Frühwirth R (2010). "Data Augmentation and MCMC for Binary and Multinomial Logit Models." In T Kneib, G Tutz (eds.), *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pp. 111–132. Physica-Verlag, Heidelberg.

Fusco A, Guio AC, Marlier E (2010). "Income Poverty and Material Deprivation in European Countries." *Eurostat - Methodologies Working papers*.

George EI, McCulloch R (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, **7**, 339–373.

Griffin J, Brown PJ (2010). "Inference with Normal-Gamma Prior Distributions in Regression Problems." *Bayesian Analysis*, **5**, 171–188.

Ishwaran H, Rao SJ (2005). "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies." *Annals of Statistics*, **33**, 730–773.

Lang S, Brezger A (2004). "Bayesian P-Splines." *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Malsiner-Walli G, Wagner H (2011). "Comparing Spike and Slab Priors for Bayesian Variable Selection." *Austrian Journal of Statistics*, **40**, 241–264.

Mohanan J, Stefanski LA (1992). "Normal Scale Mixture Approximations to $F^{\star}(z)$ and Computation of the Logistics Normal Integral." In N Balakrishnan (ed.), *Handbook of the Logistic Distribution*, pp. 529–549. Marcel Dekker, New York.

Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society, Series B, Methodological*, **64**(4), 583–616.

Stiglitz J, Sen A, Fitoussi J (2009). "Report by the Commission on the Measurement of Economic Performance and Social Progress." URL www.stiglitz-sen-fitoussi.fr.

Wagner H, Duller C (2012). "Bayesian Model Selection for Logistic Regression Models with Random Intercept." *Computational Statistics and Data Analysis*, **56**.

Wiesenfarth M, Hisgen CM, Kneib T, Cadarso-Suarez C (2012). "Bayesian Nonparametric Instrumental Variable Regression based on Penalized Splines and Dirichlet Process Mixtures." *Technical report*.

**Affiliation:**

Helga Wagner
Department of Applied Statistics and Econometrics
Johannes Kepler Universität Linz
A-4040 Linz, Austria
E-mail: helga.wagner@jku.at
URL: http://www.jku.at/ifas

Regina Tüchler
Department of Statistics
Austrian Federal Economic Chamber
A-1040 Vienna, Austria E-mail: regina.tuechler@wko.at
URL: https://www.wko.at/Content.Node/Mitarbeiterkontaktseite.html?rollenid=2279681

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 43 *Submitted:* 2013-12-19
June 2014 *Accepted:* 2014-05-01

Austrian Journal of Statistics
*June 2014, Volume 43/2, 119–131.*
*http://www.ajs.or.at/*

# The Transmuted Generalized Inverse Weibull Distribution

**Faton Merovci**
University of Prishtina

**Ibrahim Elbatal**
Cairo University

**Alaa Ahmed**
Cairo University

**Abstract**

A generalization of the generalized inverse Weibull distribution the so-called transmuted generalized inverse Weibull distribution is proposed and studied. We will use the quadratic rank transmutation map (QRTM) in order to generate a flexible family of probability distributions taking the generalized inverse Weibull distribution as the base value distribution by introducing a new parameter that would offer more distributional flexibility. Various structural properties including explicit expressions for the moments, quantiles, and moment generating function of the new distribution are derived. We propose the method of maximum likelihood for estimating the model parameters and obtain the observed information matrix. A real data set are used to compare the flexibility of the transmuted version versus the generalized inverse Weibull distribution.

*Keywords*: generalized inverse Weibull distribution, order statistics, transmutation map, maximum likelihood estimation, reliability function.

## 1. Introduction

The inverse Weibull distribution is another life time probability distribution which can be used in the reliability engineering discipline. The inverse Weibull distribution can be used to model a variety of failure characteristics such as infant mortality, useful life and wear- out periods. It can also be used to determine the cost effectiveness, maintenance periods of reliability centered maintenance activities and applications in medicine, reliability and ecology. Keller, Goblin, and Farnworth (1985) obtained the inverse Weibull model by investigating failures of mechanical components subject to degradation. Drapella (1993); Mudholkar and Kollia (1994), and de Gusmão, Ortega, and Cordeiro (2011) introduced the generalized inverse Weibull distribution, among others. The cumulative distribution function (cdf) of the generalized inverse Weibull (GIW) distribution can be defined by

$$G(x,\alpha,\gamma,\theta) = e^{-\gamma(\alpha x)^{-\beta}}, \qquad \alpha>0, \gamma>0, \beta>0, x\geq 0\,, \tag{1}$$

where $\alpha$ is a scale parameter and $\beta$, $\gamma$ are shape parameters, respectively. The corresponding probability density function (pdf) is given by

$$g(x,\alpha,\gamma,\theta) = \alpha\beta\gamma(\alpha x)^{-\beta-1}e^{-\gamma(\alpha x)^{-\beta}}\,. \tag{2}$$

In this article we present a new generalization of the generalized inverse Weibull distribution called the transmuted generalized inverse Weibull distribution. We will derive the subject distribution using the quadratic rank transmutation map studied by Shaw and Buckley (2009).

A random variable $X$ is said to have transmuted distribution if its cdf is given by

$$F(x) = (1+\lambda)G(x) - \lambda G(x)^2\,, \qquad |\lambda| \leq 1\,, \tag{3}$$

where $G(x)$ is the cdf of the base distribution, which on differentiation yields

$$f(x) = g(x)\left[1+\lambda - 2\lambda G(x)\right]\,, \tag{4}$$

where $f(x)$ and $g(x)$ are the corresponding pdf's associated with the cdf's $F(x)$ and $G(x)$, respectively. An extensive information about the quadratic rank transmutation map is given in Shaw and Buckley (2009). Observe that at $\lambda = 0$ we have the distribution of the base random variable.

Many authors deal with the generalization of some well-known distributions. Aryal and Tsokos (2009) defined the transmuted generalized extreme value distribution and they studied some basic mathematical characteristics of the transmuted Gumbel probability distribution and it has been observed that the transmuted Gumbel can be used to model climate data. Also Aryal and Tsokos (2011) presented a new generalization of the Weibull distribution called the transmuted Weibull distribution. Recently, Aryal (2013) proposed and studied various structural properties of the transmuted log-logistic distribution. Khan and King (2013) introduced the transmuted modified Weibull distribution which extended recent developments on the transmuted Weibull distribution by Aryal and Tsokos (2009). They studied the mathematical properties and the maximum likelihood estimation of the unknown parameters. In the present study we will provide the mathematical formulation of the transmuted generalized inverse Weibull distribution and some of its properties. We will also provide possible areas of applications.

The rest of the paper is organized as follows. In Section 3 we demonstrate the transmuted probability distribution. In Section 4 we find the reliability functions of the subject model. The statistical properties including quantile functions, moments and moment generating functions are derived in Section 5. The minimum, maximum, and median order statistics models are discussed in Section 6. Least squares and weighted least squares estimators are discussed in Section 7. In Section 8 we demonstrate the maximum likelihood estimates and some asymptotic confidence intervals for the unknown parameters. In Section 9, the TGIW distribution is applied to a real data set. Finally, we provide some conclusion in Section 10.

## 2. Transmutation map

In this section we consider the transmuted probability distribution. Let $F_1$ and $F_2$ be the cdf's of two distributions with a common sample space. The general rank transmutation as given in Shaw and Buckley (2009) is defined as

$$G_{R12}(u) = F_2(F_1^{-1}(u))\,, \qquad G_{R21}(u) = F_1(F_2^{-1}(u))\,.$$

Note that the inverse cdf is also known as the quantile function an is defined as

$$F^{-1}(y) = \inf_{x\in\mathbb{R}}\{F(x) \geq y\} \qquad \text{for } y \in [0,1].$$

The functions $G_{R12}(u)$ and $G_{R21}(u)$ both map the unit interval $I = [0,1]$ onto itself, and under suitable assumptions they are mutual inverses and satisfy $G_{Rij}(0) = 0$ and $G_{Rij}(1) = 1$. A quadratic rank transmutation map (QRTM) is defined as

$$G_{R12}(u) = u + \lambda u(1-u)\,, \qquad |\lambda| \leq 1\,, \tag{5}$$

from which follows that the cdf's satisfy the relationship

$$F_2(x) = (1+\lambda)F_1(x) - \lambda F_1(x)^2\,, \tag{6}$$

which after differentiation yields

$$f_2(x) = f_1(x)[1+\lambda - 2\lambda F_1(x)]\,, \tag{7}$$

where $f_1(x)$ and $f_2(x)$ are the corresponding pdf's associated with the cdf's $F_1(x)$ and $F_2(x)$, respectively. An extensive information about the quadratic rank transmutation map is given in Shaw and Buckley (2009). Observe that at $\lambda = 0$ we have the distribution of the base random variable. The function $f_2(x)$ in (7) satisfies the property of a pdf.

## 3. Transmuted generalized inverse Weibull distribution

In this section we study the transmuted generalized inverse Weibull (TGIW) distribution and submodels of this distribution. Now using (1) and (2) we have the cdf of the transmuted generalized inverse Weibull distribution

$$F_{TGIW}(x) = e^{-\gamma(\alpha x)^{-\beta}} \left[1 + \lambda - \lambda e^{-\gamma(\alpha x)^{-\beta}}\right], \tag{8}$$

where $\alpha$ is a scale parameter and $\beta$ and $\gamma$ are shape parameters representing the different patterns of the transmuted generalized inverse Weibull distribution, and $\lambda$ is the transmuted parameter. The corresponding pdf of the transmuted generalized inverse Weibull distribution is given by

$$f_{TGIW}(x) = \alpha\beta\gamma(\alpha x)^{-\beta-1} e^{-\gamma(\alpha x)^{-\beta}} \left[1 + \lambda - 2\lambda e^{-\gamma(\alpha x)^{-\beta}}\right]. \tag{9}$$

Figures 1 and 2 illustrate some of the possible shapes of the pdf and cdf of a TGIW distribution for selected values of the parameters $\beta$, $\gamma$, and $\lambda$ by keeping $\alpha = 1$, respectively.

The transmuted generalized inverse Weibull distribution is a very flexible model that approaches to different distributions when its parameters are changed. The flexibility of the transmuted generalized inverse Weibull distribution is explained in the following. If $X$ is a random variable with pdf (9), then we have the following cases:

(a) If $\gamma = 1$, we get the transmuted inverse Weibull.

(b) If $\lambda = 0$ and $\gamma = 1$, we get the inverse Weibull.

(c) If $\beta = 1$ and $\gamma = 1$, we get the transmuted inverse exponential distribution.

(d) If $\beta = 1$, $\gamma = 1$, and $\lambda = 0$, we get the inverse exponential distribution.

(e) If $\beta = 2$ and $\gamma = 1$ we get transmuted inverse Rayleigh distribution.

(f) If $\beta = 2$, $\gamma = 1$, and $\lambda = 0$ we get the inverse Rayleigh distribution.

(g) If $\alpha = 1$ we get the transmuted Frechet distribution.

(h) If $\alpha = 1$ and $\lambda = 0$ we get the Frechet distribution.

## 4. Reliability analysis

The reliability function $R(x)$, which is the probability of an item not failing prior to some time $t$, is defined by $R(x) = 1 - F(x)$. The reliability function of a transmuted generalized inverse Weibull distribution $R_{TGIW}(x)$ can be a useful characterization of life time data analysis. It is defined as

$$\begin{aligned} R_{TGIW}(x) &= 1 - F_{TGIW}(x) \\ &= 1 - e^{-\gamma(\alpha x)^{-\beta}} \left[1 + \lambda - \lambda e^{-\gamma(\alpha x)^{-\beta}}\right]. \end{aligned} \tag{10}$$

It is important to note that $R_{TGIW}(x) + F_{TGIW}(x) = 1$. The other characteristic of interest is the hazard rate function defined by $h_{TGIW}(x) = f_{TGIW}(x)/(1 - F_{TGIW}(x))$, which is an important quantity characterizing life phenomenon. It can be loosely interpreted as the conditional probability

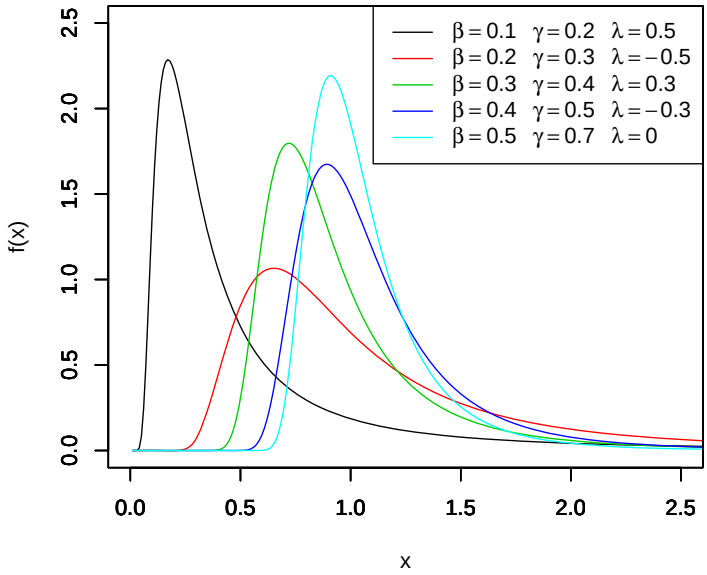


Figure 1: The pdf's of various TGIW distributions.

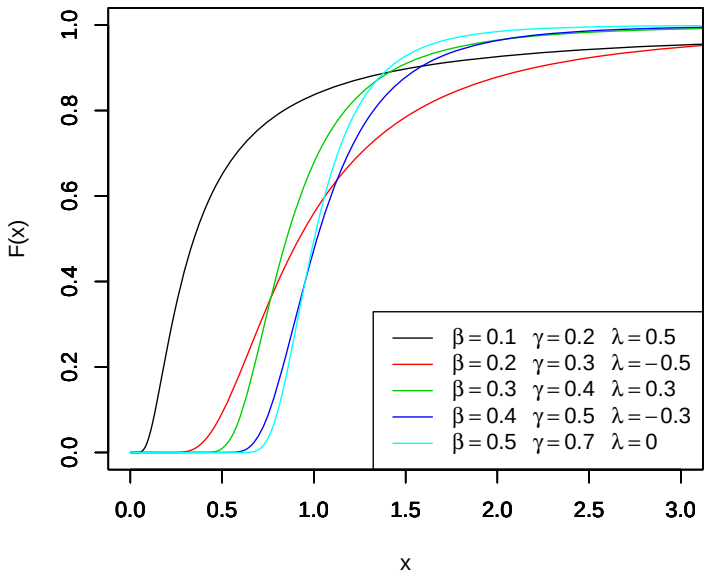


Figure 2: The cdf's of various TGIW distributions.

of failure, given it has survived to the time $t$. The hazard rate function for a transmuted generalized

inverse Weibull distribution is defined by

$$\begin{aligned} h_{TGIW}(x) &= \frac{f_{TGIW}(x)}{1-F_{TGIW}(x)} \\ &= \frac{\alpha\beta\gamma(\alpha x)^{-\beta-1}e^{-\gamma(\alpha x)^{-\beta}}\left[1+\lambda-2\lambda e^{-\gamma(\alpha x)^{-\beta}}\right]}{1-e^{-\gamma(\alpha x)^{-\beta}}\left[1+\lambda-\lambda e^{-\gamma(\alpha x)^{-\beta}}\right]}. \end{aligned} \tag{11}$$

Figures 3 and 4 illustrate some of the possible shapes of the hazard rate function and the survival function of the TGIW distribution for selected values of the parameters $\beta$, $\gamma$, and $\lambda$ by keeping $\alpha = 1$, respectively.

It is important to note that the unit for $h_{TGIW}(x)$ is the probability of failure per unit of time, distance or cycles. These failure rates are defined with different choices of parameters.The cumulative hazard function of the transmuted generalized inverse Weibull distribution is denoted by

$$H_{TGIW}(x) = -\log\left|e^{-\gamma(\alpha x)^{-\beta}}\left[1+\lambda-\lambda e^{-\gamma(\alpha x)^{-\beta}}\right]\right|. \tag{12}$$

It is important to note that the unit for $H_{TGIW}(x)$ is the cumulative probability of failure per unit of time, distance or cycles.

# 5. Statistical properties

In this section we discuss the statistical properties of the transmuted generalized inverse Weibull distribution. Specifically we are interested in quantiles, a random number generation function, moments and the moment generating function.

## 5.1. Quantiles

The quantile $x_q$ of the $T_{GIW}(\alpha,\beta,\gamma,\lambda,x)$ distribution is the solution of the equation

$$x_q = \frac{1}{\alpha}\left[\frac{1}{\gamma}\log\left(\frac{1+\lambda-\lambda e^{-\gamma(\alpha x_q)^{-\beta}}}{q}\right)\right]^{-\frac{1}{\beta}}. \tag{13}$$

The above equation has no closed form solution in $x_q$, so we have to use a numerical technique such as a Newton-Raphson method to get the quantile. If we put $q = 0.5$ in equation (13) one gets the median.

## 5.2. Random number generation

A random variate $X$ from $T_{GIW}(\alpha,\beta,\gamma,\lambda,x)$ can be generated as $x_U$ according to (13), where $q$ is replaced by $U \sim U(0,1)$.

## 5.3. Moments

The following theorem gives the $r$th moment $\mu'_r$ of the $T_{GIW}(\alpha,\beta,\gamma,\lambda,x)$ distribution.

**Theorem 4.1**. If $X$ is from the $T_{GIW}(\alpha,\beta,\gamma,\lambda,x)$ distribution with $|\lambda| \leq 1$, then the $r$th non central moments are given by

$$\mu'_r = \mathsf{E}(X^r) = \frac{\gamma^{r/\beta}\Gamma(1-r/\beta)}{\alpha^r}\left[1+\lambda-\lambda 2^{r/\beta}\right]. \tag{14}$$

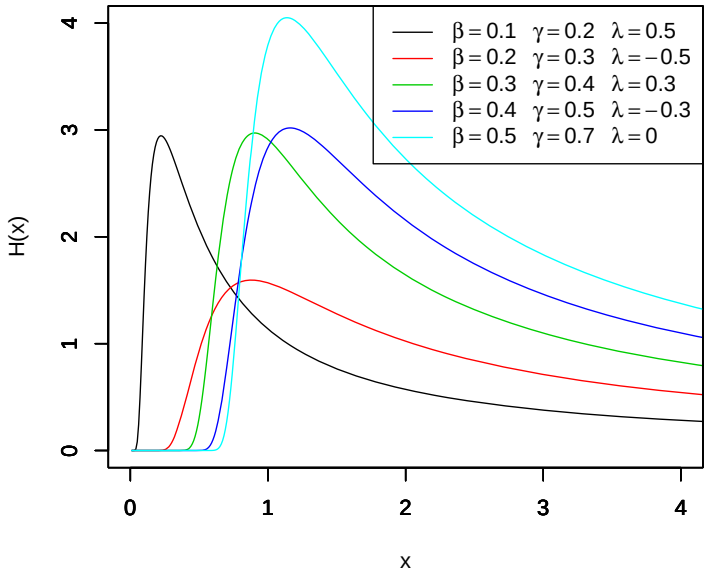


Figure 3: The hazard function of various TGIW distributions.

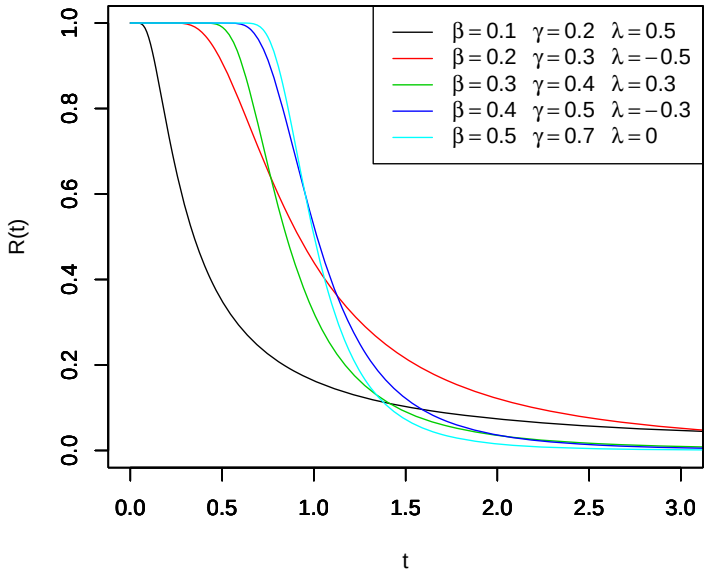


Figure 4: The survival function of various TGIW distributions.

Proof: Starting with

$$\begin{aligned}
\mu'_r &= \int_0^\infty x^r f_{TGIW}(\alpha,\beta,\gamma,\lambda,x)dx \\
&= \int_0^\infty x^r \alpha\beta\gamma(\alpha x)^{-\beta-1} e^{-\gamma(\alpha x)^{-\beta}} \left[1+\lambda-2\lambda e^{-\gamma(\alpha x)^{-\beta}}\right] dx \\
&= \frac{\alpha\beta\gamma}{\alpha^r}(1+\lambda)\int_0^\infty (\alpha x)^{r-\beta-1} e^{-\gamma(\alpha x)^{-\beta}} dx - \frac{2\lambda\alpha\beta\gamma}{\alpha^r}\int_0^\infty (\alpha x)^{r-\beta-1} e^{-2\gamma(\alpha x)^{-\beta}} dx \quad (15)
\end{aligned}$$

Now let $\gamma(\alpha x)^{-\beta} = t$, then $x = \frac{1}{\alpha}\gamma^{1/\beta}t^{-1/\beta}$, therefore

$$\begin{aligned}\mu'_r &= \frac{1+\lambda}{\alpha^r}\gamma^{r/\beta}\Gamma(1-r/\beta) - \frac{\lambda}{\alpha^r}(2\gamma)^{r/\beta}\Gamma(1-r/\beta) \\ &= \frac{\gamma^{r/\beta}\Gamma(1-r/\beta)}{\alpha^r}\left[1+\lambda-\lambda 2^{r/\beta}\right], \end{aligned} \tag{16}$$

which completes the proof.

Based on Theorem 4.1 the coefficients of variation (CV), skewness (CS), and kurtosis (CK) can be obtained according to the following well-known relations as

$$\begin{aligned}\text{CV}_{TMIW} &= \sqrt{\frac{\mu_2}{\mu_1} - 1} \\ \text{CS}_{TMIW} &= \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{(\mu_2-\mu_1)^{3/2}} \\ \text{CK}_{TMIW} &= \frac{\mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2}{(\mu_2-\mu_1^2)^2}. \end{aligned}$$

### 5.4. Moment generating function

In this subsection we derive the moment generating function (mgf) of the transmuted generalized inverse Weibull distribution.

**Theorem 4.2**. If $X$ has the $T_{GIW}(\alpha,\beta,\gamma,\lambda,x)$ distribution with $|\lambda| \leq 1$, then the moment generating function (mgf) of $X$ is given as

$$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r\gamma^{r/\beta}\Gamma(1-r/\beta)}{r!\alpha^r}\left[1+\lambda-\lambda 2^{r/\beta}\right]. \tag{17}$$

Proof:

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} f_{T_{GIW}}(\alpha,\beta,\gamma,\lambda,x)dx \\ &= \int_0^\infty \sum_{r=0}^{\infty}\frac{t^r}{r!}x^r f_{T_{GIW}}(\alpha,\beta,\gamma,\lambda,x) \\ &= \sum_{r=0}^{\infty}\frac{t^r}{r!}\mu'_r. \end{aligned} \tag{18}$$

By using equation (14) in result (18) we get

$$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r\gamma^{r/\beta}\Gamma(1-r/\beta)}{r!\alpha^r}\left[1+\lambda-\lambda 2^{r/\beta}\right],$$

which completes the proof.

## 6. Order statistics

In fact, the order statistics have many applications in reliability and life testing. The order statistics arise in the study of reliability of a system. Let $X_1,\ldots,X_n$ be a simple random sample from the $T_{GIW}(\alpha,\beta,\gamma,\lambda,x)$ distribution with cdf and pdf as in (8) and (9), respectively. Let $X_{(1:n)} \leq \cdots \leq X_{(n:n)}$ denote the order statistics obtained from this sample. In reliability literature, $X_{(i:n)}$ denotes

the lifetime of an $(n-i+1)$-out-of-$n$ system which consists of $n$ iid components. Then, the pdf of $X_{(i:n)}$, $1 = 1, \ldots, n$ is

$$f_{i:n}(x) = \frac{1}{\text{Beta}(i, n-i+1)} \left[F(x,\Phi)\right]^{i-1} \left[1 - F(x,\Phi)\right]^{n-i} f(x,\Phi)\,, \tag{19}$$

where $\Phi = (\alpha, \beta, \gamma, \lambda)$. Also, the joint pdf of $(X_{(i:n)}, X_{(j:n)})$, for $i = 1, \ldots, n$, is

$$f_{i:j:n}(x_i, x_j) = C\left[F(x_i)\right]^{i-1} \left[F(x_j) - F(x_i)\right]^{j-i-1} \left[1 - F(x_j)\right]^{n-j} f(x_i) f(x_j)\,, \tag{20}$$

where

$$C = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}\,.$$

We define the first order statistics as $X_{(1)} = \min(X_1, \ldots, X_n)$, the last order statistics as $X_{(n)} = \max(X_1, \ldots, X_n)$, and the median order as $X_{m+1}$, if $n = 2m + 1$. .

### 6.1. Distribution of minimum, maximum, and median

Let $X_1, \ldots, X_n$ be independently identically distributed random variables from the transmuted generalized inverse Weibull distribution. The first, last, and median order pdf's are given by

$$\begin{aligned}
f_{1:n}(x) &= n\left[1 - F(x,\Phi)\right]^{n-1} f(x,\Phi) \\
&= n\left\{1 - e^{-\gamma(\alpha x)^{-\beta}}\left[1 + \lambda - \lambda e^{-\gamma(\alpha x)^{-\beta}}\right]\right\}^{n-1} \\
&\quad \times \alpha\beta\gamma(\alpha x)^{-\beta-1} e^{-\gamma(\alpha x)^{-\beta}}\left[1 + \lambda - 2\lambda e^{-\gamma(\alpha x)^{-\beta}}\right] \qquad (21) \\
f_{n:n}(x) &= n\left[F(x,\Phi)\right]^{n-1} f(x),\Phi) \\
&= n\left\{e^{-\gamma(\alpha x)^{-\beta}}\left[1 + \lambda - \lambda e^{-\gamma(\alpha x)^{-\beta}}\right]\right\}^{n-1} \\
&\quad \times \alpha\beta\gamma(\alpha x)^{-\beta-1} e^{-\gamma(\alpha x)^{-\beta}}\left[1 + \lambda - 2\lambda e^{-\gamma(\alpha x)^{-\beta}}\right] \qquad (22) \\
f_{m+1:n}(x) &= \frac{(2m+1)!}{m!m!}(F(x))^m (1 - F(x))^m f(x) \\
&= \frac{(2m+1)!}{m!m!}\left\{e^{-\gamma(\alpha x)^{-\beta}}\left[1 + \lambda - \lambda e^{-\gamma(\alpha x)^{-\beta}}\right]\right\}^m \\
&\quad \times \left\{1 - e^{-\gamma(\alpha x)^{-\beta}}\left[1 + \lambda - \lambda e^{-\gamma(\alpha x)^{-\beta}}\right]\right\}^m \\
&\quad \times \alpha\beta\gamma(\alpha x)^{-\beta-1} e^{-\gamma(\alpha x)^{-\beta}}\left[1 + \lambda - 2\lambda e^{-\gamma(\alpha x)^{-\beta}}\right]\,. \qquad (23)
\end{aligned}$$

### 6.2. Joint distribution of the $i$th and $j$th order statistics

The joint distribution of the the $i$th and $j$th order statistics from the transmuted generalized inverse Weibull is

$$\begin{aligned}
f_{i:j:n}(x_i, x_j) &= C\left[F(x_i)\right]^{i-1} \left[F(x_j) - F(x_i)\right]^{j-i-1} \left[1 - F(x_j)\right]^{n-j} f(x_i) f(x_j) \\
&= C\left\{h_i\left[1 + \lambda - \lambda h_i\right]\right\}^{i-1} \\
&\quad \times \left\{h_j\left[1 + \lambda - \lambda h_j\right] - h_i\left[1 + \lambda - \lambda h_i\right]\right\}^{j-i-1} \\
&\quad \times \left\{1 - h_j\left[1 + \lambda - \lambda h_j\right]\right\}^{n-j} \\
&\quad \times \alpha\beta\gamma(\alpha x_i)^{-\beta-1} h_i\left[1 + \lambda - 2\lambda h_i\right] \\
&\quad \times \alpha\beta\gamma(\alpha x_j)^{-\beta-1} h_j\left[1 + \lambda - 2\lambda h_j\right]\,, \qquad (24)
\end{aligned}$$

where

$$h_i = e^{-\gamma(\alpha x_i)^{-\beta}}\,.$$

For the special case $i = 1$ and $j = n$ we get the joint distribution of the minimum and maximum as

$$\begin{aligned} f_{1:n:n}(x_1, x_n) &= n(n-1)\left[F(x_n) - F(x_1)\right]^{n-2} f(x_1) f(x_n) \\ &= n(n-1)\left\{h_n\left[1+\lambda-\lambda h_n\right] - h_1\left[1+\lambda-\lambda h_1\right]\right\}^{n-2} \\ &\quad \times \alpha\beta\gamma(\alpha x_1)^{-\beta-1} h_1\left[1+\lambda-2\lambda h_1\right] \\ &\quad \times \alpha\beta\gamma(\alpha x_n)^{-\beta-1} h_n\left[1+\lambda-2\lambda h_n\right]. \end{aligned} \tag{25}$$

## 7. Weighted least squares estimators

In this section we provide the regression based method estimators of the unknown parameters of the transmuted generalized inverse Weibull distribution, which was originally suggested by Swain, Venkatraman, and Wilson (1988) to estimate the parameters of the beta distribution. It can be also used for some other cases. Suppose $X_1, \ldots, X_n$ is a random sample of size $n$ from a cdf $G(\cdot)$ and suppose $X_{(i)}$, $i = 1, \ldots, n$, denotes the ordered sample. The proposed method uses the distribution of $G(X_{(i)})$, which is Beta$(i, n)$. Hence, for a sample of size $n$, we have

$$\begin{aligned} \mathsf{E}\left(G(X_{(j)})\right) &= \frac{j}{n+1} \\ \text{var}\left(G(X_{(j)})\right) &= \frac{j(n-j+1)}{(n+1)^2(n+2)} \\ \text{cov}\left(G(X_{(j)}), G(X_{(k)})\right) &= \frac{j(n-k+1)}{(n+1)^2(n+2)}, \qquad \text{for } j < k, \end{aligned}$$

see Johnson, Kotz and Balakrishnan Johnson, Kotz, and Balakrishnan (1994). Using the expectations and the variances, two variants of the least squares methods can be used.

**Method 1 (Least Squares Estimators)** Obtain the least squares estimators by minimizing

$$\sum_{j=1}^{n} \left( G(X_{(j)}) - \frac{j}{n+1} \right)^2, \tag{26}$$

with respect to the unknown parameters. Therefore, in case of the $TGIW$ distribution the least squares estimators of $\alpha$, $\beta$, and $\lambda$, say $\hat{\alpha}_{LS}$, $\hat{\beta}_{LS}$, and $\hat{\lambda}_{LS}$, can be obtained by minimizing

$$\sum_{j=1}^{n} \left[ e^{-\gamma(\alpha x_{(j)})^{-\beta}} \left[1+\lambda-\lambda e^{-\gamma(\alpha x_{(j)})^{-\beta}}\right] - \frac{j}{n+1} \right]^2,$$

with respect to $\alpha$, $\beta$, and $\lambda$.

**Method 2 (Weighted Least Squares Estimators)** The weighted least squares estimators can be obtained by minimizing

$$\sum_{j=1}^{n} w_j \left( G(X_{(j)}) - \frac{j}{n+1} \right)^2, \tag{27}$$

with respect to the unknown parameters, where

$$w_j = \frac{1}{\text{var}\left(G(X_{(j)})\right)} = \frac{(n+1)^2(n+2)}{j(n-j+1)}.$$

Therefore, in case of the $TGIW$ distribution the weighted least squares estimators of $\alpha$, $\beta$, and $\lambda$, say $\hat{\alpha}_{WLS}$, $\hat{\beta}_{WLS}$, and $\hat{\lambda}_{WLS}$ can be obtained by minimizing

$$\sum_{j=1}^{n} w_j \left[ e^{-\gamma(\alpha x_{(j)})^{-\beta}} \left[1+\lambda-\lambda e^{-\gamma(\alpha x_{(j)})^{-\beta}}\right] - \frac{j}{n+1} \right]^2$$

with respect to the unknown parameters.

## 8. Maximum likelihood estimators

Now we derive the maximum likelihood estimators (MLEs) and discuss inference under the $T_{GIW}(\alpha,\beta,\gamma,\lambda,x)$ distribution. Let $X_1,\ldots,X_n$ be a random sample of size $n$ from the $T_{GIW}$ distribution then the likelihood function can be written as

$$L(\alpha,\beta,\gamma,\lambda,x)=\prod_{i=1}^{n}\alpha\beta\gamma(\alpha x_i)^{-\beta-1}e^{-\gamma(\alpha x_i)^{-\beta}}\left[1+\lambda-2\lambda e^{-\gamma(\alpha x_i)^{-\beta}}\right]. \tag{28}$$

Taking the logarithm results in the log-likelihood function

$$\begin{aligned}\log L &= n\log(\alpha\beta\gamma)-(\beta+1)\sum_{i=1}^{n}\log(\alpha x_i)\\ &\quad -\gamma\sum_{i=1}^{n}(\alpha x_i)^{-\beta}+\sum_{i=1}^{n}\log\left[1+\lambda-2\lambda e^{-\gamma(\alpha x_i)^{-\beta}}\right].\end{aligned} \tag{29}$$

Differentiating equation (29) with respect to $\alpha,\beta,\gamma$, and $\lambda$ results in

$$\begin{aligned}\frac{\partial\log L}{\partial\alpha} &= \frac{n}{\alpha}-(\beta+1)\frac{n}{\alpha}+\gamma\beta\sum_{i=1}^{n}x_i(\alpha x_i)^{-\beta-1}\\ &\quad -\sum_{i=1}^{n}\frac{2\lambda\alpha x_i(\alpha x_i)^{-\beta-1}e^{-\gamma(\alpha x_i)^{-\beta}}}{1+\lambda-2\lambda e^{-\gamma(\alpha x_i)^{-\beta}}},\end{aligned} \tag{30}$$

$$\begin{aligned}\frac{\partial\log L}{\partial\beta} &= \frac{n}{\beta}-\sum_{i=1}^{n}\log(\alpha x_i)+\gamma\sum_{i=1}^{n}(\alpha x_i)^{-\beta}\log(\alpha x_i)\\ &\quad +\sum_{i=1}^{n}\frac{-2\lambda\gamma e^{-\gamma(\alpha x_i)^{-\beta}}(\alpha x_i)^{-\beta}\log(\alpha x_i)}{1+\lambda-2\lambda e^{-\gamma(\alpha x_i)^{-\beta}}},\end{aligned} \tag{31}$$

$$\frac{\partial\log L}{\partial\gamma} = \frac{n}{\gamma}-\sum_{i=1}^{n}(\alpha x_i)^{-\beta}+\sum_{i=1}^{n}\frac{2\lambda e^{-\gamma(\alpha x_i)^{-\beta}}(\alpha x_i)^{-\beta}}{1+\lambda-2\lambda e^{-\gamma(\alpha x_i)^{-\beta}}}, \tag{32}$$

$$\frac{\partial\log L}{\partial\lambda} = \sum_{i=1}^{n}\frac{1-2e^{-\gamma(\alpha x_i)^{-\beta}}}{1+\lambda-2\lambda e^{-\gamma(\alpha x_i)^{-\beta}}}. \tag{33}$$

We can find estimates of the unknown parameters by the maximum likelihood method equating all the above nonlinear terms to zero and solving these equations simultaneously. The solutions are the MLE's $\hat\alpha$, $\hat\beta$, $\hat\gamma$, and $\hat\lambda$. For the three parameter transmuted generalized inverted Weibull distribution all second order derivatives exist. Thus we have

$$\left(\hat\alpha,\hat\beta,\hat\gamma,\hat\lambda\right)^{T}\sim\text{Normal}\left((\alpha,\beta,\gamma,\lambda)^{T},V^{-1}\right) \tag{34}$$

with the symmetric matrix

$$V=-E\begin{bmatrix}V_{\alpha\alpha} & V_{\alpha\beta} & V_{\alpha\gamma} & V_{\alpha\lambda}\\ & V_{\beta\beta} & V_{\beta\gamma} & V_{\beta\lambda}\\ & & V_{\gamma\gamma} & V_{\gamma\lambda}\\ & & & V_{\lambda\lambda}\end{bmatrix},$$

where $V_{\theta_1\theta_2}=\partial^2L/\partial\theta_1\partial\theta_2$ denotes the second derivative of $L$ with respect to $\theta_1$ and $\theta_2$. The matrix $V^{-1}$ represents the asymptotic variance/covariance matrix of the MLE's. Based on (34), approximate $100(1-\delta)\%$ confidence intervals are determined as

$$\hat\alpha\pm z_{1-\delta/2}\sqrt{\hat V_{\alpha\alpha}},\quad \hat\beta\pm z_{1-\delta/2}\sqrt{\hat V_{\beta\beta}},\quad \hat\gamma\pm z_{1-\delta/2}\sqrt{\hat V_{\gamma\gamma}},\quad \hat\lambda\pm z_{1-\delta/2}\sqrt{\hat V_{\lambda\lambda}},$$

where $z_{1-\delta}$ is the $100(1-\delta)\%$ percentile of the standard normal distribution and $\hat{V}_{..}$ denotes the element of $V_{..}$ evaluated in the MLE's.

We can compute the maximized unrestricted and restricted log-likelihood functions to construct the likelihood ratio test (LRT) statistic for testing some transmuted GIW sub-models. For example, we can use the LRT statistic to check whether the TGIW distribution for a given data set is statistically *superior* to the GIW distribution. In any case, hypothesis tests of the type $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ can be performed using a LRT. In this case, the LRT statistic is $\omega = 2(\log L(\hat{\theta}, x) - \log L(\hat{\theta}_0, x))$, where $\hat{\theta}$ denotes the unrestricted MLE and $\hat{\theta}_0$ is the MLE under $H_0$. The LRT statistic $\omega$ is asymptotically (as $n \to \infty$) distributed as $\chi^2_k$, where $k$ is the number of parameters specified under $H_0$. The LRT rejects $H_0$ if $\omega > \chi^2_{k;1-\delta}$, where $\chi^2_{k;1-\delta}$ denotes the $100\delta\%$ quantile of the $\chi^2_k$ distribution.

## 9. Application

Now we use real data to show that the TGIW distribution might fit better than a model based on the GIW distribution. The data set given in Table 1 is taken from Murthy, Xie, and Jiang (2004) page 180 and represents 50 items put into use at $t = 0$ and failure times are in weeks.

Table 1: 50 items put into use at $t = 0$ and their failure times in weeks.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.013 | 0.065 | 0.111 | 0.111 | 0.163 | 0.309 | 0.426 | 0.535 | 0.684 | 0.747 |
| 0.997 | 1.284 | 1.304 | 1.647 | 1.829 | 2.336 | 2.838 | 3.269 | 3.977 | 3.981 |
| 4.520 | 4.789 | 4.849 | 5.202 | 5.291 | 5.349 | 5.911 | 6.018 | 6.427 | 6.456 |
| 6.572 | 7.023 | 7.087 | 7.291 | 7.787 | 8.596 | 9.388 | 10.261 | 10.713 | 11.658 |
| 13.006 | 13.388 | 13.842 | 17.152 | 17.283 | 19.418 | 23.471 | 24.777 | 32.795 | 48.105 |

Table 2: Parameter estimates of the GIW and TGIW distribution for 50 items put into use at $t = 0$ and their failure times in weeks.

| Model | Parameter Estimate | $-\log L(\cdot; x)$ |
|---|---|---|
| GIW | $\hat{\alpha} = 0.854$, $\hat{\beta} = 0.479$ | 168.638 |
| | $\hat{\gamma} = 1.044$ | |
| TGIW | $\hat{\alpha} = 2.383$, $\hat{\beta} = 0.530$ | 166.387 |
| | $\hat{\gamma} = 1.143$, $\hat{\lambda} = -0.747$ | |

Table 3: Goodness of fit criteria.

| Model | K-S | $-2\log L$ | AIC | AICC |
|---|---|---|---|---|
| GIW | 0.199 | 337.276 | 343.276 | 343.797 |
| TGIW | 0.192 | 332.774 | 340.774 | 341.662 |

The LRT statistic to test $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$ gives $\omega = 4.502 > 3.841 = \chi^2_{1;0.95}$, so we reject the null hypothesis. In order to compare the two distribution models we consider criteria like the Kolmogorov-Smirnov test statistic, $-2\log L$, the AIC (Akaike information criterion) as also the AICC (corrected AIC). The better distribution corresponds to smaller criterion values of AIC $= -2\log L + 2k$ and AICC $=$ AIC $+ \frac{2k(k+1)}{n-k-1}$, where $k$ is the number of parameters in the statistical model, $n$ denotes the sample size and $\log L$ is the maximized value of the log-likelihood function under the considered model. Table 2 shows the MLEs under both models, Table 3 contains the values the Kolmogorov-Smirnov test statistic, $-2\log L$, AIC and AICC. These values indicate that the TGIW distribution leads to a better fit than the GIW model.

## 10. Conclusion

Here we propose a new model, the so-called transmuted generalized inverse Weibull distribution which extends the generalized inverse Weibull distribution in the analysis of data with real support. An obvious reason for generalizing a standard distribution is because the generalized form provides more flexibility in modeling real data. We derive expansions for moments and for the moment generating function. The estimation of parameters is approached by the method of maximum likelihood, also the information matrix is derived. An application of the TGIW distribution to real data show that the new distribution can be used quite effectively to provide better fits than the GIW distribution.

### Acknowledgements

The author would like to thank the Editor Herwig Friedl and the referee for carefully reading the manuscript and for their comments which greatly improved the presentation.

## References

Aryal GR (2013). "Transmuted Log-Logistic Distribution." *Journal of Statistics Applications & Probability*, **2**, 11–20.

Aryal GR, Tsokos CP (2009). "On the Transmuted Extreme Value Distribution with Application." *Nonlinear Analysis: Theory, Methods and Applications*, **71**, 1401–1407.

Aryal GR, Tsokos CP (2011). "Transmuted Weibull Distribution: A Generalization of the Weibull Probability Distribution." *European Journal of Pure & Applied Mathematics*, **4**, 89–102.

de Gusmão FRS, Ortega EMM, Cordeiro GM (2011). "The Generalized Inverse Weibull Distribution." *Statistical Papers*, **52**, 591–619.

Drapella A (1993). "The Complementary Weibull Distribution: Unknown or Just Forgotten?" *Quality and Reliability Engineering International*, **9**, 383–385.

Johnson NL, Kotz S, Balakrishnan N (1994). *Continuous Univariate Distributions*. Wiley, New York.

Keller AZ, Goblin MT, Farnworth NR (1985). "Reliability Analysis of Commercial Vehicle Engines." *Reliability Engineering*, **10**, 15–25.

Khan MS, King R (2013). "Transmuted Modified Weibull Distribution: A Generalization of the Modified Weibull Probability Distribution." *European Journal of Pure and Applied Mathematics*, **6**, 66–88.

Mudholkar GS, Kollia GD (1994). "Generalized Weibull Family: A Structural Analysis." *Communications in Statistics – Theory and Methods*, **23**, 1149–1171.

Murthy DNP, Xie M, Jiang R (2004). *Weibull Models*. John Wiley & Sons.

Shaw WT, Buckley IR (2009). "The Alchemy of Probability Distributions: Beyond Gram-Charlier Expansions, and a Skew-Kurtotic-Normal Distribution from a Rank Transmutation Map." *arXiv preprint*, p. arXiv:0901.0434.

Swain JJ, Venkatraman S, Wilson JR (1988). "Least-Squares Estimation of Distribution Functions in Johnson's Translation System." *Journal of Statistical Computation and Simulation*, **29**, 271–297.

**Affiliation:**

Faton Merovci
Department of Mathematics,
University of Prishtina
Prishtinë, 10000
Kosovo
E-mail: fmerovci@yahoo.com

Ibrahim Elbatal
Institute of Statistical Studies and Research
Department of Mathematical Statistics
Cairo University
Egypt
E-mail: i_elbatal@staff.cu.edu.eg

Alaa Ahmed
Institute of Statistical Studies and Research
Department of Mathematical Statistics
Cairo University
Egypt
E-mail: Alaa_mnn@yahoo.com

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43/2
June 2014

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

*Submitted:* 2013-10-08
*Accepted:* 2014-03-01

Austrian Journal of Statistics
*June 2014, Volume 43/2, 133–149.*
*http://www.ajs.or.at/*

# The Beta Transmuted Weibull Distribution

**Manisha Pal**
Calcutta University

**Montip Tiensuwan**
Mahidol University

**Abstract**

The paper introduces a beta transmuted Weibull distribution, which contains a number of distributions as special cases. The properties of the distribution are discussed and explicit expressions are derived for the mean deviations, Bonferroni and Lorenz curves, and reliability. The distribution and moments of order statistics are also studied. Estimation of the model parameters by the method of maximum likelihood is discussed. The log beta transmuted Weibull model is introduced to analyze censored data. Finally, the usefulness of the new distribution in analyzing positive data is illustrated.

*Keywords*: reliability function, moment generating function, mean deviation, Bonferroni and Lorenz curve, reliability and entropies, maximum likelihood estimation.

## 1. Introduction

The Weibull distribution is a very popular life time probability distribution that has been extensively used for modeling in reliability, engineering and biological studies. Generalizations of the distribution have been suggested by many authors. Sarhan and Zaindin (2009) studied the modified Weibull distribution, Mudholkar and Srivastava (1993) introduced the exponentiated Weibull distribution and Pal, Masoom Ali, and Woo (2006) investigated many of its properties. Elbatal (2011) studied the exponentiated modified Weibull distribution.

A class of generalized distributions $F(x)$ has been receiving considerable attention over the last few years, in particular, after the studies by Eugene, Lee, and Famoye (2002) and Jones (2004). If $G$ denotes the baseline cumulative distribution function (cdf) of a random variable, then the beta-$G$ distribution is defined as

$$F(x) = I_{G(x)}(a,b) = \frac{1}{B(a,b)} \int_0^{G(x)} w^{a-1}(1-w)^{b-1} dw\,, \qquad 0 < a\,,\ 0 < b\,. \tag{1}$$

Here, $I_y(a,b) = B_y(a,b)/B(a,b)$ is the incomplete beta function ratio, $B_y(a,b) = \int_0^y w^{a-1}(1-w)^{b-1}dw$ is the incomplete beta function and $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function, where $\Gamma(\cdot)$ is the gamma function. The probability density function (pdf) of the above distribution has the form

$$f(x) = \frac{1}{B(a,b)} G(x)^{a-1}\{1-G(x)\}^{b-1} g(x)\,, \qquad x > 0\,. \tag{2}$$

Based on the above generalization, Lee, Famoye, and Olumolade (2007) introduced the beta-Weibull distribution. Thereafter, Silva, Ortega, and Cordeiro (2010) investigated the beta modified Weibull

distribution, and Cordeira, Gomes, da-Silva, and Ortega (2013) made a detailed study of the beta-exponentiated Weibull distribution.

Recently, Aryall and Tsokos (2011) introduced another generalization of the Weibull distribution, which they called the transmuted Weibull distribution. A random variable $T$ is said to have transmuted Weibull probability distribution with parameters $\alpha$, $\beta > 0$ and $|\lambda| \leq 1$, if it has the pdf given by

$$g_{TW}(x) = \alpha\beta x^{\beta-1} \exp(-\alpha x^{\beta})(1 - \lambda + 2\lambda \exp(-\alpha x^{\beta}))\,, \qquad x > 0\,, \tag{3}$$

where $\alpha$ and $\beta$ are the shape parameters representing the different patterns of the transmuted Weibull distribution and are positive, and $\lambda$ is the transmuted parameter.

The cdf of the transmuted Weibull distribution is obtained as

$$G_{TW}(x) = (1 - \exp(-\alpha x^{\beta}))(1 + \lambda \exp(-\alpha x^{\beta}))\,, \qquad x > 0\,. \tag{4}$$

In this paper, we introduce and study several mathematical properties of a new distribution, referred to as a beta transmuted Weibull (BTW) distribution. The distribution has two extra shape parameters which provide greater flexibility in modelling observed positive data. The paper is organized as follows. In Section 2, we introduce the distribution. In Sections 3, we obtain expansions of the cdf and pdf of the distribution using power series. Quantile function and mean deviation are derived in Sections 4 and 5. Order statistics and their moments are discussed in Sections 6 and 7. In Section 8, the stress-strength reliability is obtained. Estimation of parameters by the maximum likelihood method is discussed in Section 9. In Section 10, log beta transmuted Weibull regression model is investigated. In Section 12, the distribution is used for analyzing real life data. Finally, in Section 13, we make some concluding remarks on our study.

## 2. The beta transmuted Weibull distribution

The five-parameter BTW distribution is obtained by taking $G(x)$ in (1) to be the cdf of a transmuted Weibull distribution given by (4). The BTW cdf then becomes

$$\begin{aligned} F(x) &= I_{(1-\exp(-\alpha x^{\beta}))(1+\lambda \exp(-\alpha x^{\beta}))}(a, b) \\ &= \frac{1}{B(a,b)} \int_0^{(1-\exp(-\alpha x^{\beta}))(1+\lambda \exp(-\alpha x^{\beta}))} w^{a-1}(1-w)^{b-1} dw\,, \qquad x > 0\,, \end{aligned} \tag{5}$$

where $\alpha > 0$, $\beta > 0$, $|\lambda| \leq 1$, and $a > 0$, $b > 0$.

The cdf can be expressed in a closed form using the hypergeometric function (see Cordeiro and Nadarajah 2011) as follows:

$$\begin{aligned} F(x) &= \frac{\{(1 - \exp(-\alpha x^{\beta}))(1 + \lambda \exp(-\alpha x^{\beta}))\}^a}{aB(a,b)} \\ &\quad {}_2F_1(a, 1-b; a+1; (1 - \exp(-\alpha x^{\beta}))(1 + \lambda \exp(-\alpha x^{\beta})))\,, \end{aligned}$$

where

$${}_2F_1(c, d; e; z) = \sum_{k=0}^{\infty} \frac{(c)_k (d)_k}{k!(e)_k} z^k$$

is the Gaussian hypergeometric function with $(c)_k$ defined as

$$\begin{aligned} (c)_k &= c(c+1)(c+2)\cdots(c+k-1)\,, \qquad k = 1, 2, \ldots \\ (c)_0 &= 1\,. \end{aligned}$$

The pdf $f(x)$ and the hazard rate function $h(x)$ are obtained as

$$\begin{aligned} f(x) &= \frac{1}{B(a,b)}\alpha\beta t^{\beta-1}\exp(-\alpha t^{\beta})(1-\lambda+2\lambda\exp(-\alpha t^{\beta})) \\ &\quad (1-\exp(-\alpha x^{\beta}))^{a-1}(1+\lambda\exp(-\alpha x^{\beta}))^{a-1} \\ &\quad \{1-(1-\exp(-\alpha x^{\beta}))(1+\lambda\exp(-\alpha x^{\beta}))\}^{b-1}, \quad t>0, \end{aligned} \tag{6}$$

$$\begin{aligned} h(x) &= \frac{\alpha\beta x^{\beta-1}\exp(-\alpha x^{\beta})(1-\lambda+2\lambda\exp(-\alpha x^{\beta}))}{B(a,b)I_{1-(1-\exp(-\alpha x^{\beta}))(1+\lambda\exp(-\alpha x^{\beta}))}(b,a)} \\ &\quad (1-\exp(-\alpha x^{\beta}))^{a-1}(1+\lambda\exp(-\alpha x^{\beta}))^{a-1} \\ &\quad \{1-(1-\exp(-\alpha x^{\beta}))(1+\lambda\exp(-\alpha x^{\beta}))\}^{b-1}, \quad t>0\,. \end{aligned} \tag{7}$$

Plots of the pdf (6) and the hazard rate function (7) for some values of $\alpha$, $\beta$, $\lambda$, $a$ and $b$ are given in Figures 1 and 2, respectively. The BTW failure rate function can be monotonically decreasing or increasing and upside-down bathtub depending on the values of its parameters.

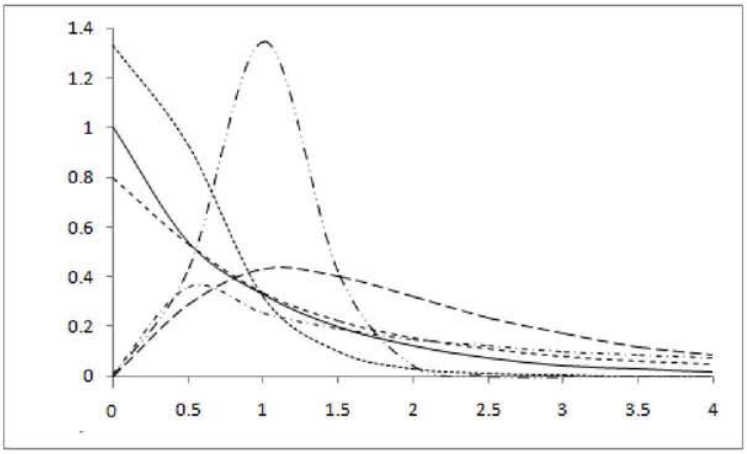


Figure 1: Pdf of beta transmuted Weibull distribution for $\alpha = 1$ and (i) $\beta = 1$, $\lambda = 0$, $a = 1$, $b = 1$, (ii) $\beta = 1$, $\lambda = 0.2$, $a = 3$, $b = 0.75$, (iii) $\beta = 0.5$, $\lambda = 0.5$, $a = 3$, $b = 0.75$, (iv) $\beta = 2$, $\lambda = 1$, $a = 3$, $b = 0.75$, (v) $\beta = 1.5$, $\lambda = -1$, $a = 3$, $b = 0.75$, (vi) $\beta = 1.3$, $\lambda = 0.7$, $a = 0.8$, $b = 1.2$.

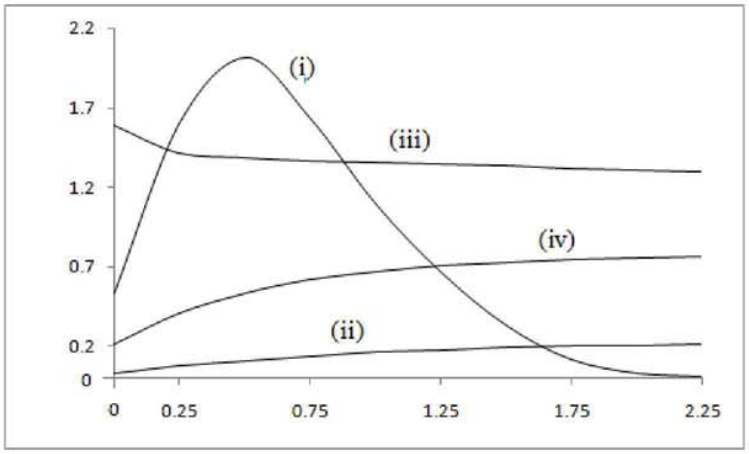


Figure 2: Hazard rate function of beta transmuted Weibull distribution for $\alpha = 1$ and (i) $\beta = 3$, $\lambda = 0.5$, $a = 1$, $b = 2$, (ii) $\beta = 0.75$, $\lambda = -1$, $a = 2$, $b = 0.75$, (iii) $\beta = 0.75$, $\lambda = 1$, $a = 3$, $b = 2$, (iv) $\beta = 0.75$, $\lambda = -0.5$, $a = 3$, $b = 2$.

The following distributions are obtained from the BTW distribution by proper choice of its parameters:

| Parameters | Distribution |
|---|---|
| $b = 1$ | exponentiated transmuted Weibull |
| $a = b = 1$ | transmuted Weibull |
| $\lambda = 0$ | beta Weibull |
| $\lambda = 0$, $b = 1$ | exponentiated Weibull |
| $\lambda = 0$, $a = b = 1$ | Weibull |
| $\beta = 1$ | beta transmuted exponential |
| $\beta = 1$, $a = b = 1$ | transmuted exponential |
| $\beta = 1$, $\lambda = 0$ | beta exponential |
| $\beta = 1$, $\lambda = 0$, $b = 1$ | exponentiated exponential |
| $\beta = 1$, $\lambda = 0$, $a = b = 1$ | exponential |

## 3. Expansions for the CDF and PDF

Here we express $F(x)$ and $f(x)$ in terms of infinite (or finite) weighted sums of cdf's and pdf's of Weibull distributions, respectively.

We note that for $b > 0$ real non-integer, we can replace $(1 - w)^{b-1}$ under the integral in (1) by the power series expansion of binomials and integrate to obtain

$$\frac{1}{B(a,b)} \int_0^{G(x)} w^{a-1}(1-w)^{b-1} dw = \sum_{j=0}^{\infty} (-1)^j \binom{b-1}{j} \frac{G(x)^{a+j}}{a+j},$$

where the binomial term

$$\binom{b-1}{j} = \frac{\Gamma(b)}{\Gamma(b-j)j!}$$

is defined for any real $b$. Then, from (5) we get

$$F(x) = \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(b)}{\Gamma(b-j)j!} \frac{\{(1-\exp(-\alpha x^\beta))(1+\lambda \exp(-\alpha x^\beta))\}^{a+j}}{B(a,b)(a+j)}, \qquad x > 0. \quad (8)$$

Using the binomial expansion another two times we have for $x > 0$

$$\begin{aligned} F(x) &= \sum_{j,k,l=0}^{\infty} (-1)^{j+k} \binom{b-1}{j} \binom{a+j}{k} \binom{a+j}{l} \lambda^l \frac{\exp(-\alpha(k+l)x^\beta))}{B(a,b)(a+j)} \\ &= \sum_{j,k,l=0}^{\infty} (-1)^{j+k} \binom{b-1}{j} \binom{a+j}{k} \binom{a+j}{l} \frac{\lambda^l \{1 - G_1(x; \alpha(k+l), \beta)\}}{B(a,b)(a+j)}, \end{aligned} \quad (9)$$

where $G_1(\alpha(k+l), \beta)$ is the Weibull cdf with scale $\alpha(k+l)$ and shape $\beta$.

Differentiating (9) with respect to $x$ gives a useful expansion of $f(x)$ as

$$f(x) = \sum_{k,l=0}^{\infty} w_{kl} g(x; \alpha(k+l), \beta), \qquad x > 0, \quad (10)$$

where

$$w_{kl} = \sum_{j=0}^{\infty} (-1)^{j+k+1} \binom{b-1}{j} \binom{a+j}{k} \binom{a+j}{l} \frac{\lambda^l}{B(a,b)(a+j)}$$

and $g(x; \alpha(k+l), \beta)$ is the Weibull pdf with scale $\alpha(k+l)$ and shape $\beta$. If $b > 0$ is an integer, the index $j$ in the sum stops at $b - 1$, and if $a$ is an integer, then the indices $k$ and $l$ in the sum stop at $a + j$.

The moments and the moment generating function of the BTW distribution can be easily expressed as functions of those quantities for Weibull distributions by using expression (10) of its pdf.

If $X$ has a Weibull distribution with scale $\theta$ and shape $\delta$, we have

$$\begin{aligned} \mathsf{E}(X^r) &= \frac{1}{\theta^{r/\delta}}\Gamma(1+r/\delta)\,, \\ M_X(t) &= \sum_{r=0}^{\infty}\frac{t^r\theta^{-r/\delta}}{r!}\Gamma(1+r/\delta)\,, \qquad \delta \geq 1\,. \end{aligned}$$

Hence, for $X \sim$ BTW with density given by (10) we get

$$\mathsf{E}(X^r) = \Gamma(1+r/\beta)\sum_{k,l=0}^{\infty} w_{kl}\frac{1}{(\alpha(k+l))^{r/\beta}}\,, \tag{11}$$

$$M_X(t) = \sum_{k,l=0}^{\infty} w_{kl}\sum_{r=0}^{\infty}\frac{t^r(\alpha(k+l))^{-r/\beta}}{r!}\Gamma(1+r/\beta) \qquad \beta \geq 1\,. \tag{12}$$

## 4. Quantile function and simulation

The quantile function corresponding to the BTW distribution with cdf (5) is

$$x = Q(y) = F^{-1}(y) = \left[-\frac{1}{\alpha}\log(z^*)\right]^{1/\beta}\,, \tag{13}$$

where $z^* \in (0,1)$ is a solution to the quadratic equation

$$\lambda z^2 + (1-\lambda)z - (1 - I_y^{-1}(a,b)) = 0\,,$$

and $I_y^{-1}(a,b)$ denotes the inverse of the incomplete beta function with parameters $a$ and $b$. Clearly,

$$z^* = \frac{1}{2\lambda}\left\{\sqrt{(1-\lambda)^2 + 4\lambda(1 - I_y^{-1}(a,b))} - (1-\lambda)\right\}\,. \tag{14}$$

The following expansion for the inverse of the beta incomplete function $I_y^{-1}(a,b)$ can be found on the Wolfram website http://functions.wolfram.com/06.23.06.0004.01

$$\begin{aligned} I_u^{-1}(a,b) &= w + \frac{b-1}{a+1}w^2 + \frac{(b-1)(a^2+3ab-a+5b-4)}{2(a+1)^2(a+2)}w^3 \\ &+ \frac{(b-1)[a^4+(6b-1)a^3+(b+2)(8b-5)a^2]}{2(a+1)^2(a+2)}w^4 \\ &+ \frac{(b-1)[(33b^2-30b+4)a+b(31a-47)+18]}{3(a+1)^3(a+2)(a+3)}w^4 + O(p^{5/a})\,, \end{aligned}$$

where $w = \{aB(a,b)y\}^{1/a}$, $a > 0$.

Simulation of $X$ is straightforward from (13) by taking

$$X = \left[-\frac{1}{\alpha}\log\left(\frac{\sqrt{(1-\lambda)^2 + 4\lambda(1-B)} - (1-\lambda)}{2\lambda}\right)\right]^{1/\beta}\,, \tag{15}$$

where $B$ is a beta variate with shape parameters $a$ and $b$.

## 5. Mean deviation

The amount of scatter in a population is evidently measured to some extent by the totality of deviations from the mean and the median. If $X$ has a BTW distribution, then we can derive the mean deviations

about the mean $\mu = \mathsf{E}(X)$ and about the median $M$ as

$$\begin{aligned} \eta_1 &= \int_0^\infty |x-\mu| f(x)dx\,, \\ \eta_2 &= \int_0^\infty |x-M| f(x)dx\,. \end{aligned}$$

The mean of the distribution is obtained from (11) by putting $r = 1$, and the median is obtained by solving the equation

$$I_{(1-\exp(-\alpha x^\beta))(1+\lambda\exp(-\alpha x^\beta))}(a,b) = \frac{1}{2}\,.$$

Thus, the above measures can be derived from the following relations:

$$\eta_1 = 2[\mu F(\mu) - J(\mu)] \quad \text{and} \quad \eta_2 = \mu - 2J(M)\,, \tag{16}$$

where $J(t) = \int_0^t xf(x)dx$. From (10) we have

$$\begin{aligned} J(t) &= \sum_{k,l=0}^{\infty} w_{kl} \int_0^t \alpha(k+l)\beta x^\beta \exp\left(-\alpha(k+l)x^\beta\right) dx \\ &= \sum_{k,l=0}^{\infty} \frac{w_{kl}}{\{\alpha(k+l)\}^{1/\beta}} \int_0^{\alpha(k+l)t^\beta} z^{1/\beta}\exp(-z)dz \\ &= \sum_{k,l=0}^{\infty} \frac{w_{kl}}{\{\alpha(k+l)\}^{1/\beta}} \gamma\left(\alpha(k+l)t^\beta, \beta^{-1}+1\right)\,, \end{aligned} \tag{17}$$

where $\gamma(x,\delta) = \int_0^x w^{\delta-1}\exp(-w)dw$, $\delta > 0$ is an incomplete gamma function. Using (9), one can easily find $\eta_1$ and $\eta_2$ from (16).

The quantity $J(t)$ can also be used to determine Bonferroni and Lorenz curves, which have applications in economics to study income and poverty, and also in other fields like reliability, demography, insurance and medicine. Bonferroni and Lorenz functions are given by $B(\pi) = J(p)/(\pi\mu)$ and $L(\pi) = J(p)/\mu$, respectively, where $p = Q(\pi)$ is calculated from (13) for a given probability $\pi$.

## 6. Order statistics

If $X_{(1)} < \cdots < X_{(n)}$ denote the ordered observations in a data set from the BTW distribution given by (5) and (6), then the pdf $f_{i:n}(x)$ of the $i$th order statistic $X_{(i)}$ is

$$f_{i:n}(x) = \frac{1}{B(i,n-i+1)} f(x)F(x)^{i-1}[1-F(x)]^{n-1}\,.$$

Using expressions (9) and (10) for $F(x)$ and $f(x)$, respectively, and applying the binomial expansion yields

$$\begin{aligned} f_{i:n}(x) &= \frac{1}{B(i,n-i+1)} f(x) \sum_{s=0}^{n-i} (-1)^s \binom{n-i}{s} F(x)^{i+s-1} \\ &= \frac{\alpha\beta x^{\beta-1}}{B(i,n-i+1)} \left(\sum_{k,l=0}^{\infty} w_{kl}(k+l)\exp(-\alpha(k+l)x^\beta)\right) \\ &\quad \sum_{s=0}^{n-i} (-1)^{s+1} \binom{n-i}{s} \left(\sum_{k,l=0}^{\infty} w_{kl}\exp(-\alpha(k+l)x^\beta)\right)^{i+s-1}\,. \end{aligned} \tag{18}$$

Writing $u = \exp(-\alpha x^\beta)$, $f_{i:n}(x)$ can be expressed as

$$\begin{aligned} f_{i:n}(x) &= \frac{\alpha\beta x^{\beta-1}}{B(i, n-i+1)} \left( \sum_{k,l=0}^{\infty} w_{kl}(k+l) u^{k+l} \right) \\ &\quad \sum_{s=0}^{n-i} (-1)^{s+1} \binom{n-i}{s} \left( \sum_{k,l=0}^{\infty} w_{kl} u^{k+l} \right)^{i+s-1} . \end{aligned} \tag{19}$$

We note that in (19) we can write

$$\sum_{k,l=0}^{\infty} w_{kl} u^{k+l} = \sum_{m=0}^{\infty} w_m^* u^m$$

and

$$\sum_{k,l=0}^{\infty} w_{kl}(k+l) u^{k+l} = \sum_{m=0}^{\infty} m w_m^* u^m ,$$

where $w_m^* = \sum_{k,l:k+l=m} w_{kl}$. Further, from (Gradshteyn and Ryzhik 2000, Section 0.314), for any positive integer $r$,

$$\left( \sum_{k=0}^{\infty} a_k u^k \right)^r = \sum_{k=0}^{\infty} d_{r,k} u^k , \tag{20}$$

where the coefficients $d_{r,k}$, for $k = 1, 2, \ldots$, can be determined from the recurrence equation

$$d_{r,k} = (k a_0)^{-1} \sum_{m=1}^{k} \{m(r+1) - k\} a_m d_{r,k-m} \tag{21}$$

and $d_{r,0} = a_0^r$. Hence, $d_{r,k}$ comes directly from $d_{r,0}, \ldots, d_{r,k-1}$ and, therefore, from $a_0, \ldots, a_k$. Using (20) and (21) it follows that

$$f_{i:n}(x) = \frac{\alpha\beta x^{\beta-1}}{B(i, n-i+1)} \left( \sum_{m=0}^{\infty} m w_m^* u^m \right) \sum_{s=0}^{n-i} (-1)^{s+1} \binom{n-i}{s} \left( \sum_{m=0}^{\infty} d_{i+s-1,m} u^m \right) ,$$

where

$$\begin{aligned} d_{i+s-1,m} &= (m w_0^*)^{-1} \sum_{q=1}^{m} [q(i+s) - m] w_m^* d_{i+s-1,m-q} , \\ d_{i+s-1,0} &= (w_0^*)^{i+s-1} = \left( \sum_{j=0}^{\infty} (-1)^{j+1} \binom{b-1}{j} \frac{1}{B(a,b)(a+j)} \right)^{i+s-1} . \end{aligned}$$

Combining terms, we obtain

$$\begin{aligned} f_{i:n}(x) &= \frac{\alpha\beta x^{\beta-1}}{B(i, n-i+1)} \sum_{s=0}^{n-i} (-1)^{s+1} \binom{n-i}{s} \sum_{m=1}^{\infty} \sum_{t=0}^{\infty} m d_{i+s-1,t} w_m^* u^{m+t} \\ &= \frac{1}{B(i, n-i+1)} \sum_{s=0}^{n-i} (-1)^{s+1} \binom{n-i}{s} \\ &\quad \sum_{m=1}^{\infty} \sum_{t=0}^{\infty} \frac{m d_{i+s-1,t} w_m^*}{m+t} \{(m+t)\alpha\beta x^{\beta-1} \exp(-(m+t)\alpha x^\beta)\} \\ &= \sum_{m=1}^{\infty} \sum_{t=0}^{\infty} c_i(m,t) g(x; (m+t)\alpha, \beta) , \end{aligned} \tag{22}$$

where $g(x;(m+t)\alpha,\beta)$ denotes the pdf of a Weibull distribution with scale parameter $(m+t)\alpha$ and shape parameter $\beta$ and

$$c_i(m,t) = \frac{1}{B(i,n-i+1)}\frac{mw_m^*}{m+t}\sum_{s=0}^{n-i}(-1)^{s+1}\binom{n-i}{s}d_{i+s-1,t}\,. \tag{23}$$

## 7. Moments of order statistics and L-moments

The moments of the order statistics of BTW distribution can be easily written in terms of those of a Weibull distribution by using the expression (22) of the pdf of the order statistic distribution. We get

$$\mathsf{E}(X_{i:n}^r) = \Gamma(\frac{r}{\beta+1})\sum_{m=1}^{\infty}\sum_{t=0}^{\infty}c_i(m,t)\{(m+t)\alpha\}^{-r/\beta}\,, \tag{24}$$

where $c_i(m,t)$ is given in (23).

As indicated by Cordeira *et al.* (2013), L-moments Hosking (1990) are summary statistics for probability distributions and data samples but have several advantages over ordinary moments. For example, they apply for any distribution having a finite mean and no higher-order moments need be finite. The $r$th L-moment is computed from linear combinations of the ordered data values by

$$\rho_r = \sum_{j=0}^{r-1}(-1)^{r-j-1}\binom{r-1}{j}\binom{r+j-1}{j}\gamma_j\,,$$

where $\gamma_j = \mathsf{E}(XF(X)^j)$. Thus, $\rho_1 = \gamma_0$, $\rho_2 = 2\gamma_1 - \gamma_0$, $\rho_3 = 6\gamma_2 - 6\gamma_1 + \gamma_0$, and $\rho_4 = 20\gamma_3 - 30\gamma_2 + 12\gamma_1 - \gamma_0$. In general, we get $\gamma_k = (k+1)^{-1}\mathsf{E}(X_{k+1:k+1})$, which can be computed from (24) by using (23) and putting $i = n = k+1$ and $r = 1$.

## 8. Reliability

A stress-strength model describes the life of a component which has a random strength $X_1$ and is subjected to a random stress $X_2$. The component functions satisfactorily as long as $X_1 > X_2$, and fails when $X_1 < X_2$. The probability $R = \Pr(X_1 > X_2)$ defines the component reliability. Stress-strength models have many applications especially in engineering concepts such as structures, deterioration of rocket motors, static fatigue of ceramic components, fatigue failure of aircraft structures and the aging of concrete pressure vessels.

Consider $X_1$ and $X_2$ to be independently distributed, with $X_1 \sim \text{BTW}(\alpha_1,\beta,\lambda_1,a_1,b_1)$ and $X_2 \sim \text{BTW}(\alpha_2,\beta,\lambda_2,a_2,b_2)$.The cdf $F_1$ of $X_1$ and pdf $f_2$ of $X_2$ are obtained from (9) and (10), respectively. Then,

$$\begin{aligned} R = \Pr(X_1 > X_2) &= \int_0^{\infty} f_2(y)[1-F_1(y)]dy \\ &= 1 + \sum_{k,l=0}^{\infty} w_{kl}^{(1)}\int_0^{\infty} f_2(y)\exp(-\alpha(k+l)y^{\beta})dy \\ &= \sum_{k,l=0}^{\infty} w_{kl}^{(1)}A(k,l)\,, \end{aligned}$$

where

$$w_{kl}^{(i)} = \sum_{j=0}^{\infty}(-1)^{j+k+1}\binom{b_i-1}{j}\binom{a_i+j}{k}\binom{a_i+j}{l}\frac{\lambda^l}{B(a_i,b_i)(a_i+j)}\,, \qquad i=1,2\,,$$

and

$$A(k,l) = \int_0^\infty f_2(y)\exp(-\alpha(k+l)y^\beta)dy\,.$$

Now,

$$\begin{aligned} A(k,l) &= \sum_{r,s=0}^{\infty} w_{rs}^{(2)} \int_0^\infty (r+s)\alpha_2\beta y^{\beta-1}\exp[-\{\alpha_1(k+l)+(r+s)\alpha_2\}y^\beta]dy \\ &= \sum_{r,s=0}^{\infty} w_{rs}^{(2)} \frac{(r+s)\alpha_2}{\alpha_1(k+l)+(r+s)\alpha_2}\,. \end{aligned}$$

Hence,

$$\begin{aligned} R &= 1+\sum_{k,l=0}^{\infty} w_{kl}^{(1)} \sum_{r,s=0}^{\infty} w_{rs}^{(2)} \frac{(r+s)\alpha_2}{(k+l)\alpha_1+(r+s)\alpha_2} \\ &= 1+\sum_{k=0}^{\infty}\sum_{r=0}^{\infty} w_k^{*(1)} w_r^{*(2)} \frac{r\alpha_2}{k\alpha_1+r\alpha_2}\,, \end{aligned} \tag{25}$$

where

$$w_m^{*(i)} = \sum_{k,l:k+l=m} w_{kl}^{(i)}\,, \qquad i=1,2\,.$$

## 9. Maximum likelihood estimation

Let $\theta = (\alpha,\beta,\lambda,a,b)$ denote the parameter vector for the BTW distribution with pdf given by (6). Then, the log-likelihood function $\ell(\theta)$ based on a single observation $x$ is

$$\begin{aligned} \ell(\theta) &= \log(\alpha)+\log(\beta)-\log B(a,b)+(\beta-1)\log(x)-\alpha x^\beta+\log(1-\lambda+2\lambda\exp(-\alpha x^\beta)) \\ &\quad +(a-1)\{\log(1-\exp(-\alpha x^\beta))+\log(1+\lambda\exp(-\alpha x^\beta))\} \\ &\quad +(b-1)\log\{1-(1-\exp(-\alpha x^\beta))(1+\lambda\exp(-\alpha x^\beta))\}\,. \end{aligned}$$

Hence, the components of the unit score vector

$$\frac{\partial\ell(\theta)}{\partial\theta} = \left(\frac{\partial\ell(\theta)}{\partial\alpha}, \frac{\partial\ell(\theta)}{\partial\beta}, \frac{\partial\ell(\theta)}{\partial\lambda}, \frac{\partial\ell(\theta)}{\partial a}, \frac{\partial\ell(\theta)}{\partial b}\right)'$$

are

$$\begin{aligned} \frac{\partial\ell(\theta)}{\partial\alpha} &= \frac{1}{\alpha}-x^\beta-\frac{2\lambda x^\beta\exp(-\alpha x^\beta)}{1-\lambda+2\lambda\exp(-\alpha x^\beta)}+\frac{(a-1)x^\beta\exp(-\alpha x^\beta)}{1-\exp(-\alpha x^\beta)}-\frac{(a-1)\lambda x^\beta\exp(-\alpha x^\beta)}{1+\lambda\exp(-\alpha x^\beta)} \\ &\quad -\frac{(b-1)x^\beta\exp(-\alpha x^\beta)}{1-(1-\exp(-\alpha x^\beta))(1+\lambda\exp(-\alpha x^\beta))}(1-\lambda+2\lambda\exp(-\alpha x^\beta)) \end{aligned}$$

$$\begin{aligned} \frac{\partial\ell(\theta)}{\partial\beta} &= \frac{1}{\beta}+\log(x)-\alpha x^\beta\log(x)-\frac{2\lambda\alpha x^\beta\log(x)\exp(-\alpha x^\beta)}{1-\lambda+2\lambda\exp(-\alpha x^\beta)} \\ &\quad +\frac{(a-1)x^\beta\log(x)\exp(-\alpha x^\beta)}{1-\exp(-\alpha x^\beta)}-\frac{(a-1)\lambda x^\beta\log(x)\exp(-\alpha x^\beta)}{1+\lambda\exp(-\alpha x^\beta)} \\ &\quad -\frac{(b-1)x^\beta\log(x)\exp(-\alpha x^\beta)}{1-(1-\exp(-\alpha x^\beta))(1+\lambda\exp(-\alpha x^\beta))}(1-\lambda+2\lambda\exp(-\alpha x^\beta)) \end{aligned}$$

$$\begin{aligned}\frac{\partial \ell(\theta)}{\partial \lambda} &= -\frac{1-2\exp(-\alpha x^{\beta})}{1-\lambda+2\lambda\exp(-\alpha x^{\beta})}+\frac{(a-1)\exp(-\alpha x^{\beta})}{1+\lambda\exp(-\alpha x^{\beta})}\\ &\quad -\frac{(b-1)\exp(-\alpha x^{\beta})(1-\exp(-\alpha x^{\beta}))}{1-(1-\exp(-\alpha x^{\beta}))(1+\lambda\exp(-\alpha x^{\beta}))}\end{aligned}$$

$$\begin{aligned}\frac{\partial \ell(\theta)}{\partial a} &= \Psi(a+b)-\Psi(a)+\{\log(1-\exp(-\alpha x^{\beta}))+\log(1+\lambda\exp(-\alpha x^{\beta}))\}\\ \frac{\partial \ell(\theta)}{\partial b} &= \Psi(a+b)-\Psi(b)+\log\{1-(1-\exp(-\alpha x^{\beta}))(1+\lambda\exp(-\alpha x^{\beta}))\}\,,\end{aligned}$$

where $\Psi(x)=\frac{d}{dx}\log\Gamma(x)$.

For a random sample $(x_1,\ldots,x_n)$ of size $n$ from $X$, distributed with pdf (6), the sample log-likelihood is $\ell(\theta)=\sum_{i=0}^{n}\ell_i(\theta)$, where $\ell_i(\theta)$ is the log-likelihood for the $i$th observation ($i=1,\ldots,n$), and the score vector is

$$\frac{\partial \ell(\theta)}{\partial \theta}=\sum_{i=0}^{n}\frac{\partial \ell_i(\theta)}{\partial \theta}\,.$$

The maximum likelihood estimate (MLE) $\hat{\theta}$ of $\theta$ is obtained by solving the system

$$\frac{\partial \ell(\theta)}{\partial \theta}=0\,.$$

Under certain regularity conditions, $\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}\text{Normal}(0,I^{-1}(\theta))$ (here $\xrightarrow{d}$ stands for convergence in distribution), where $I(\theta)$ denotes the information matrix given by

$$I(\theta)=\mathsf{E}\left(\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta'}\right)\,.$$

This information matrix $I(\theta)$ may be approximated by the observed information matrix

$$I(\hat{\theta})=\left(\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta'}\right)|_{\theta=\hat{\theta}}\,.$$

Then, using the approximation $\sqrt{n}(\hat{\theta}-\theta)\sim\text{Normal}(0,I^{-1}(\hat{\theta}))$, one can carry out tests and find confidence regions for functions of some or all parameters in $\theta$.

## 10. The log beta transmuted Weibul distribution

If $X$ is a random variable having the BTW distribution given by (8), then $Y=\log(X)$ is said to have a log beta transmuted Weibull (LBTW) distribution.

Let us define $\mu=-1/\beta\log(\alpha)$. Then, the pdf of $Y$ is given by

$$\begin{aligned}f_Y(y) &= \frac{\beta}{B(a,b)}\exp\{\beta(y-\mu)-\exp(\beta(y-\mu))\}\{1-\lambda+2\lambda\exp(-\exp(\beta(y-\mu)))\}\\ &\quad \{1-\exp(-\exp(\beta(y-\mu)))\}^{a-1}\{1+\lambda\exp(-\exp(\beta(y-\mu)))\}^{a-1}\\ &\quad [1-\{1-\exp(-\exp(\beta(y-\mu)))\}\{1+\lambda\exp(-\exp(\beta(y-\mu)))\}]^{b-1}\,,\end{aligned}\tag{26}$$

where $y,\mu\in\mathbb{R}$ and $\beta>0$. Its corresponding cdf is

$$F_Y(y)=I_{A(y)}(a,b)\,,\qquad y\in\mathbb{R}\,,\tag{27}$$

where

$$A(y)=\{1-\exp(-\exp(\beta(y-\mu)))\}\{1+\lambda\exp(-\exp(\beta(y-\mu)))\}\,.$$

The standardized random variable $Z = \beta(Y - \mu)$, therefore, has pdf

$$\begin{aligned} f_Z(z) &= \frac{1}{B(a,b)} \exp\{z - \exp(z)\}\{1 - \lambda + 2\lambda \exp(-\exp(z))\} \\ &\quad \{1 - \exp(-\exp(z))\}^{a-1}\{1 + \lambda \exp(-\exp(z))\}^{a-1} \\ &\quad [1 - \{1 - \exp(-\exp(z))\}\{1 + \lambda \exp(-\exp(z))\}]^{b-1}, \qquad z \in \mathbb{R}. \end{aligned} \tag{28}$$

For $a = b = 1$, we get the log-transmuted Weibull distribution, while for $a = b = 1$ and $\lambda = 0$ we get the log-Weibull distribution or the extreme value distribution.

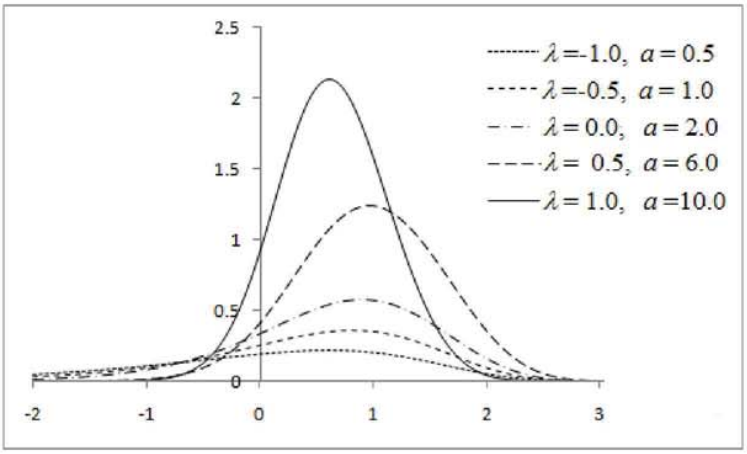


Figure 3: Plots of LBTW pdf's for increasing $\lambda$ and $a$, when $b = 0.5$, $\mu = 0$, and $\alpha = 1$.

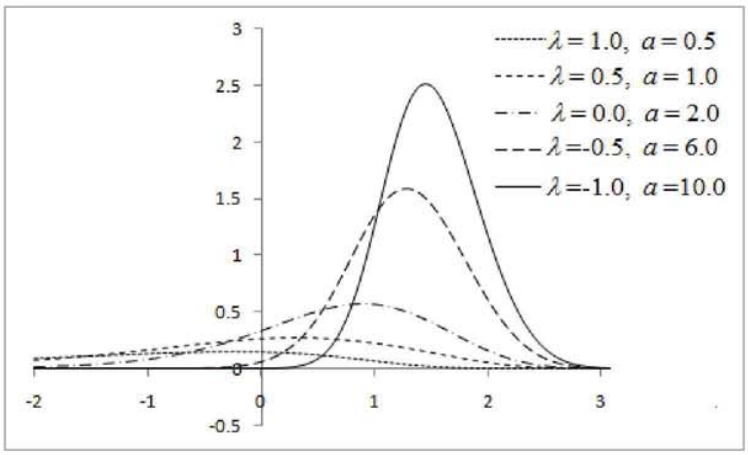


Figure 4: Plots of LBTW pdf's for $\lambda$ decreasing and $a$ increasing, when $b = 0.5$, $\mu = 0$, and $\alpha = 1$.

The $r$th moment of the standardized distribution (28) is given by

$$\begin{aligned} \mathsf{E}(Z^r) &= \int_{-\infty}^{\infty} z^r f_Z(z) dz \\ &= \frac{1}{B(a,b)} \sum_{i=0}^{\infty} (-1)^i \binom{b-1}{i} \int_{-\infty}^{\infty} z^r \exp[z - \exp(z)][1 - \lambda + 2\lambda \exp\{-\exp(z)\}] \\ &\quad [1 - \exp\{-\exp(z)\}]^{a+i-1}[1 + \lambda \exp\{-\exp(z)\}]^{a+i-1} dz, \end{aligned}$$

using the binomial expansion.

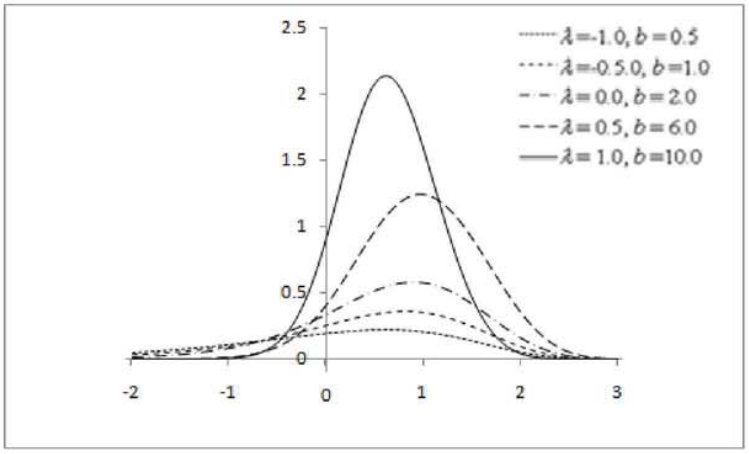


Figure 5: Plots of LBTW densities for increasing $\lambda$ and increasing $b$, when $a = 0.5$, $\mu = 0$, and $\alpha = 1$.

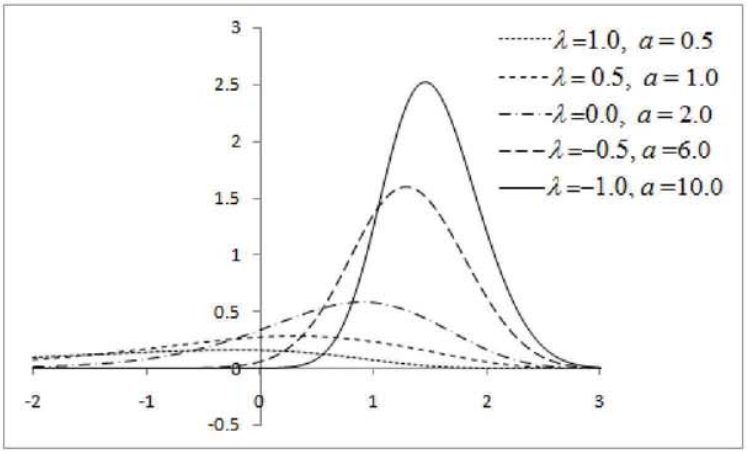


Figure 6: Plots of LBTW pdf's for decreasing $\lambda$ and increasing $b$, when $a = 0.5$, $\mu = 0$, and $\alpha = 1$.

Setting $u = \exp(z)$ we get

$$\begin{aligned}\mathsf{E}(Z^r) &= \frac{1}{B(a,b)}\sum_{i=0}^{\infty}(-1)^i\binom{b-1}{i}\\ &\int_{-\infty}^{\infty}\log(u)^r\exp(-u)[1-\lambda+2\lambda\exp\{-u\}]\\ &\qquad[1-\exp\{-u\}]^{a+i-1}[1+\lambda\exp\{-u\}]^{a+i-1}du\,.\end{aligned}$$

By further power series expansion of binomial, we then have

$$\begin{aligned}\mathsf{E}(Z^r) &= \frac{1}{B(a,b)}\sum_{i,j,k=0}^{\infty}(-1)^{i+j}\lambda^k\binom{b-1}{i}\binom{a+i-1}{j}\binom{a+i-1}{k}\\ &\int_{-\infty}^{\infty}\log(u)^r\exp[-(j+k+1)u][1-\lambda+2\lambda\exp\{-u\}]du\,. \qquad (29)\end{aligned}$$

By (Prudnikov, Brychkov, and Marichev 1986, equation 2.6.21.1), we have

$$I(r,s) = \int_{-\infty}^{\infty}\log(u)^r\exp(-su)du = \left(\frac{\partial^r s^{-p}\Gamma(p)}{\partial p^r}\right)|_{p=1}\,. \qquad (30)$$

Using (30) in (29), we thus obtain

$$\begin{aligned}\mathsf{E}(Z^r) &= \frac{1}{B(a,b)}\sum_{i,j,k=0}^{\infty}(-1)^{i+j}\lambda^k\binom{b-1}{i}\binom{a+i-1}{j}\binom{a+i-1}{k}\\ &\quad [(1-\lambda)I(r,j+k+1)+2\lambda I(r,j+k+2)]\,. \end{aligned} \tag{31}$$

In many practical situations, the value of a random variable is affected by the values of a number of other variables, called explanatory variables. For example, if $X$ denotes the lifetime of a system, then it is affected by explanatory variables like lifetimes of its sub-components, surrounding temperature, etc.

Consider a type 1 censored sample of size $n$, where $x_i$ denotes the true lifetime and $c_i$ the censoring time of the $i$th sampled unit, and $v_i = (v_{1i}, \ldots, v_{pi})'$ denotes the corresponding vector of explanatory variables, $i = 1, \ldots, n$. The $i$th response $y_i$ is defined as $y_i = \min[\log(x_i), \log(c_i)]$. Consider a linear regression model for the response variable using LBTW distribution as follows:

$$y_i = v_i'\gamma + \frac{1}{\beta}z_i\,, \qquad i = 1, \ldots, n\,, \tag{32}$$

where the $z_i$'s are independently distributed with density (28), $\gamma = (\gamma_1, \ldots, \gamma_p)'$, $|\lambda| \leq 1$, $a > 0$, $b > 0$, and the location parameter $\mu_i$ corresponding to the $i$th lifetime is modeled as $\mu_i = v_i'\gamma$. Thus, the location vector for the LBTW model has the structure $\mu = v'\gamma$, where $\mu = (\mu_1, \ldots, \mu_n)'$ and $v = (v_1, \ldots, v_n)'$. For $a = b = 1$, the model reduces to the log-transmuted Weibull model, while for $\lambda = 0$ and $a = b = 1$ it reduces to the log-Weibull (or the extreme value) model.

Denoting by $C$ and $N$ the sets of indices for the censored and uncensored observations respectively, the log-likelihood for the model parameters $\theta = (\beta, \lambda, a, b, \gamma')'$ is given by

$$\begin{aligned}\ell(\theta) &= q[\log(\beta) - \log\{B(a,b)\}]\\ &\quad +\sum_{i\in N}[\beta(y_i-\mu_i)-\exp(\beta(y_i-\mu_i))+\log\{1+\lambda\exp(-\exp(\beta(y_i-\mu_i)))\}]\\ &\quad +(a-1)\sum_{i\in N}[\log\{1-\exp(-\exp(\beta(y_i-\mu_i)))\}+\log\{1+\lambda\exp(-\exp(\beta(y_i-\mu_i)))\}]\\ &\quad +(b-1)\sum_{i\in N}[\log\{1-\{1-\exp(-\exp(\beta(y_i-\mu_i)))\}\{1+\lambda\exp(-\exp(\beta(y_i-\mu_i)))\}\}]\\ &\quad +\sum_{i\in C}\log[1-I_{\{1-\exp(-\exp(\beta(c_i-\mu_i)))\}\{1+\lambda\exp(-\exp(\beta(c_i-\mu_i)))\}}(a,b)]\,, \end{aligned} \tag{33}$$

where $q$ is the number of observed failures. The MLE $\hat{\theta}$ of $\theta$ is obtained by solving the likelihood equations $\partial\ell(\theta)/\partial\theta = 0$.

Under certain regularity conditions, the centered form of the MLE, $\sqrt{n}(\hat{\theta} - \theta)$, is asymptotically distributed as Normal$(0, K^{-1}(\theta))$, where $K(\theta)$ is the information matrix, given by

$$K(\theta) = \mathsf{E}\left(\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta'}\right)$$

and can be approximated by $K(\hat{\theta})$. Then, based on $\sqrt{n}(\hat{\theta} - \theta) \sim$ Normal$(0, K^{-1}(\hat{\theta}))$, one can carry out tests and find confidence regions for functions of $\theta$.

## 11. Simulation study

A simulation study is carried out to investigate the performance of the MLEs. We take sample sizes to be $n \in \{15, 25, 50\}$, and generate observations from a BTW distribution with parameters $\alpha = 1$, $\beta = 2$, $\lambda = 0.5$, $a = b = 2$.The MLEs and 95% confidence intervals are computed using the observed Fisher information matrix. The process is replicated 1000 times, and the average estimates, along with

Table 1: Average MLEs of the parameters and the corresponding mean squared errors (in parenthesis).

| | MLEs | | | | |
|---|---|---|---|---|---|
| $n$ | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\lambda}$ | $\hat{a}$ | $\hat{b}$ |
| 15 | 0.565 (0.201) | 1.251 (0.229) | 0.392 (0.242) | 1.121 (0.198) | 1.301 (0.213) |
| 25 | 0.799 (0.115) | 1.723 (0.099) | 0.435 (0.123) | 1.728 (0.089) | 1.591 (0.101) |
| 50 | 1.027 (0.036) | 1.923 (0.034) | 0.485 (0.041) | 2.013 (0.021) | 1.935 (0.024) |

the mean squared error are presented in Table 1. In Table 2, the average 95% confidence intervals are reported.

From Table 1 it is observed that as the sample size increases, the average biases and the mean squared errors decrease. This verifies the consistency properties of the estimates.

Table 2: Average 95% confidence intervals for the parameters.

| $n$ | $\alpha$ | $\beta$ | $\lambda$ | $a$ | $b$ |
|---|---|---|---|---|---|
| 15 | (0.375, 2.075) | (0.617, 3.675) | (0.072, 1.727) | (0.523, 4.026) | (0.576, 3.972) |
| 25 | (0.477, 1.736) | (0.760, 3.521) | (0.162, 1.216) | (0.727, 3.529) | (0.833, 3.488) |
| 50 | (0.625, 1.421) | (1.102, 2.648) | (0.199, 0.874) | (1.001, 3.135) | (0.928, 2.846) |

Table 2 shows that as the sample size increases, the average confidence lengths decrease and the intervals tend towards symmetry.

## 12. Application of the beta transmuted Weibull model

In this section we illustrate the usefulness of the beta transmuted Weibull distribution for modeling reliability data, and also give an application of the log beta transmuted Weibull regression model. We consider two real data sets, and, for the former, we compare our results with those obtained by fitting the transmuted Weibull distribution, beta exponentiated Weibull distribution, exponentiated Weibull distribution and the Weibull distribution. The cdf of the beta exponentiated Weibull distribution is given by

$$F_{BEW}(x) = \frac{1}{B(a,b)} \int_0^{[1-\exp(-\alpha x^{\beta})]^{\gamma}} w^{a-1}(1-w)^{b-1} dw\,, \qquad x, \alpha, \beta, \gamma, a, b > 0\,, \tag{34}$$

where $\gamma$ is the exponentiating parameter. For $a = b = 1$, it becomes the exponentiated Weibull distribution, which is a special case of BTW distribution with $\lambda = 0$, $a = \gamma$, and $b = 1$. For $a = b = \gamma = 1$, (34) reduces to the Weibull distribution.

### 12.1. Tensile fatigue characteristics of yarn

The first data set relates to the time-to-failure of a polyster/viscose yarn in a textile experiment for testing the tensile fatigue characteristics of yarn. It consists of a sample of 100 centimeter yarn at 2.3% strain level. This data was also studied by Quesenberry and Kent (1982), and is given in Table 3. The Weibull, exponentiated Weibull, beta exponentiated Weibull, transmuted Weibull and beta transmuted Weibull distributions are fitted to the data and the MLEs of the parameters are computed are given in Table 3. The values of maximized log-likelihoods, Akaike information criterion (AIC), Bayesian information criterion (BIC) and Kolmogorov-Smirnov statistic (K-S) for the different fitted distributions are also given. Though all the distributions considered give good fits, the beta transmuted Weibull distribution is seen to be marginally better than the others. A graphical comparison of the fitted models is displayed in Figure 7.

### 12.2. Class-H insulation data

As an application of the LBEW regression model, we consider the accelerated test data given in

Table 3: Failure time data on 100 cms. Yarn at 2.3 % strain level (sample size $n = 100$).

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 86 | 146 | 251 | 653 | 98 | 249 | 400 | 292 | 131 | 169 | 175 | 176 | 76 |
| 264 | 15 | 364 | 195 | 262 | 88 | 264 | 157 | 220 | 42 | 321 | 180 | 198 |
| 38 | 20 | 61 | 121 | 282 | 224 | 149 | 180 | 325 | 250 | 196 | 90 | 229 |
| 166 | 38 | 337 | 65 | 151 | 341 | 40 | 40 | 135 | 597 | 246 | 211 | 180 |
| 93 | 315 | 353 | 571 | 124 | 279 | 81 | 186 | 497 | 182 | 423 | 185 | 229 |
| 400 | 338 | 290 | 398 | 71 | 246 | 185 | 188 | 568 | 55 | 55 | 61 | 244 |
| 20 | 284 | 393 | 396 | 203 | 829 | 239 | 236 | 286 | 194 | 277 | 143 | 198 |
| 264 | 105 | 203 | 124 | 137 | 135 | 350 | 193 | 188 | | | | |

Table 4: Estimated parameters of the Weibull (W), exponentiated Weibull (EW), beta exponentiated Weibull (BEW), transmuted Weibull (TW) and beta transmuted Weibull (BTW) distributions, and the corresponding values of log-likelihood (LL), Akaike information criterion (AIC), Bayesian information criterion (BIC) and Kolmogorov-Smirnov statistic (K-S).

| Distribution | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\gamma}$ | $\hat{\lambda}$ | $\hat{a}$ | $\hat{b}$ | LL | AIC | BIC | K-S |
|---|---|---|---|---|---|---|---|---|---|---|
| W | $1.48 \cdot 10^{-4}$ | 1.60 | 1.00 | 0.00 | 1.00 | 1.00 | $-627.05$ | 1258.10 | 1263.31 | 0.0700 |
| EW | $2.63 \cdot 10^{-4}$ | 1.50 | 1.00 | 0.00 | 1.00 | 1.00 | $-625.58$ | 1257.16 | 1264.98 | 0.0685 |
| TW | $4.68 \cdot 10^{-5}$ | 1.72 | 1.00 | 0.75 | 1.00 | 1.00 | $-624.52$ | 1255.04 | 1264.86 | 0.0669 |
| BEW | $1.75 \cdot 10^{-3}$ | 1.04 | 1.29 | 0.00 | 2.01 | 0.26 | $-622.63$ | 1255.26 | 1268.28 | 0.0681 |
| BTW | $6.14 \cdot 10^{-4}$ | 1.45 | 1.00 | 0.99 | 1.29 | 0.27 | $-619.80$ | 1249.58 | 1262.63 | 0.0659 |

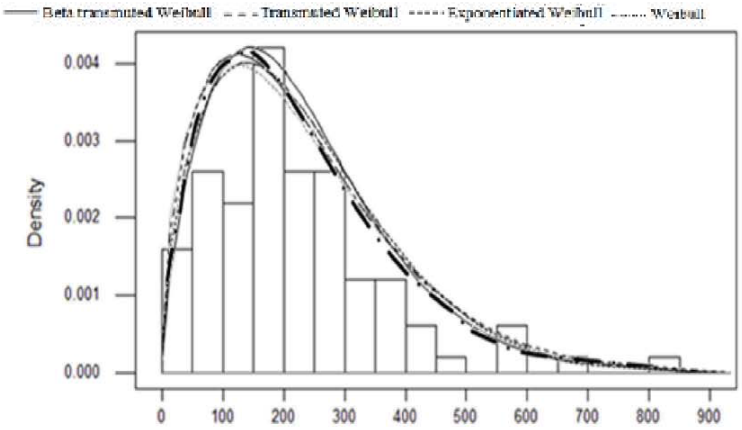


Figure 7: Beta transmuted Weibull, beta exponentiated Weibull, transmuted Weibull, exponentiated Weibull and Weibull densities fitted to the data given in Table 3.

Nelson (1982), which relates to the log time-to-failure of a class H electrical insulation for motors. Four test temperatures were considered: 190, 220, 240 and 260 °C, and a sample of 10 specimens were taken for each test temperature. The specimens were periodically inspected for failure, and the failure time (in hours) of observation $i$, viz. $t_i$, was defined as the midpoint of the interval where the failure occurred. Let, $x_{i1}$ denote the temperature at the $i$th failure. The data are given in Table 5 below:

We adopt the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \frac{1}{\beta} z_i \,,$$

where the variable $y_i = \log t_i$ follows the LBTW distribution (28) for $i = 1, \ldots, 40$. The MLEs of the model parameters are obtained as $\hat{\beta} = 1.2235$, $\hat{\lambda} = 0.2345$, $\hat{a} = 60.25$, $\hat{b} = 1.0210$, $\hat{\beta}_0 = 11.21$, $\hat{\beta}_1 = -0.0287$, and the corresponding log-likelihood value is $-3.25$. Further, it is noted that from the fitted model that that there is a significant difference between the temperatures levels 190 °C, 220 °C, 240 °C, and 260 °C for the failure times. The curves displayed in Figure 8 represent the empirical survival function and the estimated survival function obtained from (27). It shows that the

Table 5: Log life of class H specimens.

| 190 °C | 220 °C | 240 °C | 260 °C |
|---|---|---|---|
| 3.8590 | 3.2465 | 3.0700 | 2.7782 |
| 3.8590 | 3.3867 | 3.0700 | 2.8716 |
| 3.8590 | 3.3867 | 3.1821 | 2.8716 |
| 3.9268 | 3.3867 | 3.1956 | 2.8716 |
| 3.9622 | 3.3867 | 3.2087 | 2.9600 |
| 3.9622 | 3.2867 | 3.2214 | 3.0892 |
| 3.9622 | 3.4925 | 3.2214 | 3.1206 |
| 3.9622 | 3.4925 | 3.2338 | 3.1655 |
| 4.0216 | 3.4925 | 3.2458 | 3.2063 |
| 4.0216 | 3.4925 | 3.2907 | 3.2778 |

fit is considerably good.

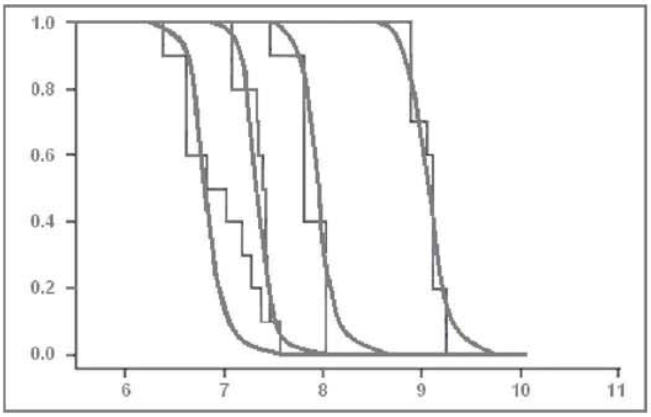


Figure 8: Estimated survival function and the empirical survival function.

## 13. Conclusions

The paper studies some general properties of a new distribution called beta transmuted Weibull distribution. The distribution is a generalization of the Weibull distribution, and includes the Weibull, exponentiated Weibull, exponentiated transmuted Weibull, transmuted Weibull, exponentiated exponential and the exponential distributions as special cases. The log beta transmuted Weibull model has also been discussed, which is appropriate for modeling censored data. Applications of the models to real-life data have been cited and shown to give considerable good fits.

### Acknowledgements

The authors thank the anonymous referee for the valuable comments, which helped immensely to improve the presentation.

## References

Aryall GR, Tsokos CP (2011). "Transmuted Weibull Distribution: A Generalization of the Weibull Probability Distribution." *European Journal of Pure and Applied Mathematics*, **4**, 89–102.

Cordeira GM, Gomes AE, da-Silva CQ, Ortega EMM (2013). "The Beta Exponentiated Weibull Distribution." *Journal of Statistical Computation and Simulation*, **83**, 114–138.

Cordeiro GM, Nadarajah S (2011). "Closed form Expressions for Moments of a Class of Beta Generalized Distributions." *Brazilian Journal of Probability and Statistics*, **25**, 14–33.

Elbatal I (2011). "Exponentiated Modified Weibull Distribution." *Economic Quality Control*, **26**, 189–200.

Eugene N, Lee C, Famoye F (2002). "Beta-normal Distribution and its Applications." *Communications in Statistics – Theory and Methods*, **31**, 497–512.

Gradshteyn IS, Ryzhik IM (2000). *Table of Integrals, Series, and Products*. Academic Press, New York.

Hosking JRM (1990). "L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics." *Journal of the Royal Statistical Society (Series B)*, **52**, 105–124.

Jones MC (2004). "Family of Distributions Arising from Distribution of Order Statistics." *Test*, **13**, 1–43.

Lee C, Famoye F, Olumolade O (2007). "Beta-Weibull Distribution: Some Properties and Applications to Censored Data." *Journal of Modern Applied Statistical Methods*, **6**, 173–186.

Mudholkar GS, Srivastava DK (1993). "Exponentiated Weibull Family for Analyzing Bathtub Failure-rate Data." *IEEE Transactions on Reliability*, **42**, 299–302.

Nelson W (1982). *Applied Life Data Analysis*. John Wiley and Sons, New York.

Pal M, Masoom Ali M, Woo J (2006). "On the Exponentiated Weibull Distribution." *Statistica*, **66**, 139–147.

Prudnikov AP, Brychkov YA, Marichev OI (1986). *Integrals and Series*, volume 1. Gordon and Breach Science Publishers, Amsterdam.

Quesenberry CP, Kent J (1982). "Selecting Among Probability Distributions used in Reliability." *Technometrics*, **24**, 59–65.

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

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

## Affiliation:

Manisha Pal
Department of Statistics
Calcutta University
India
E-mail: manishapal2@gmail.com

Montip Tiensuwan
Department of Mathematics
Faculty of Science
Mahidol University
Thailand
E-mail: montip.tie@mahidol.ac.th

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 43/2 *Submitted:* 2013-11-01
June 2014 *Accepted:* 2014-04-01

Austrian Journal of Statistics
*May 2014, Volume 43, 151–162.*
*http://www.ajs.or.at/*

# Ein Interview mit Wilfried Grossmann

**Wilfried Grossmann**
University of Vienna

**Werner Müller**
Johannes Kepler Univ.

**Matthias Templ**
TU Wien & Statistics Austria

**Abstract**

Das Interview mit Wilfried Grossmann wurde von Werner G. Müller und Matthias Templ am 17.2.2014 durchgeführt. Es beleuchtet das historische Klima des Statistikinstitutes an der Universität Wien, die Ausrichtung der Statistik als „breite“ Datenwissenschaft and der Universität Wien, die Kooperation mit der TU Wien und anderen Institutionen, sowie das Verhältnis zur Statistik Austria, und zwischen der Amtlichen Statistik und der Universitätsstatistik. Zusätzlich wird die Rolle der ÖSG und von EUROSTAT beleuchted. Das Interview widmet sich ausserdem dem Studium der Statistik im Wandel der Zeit - von Lochkarten bis zur Softwareumgebung R und Big Data.

Wilfried Grossmann war Professor für Statistik am Institut für Statistik und danach an der Fakultät für Informatik der Universität Wien Forschungsgruppenleiter der Arbeitsgruppe Data Analysis and Computing. Er hat über 100 Forschungsarbeiten publiziert im Bereich Computational Statistics, Statistisches Datenmanagement, Angewandte Statistik, Theoretische Statistik und Operations Research. Seine aktuellen Forschungsinteressen gelten dem Statistischen Datenmanagement und Informationssystemen, Statistical Computing im Bereich der Amtlichen Statistik, Statistical Knowledge Management, Lehre in der Statistik und Informatik, und Anwendungen von Methoden des Data Mining.

*Keywords*: interview, computational statistics, official statistics.

**Werner Müller:** *Vielen Dank, dass Du dich bereit erklärt hast für dieses Gespräch. Es ist die Idee gekommen in der Österreichischen Zeitschrift für Statistik durch diese Interviewreihe, die hoffentlich eine Reihe wird, auch ein bisschen den Hintergrund unserer Wissenschaft beleuchten zu können. Also nicht immer nur wissenschaftliche Beiträge sondern auch wie es dazukommt, wie man in die Lage kommt, eine wissenschaftliche Karriere in der Statistik überhaupt zu verfolgen in deinem Fall erfolgreich bis zur jetzt erfolgten Pensionierung und hoffentlich auch darüber hinaus. Wie bist du zur Statistik gekommen, bei dir war es ja über die klassische Schiene über die Mathematik?*

**Wilfried Grossmann:** Ich bin eigentlich zufällig zur Statistik gekommen. Als ich mit dem Mathematikstudium begann habe ich von Statistik gar nichts gewusst. Statistik war

damals in der Öffentlichkeit kaum präsent. Ein ehemaliger Schulkollege, der schon ein Jahr vor mir zu studieren begonnen hatte, hat mich mit Erich Neuwirth bekannt gemacht, der ebenfalls Mathematik studierte und als zweites Fach nicht wie traditionell üblich Physik, sondern Statistik im Rahmen eines Studium irregulare. Er erzählte mir von dieser Möglichkeit und ich fand das eine sehr interessante Kombination und entschloss mich ebenfalls diesen Weg zu gehen. Vertreter der Statistik war damals Gerhart Bruckmann, der auf der Wirtschafts- und Sozialwissenschaftlichen Fakultät in Wien neu berufen war. Ich habe diese Entscheidung nicht bereut, weil ich im Laufe des Studiums sehr bald erkannt habe, dass meine Art zu denken wohl eher eine statistische ist als eine mathematische. Mein Studium war natürlich stark mathematisch orientiert und ich habe bei Leopold Schmetterer eine Dissertation über asymptotische Statistik geschrieben. Schmetterer wurde damals als Nachfolger von Swlatscho Sagoroff berufen und wechselte von der Mathematik zur Statistik, also von der Philosophischen auf die Wirtschaft- und Sozialwissenschaftliche Fakultät. Er hat Mitarbeiter für die neue Stelle gesucht und so wurde ich im Jahr 1973 als wissenschaftliche Hilfskraft am damaligen Institut für Statistik angestellt. Ich bin also durch Zufall zur Statistik gekommen, aber ich glaube, dass die damalige Entscheidung für mich die Richtige war.

**Werner Müller:** *Erzähl uns etwas über das historische Klima und die Anfänge dieses Instituts.*

**Wilfried Grossmann:** Was mir gefallen hat und auch meiner Mentalität entspricht war die unerhörte Breite die am Institut für Statistik vertreten war. Ich bin jemand, der immer Interesse hatte Zusammenhänge zwischen verschiedenen Bereichen zu sehen und zu verstehen und nicht so sehr sich in einem engen Bereich zu spezialisieren. Das Klima am damaligen Statistikinstitut war primär von den Persönlichkeiten Gerhart Bruckmann und Leopold Schmetterer geprägt. Bruckmann hatte zwar Mathematik studiert, hat sich aber immer mehr als sozialwissenschaftlicher Statistiker verstanden, während Schmetterer immer Mathematiker und mathematischer Statistiker war. Dann gab es in dieser Zeit an diesem Institut auch den Beginn der Informatik an Österreichischen Universitäten, was mir ganz wichtig erscheint. Es ist interessant, dass die Informatik eigentlich an den Universitäten in Wien aus der Statistik heraus entstanden ist. Swlatscho Sagoroff, der Vorgänger von Schmetterer, hat den ersten Rechner auf einer Österreichischen Universität initiiert. Die Mitarbeiter des Rechenzentrums an der Unversität Wien hatten daher damals eine enge Verbindung mit der Statistik. Eine Person, die in dieser Entwicklung unterschätzt und selten genannt wird, ist Gerhard Derflinger, der dann Statistikprofessor an der Wirtschaftsuniversität wurde. Derflinger hat damals Programme für die Faktorenanalyse erstellt und war weltweit führend in der Entwicklung algorithmischer Lösungen der Faktorenanalyse. Es gab also am Institut die sozialwissenschaftlichen und wirtschaftliche Anwendungen auf der einen Seite, die mathematische Grundlage auf der anderen Seite und natürlich auch die Idee, dass man in Anwendungen die Statistik mit Hilfe der Informatik umsetzen muss. Die Informatiker selbst beschäftigten sich sowohl mit Fragen der Datenorganisation und -speicherung, eine traditionelle Anwendung in der amtlichen Statistik, als auch mit computationalen-algorithmischen Lösungen von statistischen Problemen, das war ein sehr breites Klima.

Bei aller Spezialisierung waren sowohl Schmetterer als auch Bruckmann generell vielseitig interessiert. Bruckmann war der Meinung, dass Statistik das Studium für Generalisten sei, das hat er immer propagiert und in gewissen Sinne hat dieses Institut den Anspruch erhoben, alle Art von Anwendungen formaler Methoden in den Bereichen der Fakultät betreuen zu können. Es gab hier nicht nur die Statistik, es gab Operations Research und Ökonometrie und die Informatik selbst, also die praktische Umsetzung. An diesem Institut herrschte die Vorstellung, „Wir machen das alles“. Im Gegensatz zur TU, da gab es schon sehr früh ein Institut für Operations Research, ein Institut für Ökonometrie, wo sehr gezielt in diesen Bereichen geforscht wurde und natürlich ein

Institut für Statistik und Wahrscheinlichkeitstheorie. Das war für die TU ganz selbstverständlich. Das Institut an der Universität Wien hat für sich immer in Anspruch genommen, das alles mit sehr knappen personellen Ressourcen abzudecken. Es gab 3 Professoren, vielleicht 10 Assistenten und Assistentinnen, und die Unterstützung von externen Lektoren, besonders vom Rechenzentrum, das wie gesagt zu Beginn de facto in Personalunion mit dem Institut für Statistik betrieben wurde. Daher war es so, dass man als Mitarbeiter an diesem Institut nicht nur Statistikvorlesung betreuen musste, sondern auch Mathematik-Lehrveranstaltungen, die Lehre in Operations Research und Ökonometrie, es wurde alles abgedeckt. Das war natürlich sehr interessant und eine Herausforderung.

**Werner Müller:** *Du hast das jetzt angesprochen mit der Informatik, es war ja in Linz auch so ähnlich. Da hat der Kollege Adam das Informatikstudium gegründet, es war wahrscheinlich die Tendenz der Zeit. So gesehen wollen wir darüber sprechen über den Zweig der Statistik der sich wahrscheinlich am stärksten an dieser Entwicklung der computationale Technik entsponnen hat und über die computationale Statistik, wo du eine führende Rolle gespielt hast in den Anfängen. Dann ist da auch die CompStat-Reihe, vielleicht möchtest du auch darüber reden.*

**Wilfried Grossmann:** Nicht nur zur Statistik bin ich durch Zufall gekommen sondern auch viele andere Entscheidungen in meinem Leben sind durch äußere Einflüsse zufällig zustande gekommen. Wenn mich etwas interessierte und der Meinung war, das könnte man sich ansehen und etwas machen, versuchte ich es umzusetzen. So war das auch bei der Entstehung von CompStat. Als ich im Jahr 1974 das Studium abgeschlossen hatte und fix angestellt wurde, gab es im Herbst 1974 den ersten CompStat Kongress. Die Hauptinitiatoren dieses Kongresses waren Peter Paul Sint und Johannes Gordesch, die haben das Ganze initiiert. Wohl in der Tradition, dass an diesem Institut immer computationale Statistik betrieben wurde, insbesondere Anwendungen in der Psychologie, der Chemie und der Physik, hatten sie die Idee einen Kongress zu organisieren. Gordesch ist kurz vor dem Kongress nach Berlin berufen worden, und Sint musste alleine die Organisation übernehmen. Peter Paul Sint ist ein hochintelligenter und unerhört vielseitig interessierter Mann, aber allein war die Organisation für ihn nicht zu schaffen. Da die zentrale Postadresse des Kongresses das Institut war und Sint selbst nicht mehr am Institut beschäftigt war, haben Georg Pflug und ich begonnen Sint bei der Organisation von CompStat zu unterstützen. Der Kongress war dann auch ein Erfolg und ist heute noch eine wesentliche wissenschaftliche Veranstaltung in der computationalen Statistik. Dadurch ist meine Nähe zu der computationalen Statistik gekommen und ist auch ein wesentliches wissenschaftliches Interesse für mich geblieben. Anfangs habe ich mich für algorithmische Fragen interessiert. Das war die algorithmische Lösung für nichtlineare Regression, also eher numerische Probleme. Da am Institut damals die Informatik immer mehr an Bedeutung gewonnen hat, haben wir uns im Laufe der Zeit für Fragen der statistischen Softwareentwicklung, der Simulation, der statistischen Expertensystemen und Fragen der statistischen Datenorganisation interessiert. Dadurch hat es immer eine gewisse Verbundenheit mit den CompStat Entwicklungen gegeben und auch ein Interesse, dass diese Idee weiter Bestand hat. Wir haben dann gemeinsam mit Kollegen Rudi Dutter von der TU im Jahr 1994, den Kongress anlässlich 20 Jahre CompStat an der TU organisiert. Bei diesem Kongress ist meines Wissens nach erstmals SPlus® im wissenschaftlichen Programm und als Aussteller verstärkt aufgetreten. SPlus® wurde ja dann bald von R abgelöst und die heute dominante Rolle von R in der Statistik ist sicher zu einem guten Teil der Verdienst von Kurt Hornik und Friedrich Leisch. Ihre Leistungen für die computationale Statistik ist weit bedeutender als meine eigene.

Weil du das angesprochen hast, möchte ich vielleicht noch einen Punkt über die Verbindung zwischen Statistik und Informatik ansprechen, der aus heutiger Sicht wahrscheinlich eine Fehleinschätzung war. Das ist die Einführung und Planung der Wirtschaftsin-

formatik die an der Universität besonders von Bruckmann sehr stark forciert wurde. Das Engagement für die Wirtschaftsinformatik war sicher richtig auf Grund der Bedeutung und Wichtigkeit der Informatik. Man hat bei der Planung der Wirtschaftsinformatik aber übersehen, dass man im Studium die computationale Statistik als einen essentiellen Bestandteil des Studiums etabliert. Man hat anfangs eine Betriebsinformatik mit Operations Research und eine Wirtschaftsinformatik mit Ökonometrie eingerichtet, das war klar. Aber dass natürlich die Statistik selbst für beide Zweige ein essentieller Bestandteil ist, hat man zu wenig beachtet. Aus heutiger Sicht ist das sicher nicht richtig gewesen, insbesondere wenn man an die heutige Bedeutung von Data Mining im Zusammenhang mit Business Intelligence denkt. Durch die Wirtschaftsinformatik ist das Statistikstudium für Studierende ein bisschen uninteressanter geworden und in eine Nische gewandert. Viele sahen die Statistik primär als eine Verwaltungswissenschaft zur Unterstützung der BWL und der Ökonomie. Erst als du 1983 studiert hast ist es etwas besser geworden, weil sich immer mehr Anwendungen der Statistik ergeben haben. Aus heutiger Sicht würde ich die damalige Entscheidung als Planungsfehler bezeichnen. Aber man kann nicht alles im Vorhinein wissen. Bruckmann hat, wie damals wohl weltweit die meisten Wissenschaftler, mehr auf Operations Research als wesentliches Planungsinstrument für Management Science gesetzt und nicht auf eine datenzentrierte Management Science, die heute als Business Intelligence im Vordergrund steht.

**Matthias Templ:** *Für die jungen Leser des AJS. Wie wurde damals gerechnet? Wenn man z.B. nichtlineare Regression denkt; es war irrer Aufwand dies zu implementieren, sich die LAPACK-Routinen zu besorgen.*

**Wilfried Grossmann:** Genauso war es. Es gab ein Rechenzentrum an der Universität Wien, das war im Keller des neuen Institutsgebäudes, da gab es einen großen Raum, mit Lochkartendruckern.

Da hat man die Jobs gestanzt. Dann hat man die fertigen Jobs fertig in den Kartenleser gesteckt, die sind gelesen und verarbeitet worden, und dann musste man warten, bis auf diesem Endlospapier ein Output gekommen ist. Wenn man einen Fehler gemacht hat, hat man wieder von vorne begonnen. Ein schnelles effektives Arbeiten in heutigem Sinn war undenkbar, dafür war es kommunikativ. Wenn man gewartet hat bis das Programm fertig ist, hat man draußen mit anderen Leuten getratscht, es waren hauptsächlich Chemiker, Physiker und Psychologen dort. Es war alles eine Batchverarbeitung. Meist waren das Fortran-Programme, die Libraries wie die NAG Library für das numerische Rechnen verwendet haben, später kam dann erst PASCAL. Als wir uns dann mit Verkehrssimulation beschäftigten sind auch noch andere Sprachen wie SIMULA dazu gekommen.

Langsam setzten sich erst Terminals durch, wo man Plätze reservieren musste. Dass jeder einen eigenen Rechner mit einer geeigneten Arbeitsumgebung hat, wie es heute der Fall ist, war unvorstellbar. Auch der Speicherplatz und die Rechenkapazität waren begrenzt. Als ich in den späten 80er Jahren und frühen 90er Jahren an Projekten zur Ökosystemsimulation arbeitete haben wir Ozonkonzentrationen analysiert. Wenn ich da eine Clusteranalyse für die Tagesgänge machen wollte musste ich extra eine Erweiterung des Speicherplatzes anfordern und man musste die Daten selbständig partitionieren. Problemlösungen für heute selbstverständliche Datenvolumina waren oft illusorisch.

**Werner Müller:** *Weil du es vorher schon angesprochen hast: die amtliche Statistik. Es gibt ja ein gesundes Spannungsverhältnis zwischen der amtlichen und der akademischen Statistik. Vielleicht kannst du uns darüber etwas erzählen, wie sich das im Laufe der Zeit*

*entwickelt hat.*

**Wilfried Grossmann:** Das ist ein interessanter Punkt und bin über die derzeitige Entwicklung sehr froh, wenn ich sie mit der Vergangenheit vergleiche, als noch Leopold Schmetterer und Lothar Bosse die Vorsitzenden der Statistischen Gesellschaft waren. Ich kann mich an eine Gespräch mit Bosse bei einer Veranstaltung erinnern, als er sagte, es gibt zwei Bereiche in der Statistik, das eine ist die mathematische Statistik das andere ist die amtliche Statistik, das sind zwei Welten die wenig miteinander zu tun haben. Damals war das auch so, dass diese zwei Bereiche eher getrennt voneinander in der Statistischen Gesellschaft lebten. Jede Gruppe respektierte die andere, aber es war mehr ein „teile und herrsche Prinzip". Bei Veranstaltungen trat man gemeinsam auf und Veranstaltungen einer der beiden Gruppen wurde immer finanziell gefördert. Für die universitäre Statistik war dies auch eine große Hilfe, da die Einnahmen der Gesellschaft zu einem überwiegenden Teil aus dem Bereich der amtlichen Statistik kamen. Wissenschaftlich war die amtliche Statistik damals sicher näher an der Informatik, insbesondere Fragen Datenorganisation und Datenspeicherung spielten eine zentrale Rolle, da sie klarerweise sehr eng mit den Aufgaben amtlichen Statistik, die ja statistische Information verwalten und bereitstellen muss, in Verbindung stehen. Hier muss man Präsident Bosse im Nachhinein große Weitsicht zugestehen und auch und großes Lob aussprechen. Er erkannte sehr früh, dass in der amtlichen Statistik statistische Datenbanken eine zentrale Rolle spielen und er hat dementsprechend die Entwicklung im statistischen Zentralamt sehr gefördert und eine Gruppe unter Lutz arbeiten lassen. Diese Gruppe hat ein sehr fortschrittliches Modell für statistische Datenbanken entwickelt, es war damals weltweit eines der besten Systeme. Die Ideen wurden dann von anderen Ländern aufgegriffen, in Österreich ist die Entwicklung leider etwas eingeschlafen.

An methodischer Statistik gab es in der amtlichen Statistik nur geringes Interesse und beschränkte sich auf einfache summarische Statistiken wie Summe, Mittelwerte, Indizes und eventuell noch Varianz. Dazu hat man einfache Grafiken gemacht, mehr gab es nicht. Daher gab es immer ein gewisses Spannungsfeld, das durch das persönliche Geschick und die Toleranz der beiden Vorsitzenden, insbesondere Schmetterer und Bosse aber auch danach Bruckmann, Josef Schmidl und Reinhold Viertl, ausgeglichen werden konnte. Dass man in der amtlichen Statistik auch komplexere Methoden verwendet war damals unüblich. Die wesentliche Methode der amtlichen Statistik war die Stichprobenziehung. Die Stichprobentheorie wurde ja vom Neyman als Grundlage für alle statistischen Verfahren entwickelt und das Prinzip der Randomisierung spielt ja auch für statistische Modellen eine zentrale Rolle. Dann hat sich aber die methodische Statistik rasch weiter entwickelt und in den österreichischen Statistikcurricula ist amtliche Statistik kaum vorgekommen. Im Statistikstudium standen statistische Modelle und deren Fundierung durch die Wahrscheinlichkeitstheorie im Vordergrund.

Stichprobentheorie war nur ein Thema am Rande, mit Ausnahme der Biometrie, aber die Versuchspläne für statistische Experimente setzen doch einen anderen Schwerpunkt als die amtliche Statistik. Modelle, die für die amtliche Statistik interessant sind, sind erst später entwickelt worden, aber da hat es in Österreich kaum Beiträge gegeben. In Österreich waren es zwei Welten. In anderen Ländern haben Statistiker mit Modellen für Modell Assisted Survey Information Collection, oder Model Based Survey Information neue Impulse gesetzt. Besonders in England, Schweden und den USA und Canada wurde diese Entwicklung vorangetrieben. Das begann Ende der 70-iger Jahre. Es gab dann auch die Diskussion über die den Design Based und den Model Based Zugang zur Statistik. Ein Meilenstein für die amtliche Statistik und auch für die computationale Statistik ist meiner Meinung nach die Entwicklung des EM-Algorithmus 1977 durch Dempster Laird und Rubin und dessen Anwendung zur Behandlung von fehlenden Werten.

Heute gibt es erfreulicherweise eine Vielzahl von Methoden die ganz wichtig für die amtliche Statistik sind. Auszählen und summarische Statistiken allein sind nicht mehr

ausreichend. Daten müssen sorgfältig vorverarbeitet werde, z.B. bei fehlenden Werten, damit die Ergebnisstatistiken eine entsprechende Qualität haben. Aber auch für die Modellierung von Zusammenhängen von Daten in der amtlichen Statistik spielen heute Modelle eine größere Rolle. Die Glättung von Zeitreihen ist ein klassisches Beispiel, neuere Anwendungen sind Small Area Estimation um globale Zahlen auf lokale Ebene umzulegen. Es gibt also heute eine Reihe von Bereichen der methodischen Statistik, die für die amtliche Statistik interessant sind. Heute ist die amtliche Statistik mehr als Datenverarbeitung und nicht mehr von der methodische Statistik zu trennen. Das wird auch in den Themen der NTTS-Konferenzen (New Techniques and Technologies for Official Statistics) deutlich, die von EUROSTAT initiiert wurden. Anfangs standen Fragen der Datenorganisation und Metadaten im Vordergrund, heute findet man mehr Beiträge zur statistischen Methodik. Ich empfinde das als eine positive Entwicklung, weil sie zeigt, dass man zu Recht von einer Statistik sprechen kann.

**Werner Müller:** *Dass das Verhältnis von Statistik Austria und der akademischen Statistik nicht vollständig auseinandergedriftet ist in den 70-iger und 80-igern war vielleicht ein Verdienst der ÖSG und der Protagonisten damals.*

**Wilfried Grossmann:** Ja, das ist zweifelsohne ein Verdienst der ÖSG. Weil die ÖSG mit dieser Konstruktion der Doppelvorsitzenden (eine Person aus dem amtlichen Bereich, eine Person aus dem universitärem Bereich) immer versucht hat die Balance aufrecht zu halten. Man war sich bewusst, dass es schwierig ist eine gemeinsame Sprache zu entwickeln aber beide Seiten haben einander respektiert, das ist ein ganz wichtiger Punkt. Innerhalb der ÖSG war immer ein Respekt der beiden Bereiche in einem hohen Maße vorhanden. Besonders erwähnen möchte ich in diesem Zusammenhang Alfred Franz, der sich in seiner Zeit als Sekretär der Gesellschaft sehr um eine stärkere wissenschaftliche Ausrichtung bemüht hat und an der methodischen Statistik immer sehr interessiert war. Die ÖSG hat viel zu einem gemeinsamen Verständnis der Statistik beigetragen. Als es dann Ende der 80-iger Jahre zu Schwierigkeiten im gegenseitigen Verständnis kam wurde die Gesellschaft ja neu strukturiert. Das kann man in dem Artikel über die Geschichte der ÖSG, auf der Homepage nachlesen. Mitte der 90-iger Jahre wurde die Gesellschaft unter der Leitung von Peter Hackl reorganisiert und das neue Modell ist seither sehr erfolgreich. Präsident Joachim Lamel hat das Motto, Triple A ausgegeben: Amtliche Statistik, Angewandte Statistik und Akademische Statistik. Die Statistische Gesellschaft deckt alle diese Bereiche ab und ein auseinanderdriften muss verhindert werden. Das scheint mir sehr wichtig und scheint auch in Österreich gut gelungen.

**Matthias Templ:** *Vielleicht ist doch das ein bisschen zu kritisieren bezgl. den Universitäten in Österreich. Du hast Stichwörter genannt wie Rubin, missing values, design-basierte Verfahren, model-assisted Verfahren. Mit Ausnahme von Linz, wo Quatember und Bacher diesen Bereich betreuen, wird das in Österreich nicht mehr gelehrt. Die ganze Problematik der Methoden in der Offiziellen Statistik wird weitgehend nicht behandelt. Wenn man andere Ländern vergleicht, wie z.B. Deutschland, gibt es sehr wohl Lehrstühle welche die Methoden der Offiziellen Statistik und Stichprobentheorie abdecken. Siehst du das befremdlich?*

**Wilfried Grossmann:** Das Problem sehe ich, aber es ist vielleicht eine Art von Pendelbewegung, wenn man die Geschichte des Statistikstudiums an der Universität Wien ansieht. Am Beginn war das ein sozial- und wirtschaftswissenschaftliches Studium, der Anteil der methodischen Statistik im Studium war eher bescheiden. Es gab anfangs Lehrveranstaltungen aus der amtlichen Statistik. Mitte der 80-iger Jahre gab es die erste Studienreform, die stärker den methodischen Bereich der Statistik in das Statistikstudium einbringen sollte, ich war auch der Meinung, dass das notwendig war. Gleichzeitig wollte man die Bereiche Volkswirtschaft und die Betriebswirtschaft reduzieren. Werner, hast du noch Buchhaltung lernen müssen?

**Werner Müller:** *Natürlich, ich bin noch aus dieser Vorgeneration.*

**Wilfried Grossmann:** Eben, es war ein wirtschaftswissenschaftliches Studium. Buchhaltung und Kostenrechnung waren essentiell, weil man der Meinung war, das müssen die Absolventen können. Mit der Reduktion der Wirtschaftswissenschaften ist auch die amtliche Statistik, die ja einen starken Bezug zur Volkswirtschaft hat, im Studienplan reduziert worden. Die Lehrveranstaltung für Amtliche Statistik hat dann Alfred Franz betreut. Ich habe auch einige Male gemeinsam mit Hofrat Franz eine Lehrveranstaltung zum Thema Amtliche Statistik gemacht. Dann habe ich gemeinsam mit Karl Fröschl und Marcus Hudec diesen Bereich betreut. Wir haben uns in Kenntnis der internationalen Entwicklung bemüht die Verwendung von Methoden in der amtlichen Statistik in das Vorlesungskonzept einzubeziehen. In der letzten Studienreform ist die amtliche Statistik ganz rausgefallen. Leider, denn ich sehe amtliche Statistik als eine wichtige und spezifische Anwendung der Statistik mit eigenständigen Fragestellungen.

Es gab dann noch im Informatikstudium den Zweig „Data Engineering and Statistics“, da gab es auch eine Lehrveranstaltung zur Amtlichen Statistik. Leider gibt es jetzt das Studium nicht mehr, da es sich nicht durchgesetzt hat, vermutlich weil die Organisation im Rahmen eines Informatikstudiums schwierig war.

Aber das ist sicherlich ein Fehler, dass man bei der Planung und Adaption der Statistikstudien, wie ich vorher gesagt habe, jene internationalen Entwicklungen nicht berücksichtigt hat, die helfen die Kluft zwischen amtlicher Statistik und methodischer Statistik zu schließen. Ich persönlich sehe die amtliche Statistik als einen speziellen Bereich der angewandten Statistik. Amtliche Statistik ist eine spezielle Anwendung der Statistik, traditionell in Verbindung mit den Wirtschaftswissenschaften, insbesondere in der Makroökonomie und den Sozialwissenschaften. Der Schwerpunkt war historisch die Bereitstellung der Daten, es ist also eine spezielle Anwendung. Und jede spezielle Anwendung der Statistik erfordert spezielle Methoden, wir haben über die schon vorhin gesprochen. Es ist im Grunde genommen ähnlich zu sehen, wie andere Anwendungen, z.B. im Marketing, in der Betriebswirtschaft, in der Technik oder in der Medizin. Natürlich spielen in der Medizin andere Fragen eine Rolle. Andere Modelle, die scheinbar nichts mit Amtlicher Statistik zu tun haben wurden aber in letzter Zeit interessant. Matthias, Du hast ja in dieser Richtung viel gemacht, z.B. die robuste Statistik ist für viele Fragen der amtlichen Statistik interessant. Nicht im Sinne des Outlier-Modells, dass man fehlerhafte Daten hat, sondern dass man kleine exzeptionelle Gruppen hat, das ist ein essentielles Problem in der Wirtschaftsstatistik, und die müssen geeignet berücksichtigt und dargestellt werden. Es ist auch interessant dass es im Bereich der Registerzählung methodische Anwendungen gibt. Wir haben kürzlich mit Kollegin Lenk eine Anwendung von Klassifikationsverfahren für die Frage der Zuordnung von Personen zu bestimmten Gruppen gemacht, wenn man diese Information fehlt. Wir haben logistische Regression angewendet und auch Boosting, um von Personen festzustellen ob sie in Österreich wohnhaft sind oder nicht. Eine weitere Anwendung sind Belief-functions um zum Beispiel den plausibelsten Familienstand zu bestimmen, wenn in verschiedenen Registern unterschiedliche Einträge sind.

Es ist also heute nicht mehr so, dass der Bereich der Methoden in der Amtlichen Statistik auf wenige einfache Verfahren beschränkt ist. Auf der anderen Seite wäre es vielfach vorteilhaft, wenn die methodischen Statistiker in ihren Anwendungen die Genauigkeit der amtlichen Statistik hinsichtlich der Dokumentation der Datenerhebung übernehmen würden. Das gilt für viele Anwendungen, die unter dem Titel Data Mining gemacht werden, insbesondere wenn Daten vom Internet verwendet werden. Es wäre nicht schlecht, sich zu fragen, woher diese Zahlen kommen, wie valide sie sind, und wie die Grundgesamtheit aussieht die sie repräsentieren sollen. Vielfach glaubt man solche Daten seien eine Vollerhebung, dabei weiß man gar nicht wie die Grundgesamtheit aussieht. Also diese ganzen klassischen Fragen der Datenqualität die in der amtlichen Statistik zentral

sind, die sollte man sich auch bei der Anwendung der Statistik in anderen Bereichen immer wieder stellen.

**Matthias Templ:** *Big Data zum Beispiel...*

**Wilfried Grossmann:** Gerade bei Big Data sind solche Fragen eine Herausforderung. Ich habe von Ralf Münnich bei den letzten Statistiktagen gehört, dass es jetzt bei EUROSTAT einen Arbeitskreis gibt, der sich genau mit diesen Fragen beschäftigt. Eine Lösung des Problems ist sicher nicht einfach und erfordert gute Ideen.

**Werner Müller:** *Es soll ein Master für Offical Statistics, European Master kreiert werden, darüber wird schon längere Zeit gesprochen. In Linz, in unseren Studienplänen ist die Amtliche Statistik sehr wohl noch vorhanden.*

**Wilfried Grossmann:** Das ist sicher ein möglicher Schwerpunkt und das würde ich positiv sehen. Wenn ich heute die Statistikstudien in Österreich ansehe, haben sie sich generell sehr positiv entwickelt. Der Aufbau des Studiums in Linz, was dort den Studierenden geboten wird, das halte ich für sehr gut, das gefällt mir sehr gut. Es trifft vielleicht am besten meine heutige Sicht zur Statistik die eben von der Vorstellung geprägt ist, dass es eine Einheit gibt zwischen den unterschiedlichen Bereichen und das diese gemeinsam behandelt werden sollten. In Wien sind etwas andere Schwerpunkte und die Entwicklung der Studierendenzahlen ist sehr gut. Auch an der WU hat sich die Statistik sehr gut entwickelt und der Bereich der Statistik ist jetzt dort stark vertreten und sehr aktiv. In Salzburg ist jetzt mit Kollegen Arne Bathke auch ein neuer Schwung in die Statistikausbildung gekommen. Dort ist die Statistik soweit ich weiß enger mit der Mathematikausbildung verbunden und es ist wichtig auch in diesem Bereich präsent zu sein.

Weil ihr vorhin gefragt habt, wie ich zur Statistik gekommen bin. Heute ist es ja ganz anders, weil Statistik in aller Munde ist und auch der Stellenmarkt ist viel besser. Früher war der Stellenmarkt doch sehr beschränkt auf den Verwaltungsbereich oder den Akademischen Bereich. Aber heute finden StatistikerInnen unter dem Titel Data Scientist in der Wirtschaft, bei Versicherungen oder bei Banken sehr gute Berufschancen. Es ist positiv, dass es so viele Möglichkeiten zur Spezialisierung gibt. Es gibt ja auch den riesigen Bereich der Bioinformatik, der in Wien ja durch Andreas Futschik vertreten war. Da sind ja völlig neue Anwendungsfelder gekommen, die nicht Statistik genannt werden sondern Bioinformatik, Machine Learning oder Data Mining. Die Statistik ist oft nicht so clever, dass sie ihren Beitrag richtig vermarkten kann, vielleicht auch, weil sie traditionell ein bisschen akribischer mit der Qualität der Information umgeht, dass muss man auch positiv sehen. Ich sehe es nicht negativ, dass man bei der Vermarktung etwas ins Hintertreffen kommt, aber generell sind die Möglichkeiten sehr gut.

Und der Master in Official Statistics, das ist eine neue Spezialisierung in einem interessanten Bereich, weil die internationalen Organisationen in ihren Analysen sehr viel statistische Modellierung verwenden. Man darf nicht übersehen, dass die PISA Studie ein hochkomplexes statistisches Modell ist mit einem psychometrischen Modell im Hintergrund. Die Wirklichkeit ist von der Idee, da fragen und zählen wir, wieviel richtig und wieviel falsch, meilenweit entfernt. Es steckt soviel statistische Methodik dahinter, dass zum Verständnis des Details eine fundierte Statistikausbildung notwendig ist.

**Matthias Templ:** *Darf ich etwas außerhalb des Protokolls fragen oder ist es zu brisant? Die Analyse der PISA Studie war in deinen und Erich Neuwirth's Händen und ich glaube auch Fritz Leisch hat etwas beigetragen, danach wurde der Auftrag aus wenig nachvollziehbaren Gründen anderweitig vergeben. Ich habe dann verwundert weniger professionelle Vorträge gehört, und z.B bekreidet dass selbst die Problematik der fehlenden Werte nicht berücksichtigt wurde. Ist dass ein Politikum geworden?*

**Wilfried Grossmann:** PISA ist leider ein Politikum geworden. Wir haben ja etwas auf Initiative von Erich Neuwirth gemacht. Das ist im Zusammenhang mit PISA 2003 zustande gekommen. Da hat es die große Aufregung gegeben hat, weil die Österreicher plötzlich so schlecht abgeschnitten haben. Da war besonders Erich Neuwirth aktiv und hat begonnen sich die Daten herunterzuholen und anzusehen. Dann haben wir begonnen zu diskutieren und haben uns die ganzen Unterlagen angesehen. Wir haben versucht das Modell zu verstehen und nachzuvollziehen. Da haben wir festgestellt, dass der Datenerhebungsaspekt und der psychometrische Aspekt nur ein Teil sind. Von der Psychologie hat ja auch Ivo Ponocny mitgearbeitet. Das Modell für die Skalierung der Items ist ein psychometrisches Modell, dass auf dem Rasch Modell beruht. Wenn man das statistisch sieht ist es in Wirklichkeit ein multivariates logistisches Regressionsmodell, also ein komplexes Generalized Linear Model. Und das Ganze wird dann noch eingebettet in einen Bayesianischen Kontext um die Effekte der Schulen und den individuellen Effekt zu berücksichtigen. Und dann wird aus diesem Modell mit Methoden der Analyse von fehlenden Werten eine Vorhersage der Scores mit Hilfe von multiplen Imputation gemacht. Da stecken also in Wirklichkeit sehr viele Modelle drin und die Schätzung der Varianzen ist dann noch ein eigenes Problem, aber da kennst du dich besser aus mit der Methode von Fay zur Varianzschätzung.

Wir haben also die Struktur des Modells analysiert und die einzelnen Bestandteile im Detail angesehen. Dabei haben wir festgestellt, dass bei der Berechnung der Scores, defacto sind die Scores die man erhält Posterior-Mittelwerte, eine gewisse Ungenauigkeit ist. Wir haben dann eine Verbesserung vorgeschlagen, die auch von der PISA akzeptiert wurde, da waren wir interessiert dran. Es ist nicht leicht, das Modell statistisch zu analysieren. Daher gibt es immer wieder Aufträge, Andreas Quatember hat jetzt wieder etwas gemacht, es gibt so viele Aspekte. Du kannst die Stichprobenkonzeption hinterfragen, man kann das Rechenmodell hinterfragen, man kann psychometrische Kalibrierung hinterfragen, es gibt so viele Komponenten die alle hinterfragbar sind. Ob das Ministerium das will, ist wiederum eine andere Frage, aber vom statistischen Standpunkt ist das das Interessante bei der ganzen Sache. Ich glaube PISA ist ein sehr gutes Beispiel, wo heute Survey Statistik hingeht. Es reicht nicht mehr aus, dass du einen Fragebogen ausgibst und dann zählst wieviel Angestellte ein Unternehmen hat.

**Werner Müller:** *Diese Schilderung scheint mir auch eine gute Illustration der Entwicklung der Statistik im Allgemeinen zu sein: dass immer mehr Layer kreiert werden. Auch in der Methodik und dass das dann an und für sich relativ undurchschaubar wird, für jemanden der das vielleicht analysieren muss. Es gibt ja schon Bereiche, wo man wieder dazu übergeht Metamodelle zu bauen, weil man die eigentlichen Modelle nicht mehr versteht. Hast du da Gedanken zu dieser Entwicklung?*

**Wilfried Grossmann:** Was du ansprichst, ist eine sehr schwierige Frage. Ich glaube man muss jedes Modell immer in seiner praktischen Anwendbarkeit hinterfragen. Man muss sich fragen, was gewinne ich durch das komplexe Modell oder wird durch das komplexe Modell in Wirklichkeit alles noch schwieriger zu interpretieren. Diese komplexen Modelle kommen oft dadurch zustande, dass wir immer mehr Informationen haben und wir wollen in einem Modell alle verfügbaren Informationen reinrechnen. Wir machen jetzt ein Projekt gemeinsam mit der Medizinuni, wo es um evidenzbasierte Medizin geht. Da ist die Hoffnung vieler Leute, dass durch ein komplettes Monitoring mehr Information zur Verfügung steht und dieses mehr an Information könnte dann zu besseren und richtigeren Entscheidungen führen. Wir haben eine kleine Diskussionsrunde gehabt, da hat Georg Heinze von der MedUni sehr klar und schön argumentiert, dass im Fall der klassischen Biometrie durch mehr Kovariaten und mehr Confounder die Lage für eine rationale Beurteilung immer schwieriger wird, weil da natürlich oft keine kontrollierten Experimente mehr möglich sind. Es gibt dann nur mehr ganz wenige Fälle und die Interaktion zwischen den einzelnen Confounder sind schwer abzuschätzen. Das Prinzip

eines Parsimonious Models ist auch heute noch wichtig. So gesehen ist Big Data nicht nur ein Segen sondern auch ein Fluch. Big Data ist ein Fluch, weil man nie weiß welche Information man für ein Modell selektieren soll. Zu all der Information kommt noch wahnsinnig viel zeitliche Information dazu. Es gibt ja nur mehr Zeitreihen aber die Zeitreihen von logfiles muss man sehr wohl hinterfragen und nachdenken darüber was man damit machen kann. Leichter wird es dadurch nicht. Die inhaltliche Beurteilung ist ein zentraler Punkt und da ist meiner Meinung nach die Statistik besser aufgestellt als die Informatik. Die Statistik ist es gewohnt für die Daten auch eine inhaltliche Beschreibung zu geben und inhaltlich darüber nachzudenken, da hat die Statistik einen Vorsprung gegenüber anderen Wissenschaften.

**Matthias Templ:** *Wie ich gelesen habe bist noch involviert bei EUROSTAT. (Anm.: jetzt nicht mehr.) Meiner Meinung wird EUROSTAT in letzter Zeit immer mehr zu einer Verwaltung, immer mehr Administration steht im Vordergrund. Da keine Statistiker mehr dort sitzen passiert es z.B. auch bei Big Data das Consultingfirmen beauftragt werden welche das Thema sehr puschen.*

**Wilfried Grossmann:** Big Data und EUROSTAT ist ein schwieriger Punkt. Wir haben voriges Jahr bei den Statistiktagen diese Sektion Big Data gehabt, du wolltest ja dass ich das organisiere. Es war für mich klar, man sucht Vortragende die Big Data aus Sicht des Data Mining repräsentieren. Aber dann habe ich mir gedacht, man sollte auch jemand einladen der Big Data aus der Sicht der Amtlichen Statistik präsentiert und habe mir angesehen, was es im Bereich von EUROSTAT in Big Data gibt. Die meisten Projekte werden nicht an Statistikinstitute oder an StatistikexpertInnen vergeben, sondern die gehen alle an Informatikfakultäten. Für die ist Big Data primär die Frage, wie kann ich die Daten manipulieren, wie kann ich Big Data managen. Zentral ist dass die Daten in ein System eingepackt werden, in Ontologien, xml-Schemata oder so etwas und das muss dann schnellsten verarbeitet werden. Inhaltlich interessieren die Daten fast gar nicht. Das ist sicherlich ein Problem bei der ganzen Sache und EUROSTSAT beschränkt sich hier zu stark auf die Rolle der Verwaltung. Es wird dann nicht mehr kritisch hinterfragt woher die Zahlen kommen, dieser Hype ist eine große Herausforderung, weil man muss das für jeden einzelnen Fall entscheiden. Das Problem ist etwas anders als bei Daten von physikalischen Messungen durch Satelliten oder Genomdatenbanken in der Bioinformatik. Die Probleme der Speicherung und Datenhaltung sind natürlich ähnlich, aber hier gibt es von der Substanzwissenschaft klarere Vorstellungen was gespeichert werden soll und wie diese Information weiter verarbeiten soll. Im Sozialbereich und in der Wirtschaft ist dies oft viel schwieriger, besonders bei Daten über den Internetverkehr. Auf Vorrat nur zu organisieren und das vom rein informatischen Standpunkt her zu betrachten löst zwar das Verwaltungsproblem aber nicht das inhaltliche Problem. Es ist meiner Meinung nach ein Problem, dass die Statistik in diesem Schema derzeit zu kurz kommt, aber wie schon gesagt es besteht die Hoffnung dass sich hier in nächster Zeit etwas bewegen wird.

**Werner Müller:** *Vielleicht sind wir eh schon an dem Punkt, wo wir uns ein bisschen über die Zukunft unterhalten können. Es ist nichts schwerer, als die die Zukunft vorherzusagen, aber hast du irgendwelche Perspektiven, wie du glaubst, wie sich die Statistik entwickeln wird?*

**Wilfried Grossmann:** Ich glaube die angewandte Statistik und die computationale Statistik werden in der Zukunft weiterhin sehr wichtig sein. Die Rolle der Statistik in der Gesellschaft ist derzeit sehr gut und Statistik gilt als eine sehr wichtige Wissenschaft. Ich glaube nicht, dass die Statistik diese Rolle in nächster Zeit verlieren wird. Statistik ist die Wissenschaft, die Information vom inhaltlichen Standpunkt her am besten verarbeiten kann, aber es wird wohl kaum eine Statistik ohne die Informatik geben. Viele Fragen der theoretischen Statistik sind heute schon im Zusammenhang mit Informatik

zu sehen. Die Handhabung von neuen und komplexeren Datenstrukturen ist dabei ganz wesentlich. Für mich ist das schon schwierig, das sage ich ganz offen, aber für die jungen StatistikerInnen ist das Manipulieren von Big Data und komplexer Anwendungssoftware kein Problem mehr, die beherrschen das ja, weil sie Digital Natives sind. Besonders für eine intelligente Darstellung der Daten wird man auch in Zukunft viel Statistik brauchen. Die Visualisierung wird eine zentrale Rolle spielen, es wird auch die Statistik eine große Rolle spielen. Ein Bereich wo noch viel zu machen ist, ist wie man mit Textdaten umgeht. Soweit ich das verstanden habe ist bei Text Mining die methodische Komponente noch nicht sehr entwickelt. Ich glaube sehr wohl, dass man hier noch eine Menge erreichen kann, nur hat man die Modelle noch nicht ganz im Griff. Nur die Häufigkeiten auszuzählen und diese dann als topic maps darzustellen ist sicher erst ein Anfang, da wird man noch viel Statistik machen können, vielleicht eine andere Art. Das ist noch am Stand wie die amtliche Statistik früher war, es geht nur um Häufigkeiten, aber a la longue wird man auch hier mit statistischen Methoden mehr zustande bringen.

Es gibt viele andere Bereiche die derzeit ein Randgebiet der Statistik sind, zum Beispiel Bildverarbeitung und automatische Übersetzung. Es ist interessant, dass viele Übersetzungsprogramme jetzt eher statistisch orientiert sind. Sie suchen nach statistischen Ähnlichkeitsstrukturen in den sprachlichen Konstrukten. Das wird nur von wenigen Statistikern gemacht und spielt in den Curricula nur am Rande eine Rolle. Aber das sind sicherlich Bereiche wo Statistik sehr wichtige Beiträge leisten kann. Auch dynamische Grafiken werden immer bedeutender und vielleicht sollte sich die Ausbildung für solche Bereiche öffnen. Das wird sicherlich in Zukunft ein wichtiger Markt sein und solche Anwendungen sollten von der Statistik nicht außer Acht gelassen werden.

In der Zeitschrift „Significance“ gab es kürzlich einen interessanten Artikel, wo „Dr. Fisher“, schreibt, irgendwie ist es erstaunlich, das die statistische Methodik, die ja eine Methodik für Datensätze der Größenordnung von 100 Beobachtungen ist und mit einem Taschenrechner betrieben werden kann, auch für die neuen Probleme angewendet werden kann. Die Grundlagen dieser Methodik funktionieren also auch bei großen Datensätzen und werden immer angewendet, es hat sich in diesem Sinne ja nicht viel geändert. Auch in der Bioinformatik verwendet man klassische Methoden. Also die Kunst wird wahrscheinlich wirklich sein, wie man Strukturierung in diese großen Datenmengen mit statistischen Methoden erzeugt. Vielleicht wird man dann zu dem Ergebnis kommen, dass weniger Information nützlicher ist, das ist der Erfolg der Statistik. Dieser Erfolg der Statistik beruht auf statistischen Prinzipien wie der Randomisierung, dem Likelihoodprinzip, oder statistischer Modellierung. Sie erlauben eine Verdichtung der Information die inhaltlich interpretiert werden kann. Und wie diese Prinzipien für diese neue Datenwelt vernünftig umgesetzt werden, das halte ich für eine interessante Frage.

**Werner Müller:** *Was interessiert dich noch, in welcher Rolle werden wir dich noch sehen?*

**Wilfried Grossmann:** Naja, ich habe ja in verschiedenen Bereichen Statistik betrieben, weil ich durch die anfangs genannte Vorstellung, dass Statistik eine Generalisten-Disziplin ist geprägt wurde und ich mir immer wieder gesagt habe, wenn etwas interessantes Neues kommt dann soll man sich damit auseinander setzen. Jetzt ist der Bereich Business Intelligence dazu gekommen, das interessiert mich im Moment. Da sehe ich auch, dass hier vieles von der Informatik gemacht wird, z.B. Modelle für work flows. Diese Prozessmodellierung verwendet hauptsächlich Modelle, die an der Logik orientiert sind, die nur wenige Variable berücksichtigen. Da bin ich auch der festen Überzeugung, dass, wenn man genau schaut, für die Analyse dieser Prozessdaten, statistische Modellierungen von Longitudinaldaten oder Markovprozesse vielfach besser geeignet sind, weil sie eine inhaltliche Komponente einbeziehen können. Ein Regressionsmodell ist immer ein inhaltliches Modell, es ist nicht nur ein Ablaufmodell oder ein logisches Modell und hat nicht nur eine Interpretation der Korrektheit sondern eine inhaltliche Interpretation. Die Idee ein inhaltliches Modells mit einem Modell für die Variabilität zu kombinieren,

das ist ein genuin statistischer Zugang und das ist wichtig und das interessiert mich auch.

**Matthias Templ:** *Eine letzte persönliche Frage: Du bist philosophisch sehr interessiert, habe ich beim gemeinsamen Zugfahren mitbekommen. Hast du die Vorstellung von einem breiten Bildungsbürgertum, dass ein Professor sich praktisch auch anderweitig ...*

**Wilfried Grossmann:** Das kommt von dieser Vorstellung die am Institut für Statistik vorherrschte. Die Vorstellung dass man nicht eine Spezialwissenschaft hat sondern dass man eine gewisse Generalistenhaltung hat, das finde ich wichtig.

Die Interviewer bedanken sich herzlich bei Gabriele Mack-Niederleitner für die Transkription.

## Affiliation:

Wilfried Grossmann
Faculty of Computer Science
University of Vienna
A-1010 Vienna, Austria
E-mail: wilfried.grossmann@univie.ac.at
URL: http://cs.univie.ac.at/ke-team/infpers/Wilfried_Grossmann/

Werner Müller
Department of Applied Statistics and Econometrics
Johannes Kepler Universität Linz
A-4040 Linz, Austria
E-mail: werner.mueller@jku.at
URL: http://www.jku.at/ifas

Matthias Templ
Vienna University of Technology &
Statistics Austria
A-1040 Vienna, Austria
E-mail: matthias.templ@gmail.com
URL: https://www.statistik.tuwien.ac.at/public/templ

# News and Announcements

AStA (Wirtschafts- und Sozialstatistisches Archiv) is a journal published by the German Statistical Society and Springer in German language (Editor: Ralf Münnich). Topics on economic and social statistics as well as topics related with the importance of statistics in the society are in main focus. Contributions from authors from Austria are highly welcome and close cooperation with the Austrian Journal of Statistics is in discussion and contemplated.

More information can be found at

http://www.springer.com/statistics/business2C+economics+26+finance/journal/11943

The Austrian Society of Statistics supports GeoMap, the first international Workshop on Practical Aspects of Geochemical Exploration and Mapping with Logratio Techniques, offers a practical forum of discussion for people concerned with the statistical treatment, modelling and interpolation of compositional data in geochemical applications, particularly focused on geochemical exploration and mapping. The workshop will mainly consist of a series of invited lectures on the problems of geochemical mapping, followed by discussions on each of the compositional topics raised. The goal of the workshop is to build teams to attach each of these specific topics and to provide enough time space for panel discussions.

GeoMap will touch a wide variety of problems and oppportunities that the log-ratio approach to compositional data analysis brings to regional geochemistry contexts. Particularly, the main specific feature of the workshop will be discussions for concrete problem solving and team building. Complementarily, also contributions of participants from the fields of interest of the workshop are warmly welcome, specially if they portray unsolved problems. All participants should be preliminary familiar with the log-ratio approach to compositional data analysis (otherwise, an introductory or an intermediate course may be provided).

More information can be found at

http://geomap.data-analysis.at/

ISSN 1026 - 597X

# Contents

**Page**

*Herwig FRIEDL and Matthias TEMPL:* Editorial . . . . . 91

*Andreas QUATEMBER:* The Finite Population Bootstrap – from the Maximum Likelihood to the Horvitz-Thompson Approach . . . . . 93

*Helga WAGNER and Regina TÜCHLER:* A Comparison of Bayesian Mixed Data Models for Austrian SILC Data . . . . . 103

*Faton MEROVCI, Ibrahim EELBATAL and Alaa AHMED:* The Transmuted Generalized Inverse Weibull Distribution . . . . . 119

*Manisha PAL and Montip TIENSUWAN:* The Beta Transmuted Weibull Distribution . . . . . 133

*Wilfried GROSSMANN, Werner MÜLLER and Matthias TEMPL:* Ein Interview mit Wilfried Grossmann . . . . . 151

News and Announcements . . . . . 163