Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 45, Number 3, 2016

ISSN: 1026597X, Vienna, Austria

Österreichische Zeitschrift für Statistik

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Austrian Journal of Statistics; Information and Instructions

GENERAL NOTES

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

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

All published manuscripts are available at

http://www.ajs.or.at

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

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

PEER REVIEW PROCESS

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

OPEN ACCESS POLICY

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

ONLINE SUBMISSIONS

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

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

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

AUTHOR GUIDELINES

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

SUBMISSION PREPARATION CHECKLIST

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

COPYRIGHT NOTICE

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

Austrian Journal of Statistics

Volume 45, Number 3, 2016

Editor-in-chief: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

Österreichische Zeitschrift für Statistik

Jahrgang 45, Heft 3, 2016

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Impressum

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

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

Contents

Editorial

This volume includes five scientific papers and a honorific speech for the Bruckmann Award of the Austrian Statistical Society.

The first paper contains some *explosive* results. The Austrian Microcensus is one of the most important and largest sample surveys in Austria. Furthermore, it is also used to estimate the unemployment rates in Austria. In this first contribution, a new weighting scheme is introduced, by using additional sources of information that was not available in the past. However, using this state-of-the-art weighting scheme let the unemployment rates increase slightly. The new weighting scheme has already been applied and was presented to the media approx. one year ago.

The next three papers have practical relevance especially in life time analysis. The second contribution illustrates a new test for two-parameter Weibull distributions, especially applicable in life time analysis when data are left-truncated.

The Weibull distribution with application to life time data is also the main topic of the third contribution. Applications and a simulation study shows advantages of the proposed log-logistic Weibull distribution. The paper also provides the R code – a best practice that should be considered also for further submissions to the Austrian Journal of Statistics.

Especially in life time analysis, the content of the fourth paper is important to reach a greater flexibility in modelling data in practice. A family of distributions is summarized and described in this contribution.

As so with the first article, also the last article of this issue has high impact on our society and could have major impact on jurisdiction. If the constitutional court of justice would also take statistical analysis and probabilities into account for their decisions, the decision on a repetition of the presidential election in 2016 in Austria must only be negative. The authors show that the (estimated) probability of that the irregular counting of those votes reversed the result is 0,000000000132.

The Austrian Statistical Society awards prizes on yearly basis at it's general assembly. The most famous award is the Bruckmann Award (2015 awarded the first time to a natural person). This award is dedicated to the kind of honorably work which aims to improve the status of statistics in public. Andreas Quatember is the second winner of this award. Johann Bacher's honorific speech is printed in this issue. It gives insights into Andreas Quatembers profession. Andreas Quatember reports in a humorous manner about nonsense in the media, have a look at http://www.jku.at/ifas/content/e101235.

Matthias Templ
(Editor-in-Chief)

Statistics Austria, Guglgasse 13, 1110 Vienna, Austria
E-mail: matthias.templ@gmail.com

Vienna, June 2016

Austrian Journal of Statistics
September 2016, Volume 45, 3–14.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i3.120

Weighting Procedure of the Austrian Microcensus using Administrative Data

Angelika Meraner
Statistics Austria

Daniela Gumprecht
Statistics Austria

Alexander Kowarik
Statistics Austria

Abstract

The Austrian microcensus is the biggest sample survey of the Austrian population, it is a regionally stratified cluster sample with a rotational pattern. The sampling fractions differ significantly between the regions, therefore the sample size of the regions is quite homogeneous. The primary sampling unit is the household, within each household all persons are surveyed. The design weights are the input for the calibration on population counts and household forecasts. It is performed by iterative proportional fitting. Until the third quarter of 2014 only demographic, regional and household information were used in the weighting procedure. From the fourth quarter 2014 onwards the weighting process was improved by adding an additional dimension to the calibration, namely a labour status generated from administrative data and available for the whole population. Apart from that some further minor changes were introduced. This paper describes the methodological and practical issues of the microcensus weighting process and the variance estimation applied from 2015 onwards. The new procedure was used for the first time for the forth quarter of 2014, published at the end of March 2015. At the same time, all previous microcensus surveys back to 2004 were reweighted according to the new approach.

Keywords: microcensus, weighting, iterative proportional fitting, variance, bootstrap.

1. Introduction

The Austrian microcensus (MC) is regulated by law[1] and carried out by Statistics Austria (STAT) since the 1970s. Over the years, the survey was modified on several occasions to better estimate reality. A principal reformation took place in 2004 when the MC was completely reorganised. Since that time the Austrian Labour Force Survey (LFS) is part of the Austrian microcensus, see Kytir and Stadler 2004. The LFS is an important data source for main economic and social indicators focusing on the labour market. It is based on definitions on employment and unemployment stated by the International Labour Organisation (ILO). Results are comparable with other countries.

For ten years the practices remained unchanged, until the utilisation of anonymised individual-related administrative data for official statistics was well developed and represented a great opportunity for quality improvement of the MC. The availability of administrative data and

[1]Erwerbs- und Wohnungsstatistikverordnung, BGBl. II Nr. 111/2010.

the possibility of micro data linkage, i.e. to connect them with other administrative data as well as survey data anonymously, so without any knowledge of names or social insurance numbers etc., and the opportunity for use in official statistics lead to a wide range of innovative methods and proceedings.

The preparation and implementation of the latest census 2011 as register based census also lead to a number of new possibilities and applications for survey data like the microcensus. In this context, the new weighting of the MC was developed. Linking MC survey data with administrative data on the labour status showed a small but non-negligible bias in the MC data concerning the labour status. Employed persons were slightly over-covered and by contrast unemployed and persons out of labour force were under-covered. The calibration (used until reference period q3_2014, see Haslinger and Kytir 2006) reduced this bias, but not sufficiently, i.e. employed persons were still overestimated and non-employed were underestimated. As a consequence, administrative data on the labour status were included in the weighting process to correct this bias.

Beside the changes of the calibration procedure there was another reason for a revision. With the results of the latest census, which took place in 2011, new population numbers and household numbers were available to which the weighting procedure had to be adapted retroactively. To avoid breaks in the time series simply based on methodological changes and to allow for time series analysis, the whole MC back to the first quarter of 2004 was reweighted with the new procedure and the results were revised. The reweighting lead to a decrease of the estimated number of persons employed and to an increase of the estimated number of persons unemployed and out of labour force, e.g. the estimated yearly unemployment rate of 2013 (persons aged between 15 to 74) increased from 4.9% to 5.4% due to the change of the weighting, whereas the estimated employment rate of persons aged between 15 and 64 decreased from 72.3% to 71.4%. These changes are along the same lines for men and women, although the differences are usually bigger for men. Regarding the whole time series from 2004 onwards the reweighting lead to level shifts but to no change of the trends and seasonal patterns.

The latest renewals and changes of the Austrian microcensus weighting procedure are also described in a working paper (German-language only) published at the website of Statistics Austria, see Meraner, Gumprecht, and Kowarik 2015.

In the following, this paper presents the sample of the Austrian microcensus, the weighting and the error estimation. Section 2 describes design and sampling frame of the Austrian microcensus as well as the problems with non-response which is on a low level but nevertheless lead to biased results which is shown in a non-response analysis using administrative data (see 2.3.1). Section 3 describes the new weighting procedure including among other things also calibration specifications and the steps of the iterative proportional fitting process. Section 4 deals with the calculation of the standard errors and confidence intervals via a bootstrap procedure as well as an approximation of the errors. As usual, the last section 5 gives a summary of the paper as well as an outlook to further topics of research.

2. Sample design and data collection

The MC is a stratified random sample of private households or rather addresses by NUTS-2 region, therefore results on NUTS-2 level are of (controlled) high quality. Due to organisational reasons the sample selection itself is done at the level of the political districts and within districts with a small population density the sampling fraction is doubled. All other districts have the same sampling fraction as its NUTS-2 region. The MC sample contains approximately 20,000 households per quarter, these are about 44,000 persons or 0.5% of the whole Austrian population.

2.1. Sampling frame

The sample frame is the Austrian central population register. All non-private households, households which were part of the MC within the last ten years (a household can participate only once in ten years), and households being no main residence of at least one person are excluded. About four to six months before the first interview takes place, the households are selected from a current copy of the central population register. Due to this time lag between sampling and interview it may happen that situations change in the meanwhile, e.g., persons move out and others move in, some die and others are born etc. Regardless of the situation in the central population register at the time of the sampling, also called "register reality", the persons who are actually living at the selected household, i.e. at this address, at the time of the interview are subjects of the MC. If e.g., the address does not exist anymore it drops out of the MC sample.

2.2. Sample design

Once a household is selected, it stays in the sample for five consecutive quarters and is questioned in each of these five quarter, starting as wave 1 in the very first quarter, becoming wave 2 in the next quarter and so on. The MC is a rotating sample, each quarter one fifth of the total sample rotates in and out. Each quarter all persons who are currently living in the household are interviewed. The first interview is a computer assisted personal interview (CAPI), the following four interviews are usually computer assisted telephone interviews (CATI). Each household is assigned to a reference week and most questions refer to that week. Moreover, the households are evenly distributed across all the reference weeks of a quarter. The interviews should be done shortly after the reference week and not later than five (in summer six) weeks after the reference week. There is a legal obligation to participate in the microcensus, but if a person is not able or willing to do the interview by herself, any other adult person living in the same household is allowed to answer by proxy.

This and more information about the MC sample design and sampling frame can be found in Haslinger and Kytir 2006.

2.3. Non-response

In general there is a very low level of non-response, mainly due to the legal obligation to participate, nevertheless there is a certain degree of non-response due to different reasons, like denial, non-reachability, language problems, wrong addresses, etc. The degree of non-response is usually given by the non-response rate or - it's counterpart - the response rate, which both can be calculated in different ways. In Austria it is the share of non-neutral non-response in the gross-sample, which does not include neutral non-response cases (the whole sample is called gross-gross-sample; the sample without neutral non-response and non-neutral non-response is called net-sample). Thus, the non-response share depends on the classification of cases as neutral and non-neutral non-response. Neutral non-response cases are e.g., if a household is absent during the quarter, if household-members are incapable of being interviewed, if it is no private household, and the like; whereas non-neutral non-response is e.g., refusal. As the MC is a sample of households (not persons) non-response statistics also refer to households. In the first quarter 2013 the gross-gross-sample was 22,499 households, 1,188 or 5.3% of them were classified as neutral non-response, this lead to a gross-sample of 21,311 households and 19,931 or 93.5% of these were response households. The non-response rate is calculated by the non-response cases of the sample divided by the whole sample cases. The size depends on the definition of non-response cases as well as whole sample cases. Yet, even if the share of non-response is quite small, it might lead to problems if non-response is not at random, i.e. if there are structural differences between response and non-response cases, a bias can be introduced.

2.3.1. Non-response analysis using administrative data

In labour market statistics, the persons are the main subjects of interest, therefore the non-response analysis has to be done on the level of persons instead of households. Structural differences in the labour status between non-response and response persons can lead to a biased estimation of certain labour market groups. On the level of persons a non-response analysis is more complex than for households, for example, there might be persons living in "response" households who are non-respondents after all, because their existence is not disclosed by the interviewed household member(s). If a whole household does not answer, it is not easily possible to know the number of persons actually living there, therefore not even the number of non-response persons is easy to determine. Beside the pure number of non-response persons also their labour status is of interest, so that the assumption of a bias concerning the labour status in the MC net sample can be checked. A special non-response analysis was done for reference period 2012 using data from the central population register, the central social security register and the labour market service (see Gumprecht and Oismüller 2013), this information from administrative data was used as a proxy for the missing MC information. Correlation of ILO- and administrative labour status is very high for persons employed, more than 90% of persons employed according to ILO definition are also employed according to administrative data. However, correlation between ILO- and administrative definitions of unemployment is considerably lower with about 70% of persons unemployed according to ILO definition also being unemployed according to administrative data and about 42% of registered unemployed persons also being unemployed according to ILO definitions. Nevertheless, correlation between the labour status according to these two definitions is quite high and administrative data are good proxies for missing values in MC data. Non-response analysis showed that persons employed according to administrative information tend to be overrepresented in the MC sample whereas unemployed and persons out of labour force are rather underrepresented, and the calibration procedure used at that time (until the third quarter 2014) could not entirely correct this bias. The findings of this non-response analysis were the motivation to change the weighting of the Austrian microcensus. Administrative employment status is a good candidate for an additional weighting specification, because it is known in the sample as well as in the whole population and it is highly correlated with the main variable of interest, the ILO labour status.

3. Weighting

Drawing a random sample means reducing the target population to a subset that should represent this population in an unbiased way, therefore, computing population statistics requires a reversal of the reduction which is done by projection. The data collected from the sample is used to estimate the unknown population parameters of interest (totals, means, medians, ratios, ...) with the respective weighted estimator. The inverse of the selection probabilities is the so-called design weight, it is the most basic weighting scheme. The Austrian microcensus has different selection probabilities for each stratum (NUTS-2 region), so each region has to be projected separately. In addition to that, known population totals, such as age and gender, retrieved from sources like the statistical population register, are used to calibrate the weights. The calibration is performed with the iterative proportional fitting procedure, which basically computes multipliers known as calibration factors that adjust the sampling weights to make the population estimates agree with the known totals. Theoretically, every characteristic known for the units in the sample and for the whole population can be used as a calibration variable, but only variables correlated with target variables enhance precision. Practically of course there are some restrictions, e.g., the sample size.

The following description of the calibration specifications, computation of base weights and the steps within the iterative proportional fitting procedure is strongly based on Haslinger and Kytir 2006 with the main difference being the usage of an additional calibration specification,

the administrative employment status.

3.1. Calibration specifications

The calibration specifications for the iterative proportional fitting procedure are the following:

- N_{rga} ... Total number of persons in private households in NUTS-2 region r (=1,...,9), of gender g (=1,2) and in age class a (1=0-2 years, 2=3-5 years, 3=6-9, ... 5-year classes ..., 18=80-84, 19=85+).
- N_{rn} ... Total number of persons in private households in NUTS-2 region r (=1,...,9) with nationality n (=1,...,6) comprising the groups "Austria", "EU-15 (w.o. Austria)", "EU 2004+ (joined the EU between 2004 and 2014)", "European non-EU countries", "Turkey", "Other".
- N_{rge} ... Total number of persons in private households in NUTS-2 region r (=1,...,9), of gender g (=1,2) and with administrative employment status e (=1,...,5) which consists of the groups "Employee standard", "Employee non-standard", "Self-Employed", "Unemployed" and "Inactive".
- M_{rh} ... Total number of private households in NUTS-2 region r (=1,...,9) of household size h (=1,...,5) with values 1 to 5+.

Information about variables gender, age, and nationality come from the statistical population register "POPREG", it includes all persons living in Austria at the beginning of a quarter. Information about the labour status from administrative data is available from the Austrian central social security register and the labour market service, see Section 3.1.3. The number of private households with h members stem from the household projection of STAT, see Section 3.1.2 which is based on the register based census for the census years (2011, 2021, etc.) and the register-based labour market statistics for the years in between.

3.1.1. Private and institutional households

To compute the number of persons in private households it is essential to segregate institutional households like prisons, care homes, residential schools etc. that are also included in the POPREG. Persons in institutional households are removed via so-called institutional rates (share of persons in institutional households) for all relevant combinations of the characteristics used as weighting specifications. These rates stem from a special survey done by STAT, called "institutional survey", which is available for each year for reference date October 31^{st} (from 2011 onwards), but with a time lag of up to two years. Institutional rates are held constant until new rates are available.

3.1.2. Household projection of STAT

In addition to the weighting specifications concerning the number of persons living in private households, the number of private households (by size and NUTS-2 region) is used for calibration. Since the real numbers are not known for all quarters, results from the household projection of STAT are used instead. For the fourth quarter of each year, real data is available, either from the census or the register-based labour market statistics, however with a time lag of about two years. For the quarters in between and thereafter, a two-step projection following Ediev 2007 is used.

3.1.3. Labour status from administrative data

Similar to the use in the non-response analysis, data from the Austrian central social security register and the labour market service as well as the statistical population register POPREG

from STAT is used to generate an administrative labour status without any survey information. It can be calculated on individual level for approximately 94% of all MC respondents, as well as for the whole population, i.e., all persons living in private households in Austria at the beginning of the MC reference quarter, for reference period end of a month. For about 6% of the MC-persons, administrative data cannot be linked to survey data because no area specific personal identifier (bPK) is available. For them, the administrative employment status is imputed using random hot-deck within a domain, which is the default procedure for most of the MC variables. In the imputation process ILO labour status, region and gender are used as explanatory variables, i.e. the correlation between administrative employment status and ILO labour status (see Section 2.3.1) can be used again.

The variable "administrative employment status" has the following five values:

- Employee standard: Persons with dependent employment concerning to the social security register, e.g., white- and blue-collar workers, civil servants.
- Employee non-standard: Persons with non-standard dependent employment concerning to the social security register, e.g., holder of non-standard contract, marginal part-timers, persons in parental leave, etc.
- Self-Employed: Self-employed persons concerning to the social security register, e.g., freelancers, self-employed and family workers in agriculture.
- Unemployed: Persons unemployed concerning to labour market service, e.g., job seeking persons, persons in training measures.
- Inactive: Persons living in Austria and being neither employed nor unemployed.

Data from the social security register and the labour market service can contain several cases per person, therefore, the dominant case has to be selected for every person, e.g., employment always is prevailing compared to unemployment and inactivity. To guarantee a complete coverage of the reference quarter, this is performed for the three end-of-month administrative data deliveries pertaining to the reference quarter as well as the end-of-month delivery of the month preceding the reference quarter, but only for persons corresponding to the population in private households at the beginning of the respective reference quarter. A weighted mean of these four final monthly results m_{t-1}, m_t, m_{t+1} and m_{t+2} with the first month t of a quarter gives the quarterly results N_{rge} as defined above which are used as weighting specifications:

$$N_{rge} = \frac{1}{3}\left(\frac{m_{t-1}+m_t}{2} + \frac{m_t+m_{t+1}}{2} + \frac{m_{t+1}+m_{t+2}}{2}\right). \tag{1}$$

3.2. Base weights

The final sampling weights are computed in an iterative process starting with the base weights which are then calibrated against the population totals defined in Section 3.1.

The base weights are basically the inverse of the selection probabilities and are determined for every person and household as

$$\frac{M_r}{m_r} \tag{2}$$

where M_r is the total number of inhabited addresses in NUTS-2 region r and m_r is the number of addresses in the net sample of NUTS-2 region r. As mentioned in Section 2, the sample selection itself is done at the level of the political districts and the sampling fraction is doubled within districts with small population density. For NUTS-2 regions containing districts like these, the base weights for persons and households corresponding to districts with "normal" sampling fraction are computed as

$$\frac{M_r}{m_{r1} + m_{r2}/2} \tag{3}$$

where m_{r1} is the number of sampled addresses belonging to districts with "normal" sampling fraction while m_{r2} denotes the number of those with double sampling fraction. The weights for persons and households with double sampling fraction are calculated as half the quotient (3).

3.3. Iterative proportional fitting

Projection with the base weights computed as the quotients (2)-(3) yields population statistics that differ from the known population figures according to age, gender, nationality and administrative labour status. Many microcensus variables used for analyses are highly correlated with these characteristics. As mentioned above, we can use the known population totals to adjust the base weights accordingly. The iterative proportional fitting procedure applied for this purpose consists of 5 iteration steps and is explained in the following:

For every person i in the sample the base weights which shall be referred to as w_i^0 and the values of the variables region r (=1,...,9), gender g (=1,2), age group a (=1,...,19), nationality n (=1,...,6) and administrative employment status e (=1,...,5) are needed. The running index k is set to $k = 0$ at the outset. If the constraint after step 5 is satisfied, k is raised by 1 and the procedure starts again with step 1.

3.3.1. Steps 1-3

The base weight w_i^0 of each person is modified by multiplication with a factor so that the projected distribution of the population matches the respective calibration specification $N_v \in \{N_{rga}, N_{rn}, N_{rge}\}$ (see Section 3.1) in every iteration step. That is, the calibration specification N_v with $v \in \{rga, rn, rge\}$ varies depending on the iteration step t:

- if $t = 1$, $N_v = N_{rga}$,
- if $t = 2$, $N_v = N_{rn}$,
- if $t = 3$, $N_v = N_{rge}$.

In every step t, the weight calibrated against N_v is computed as

$$w_i^{5k+t} = w_i^{5k+t-1} \frac{N_{v_i}}{\sum_l w_l^{5k}} \qquad \forall i. \tag{4}$$

For $v_i = r_i g_i a_i$ the summation in the equation expands over all observations l with the same values of the characteristics r,g and a as observation i. This applies to $v_i = r_i n_i$ and $v_i = r_i g_i e_i$ in analogue form.

Weights w_i^{5k+t} outside of $\frac{w_i^0}{4} \leq w_i^{5k+t} \leq 4w_i^0$ are recoded to the nearest of these two boundaries. The constraints are based on expert opinion and should in general restrict the variance which has a positive effect on the sampling error. Restricting the change of the base weights is also general practice in other countries, i.e. it is a common form of weight trimming where very large/small weights are trimmed back to an upper/lower boundary (see e.g., Potter 1990, 1993) in order to reduce the variance but with the possible side effect of introducing a bias into the estimates.

3.3.2. Step 4

Due to modifications of the sampling weights in iteration steps 1-3, the weights of the persons in a household are no longer identical unless the household members all show the same values for the characteristics age group, gender, nationality and administrative labour status. Such heterogeneous weights lead to inconsistencies between results projected with household and person weights if e.g., the weight of the household reference person is used as the household weight. To avoid this, every person in a household is assigned the mean of the person weights

corresponding to the household. For every person i who is a member of household j with N_j houshold members l, it holds that:

$$w_i^{5k+4} = \frac{\sum_{l=1}^{N_j} w_l^{5k+3}}{N_j} \qquad \forall i. \tag{5}$$

As a result, the adaptation to the population structure performed in steps 1-3 may be lost again.

3.3.3. Step 5

The weights from step 4 are modified to make the projected distribution of household size by region approximate the calibration specification M_{rh}. We do not aim at an exact match since the specifications themselves are subject to uncertainty. The uncertainty p_h for the number of households of size h per region is assumed to be 0.005 for $h = 1, \ldots, 4$ and 0.2 for $h = 5$. The adjusted weights are computed as

$$w_i^{5k+5} = \begin{cases} w_i^{5k+4} \frac{M_{r_i h_i}}{\sum_l w_l^{5k+4}} & \text{if} \quad \sum_l w_l^{5k+4} \notin ((1-p_h)M_{r_i h_i}, (1+p_h)M_{r_i h_i}) \\ w_i^{5k+4} & \text{otherwise} \end{cases} \tag{6}$$

where the summation expands over all households l with the same values of the characteristics r and h as observation i. Again, the new weights w_i^{5k+5} should not exceed 4 times the base weights w_i^0 nor should they fall below one quarter of w_i^0, i.e. if $\frac{w_i^{5k+5}}{w_i^0} > 4$, we set $w_i^{5k+5} = 4w_i^0$ and if $\frac{w_i^{5k+5}}{w_i^0} < \frac{1}{4}$ we set $w_i^{5k+5} = \frac{w_i^0}{4}$.

3.3.4. Check

Finally, we check whether the deviation of the projected results from the known totals is greater than 0.01% in any cell:

$$\max_v \left| \frac{\sum_l w_l^{5k+5} - N_v}{N_v} \right| > 0.0001 \qquad \text{for } v \in \{rga, rn, rge\}\,. \tag{7}$$

For $v = rga$ the summation in the equation expands over all observations l sharing the same values of the characteristics r,g and a. This applies to $v = rn$ and $v = rge$ in analogue form. If the maximum deviation is exceeding this limit, k is raised by 1 and the procedure continues at step 1 with the weights computed in step 5 as initial weights w_i^{5k}. It should be noted, that the constraints restricting the variance of the weights in steps 1-3 and 5 always refer to the "original" base weights w_i^0.

Convergence is reached if the maximum deviation falls below the threshold, in that case, w_i^{5k+5} are the final calibrated weights w_i. Usually, the method converges after approximately 130 iteration steps.

3.4. Monthly weights and yearly results

Monthly weights are computed almost the same way as quarterly weights with the difference of using the total number N_n of persons in private households with nationality n instead of the total number N_{rn} of persons in private households with nationality n in NUTS-2 region r as a calibration specification (see Section 3.1). Also, the specification N_{rge} is not computed as the mean of four months (see Section 3.1.3) but as the mean of two months, i.e. the end-of-month values of the reference month as well as the end-of-month values of the previous month.

For yearly data, the quarterly data sets corresponding to a year are aggregated and the quarterly weights are divided by four.

4. Error estimation

4.1. Bootstrapping

4.1.1. The naïve approach

Standard errors and confidence intervals are estimated with the help of a bootstrap procedure (see Efron 1979).

In addition to the calibrated sampling weight, each person and household in the sample is assigned a certain number c of bootstrap weights. Currently $c = 500$ bootstrap draws seem sufficient to estimate the sampling error with high quality, however, an increase to $c = 1000$ is under consideration.

Bootstrap weights are computed based on bootstrap samples. A bootstrap sample is a sample with replacement of size m taken from the original sample of the same size. This means that every original sample unit can appear 0- to m times in the bootstrap sample. In our case, the sampling units are households. The frequency of occurrence f_i^j of observation i in bootstrap sample j, where $j = 1, \ldots, c$, multiplied with the calibrated sampling weights w_i of the original sample (see Section 3.3.3) renders the uncalibrated bootstrap weights

$$b_i^{0,j} = f_i^j w_i \tag{8}$$

which are identical for every person in a household as the bootstrap samples are drawn at household level. These weights are then calibrated using the iterative proportional fitting procedure from Section 3.3 but with initial weights $b_i^{0,j}$ instead of the base weights w_i^0, returning the adapted bootstrap weights b_i^j.

4.1.2. The rescaling bootstrap

Since the microcensus is a sample without replacement drawn from a finite population, the "naïve" bootstrap procedure described above can not be applied in exactly this form. Instead, the "rescaled" bootstrap procedure introduced by Rao and Wu (1988) with the adjustment of using rescaled weights instead of rescaled survey data values (see Rao, Wu, and Yue 1992) is used with the additional modification of selecting bootstrap samples without replacement (see Chipperfield and Preston 2007; Preston 2009), also incorporating the stratification by region r (see Section 3.2). To be more specific, instead of drawing c bootstrap samples with replacement of the same size m_r as the original sample, subsamples without replacement of size $m_r^j = \lfloor m_r/2 \rfloor$ are drawn.

The uncalibrated bootstrap weights for every observation i belonging to region r are then computed for $j = 1, \ldots, c$ as

$$b_i^{0,j} = w_i \left(1 - \lambda_r + \lambda_r \frac{m_r}{m_r^j} \delta_{r_i}\right) = w_i f_i^j \qquad \forall i \in m_r \tag{9}$$

with

$$\lambda_r = \sqrt{\frac{m_r^j \left(1 - \frac{m_r}{M_r}\right)}{m_r - m_r^j}} \tag{10}$$

where w_i are the calibrated sampling weights of the original sample and $\delta_{r_i} = 1$ if observation i is selected in the subsample m_r^j and 0 otherwise. The $b_i^{0,j}$ are then calibrated to render the adapted bootstrap weights b_i^j as mentionend above in Section 4.1.1.

To determine the standard error and the approximate 95% confidence interval of the population estimate $\hat{\theta}$ of some population parameter θ we use the c bootstrap weights b_i^j to compute c population estimates $\hat{\theta}^j$. The estimated standard error of the population estimate $\hat{\theta}$ is then the standard deviation of these c estimates $\hat{\theta}^j$ with mean $\bar{\hat{\theta}}$:

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\frac{\sum_j (\hat{\theta}^j - \bar{\hat{\theta}})^2}{c-1}}. \tag{11}$$

The 2.5% and 97.5% quantiles of the c bootstrap replicates $\hat{\theta}^j$ correspond to the lower and upper boundaries of the approximate 95% confidence interval (see Efron 1981).

4.1.3. Rotation

The Austrian microcensus is a rotating quarterly sample survey (see Section 2.2) where one fifth of the sample is replaced by a new random sample every quarter. The bootstrap procedure takes this into account by drawing bootstrap samples and computing the occurrence frequencies f_i^j for this new sample only while retaining the f_i^j from the previous quarter for the remaining four fifths of the sample.

This way, the overlap of two microcensus samples is considered in cases where the objective is to the estimate standard error and confidence interval of change.

4.2. Error approximation

In cases where no bootstrap results are provided by STAT, users can resort to tables containing rough estimates of the relative sampling errors.

The relative sampling error, i.e. the sampling error divided by the estimate, can be expressed for frequency counts of persons and households by means of the binomial approach (see e.g., Cochran 1977) where the variance of $x = Np$ with sample proportion p and population size N is estimated as

$$\mathrm{Var}(x) = \frac{N(N-n)}{n} p(1-p) \tag{12}$$

Strictly speaking, the frequency counts resulting from a sample without replacement are hypergeometrically distributed, however due to the small sampling fraction, the use of the binomial distribution as an approximation is justified.

Calibration (see Section 3.3) is ignored by this approximation, only the sampling fractions corresponding to the NUTS-2 regions are considered. For a population frequency count x of persons or households the relative sampling error at confidence level 95%, i.e. with the factor 2 as an approximation of the respective quantile of the normal distribution, is computed as

$$2 * 100 \sqrt{\frac{(N_r - n_r)(N_r - x)}{N_r n_r x}} \tag{13}$$

at NUTS-2 level r and as

$$2 * 100 \sqrt{\sum_r \frac{(N_r - n_r)(N - x) N_r}{n_r N^2 x}} \tag{14}$$

for frequency counts x corresponding to the whole population where the contributions of the NUTS-2 regions to the value x are assumed to be proportional to the size of the respective regions. N is the total population size while N_r and n_r are population and sample size of NUTS-2 region r. If x refers to households, (13) and (14) are computed with M, M_r and m_r instead (see Section 3.2).

This simple calculation of the sample error is useful for a rough estimation. It gives an idea of the magnitude of the error, however it does not give the exact values. Neither calibration which leads to a reduction of the error nor sample clusters which lead to an increase of the error are taken into account. For both effects no general rule can be given, e.g. sample clusters

lead to bigger errors but the magnitude depends very much on the variable, e.g. errors of variables with a very low cluster correlation are nearly unaffected by cluster sampling. In this way, differences between the bootstrap-errors and the simple approximation vary from estimate to estimate. As an example of the differences between the error approximation and the bootstrap errors, the estimated number of persons unemployed at the age of 15 to 74 in year 2014 are used. This estimated number is 244,883 and the estimated relative standard error calculated on the bootstrap weights is 1.63% whereas the error based on the simple approximation is estimated to be 5.2%. This shows that for the number of unemployed the positive effect of calibration exceeds the negative effect of the cluster sampling.

5. Summary

One decade after the last substantial renewal of the Austrian microcensus the weighting procedure was again revised. Although the method itself is unchanged, i.e. calibration of weights via iterative proportional fitting, the external information used for calibration changed significantly. Some were updated and a completely new specification was launched, the employment status from administrative data. This led to a big improvement of the quality of the MC due to a reduction of the non-response bias as well as the standard error of variables of great interest.

For the near future there are two important tasks to implement: Firstly, the aim is to enable users to compute the exact standard errors and confidence intervals on their own, i.e. to provide the required tools to make use of the bootstrap weights and eliminate user's dependency on the error approximation.

Secondly, the focus on the longitudinal dimension of the MC, to provide information on labour market dynamics and flows, is of increasing importance. Therefore, the weighting of longitudinal data, with its related issues such as panel attrition, consistency between cross sectional and longitudinal results, plausibility along the time line, is an area of constant improvement.

Acknowledgement

The authors would like to thank the editor and the anonymous reviewer for their careful work and their helpful and constructive comments.

References

Chipperfield J, Preston J (2007). "Efficient Bootstrap for Business Surveys." *Survey Methodology*, **33**(2), 167–172.

Cochran W (1977). *Sampling Techniques, 3rd Edition*. John Wiley.

Ediev D (2007). "On Projecting the Distribution of Private Households by Size." *Vienna Institute of Demography of Austrian Academy of Sciences. Working Paper*, **4**.

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

Efron B (1981). "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods." *Biometrika*, **68**(3), 589–599.

Gumprecht D, Oismüller A (2013). "Non-Response im Mikrozensus." *Statistische Nachrichten*, **11**, 1046–1061 [in German].

Haslinger A, Kytir J (2006). "Stichprobendesign, Stichprobenziehung und Hochrechnung des Mikrozensus ab 2004." *Statistische Nachrichten*, **6**, 510–519 [in German].

Kytir J, Stadler B (2004). "Die kontinuierliche Arbeitskräfteerhebung im Rahmen des neuen Mikrozensus." *Statistische Nachrichten*, **6**, 511–520 [in German].

Meraner A, Gumprecht D, Kowarik A (2015). "Die neue Hochrechnung des Mikrozensus - Methodenbeschreibung." *Technical report*, Statistik Austria [in German]. URL http://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/arbeitsmarkt/index.html.

Potter FJ (1990). "A Study of Procedures to Identify and Trim Extreme Sampling Weights." In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, volume 225230.

Potter FJ (1993). "The Effect of Weight Trimming on Nonlinear Survey Estimates." In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, volume 758763.

Preston J (2009). "Rescaled Bootstrap for Stratified Multistage Sampling." *Survey Methodology*, **35**(2), 227–234.

Rao J, Wu C, Yue K (1992). "Some Recent Work on Resampling Methods for Complex Surveys." *Survey methodology*, **18**(2), 209–217.

Rao JN, Wu C (1988). "Resampling Inference with Complex Survey Data." *Journal of the american statistical association*, **83**(401), 231–241.

Affiliation:

Angelika Meraner, Alexander Kowarik
Methods Unit
Statistics Austria
Guglgasse 13, 1110 Vienna, Austria
E-mail: angelika.meraner@statistik.gv.at, alexander.kowarik@statistik.gv.at
URL: http://www.statistik.at

Daniela Gumprecht
Department of Social Statistics
Statistics Austria
Guglgasse 13, 1110 Vienna, Austria
E-mail: daniela.gumprecht@statistik.gv.at
URL: http://www.statistik.at

Austrian Journal of Statistics
September 2016, Volume 45, 15–42.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i3.106

Goodness-of-fit Testing for Left-truncated Two-parameter Weibull Distributions with Known Truncation Point

Ayşe Kızılersü
Department of Physics
University of Adelaide
5005, Adelaide, Australia

Markus Kreer
CAMPUSERVICE GmbH
Servicegesellschaft der
Johann Wolfgang
Goethe-Universität Frankfurt
60323, Frankfurt am Main, Germany

Anthony W. Thomas
Department of Physics
University of Adelaide
5005, Adelaide, Australia

Abstract

The left-truncated Weibull distribution is used in life-time analysis, it has many applications ranging from financial market analysis and insurance claims to the earthquake inter-arrival times. We present a comprehensive analysis of the left-truncated Weibull distribution when the shape, scale or both parameters are unknown and they are determined from the data using the maximum likelihood estimator. We demonstrate that if both the Weibull parameters are unknown then there are sets of sample configurations, with measure greater than zero, for which the maximum likelihood equations do not possess non trivial solutions. The modified critical values of the goodness-of-fit test from the Kolmogorov-Smirnov test statistic when the parameters are unknown are obtained from a quantile analysis. We find that the critical values depend on sample size and truncation level, but not on the actual Weibull parameters. Confirming this behavior, we present a complementary analysis using the Brownian bridge approach as an asymptotic limit of the Kolmogorov-Smirnov statistics and find that both approaches are in good agreement. A power testing is performed for our Kolmogorov-Smirnov goodness-of-fit test and the issues related to the left-truncated data are discussed. We conclude the paper by showing the importance of left-truncated Weibull distribution hypothesis testing on the duration times of failed marriages in the US, worldwide terrorist attacks, waiting times between stock market orders, and time intervals of radioactive decay.

Keywords: maximum likelihood estimation, Kolmogorov-Smirnov goodness-of-fit test, left-truncated data, Monte Carlo simulations, asymptotic analysis, quantiles, Brownian bridge.

1. Introduction and preliminaries

The Weibull distribution with scale and shape parameters, $\alpha > 0$ and $\beta > 0$ respectively, is widely used in areas such as statistics, engineering, finance, insurance and biology (e.g. Weibull (1951), Balakrishnan and Cohen (1991), Rinne (2009)), mainly in the context of life-time analysis (survival analysis in medical studies and reliability analysis in engineer-

ing). In practical applications, very often truncated statistical distributions must be used (see also Nadarajah and Kotz (2006)) these truncated statistical distributions arise when a random variable τ follows a known distributional model, except that a portion of the sample space cannot be observed or is removed (for example in radioactive decay phenomena a Geiger-Müller counter does not permit detection of decays events within its dead time). An independent identically distributed (i.i.d.) left-truncated data set $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_n)$ of sample size n has the property that $\tau_L < \tau_i, i = 1, ..., n$ for a given non-negative parameter τ_L, the truncation point (Kendall and Stuart (1979), pp. 551, section 32.15). The left-truncated cumulative Weibull distribution function (cdf) is given by Wingo (1989)

$$F(\tau|\alpha, \beta, \tau_L) = 1 - \exp\left[\left(\frac{\tau_L}{\alpha}\right)^{\beta} - \left(\frac{\tau}{\alpha}\right)^{\beta}\right] \quad \text{for} \quad \tau > \tau_L \tag{1}$$

and the left-truncated probability density function (pdf) is

$$f(\tau|\alpha, \beta, \tau_L) = \frac{\beta}{\alpha}\left(\frac{\tau}{\alpha}\right)^{\beta-1} \exp\left[\left(\frac{\tau_L}{\alpha}\right)^{\beta} - \left(\frac{\tau}{\alpha}\right)^{\beta}\right] \quad \text{for} \quad \tau > \tau_L\,. \tag{2}$$

Putting $\tau_L = 0$ in Equation (1) and Equation (2), cdf and pdf of the complete Weibull distribution will be recovered, respectively. Throughout this paper we use the term complete Weibull distribution to refer to the untruncated Weibull distribution and in our investigation we assume that the truncation point τ_L is known or can be set. The literature on data analysis tends to focus either on complete or censored data, with much less attention paid to truncated data, moreover truncation formally defined as in Kendall and Stuart (1979), (pp. 551, section 32.15) is sometimes confused by censoring. In the literature confusingly Type I censoring is sometimes called truncation and Type II censoring is sometimes known simply as censoring, see for example Koziol and Byar (1975), Dufour and Maag (1978), Barr and Davidson (1973). We define censoring as when all of the data is used to generate the empirical CDF, but only the uncensored data is used estimated the parameters and calculate the goodness of fit statistics. In this paper we concentrate on left-truncated (as defined in the above paragraph) data only.

When dealing with a sample data obtained from observations one may wish to test the hypothesis that these data are drawn from a left-truncated Weibull distribution, even if the scale parameter α and shape parameter β are unknown. A common method for estimating the parameters of a pdf from a sample data set is maximum likelihood estimation (MLE). Note that the left-truncated Weibull pdf, Equation (2), is continuously differentiable in the argument τ and its two parameters, $0 < \alpha, \beta < \infty$, to any order and thus $f \in C^{\infty}((\tau_L, \infty) \times (0, \infty) \times (0, \infty))$. Also f and all its derivatives with respect to τ, α, β vanish for $\tau \to \infty$, at least like $\exp\left[-(\tau/\alpha)^{\beta'}\right]$ for $\alpha > 0$ and any $\beta' \in (0, \beta)$. These regularity conditions are essential for the "well-behaviour" of MLE.

To determine how well the sampled data fits the hypothesized distribution one must measure the goodness-of-fit (gof). Studies using Kolmogorov-Smirnov gof test to determine whether the sampled data belong to an untruncated Weibull distribution began in the late 1970s by Littell, McClave, and Offen (1979), Chandra, Singapurwalla, and Stephens (1981); Parsons and Wirsching (1982). In performing the hypothesis test it is crucial to use the correct critical values. When the Weibull parameters are estimated from the sample data, the standard Kolmogorov-Smirnov test tables Smirnov (1948); Miller (1956) for the case where the parameters are known cannot be used, because the probability integral transform using the estimated parameters destroys the independence of the transformed random variables as demonstrated by David and Johnson (1948).

In the literature there are very few studies dedicated to the left-truncated Weibull distributions (LTWD) is Wingo (1989), Balakrishnan and Mitra (2012). However, the MLE-approach in the first reference is rather heuristic level whereas the second reference is more concerned with a maximisation-expectation approach to handle left-truncation and right-censoring. For

theoretical investigations of the Weibull distribution the reader is referred to Agostino and Stephens (1986) and Lehmann and Casella (1998).

For the left-truncated 2-parameter Weibull distribution we shall distinguish four cases throughout this article :

Case I: Both parameters, the scale parameter, $\alpha > 0$, and the shape parameter, $\beta > 0$, are known a-priori.

Case II: Both parameters, the scale parameter, $\alpha > 0$, and the shape parameter, $\beta > 0$, are unknown a-priori and need to be estimated from the sample data.

Case IIIa: The scale parameter, $\alpha > 0$, is unknown and needs to be estimated from the sample data, but $\beta > 0$ is known.

Case IIIb: The shape parameter, $\beta > 0$, is unknown and needs to be estimated from the sample data, but $\alpha > 0$ is known.

In the next section we briefly review the maximum likelihood estimation for Cases II - III and comment on the consistency, asymptotic normality and efficiency of the MLE when applied to data sampled from a left-truncated Weibull distribution. Details on these issues have been discussed in Kreer, Kizilersu, Thomas, and dos Reis (2015). In Section 3 we discuss and develop the Kolmogorov-Smirnov (KS) goodness-of-fit (gof) statistics for the left-truncated Weibull distribution to decide whether the sample data could belong to the hypothetical distribution. In Section 4 we present an asymptotic analysis exploiting the Brownian bridge character of the KS statistics following some prior work of Durbin (1973) and Stephens (1977) on untruncated distributions and give our results for the left-truncated Weibull distribution for all cases. The quantile analysis to determine the modified critical values using Monte Carlo simulations is given in Section 5, where we discuss our numerical algorithm and present our results on the left-truncated data for the four cases listed above. All the results obtained on modified critical values are discussed and analysed in Section 6. In Section 7 we give a procedure for interpreting the results and a power study for Case I and Case II. Section 8 discusses the application of the methods discussed throughout the paper to failed US marriages, worldwide terrorist attacks, a sample of stock market data from New York stock exchange, and the radioactive α-decay of Americium-241. All the results are discussed in the concluding section.

2. Maximum likelihood estimation of left-truncated Weibull parameters

The maximum likelihood estimates of the left-truncated Weibull parameters differ from the complete ones because the left-truncated pdf $f(\cdot)$ with left-truncation point $\tau_L > 0$ has an additional multiplicative factor $\exp\left(\frac{\tau_L}{\alpha}\right)^\beta$ in comparison to the complete one. In this paper, the left-truncation point τ_L is assumed to be known. From Equation (2) we determine that the likelihood function for the left-truncated Weibull distribution as

$$L_{trunc}(\tau_1, \tau_2, ..., \tau_n|\alpha, \beta, \tau_L) = \prod_{i=1}^{n} \frac{\beta}{\alpha} \left(\frac{\tau_i}{\alpha}\right)^{\beta-1} e^{\left(\frac{\tau_L}{\alpha}\right)^\beta - \left(\frac{\tau_i}{\alpha}\right)^\beta}, \tag{3}$$

and consequently the logarithm of the likelihood as

$$\begin{aligned} \log L_{trunc}(\tau_1, \tau_2, ..., \tau_n|\alpha, \beta, \tau_L) &= \sum_{i=1}^{n} \log\left[\frac{\beta}{\alpha}\left(\frac{\tau_i}{\alpha}\right)^{\beta-1} e^{-\left(\frac{\tau_i}{\alpha}\right)^\beta}\right] + n\left(\frac{\tau_L}{\alpha}\right)^\beta \\ &= n\log\beta - n\beta\log\alpha + (\beta-1)\sum_{i=1}^{n}\log\tau_i - \sum_{i=1}^{n}\left(\frac{\tau_i}{\alpha}\right)^\beta + n\left(\frac{\tau_L}{\alpha}\right)^\beta, \\ &= \log L\left(\boldsymbol{\tau}|\alpha, \beta, 0\right) + n\left(\frac{\tau_L}{\alpha}\right)^\beta \end{aligned} \tag{4}$$

where $L(\boldsymbol{\tau}|\alpha,\beta,0)$ is the likelihood function for the untruncated distribution.

The Weibull parameters that maximize the likelihood function, Equation (3), are the same as those that maximise the log-likelihood function, Equation (4), and are obtained by calculating the partial derivatives with respect to α and β :

$$\frac{\partial}{\partial\alpha}\log L_{trunc}(\tau_1,\tau_2,...,\tau_n|\alpha,\beta,\tau_L) = \frac{\partial}{\partial\alpha}\log L(\tau_1,\tau_2,...,\tau_n|\alpha,\beta) + n\frac{\partial}{\partial\alpha}\left(\frac{\tau_L}{\alpha}\right)^{\beta} = 0\,,$$

$$\Longrightarrow -n\beta\frac{1}{\alpha} + \beta\sum_{i=1}^{n}\tau_i^{\beta}\alpha^{-\beta-1} - n\beta\left(\frac{\tau_L}{\alpha}\right)^{\beta}\frac{1}{\alpha} = 0\,. \tag{5}$$

$$\frac{\partial}{\partial\beta}\log L_{trunc}(\tau_1,\tau_2,...,\tau_n|\alpha,\beta,\tau_L) = \frac{\partial}{\partial\beta}\log L(\tau_1,\tau_2,...,\tau_n|\alpha,\beta) + n\frac{\partial}{\partial\alpha}\left(\frac{\tau_L}{\beta}\right)^{\beta} = 0\,,$$

$$\Longrightarrow \frac{n}{\beta} - n\log\alpha + \sum_{i=1}^{n}\log\tau_i - \sum_{i=1}^{n}\log\left(\frac{\tau_i}{\alpha}\right)\cdot\left(\frac{\tau_i}{\alpha}\right)^{\beta} + n\log\left(\frac{\tau_L}{\alpha}\right)\cdot\left(\frac{\tau_L}{\alpha}\right)^{\beta} = 0\,. \tag{6}$$

Rearranging Equation (5) we get

$$\alpha \;=\; \left(\frac{1}{n}\sum_{i=1}^{n}\left[\tau_i^{\beta} - \tau_L^{\beta}\right]\right)^{1/\beta}\,. \tag{7}$$

Note that Equation (7) is one of two MLE equations in Case II but is the only MLE equation in Case IIIa. There always exists a solution for α in Case IIIa for a given β. Rewriting Equation (6) we obtain the following

$$n\frac{1}{\beta} + \sum_{i=1}^{n}\log\left(\frac{\tau_i}{\alpha}\right) - \sum_{i=1}^{n}\left(\frac{\tau_i}{\alpha}\right)^{\beta}\log\left(\frac{\tau_i}{\alpha}\right) + \sum_{i=1}^{n}\left(\frac{\tau_L}{\alpha}\right)^{\beta}\log\left(\frac{\tau_L}{\alpha}\right) \;=\; 0\,. \tag{8}$$

Equation (8) is the second MLE equation in Case II but the only MLE equation in Case IIIb, where α is known. Eliminating α in Equation (8) using Equation (7), we obtain (after some algebraic manipulation) the following equation for β (for Case II) (Wingo (1989) and arxiv-version of Malevergne, Pisarenko, and Sornette (2005))

$$0 = \frac{1}{\beta} - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tau_i}{\tau_L}\right)^{\beta}\log\frac{\tau_i}{\tau_L}}{\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\tau_i}{\tau_L}\right)^{\beta} - 1\right]} + \frac{1}{n}\sum_{i=1}^{n}\log\frac{\tau_i}{\tau_L}\,. \tag{9}$$

Equations (7) and (9), reduce, in the limit $\tau_L \to 0$, to those given in Cohen (1965) for untruncated MLE equations. The solutions for α and β to the simultaneous Equations (7) and (9) are denoted by $\hat{\beta}_n = \hat{\beta}_n(\tau_1,\ldots,\tau_n|\tau_L)$ and $\hat{\alpha}_n = \hat{\alpha}_n(\tau_1,\ldots,\tau_n|\tau_L)$. For convenience we shall suppress the dependence on the sample $\tau_1,...,\tau_n$ and the left-truncation value τ_L and simply write $\hat{\alpha}$ and $\hat{\beta}$. The existence and uniqueness of a non-trivial MLE solution is almost trivial for Case IIIa, whereas the Case II and Case IIIb are dealt with the Lemma I given in Kreer *et al.* (2015). To assert the existence of a non-vanishing MLE-solution, the sample data need to satisfy the following inequality

$$2\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{\tau_i}{\tau_L}\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{\tau_i}{\tau_L}\right)^2 > 0\,. \tag{10}$$

If the condition given Equation (10) is not satisfied then the only solution to the MLE equation for β, Equation (9) is the trivial solution $\alpha = \beta = 0$. This can be shown by inserting $\alpha = \beta^{1/\beta}$

and taking the limit $\beta \to 0$ in Equation (3). Only in this case the likelihood Equation (3) is positive and non vanishing[1].

Table 1 was generated using a Monte Carlo Simulation with 10,000 steps, and without loss of generality the parameters were chosen as $\alpha = 1$ and $\beta = 1$. It gives the percentage of left-truncated Weibull distributed random samples of size n satisfing Equation (10) for which the MLE provides a non-trivial solution. Note that the various truncation points τ_L were chosen in such a way, that $\eta = (\tau_L/\alpha)^\beta$ yields the desired truncation probability $p = 1 - \exp(-\eta)$ of $10\%, ..., 90\%$ respectively. The truncated Weibull numbers were generated by Equation (30).

Table 1: Percentage of left-truncated Weibull distributed random samples for which there exists a solution to the MLE equations for Case II.

Percentage Removed	0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %
n=30	100±0	100±0	100±0	98 ±1	94±1	89±1	85±1	83±1	82 ±0	83±0
n=50	100±0	100±0	100±0	100±0	98±0	95±1	90±1	87±1	85±1	83±1
n=100	100±0	100±0	100±0	100±0	100±0	99±0	98±0	95±1	92±1	87±1

The consistency, asymptotic normality and efficiency of the MLE method for left-truncated Weibull distribution are discussed in Theorem 1 in Kreer *et al.* (2015) and the relevant proofs are provided as well. Key for the proof is the smoothness property of the left-truncated Weibull distribution. Denoting the true parameter vector by (α^0, β^0), we note in particular that all the asymptotic properties follow in this case from the asymptotic normality, i.e. $\sqrt{n}\left((\hat{\alpha}_n, \hat{\beta}_n) - (\alpha^0, \beta^0)\right)$ is asymptotically normal with vector mean zero and covariance matrix $[Z((\alpha^0, \beta^0))]^{-1}$ being the inverse of the Fisher information matrix

$$Z(\alpha^0, \beta^0) = -\mathbb{E}\left[\begin{array}{cc} \frac{\partial^2 \log f(\tau|\alpha,\beta,\tau_L)}{\partial\alpha^2} & \frac{\partial^2 \log f(\tau|\alpha,\beta,\tau_L)}{\partial\alpha\partial\beta} \\ \frac{\partial^2 \log f(\tau|\alpha,\beta,\tau_L)}{\partial\beta\partial\alpha} & \frac{\partial^2 \log f(\tau|,\alpha,\beta,\tau_L)}{\partial\beta^2} \end{array}\right]_{\alpha=\alpha^0,\beta=\beta^0} \tag{11}$$

The elements of the Fisher information matrix, Equation (11), are calculated as

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial^2}{\partial\alpha^2}\log f(\tau|\alpha,\beta,\tau_L)\right) &= -\frac{\beta^2}{\alpha^2}\,, \\
\mathbb{E}\left(\frac{\partial^2}{\partial\alpha\partial\beta}\log f(\tau|\alpha,\beta,\tau_L)\right) &= \frac{1}{\alpha}\left\{1 + [\log\eta + e^\eta E_1(\eta)]\right\}\,, \\
\mathbb{E}\left(\frac{\partial^2}{\partial\beta^2}\log f(\tau|\alpha,\beta,\tau_L)\right) &= -\frac{1}{\beta^2}\left\{1 + 2\left[\log\eta + e^\eta E_1(\eta)\right] + \left[(\log\eta)^2 + 2e^\eta E_2(\eta)\right]\right\}.
\end{aligned}$$

where we have used the functions $E_1(s) = \int_s^\infty dy \exp(-y)/y$ (i.e. the exponential integral) and $E_2(s) = \int_s^\infty dy \exp(-y)\log(y)/y$.

3. Kolmogorov-Smirnov goodness-of-fit test for the left-truncated Weibull distribution

Let us test the following null hypothesis H_0: The i.i.d. sample $\tau_1, \tau_2, \cdots, \tau_n$ satisfying $\tau_L < \tau_i$ for $i = 1, 2, ..., n$ for some positive τ_L and some integer n, is drawn from a left-truncated

[1] An example of the violation of the second MLE equation Equation (9) is for $n = 30$ the random sample $\tau_i = \tau_L + \epsilon \cdot i$ for $i = 1, 2, ..., 25$, a sufficiently small $\epsilon > 0$ and $\tau_i = \ell \cdot \tau_L + \epsilon \cdot i$ for $i = 26, 27, ..., 30$ and some $\ell \gg 1$. In this case we see that Equation (10) is not satisfied.

Weibull distribution $F(\boldsymbol{\tau})$ as given in Equation (1) with estimated parameters $(\hat{\alpha}, \hat{\beta})$ obtained from MLE as discussed in the previous section[2]. Using the empirical distribution function $F_n(\boldsymbol{\tau})$, defined as the proportion of the values of the order statistics $\tau_{(1)}, \tau_{(2)}, ..., \tau_{(n)}$ smaller than $\tau \in (\tau_L, \infty)$, the Kolmogorov-Smirnov (KS) test statistic is given (e.g. Kendall and Stuart (1979), sect. 30.49 and Shorack and Wellner (2009)),

$$\begin{aligned} D_n &\equiv \sup_{-\infty<\tau<+\infty} \|F_n(\boldsymbol{\tau}) - F(\boldsymbol{\tau})\|, \qquad (12) \\ &= \sup_{\tau_L<\boldsymbol{\tau}<\infty} [F_n(\boldsymbol{\tau}) - F(\boldsymbol{\tau}), F(\boldsymbol{\tau}) - F_n(\boldsymbol{\tau})] \\ &= \max_{1\le i\le n} \left[\frac{i}{n} - F(\tau_i), F(\tau_i) - \frac{i-1}{n}\right]. \qquad (13) \end{aligned}$$

Here D_n is the KS distance which is compared with a critical value $D_{cv}(n, p, 0.05)$, that depends on the sample size n, the truncation level p (the theoretical percentage removed from the untruncated distribution) and significance level 0.05 used throughout the paper. If the value of $\sqrt{n}D_n$ is greater than some critical value $D_{cv}(n, p, 0.05)$ then the hypothesis that $F_n(\boldsymbol{\tau})$ and $F(\boldsymbol{\tau})$ come from the same distribution is rejected, i.e.,

H_0 is the hypothesis that the set of values $\boldsymbol{\tau}$ is sampled from a random distribution with a known cdf $F(\boldsymbol{\tau})$,

$$H_0 \quad \text{is not rejected if} \quad \sqrt{n}D_n < D_{cv}(n, p, 0.05) \quad . \qquad (14)$$

The critical values used in the hypothesis test, Equation (14), depend on whether the parameters, α, β, are known or unknown and are estimated from the data itself. The cases introduced earlier in section 1 can be grouped under two the categories for the purpose of KS statistics.

Out-sample KS statistics If the parameters of the distribution from which the sampled data is to be tested against are known precisely, i.e. if $F(\boldsymbol{\tau})$ in Equation (12) is known this referred to as an *out-sample* KS statistic. In this study this statistics is named as Case I where the critical values (CVs) of Kolmogorov and Smirnov are recovered. Moreover the CVs are independent of the distribution and the range of parameters.

In-sample KS statistics If the parameters of the distribution must be estimated from the sampled data to construct the theoretical cdf ($F(\boldsymbol{\tau})$ in Equation (12)), then D_n is referred as an *in-sample* KS statistic. It is well known, when the parameters are estimated from the sample and then the goodness-of-fit test is performed, that the probability integral transformation of the sample variables destroys their independence (see e.g. David and Johnson (1948)). Thus Kolmogorov's argument leading to Equation (13) for his universal critical values becomes invalid. We expect for each of our three cases to have different critical values, and in Case I we should recover Kolmorogorov's values.

Making use Equation (1) for F and of the τ_i's representation as given by Equation (30), Equation (13) can be written as :

$$\begin{aligned} D_n &= \max_{1\le i\le n} \left\{\frac{i}{n} - 1 + \exp\left[\left(\frac{\tau_L}{\hat{\alpha}}\right)^{\hat{\beta}} - \left(\frac{\tau_i}{\hat{\alpha}}\right)^{\hat{\beta}}\right], 1 - \exp\left[\left(\frac{\tau_L}{\hat{\alpha}}\right)^{\hat{\beta}} - \left(\frac{\tau_i}{\hat{\alpha}}\right)^{\hat{\beta}}\right] - \frac{i-1}{n}\right\} \\ &= \max_{1\le j\le n} \left\{\frac{i-n}{n} + \exp\left[\hat{\eta} - (\eta + y_i)^{\hat{\beta}/\beta^0}\left(\frac{\alpha^0}{\hat{\alpha}}\right)^{\hat{\beta}}\right], \frac{n+1-i}{n} - \exp\left[\hat{\eta} - (\eta + y_i)^{\hat{\beta}/\beta^0}\left(\frac{\alpha^0}{\hat{\alpha}}\right)^{\hat{\beta}}\right]\right\} \\ &= \max_{1\le i\le n} \left\{\frac{i-n}{n} + \exp\left[\left\{\eta^{\hat{\beta}/\beta^0} - (\eta + y_i)^{\hat{\beta}/\beta^0}\right\}\left(\frac{\alpha^0}{\hat{\alpha}}\right)^{\hat{\beta}}\right], \right. \\ &\qquad \left. \frac{n+1-i}{n} - \exp\left[\left\{\eta^{\hat{\beta}/\beta^0} - (\eta + y_i)^{\hat{\beta}/\beta^0}\right\}\left(\frac{\alpha^0}{\hat{\alpha}}\right)^{\hat{\beta}}\right]\right\}, \qquad (15) \end{aligned}$$

[2]From this point onwards we will drop the index n and use $\hat{\alpha}$ and $\hat{\beta}$.

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated parameters while α^0 and β^0 are the true ones, $\eta \equiv (\tau_L/\alpha^0)^{\beta^0}$ and likewise $\hat{\eta} \equiv (\tau_L/\hat{\alpha})^{\hat{\beta}}$ and y_i's are standard exponential random variates, as described in Appendix A. Equation (15) describes the modified critical values for all four cases above.

The critical values in general are a function of the sample size n only when the untruncated data set is considered. But clearly, they also depend on the truncation parameters, such as the truncation level p or truncation parameter η, when truncated data is considered. However, for two cases we find simplified relations for D_n, which are independent of the truncation parameter τ_L or η and also independent of the true values of α^0 and β^0 :

Case I : ($\hat{\eta} = \eta$, $\hat{\alpha} = \alpha^0$ and $\hat{\beta} = \beta^0$)

$$D_n \quad = \quad \max_{1\le i\le n} \left\{ \frac{i-n}{n} + \exp\left(-y_i\right),\ \frac{n+1-i}{n} - \exp\left(-y_i\right) \right\} . \tag{16}$$

Case IIIa : ($\hat{\beta} = \beta^0$)

$$D_n = \max_{1\le i\le n} \left\{ \frac{i-n}{n} + \exp\left[-y_i \left(\frac{\alpha^0}{\hat{\alpha}}\right)^{\beta^0}\right], \frac{n+1-i}{n} - \exp\left[-y_i \left(\frac{\alpha^0}{\hat{\alpha}}\right)^{\beta^0}\right] \right\} . \tag{17}$$

One observes that in Case IIIa when the shape parameter β^0 is known, Equation (15) simplifies to Equation (17) and becomes independent of truncation, τ_L (or η), because of $\hat{\beta}/\beta^0 = 1$ the η-terms cancel each other out.

To construct confidence tables without loss of generality one may assume $(\alpha^0, \beta^0) = (1, 1)$ and hence $\eta = \tau_L$. Following Thoman, Bain, and Antle (1969) we denote for general Weibull distributions with any positive (α^0, β^0) the random variables $(\alpha^0/\hat{\alpha})^{\hat{\beta}}$ and $\hat{\beta}/\beta^0$ as *pivotal functions*. Note that the KS distance D_n in Equation (15) depends on these pivotal functions, n and η. Consequently, D_n is "universal" for different combinations of (α^0, β^0) for the same n and η, provided the following holds true

$$\left(\frac{\alpha^0}{\hat{\alpha}_{(\alpha^0,\beta^0)}}\right)^{\hat{\beta}_{(\alpha^0,\beta^0)}} \overset{distrib.}{=} \left(\frac{1}{\hat{\alpha}_{(1,1)}}\right)^{\hat{\beta}_{(1,1)}} \quad , \quad \left(\frac{\hat{\beta}_{(\alpha^0,\beta^0)}}{\beta^0}\right) \overset{distrib.}{=} \hat{\beta}_{(1,1)} \tag{18}$$

where $\hat{\alpha}_{(1,1)}$ and $\hat{\beta}_{(1,1)}$ are the MLE estimates originating from the simplest choice of a Weibull distribution with $(\alpha^0, \beta^0) = (1, 1)$. Likewise $\hat{\alpha} = \hat{\alpha}_{(\alpha^0,\beta^0)}$ and $\hat{\beta} = \hat{\beta}_{(\alpha^0,\beta^0)}$ are the MLE estimates originating from a Weibull distribution for arbitrary positive (α^0, β^0). The latter equality in distribution, Equation (18), was demonstrated in Appendix 3 of Kreer *et al.* (2015). For untruncated data where $x_L = 0$ (thus η and p vanish), $D_{cv}(n, 0, 0.05)$ will only depend on n. This was observed by Thoman *et al.* (1969) and allowed Littell *et al.* (1979) and Parsons and Wirsching (1982) the production of confidence tables for in-sample KS tests with MLE equations solved for exponential random variates. Similarly, in Case IIIa when the shape parameter β is known, Equation (15) simplifies and becomes independent of truncation, x_L (and hence of the truncation level $p = 1 - e^{-\eta}$), because $\hat{\beta}/\beta^0 = 1$ and the η-terms cancel out. Only for Case II and Case IIIb do we need to investigate the dependence of $D_{cv}(n, p, 0.05)$ on the parameter η ($\eta = x_L$ if $(\alpha^0, \beta^0) = (1, 1)$) and n in greater detail.

4. Brownian bridge asymptotics for Kolmogorov-Smirnov goodness of fit tests

4.1. Brownian bridge and Donsker's theorem

As in the discussion of the MLE in section 2 it will be interesting to consider what happens to the KS test when $n \to \infty$. The asymptotic behaviour of the KS-test has been of interest

from the 1940s onwards, Durbin (1973), Stephens (1977), and Shorack and Wellner (2009), in calculating the asymptotic critical values. For a random variable $\boldsymbol{\tau}$ distributed according to a theoretical Weibull distribution function $F(\boldsymbol{\tau}|\theta^0)$, one may define the difference between the theoretical (with or without estimated parameters $\hat{\boldsymbol{\theta}}_n = (\hat{\alpha}_n, \hat{\beta}_n)$) and empirical distributions as (Durbin (1973), Equation (2))

$$G_n(t) \quad = \quad \sqrt{n}\left[\hat{F}_n(t) - t\right] \tag{19}$$

where $\hat{F}_n(t)$ is the proportion of $\tau_1, \tau_2, \cdots, \tau_n$, i.i.d. for which $F(\tau_i|\hat{\theta}_n) \leq t$, $t \in [0,1]$, and $\hat{\boldsymbol{\theta}}_n$ is the MLE estimate for the true parameter $\boldsymbol{\theta^0} = (\alpha^0, \beta^0)$. Note that taking the absolute value of the supremum in Equation (19) would yield the KS-distance in Equation (13). Viewing Equation (19) as a stochastic process in $t \in [0,1]$, Doob's Theorem (also known as the functional central limit theorem) asserts the convergence in distribution against a limiting stochastic process which is Gaussian with zero mean and the covariance structure of a Brownian bridge (see Shorack and Wellner (2009)). For the case where the parameters are estimated from the sample itself a modification (due to Durbin (1973)) has to be made. We may apply Theorem 2 of Durbin (1973) (here $\boldsymbol{\theta} = (\alpha, \beta)$), where the limiting Gaussian process is denoted, in analogy from above, by $G_n(t)$ with mean of 0, i.e. $\mathbb{E}\left(G_n(t)\right) = 0$ and a covariance structure given by

$$C(s,t) = \mathbb{E}\left(G_n(s)G_n(t)\right) = \min(s,t) - s \cdot t - u^T(s)\,\Sigma\,u(t), \quad 0 \leq s \leq t \leq 1 \tag{20}$$

where $\Sigma = Z^{-1}$ is the inverse of the Fisher information matrix $Z(\alpha, \beta)$ given in Equation (11), and $u(\cdot)$ are certain vector-valued functions given by Equation (21) below. Note that the supremum of this Gaussian process using only n points will converge to the asymptotic value of the KS-distance D_n, as given in Equation (13) of the previous section, when $n \to \infty$. This will be key in deriving the asymptotic values. We readily check that Durbin's assumptions (A2) and (A3) in Durbin (1973) are also satisfied for the truncated case with truncation point $\tau_L > 0$, so that Theorem 2 of Durbin (1973) may be applied. Stephens (1977) studies the Brownian bridge with the covariance structure given in Equation (20) for complete data, (i.e. $\tau_L = 0$). The vector-valued function $u(s)$ in Equation (20) for left-truncated Weibull distributions with $\tau_L > 0$ is

$$u(s) \equiv \begin{pmatrix} \frac{\partial s}{\partial \alpha} \\ \frac{\partial s}{\partial \beta} \end{pmatrix} = \begin{pmatrix} \frac{\beta}{\alpha} s \log s \\ -\frac{s}{\beta}\left\{\eta \log \eta - (\eta - \log s)\log(\eta - \log s)\right\} \end{pmatrix}, \tag{21}$$

where $s = F(\tau) = F(\tau|\alpha, \beta, \tau_L)$. In the following calculations, without loss of generality, we may choose for convenience $(\alpha, \beta) = (1,1)$. Using the covariance equations, Equation (20), together with Equation (21) and the matrix Σ as the inverse of Z, from Equation (11), we can now for any $m \in \mathbb{N}$ simulate a Brownian bridge with discrete values $t_i = i/m$ with the given discrete covariance structure $C_{i,j} = C(s = i/m, t = j/m)$, for $i, j = 0, 1, ..., m$.

4.2. Numerical implementation of the Brownian bridge

We perform the following procedure as described in Anderson and Stephens (1997) to calculate the critical values in the Brownian bridge approach :

1. Discretise the interval $[0,1]$ for given $m \in \mathbb{N}$, in discrete values $s = i/m, t = j/m$, $i, j = 0, 1, ..., m$.

2. The discrete covariance matrix $C^{(m+1)} = (C_{i,j})_{i,j}$ from Equation (20) now has entries

$$C_{i,j} = \min(i,j)/m - i/m \cdot j/m \; - \; u^T(i/m)\,\Sigma\,u(j/m) \tag{22}$$

and is symmetric and positive definite.

3. Calculate the Cholesky decomposition $C^{(m+1)} = BB^T$, where $B = B^{(m+1)}$ is a triangular matrix of dimension $(m+1) \times (m+1)$.

4. Draw $(\zeta_0, \zeta_1, ..., \zeta_m)$ standard normally distributed numbers (i.e. mean 0 and variance 1). Set $\zeta_0 = 0$ and $\zeta_m = 0$ and define the vector $z = (0, \zeta_1, ..., \zeta_{m-1}, 0)$.

5. The transformed $(m+1)$-vector Bz is a discrete representation of a Brownian bridge $G_m(t)$ starting at $t = 0$ with $G_m(0) = 0$ and ending at $t = 1$ with $G_m(1) = 0$. Find the following statistics
$D^+(m) = \max(Bz)$, $D^-(m) = -\min(Bz)$ and then set $D_m = \max\{D^+(m), D^-(m)\}$.

6. Keep D_m in a list and sort in ascending order. Take the 95% as a critical $D_m^{BB}(95\%)$.

7. Repeat procedure for $m = 30, 50, 100, 200, ...$ and fit D_m against the function $A + B/\sqrt{m}$ (see also Chandra *et al.* (1981)). The value A is the asymptotic value of the Kolmogorov-Smirnov statistic, $A = D_{cv}(\infty, 0.05)$.

4.3. Results: asymptotical critical values from Brownian bridge

We apply the Brownian Bridge (BB) approach to find the asymptotic critical values for the following cases and present the results in Table 2.

Case I Out-sample testing: Put $\Sigma = 0$ (because α and β are known precisely therefore Fisher information matrix is irrelevant here) and sample a pure Brownian bridge.

Case II In-sample testing for two unknown parameters (with truncation).

Case IIIa - IIIb In-sample testing with one-parameter known (with truncation): Get a one-dimensional Fisher Information matrix from Equation (11) with the unknown parameter and invert this element to obtain the corresponding Σ-matrix.

5. The quantile analysis for determining the critical values

5.1. The Monte-Carlo algorithm

The quantile procedure to calculate the critical values is described below.

Algorithm 1: Procedure for calculating the mean and variance of the critical values of the KS-test

Input:
The values of α and β are both set to 1
Output: The mean and standard deviation of the critical values of the KS-test for a range of sample sizes n and truncation levels p, $\eta = \tau_L = \alpha\left(-\log\left(1-p\right)\right)^{\beta}$.

for $p = 0$ **to** 0.9 *-STEP 0.1* **do**
  for $n = 30, 50, 100, 200, 500, 1000, 10000$ **do**
    for $j = 1$ **to** 100 **do**
      for $k = 1$ **to** 1000 **do**
        • Draw n random numbers u_i from a uniform distribution $u_i \sim \mathcal{U}(0,1)$. It follows directly from the discussion in appendix A that the left-truncated Weibull distributed random variables are $\tau_i = \tau_L - \log u_i$
        • Estimate $\hat{\alpha}$ and $\hat{\beta}$ using MLE equations Equations (7) and (9).
        • Calculate the Kolmogorov-Smirnov statistic using Equation (13) and store it as $D(n,p,j,k)$
        • Sort $D(n,p,j,:) \quad \forall \quad k$ in ascending order. The 95% confidence interval, i.e. $(\alpha_H = 0.05)$ is $D^q_{cv}(n,p,j) = \frac{1}{2}\left(D(n,p,j,950) + D(n,p,j,951)\right)$.
      end
    end
    • Calculate the mean $D^q_{cv}(n,p)$ and variance $\sigma^2_{D^q_{cv}(n,p)}$ from the 100 values.
  end
end

Table 2: The asymptotical critical values from BB approach for all cases.

Truncation Level	Truncation Parameter	Case I	Case II	Case IIIa	Case IIIb
p	η	D^{BB}_{cv}	D^{BB}_{cv}	D^{BB}_{cv}	D^{BB}_{cv}
0	0	1.356±0.008	0.901 ±0.007	1.093±0.005	1.317±0.006
0.1	0.1	1.359±0.004	0.862 ±0.002	1.094±0.003	1.329±0.006
0.2	0.2	1.358±0.005	0.860 ±0.007	1.095±0.003	1.321±0.006
0.3	0.35	1.358±0.008	0.860 ±0.006	1.094±0.003	1.291±0.005
0.4	0.5	1.359±0.005	0.874 ±0.003	1.095±0.005	1.260±0.006
0.5	0.7	1.358±0.003	0.880 ±0.003	1.094±0.005	1.234±0.006
0.6	0.9	1.361±0.004	0.879 ±0.006	1.094±0.002	1.198±0.003
0.7	1.2	1.357±0.005	0.892 ±0.007	1.094±0.002	1.183±0.004
0.8	1.6	1.358±0.006	0.900 ±0.007	1.093±0.001	1.163±0.004
0.9	2.3	1.359±0.005	0.909 ±0.007	1.093±0.006	1.142±0.003

5.2. Results: critical values from Monte-Carlo simulations

Our results obtained for the modified critical values using the quantile analysis (outlined in Algorithm 1) for each sample size $n = (30, 50, 100, 500, 1000, 10000)$ and truncation parameter η are summarised in Table 3 for Case I, in Table 4 for Case II, in Table 5 for Case IIIa, and in Table 6 for Case IIIb.

Table 3: The critical values, D^q_{cv}, calculated from the quantile analysis for Case I.

p	η	$D^q_{n=30}$	$D^q_{n=50}$	$D^q_{n=100}$	$D^q_{n=200}$	$D^q_{n=500}$	$D^q_{n=1000}$	$D^q_{n=10000}$
0	0	1.322±0.025	1.329±0.024	1.336±0.024	1.343±0.024	1.346±0.024	1.346±0.023	1.354±0.027
0.1	0.1	1.321±0.024	1.333±0.023	1.339±0.026	1.345±0.026	1.348±0.027	1.351±0.024	1.352±0.021
0.2	0.2	1.321±0.025	1.327±0.024	1.339±0.025	1.345±0.023	1.344±0.025	1.352±0.027	1.351±0.026
0.3	0.35	1.322±0.023	1.335±0.028	1.341±0.025	1.349±0.025	1.349±0.024	1.350±0.025	1.359±0.026
0.4	0.5	1.319±0.026	1.330±0.027	1.338±0.026	1.345±0.026	1.347±0.024	1.356±0.025	1.352±0.026
0.5	0.7	1.322±0.024	1.331±0.024	1.334±0.024	1.345±0.028	1.349±0.022	1.356±0.025	1.353±0.028
0.6	0.9	1.322±0.025	1.331±0.024	1.340±0.023	1.343±0.024	1.349±0.028	1.352±0.026	1.357±0.026
0.7	1.2	1.322±0.027	1.330±0.023	1.339±0.027	1.345±0.024	1.346±0.023	1.350±0.025	1.359±0.025
0.8	1.6	1.319±0.029	1.330±0.025	1.338±0.021	1.345±0.029	1.348±0.024	1.351±0.025	1.355±0.026
0.9	2.3	1.323±0.024	1.328±0.023	1.340±0.026	1.348±0.024	1.346±0.023	1.349±0.024	1.354±0.026

Table 4: The critical values, D^q_{cv}, from the quantile analysis for Case II.

p	η	$D^q_{n=30}$	$D^q_{n=50}$	$D^q_{n=100}$	$D^q_{n=200}$	$D^q_{n=500}$	$D^q_{n=1000}$	$D^q_{n=10000}$
0	0	0.858±0.011	0.865±0.012	0.874± 0.012	0.881±0.013	0.887±0.012	0.890±0.015	0.893±0.015
0.1	0.1	0.817±0.012	0.829±0.011	0.838±0.012	0.843±0.013	0.850±0.013	0.851±0.013	0.857±0.012
0.2	0.2	0.815±0.012	0.824±0.011	0.838±0.012	0.842±0.012	0.847±0.013	0.852±0.012	0.856±0.011
0.3	0.35	0.818±0.013	0.830±0.012	0.840±0.010	0.848±0.013	0.854±0.013	0.856±0.012	0.859±0.012
0.4	0.5	0.821±0.012	0.832±0.011	0.846±0.011	0.853±0.012	0.857±0.012	0.862±0.013	0.866±0.013
0.5	0.7	0.824±0.013	0.840±0.013	0.852±0.013	0.860±0.012	0.863±0.011	0.868±0.014	0.872±0.012
0.6	0.9	0.830±0.012	0.844±0.013	0.857±0.012	0.866±0.012	0.873±0.011	0.876±0.013	0.881±0.011
0.7	1.2	0.835±0.012	0.853±0.012	0.864±0.013	0.873±0.012	0.878±0.013	0.882±0.013	0.888±0.012
0.8	1.6	0.839±0.012	0.855±0.012	0.871±0.013	0.880±0.013	0.886±0.014	0.890±0.015	0.894±0.014
0.9	2.3	0.843±0.012	0.864±0.011	0.880±0.014	0.890±0.014	0.897±0.014	0.897±0.013	0.904±0.014

Table 5: The critical values, D^q_{cv}, from the quantile analysis for Case IIIa.

p	η	$D^q{}_{n=30}$	$D^q{}_{n=50}$	$D^q{}_{n=100}$	$D^q{}_{n=200}$	$D^q{}_{n=500}$	$D^q{}_{n=1000}$
0	0	1.055±0.020	1.064±0.018	1.072±0.020	1.080±0.021	1.083±0.016	1.086±0.018
0.1	0.1	1.054±0.019	1.064±0.019	1.074±0.019	1.078±0.018	1.085±0.017	1.085±0.018
0.2	0.2	1.058±0.018	1.064±0.018	1.074±0.018	1.080±0.019	1.085±0.018	1.086±0.022
0.3	0.35	1.057±0.016	1.064±0.019	1.075±0.016	1.081±0.019	1.082±0.017	1.085±0.019
0.4	0.5	1.054±0.018	1.066±0.019	1.072±0.019	1.079±0.019	1.083±0.017	1.085±0.016
0.5	0.7	1.054±0.017	1.066±0.021	1.074±0.021	1.080±0.021	1.084±0.018	1.083±0.019
0.6	0.9	1.057±0.017	1.065±0.018	1.077±0.020	1.077±0.018	1.086±0.020	1.086±0.019
0.7	1.2	1.056±0.018	1.065±0.018	1.075±0.019	1.083±0.018	1.084±0.017	1.085±0.019
0.8	1.6	1.056±0.021	1.064±0.018	1.076±0.020	1.080±0.021	1.082±0.019	1.086±0.017
0.9	2.3	1.057±0.017	1.065±0.017	1.074±0.018	1.077±0.017	1.084±0.020	1.088±0.019

Table 6: The critical values, D^q_{cv}, from the quantile analysis for Case IIIb.

p	η	$D^q{}_{n=30}$	$D^q{}_{n=50}$	$D^q{}_{n=100}$	$D^q{}_{n=200}$	$D^q{}_{n=500}$	$D^q{}_{n=1000}$
0	0	1.281±0.024	1.289±0.022	1.301±0.024	1.302±0.028	1.310±0.025	1.310±0.023
0.1	0.1	1.301±0.026	1.307±0.023	1.314±0.023	1.320±0.024	1.322±0.025	1.323±0.027
0.2	0.2	1.283±0.026	1.293±0.023	1.299±0.026	1.308±0.025	1.306±0.026	1.307±0.026
0.3	0.35	1.255±0.023	1.262±0.025	1.270±0.023	1.273±0.024	1.284±0.023	1.281±0.024
0.4	0.5	1.224±0.022	1.233±0.021	1.241±0.022	1.246±0.024	1.250±0.027	1.257±0.022
0.5	0.7	1.194±0.021	1.203±0.022	1.212±0.020	1.214±0.021	1.223±0.022	1.227±0.023
0.6	0.9	1.171±0.020	1.177±0.019	1.189±0.023	1.193±0.020	1.194±0.023	1.198±0.023
0.7	1.2	1.144±0.020	1.154±0.021	1.162±0.023	1.169±0.021	1.174±0.021	1.179±0.022
0.8	1.6	1.122±0.019	1.136±0.021	1.142±0.024	1.148±0.020	1.153±0.022	1.154±0.021
0.9	2.3	1.100±0.019	1.110±0.021	1.125±0.022	1.124±0.020	1.131±0.018	1.136±0.019

6. Discussion of the results

The truncation in the analysis can be defined three equivalent ways: 1) τ_L, the value below which all data is removed/absent, 2) p the percentage of data removed/absent by the truncation procedure, and 3) the generalised truncation parameter $\eta \equiv \left(\frac{\tau_L}{\alpha}\right)^\beta$. These three parameters are related by the equations

$$p = 1 - e^{-\left(\frac{\tau_L}{\alpha}\right)^\beta} = 1 - e^{-\eta} \quad . \tag{23}$$

All of these parameters will be used throughout this paper, depending on which is the most convenient.

6.1. Estimation of the Weibull parameters

Identifying and analyzing the distribution which represents the data set is our main focus, since it is the source of the predictability. Figure (1) is an errorbar plot of the MLE estimates of the parameters $\hat{\alpha}$ (Upper left, in Case II and Lower left in Case IIIa) and $\hat{\beta}$ (Upper right in Case II, Lower right in Case IIIb) for various sample sizes n and truncation levels p. Here the true values are taken as $\alpha^0 = 400$ and $\beta^0 = 0.58$. From these plots one can see that as the sample size increases the variance in the estimation of $\hat{\alpha}$ and $\hat{\beta}$ decreases in all cases. Furthermore as the truncation level increases the variance in estimation of $\hat{\alpha}$ and $\hat{\beta}$ increases continuously in Case II, while in Case IIIb it increases initially then decreases. Finally, in Case IIIa the estimation of the parameter is totally insensitive to the truncation, see Figure(1c).

In summary, the estimation is better in Cases II and IIIb when the sample size is larger and the truncation is smaller. The single parameter estimates are far better than the double parameter estimates as expected. Case IIIa, where the shape parameter β is known, and the scale parameter α is unknown, is superior to Case IIIb with the unknown shape parameter β and known scale parameter α, since in Case IIIa the CVs are independent of truncation and the estimation of α is more precise, which is the optimum scenario. Comparisons on estimation of parameters show that the variance is reduced by 75% in α and by 50% in β between the two parameter and one parameter cases.

6.2. Critical values as a function of sample size n

Figure 2 depicts, for Cases II and IIIb, the dependence of the critical values, (given in Table 4 and Table 6) on n for a range of truncation levels, p (or truncation parameter η). For clarity the $x-$axis is plotted on a log scale. Both the cases show a distinctive separation between the lines for different truncation levels, indicating a dependence on the truncation level p.

On the other hand in Case I, the critical values are independent of truncation and only depend on n. As predicted by the theory, Equation (17), we also note that truncation has no noticeable effect on the critical values in Case IIIa as well. According to Miller's formula, Miller (1956), which was derived for the out sample, untruncated case namely Case I, the critical values are quadratic in $1/\sqrt{n}$

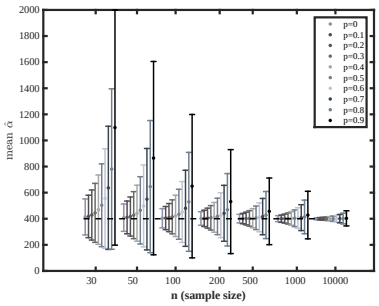
$$D_{cv}(n) = \sqrt{-\frac{1}{2}\log\frac{\alpha_H}{2}} - \frac{0.167}{\sqrt{n}} - \frac{\mathcal{A}}{n} \quad \text{for } n > 20 \quad (\alpha_H = 0.05,\ 95\%\ \text{confidence level})\,. \tag{24}$$

where the first term in above expression is Simirnov's asymptotic formula and calculated as 1.358 and

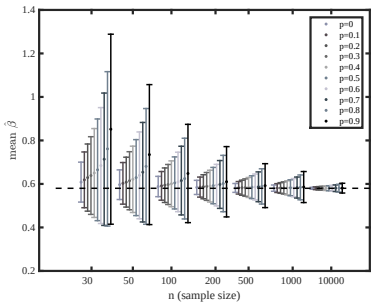
$$\begin{aligned} \mathcal{A} &\equiv 0.090\left(-\log_{10}\frac{\alpha_H}{2}\right)^{3/2} + 0.015\left(\log_{10}\frac{\alpha_H}{2}\right)^{2} - 0.085\frac{\alpha_H}{2} - 0.111 \\ &= 0.109\,. \end{aligned}$$

Although Miller's formula, Equation (24), is designed to be used for only Case I, where both the parameters are known a priori, we will however use it as a guide to investigate the functional dependence of the critical values on the sample size n for all cases. This can be achieved by fitting the critical values given in Tables 3 - 6 for each value of p to the function

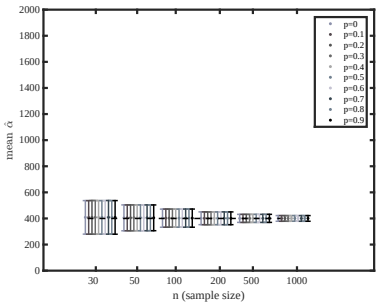
$$D_{cv}(p|n) = \tilde{A}(p) + \frac{\tilde{B}(p)}{\sqrt{n}} + \frac{\tilde{C}(p)}{n} \quad . \tag{25}$$



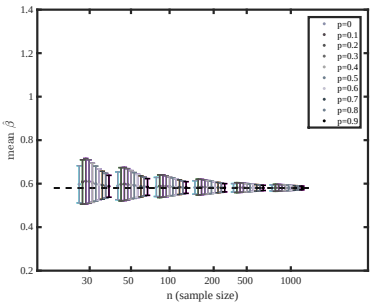
(a) $\hat{\alpha}$ – Case II.



(b) $\hat{\beta}$ – Case II.



(c) $\hat{\alpha}$ – Case IIIa.



(d) $\hat{\beta}$ – Case IIIb.

Figure 1: The mean value of the MLE of Weibull parameters α and/or β as a function of n and p (the percentage data removed by truncation). The error bars show one standard deviation in the estimated values of the parameters. The horizontal dashed line shows the true value of the parameters that was used to generate the data for $\alpha^0 = 400$ and $\beta^0 = 0.58$.

Table 7: Results of fitting $D^q_{cv}(p|n)$ to quadratic and linear functions in $1/\sqrt{n}$. The critical values, D^q_{cv}, obtained as a function of sample size n from the quadratic fit on to left-truncated data of Case I for each truncation level p, truncation parameter η. $\tilde{A}(p|n)$, $\tilde{B}(p|n)$ and $\tilde{C}(p|n)$ are the fit parameters in Equation (25), $\tilde{A}_1(p|n)$, $\tilde{B}_1(p|n)$ in Equation (26).

p	η	$\tilde{A}(p)$	$\tilde{B}(p)$	$\tilde{C}(p)$	$\tilde{A}_1(p)$	$\tilde{B}_1(p)$
0	0	1.355±0.004	-0.193±0.117	0.087±0.588	1.354±0.002	-0.177±0.024
0.1	0.1	1.353±0.003	-0.086±0.081	-0.484±0.408	1.356±0.003	-0.179±0.032
0.2	0.2	1.354±0.009	-0.128±0.223	-0.297±1.124	1.355±0.005	-0.185±0.048
0.3	0.35	1.357±0.008	-0.125±0.203	-0.354±1.026	1.359±0.005	-0.193±0.046
0.4	0.5	1.355±0.008	-0.116±0.221	-0.468±1.116	1.358±0.005	-0.206±0.052
0.5	0.7	1.358±0.010	-0.200±0.265	0.004±1.334	1.358±0.005	-0.199±0.054
0.6	0.9	1.359±0.004	-0.207±0.094	0.043±0.475	1.359±0.002	-0.199±0.019
0.7	1.2	1.359±0.007	-0.228±0.170	0.178±0.858	1.358±0.004	-0.194±0.036
0.8	1.6	1.356±0.002	-0.154±0.046	-0.266±0.234	1.358±0.002	-0.205±0.018
0.9	2.3	1.355±0.008	-0.128±0.218	-0.290±1.100	1.356±0.005	-0.184±0.047

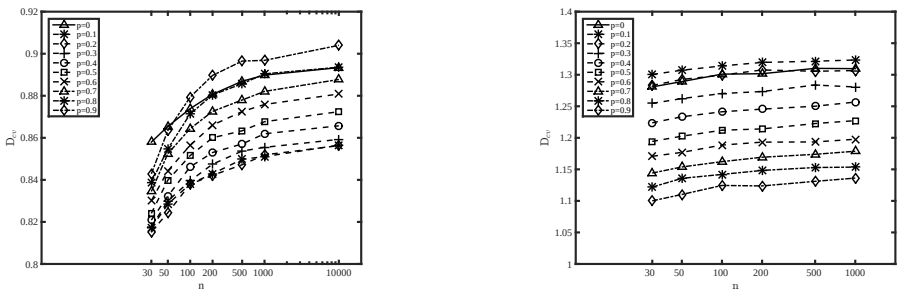
The fit results are tabulated in Table 7 for Case I where the values of $\tilde{C}(p)$ are quite variable and the standard deviation in $\tilde{C}(p)$ is greater than the value itself. This suggests that $D^q_{cv}(p|n)$ is better approximated by a function that is linear in $1/\sqrt{n}$ instead of quadratic, i.e.,

$$D_{cv}(p|n) = \tilde{A}_1(p) + \frac{\tilde{B}_1(p)}{\sqrt{n}}\,. \tag{26}$$

Table 8: The critical values obtained from the quantile analysis fitted to the linear function for a range of truncation level p for Case II, Case IIIa and Case IIIb. $\tilde{A}_1(p|n)$, $\tilde{B}_1(p|n)$ are the fit parameters defined in Equation (26).

		Case II		**Case IIIa**		**Case IIIb**	
p	η	$\tilde{A}_1(p)$	$\tilde{B}_1(p)$	$\tilde{A}_1(p)$	$\tilde{B}_1(p)$	$\tilde{A}_1(p)$	$\tilde{B}_1(p)$
0	0	0.896±0.001	-0.211±0.012	1.093±0.002	-0.204±0.022	1.318±0.005	-0.197±0.047
0.1	0.1	0.859±0.002	-0.223±0.018	1.093±0.004	-0.207±0.034	1.329±0.002	-0.154±0.020
0.2	0.2	0.859±0.002	-0.238±0.024	1.093±0.002	-0.194±0.021	1.314±0.007	-0.160±0.065
0.3	0.35	0.864±0.002	-0.243±0.023	1.092±0.004	-0.188±0.034	1.288±0.005	-0.186±0.048
0.4	0.5	0.870±0.003	-0.262±0.029	1.093±0.004	-0.203±0.038	1.261±0.004	-0.201±0.034
0.5	0.7	0.877±0.004	-0.273±0.039	1.093±0.006	-0.200±0.051	1.232±0.004	-0.208±0.040
0.6	0.9	0.885±0.002	-0.295±0.021	1.093±0.005	-0.196±0.041	1.204±0.005	-0.179±0.047
0.7	1.2	0.892±0.004	-0.298±0.039	1.093±0.005	-0.200±0.050	1.185±0.002	-0.222±0.016
0.8	1.6	0.900±0.005	-0.322±0.047	1.093±0.005	-0.195±0.045	1.162±0.005	-0.205±0.047
0.9	2.3	0.911±0.006	-0.346±0.062	1.093±0.004	-0.200±0.033	1.143±0.007	-0.227±0.061

The linear function is a better fit to the data $D^q_{cv}(p|n)$, in the sense that there is no significant change in the *adjusted r-squared* goodness of fit statistic, but the standard deviation in $\tilde{B}$ over all values of p is an order of magnitude better when Equation (26) is used instead of Equation (25). The results for $\tilde{A}_1(p|n)$ and $\tilde{B}_1(p|n)$ given by fitting $D_{cv}(p|n)$ are in very good agreement with Miller's formula, Equation (24). The fit results are given in Table 8 for Case II, Case IIIa and Case IIIb, respectively.



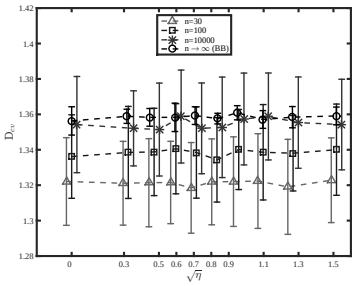
(a) Both α and β are unknown – Case II.

(b) β is unknown and α is known – Case IIIb.

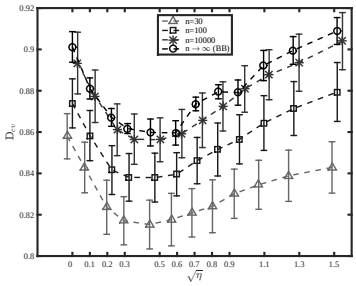
Figure 2: Critical values as function of n for a range of truncation level p.

6.3. Critical values as a function of left-truncation parameter eta

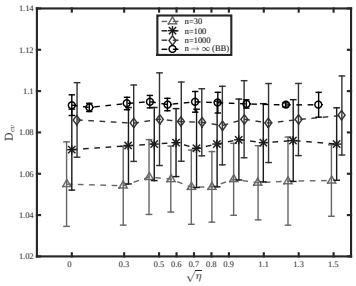
To determine the relationship between the critical values and the truncation parameter η, we plot in Figure 3 the critical values given in Tables 3-6, as a function of $\sqrt{\eta}$ for $n = (30, 100, 1000, 10000)$ for all cases. We have also included a plot of the Brownian Bridge



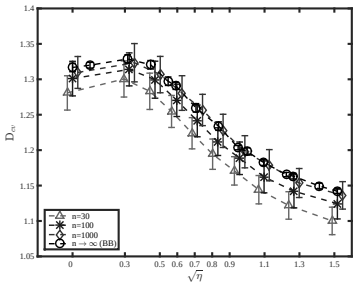
(a) Both α and β are known – Case I.



(b) Both α and β are unknown – Case II.



(c) β is known and α is not known – Case IIIa.



(d) β is unknown and α is known – Case IIIb.

Figure 3: The critical values as a function of $\sqrt{\eta}$ for a range of n values. The circled dashed line with the error bars are the Brownian Bridge calculation.

results (with error bars) for all cases, since that provides an alternative way of estimating D_{cv} in the limit $n \to \infty$. For out-sample data the critical values are independent of truncation and this is verified in Figure 3a. We see that there is no variation in the critical values as a function of $\sqrt{\eta}$. On the other hand Figure 3b for Case II shows that the critical values initially decrease but then increase as the truncation level increases (boomerang shape), which is totally different behaviour from the out-sample case (Case I). In Figure 3c for Case IIIa the CV's do not change as η increases, similar to Case I in that for a fixed value of n the critical values are independent of the truncation. These results are consistent with the theory we outlined in Equation (17). Case IIIb in Figure 3d, on the other hand, shows that CV's initially slightly increase then decrease as the truncation level increases.

In summary, the CV's in Cases I and IIIa are truncation independent while in Cases II and IIIb they are not. For all cases the asymptotic critical value analysis from the Brownian Bridge confirms the same η dependence as we found in the quantile analysis. This section

deals with formulating the critical values as a function of truncation parameter η. In both Case II and Case IIIb, the CV's are truncation dependent and among the many fit functions tried to describe the data we found that the quadratic ratio function

$$D^q_{cv}(\eta|n) \quad = \quad \frac{C(n) + B(n)\sqrt{\eta} + A(n)\eta}{E(n) + D(n)\sqrt{\eta} + \eta}, \tag{27}$$

fited best. Its parameters are given in Tables 9 and plotted in Figs. 4c,g and 4d,h for $n = 30$ and $n = 10000$, respectively. In the figure the light shaded grey, tick band shows the error range on $D^q_{cv}(\eta|n)$ values whereas the darker shaded grey area between the dashed lines is the error band on the fit values. In addition the asymptotic critical values from the Brownian Bridge analysis (squares) are shown in the figures for only $n = 10,000$.

Table 9: The critical values obtained by fitting the ratio function to the data from the quantile analysis for various sample sizes, n. The fit parameters defined in Equation (27) are given for each n values and for Cases II and IIIb.

	Case II				
n	$A(n)$	$B(n)$	$C(n)$	$D(n)$	$E(n)$
30	0.870±0.008	-0.197±0.102	0.207±0.040	-0.182±0.131	0.241±0.046
50	0.902±0.017	-0.218±0.196	0.295±0.079	-0.184±0.252	0.340±0.091
100	0.933±0.026	-0.080±0.306	0.372±0.108	0.000±0.387	0.425±0.124
200	0.934±0.014	-0.279±0.150	0.359±0.060	-0.240±0.189	0.407±0.068
500	0.955±0.032	-0.118±0.340	0.407±0.126	-0.035±0.426	0.458±0.142
1000	0.940±0.017	-0.250±0.200	0.332±0.070	-0.209±0.247	0.374±0.079
10000	0.954±0.015	-0.207±0.169	0.383±0.062	-0.148±0.208	0.428±0.069
	Case IIIb				
n	$A(n)$	$B(n)$	$C(n)$	$D(n)$	$E(n)$
30	1.096±0.047	-1.079±0.395	0.896±0.156	-0.913±0.291	0.700±0.120
50	1.115±0.040	-1.150±0.353	0.910±0.146	-0.957±0.260	0.706±0.112
100	1.139±0.033	-1.274±0.306	0.955±0.137	-1.040±0.223	0.737±0.104
200	1.125±0.043	-1.135±0.380	0.918±0.154	-0.939±0.277	0.707±0.117
500	1.138±0.028	-1.209±0.252	0.952±0.107	-0.988±0.183	0.730±0.081
1000	1.137±0.023	-1.193±0.211	0.992±0.088	-0.973±0.153	0.759±0.067

6.4. The modified critical values as a function of n and eta

In this section both the sample size, n, and truncation dependence, η, are combined to give one formula for the critical values as a function of n and η.

Case II and Case IIIb that both are sensitive to the truncation parameters. The critical values in Tables 4 and 6 can be fitted to the two dimensional function

$$D^q_{cv}(\eta, n) \quad = \quad A + \frac{B}{\sqrt{n}} + C\,\sqrt{\eta} + D\,\frac{\sqrt{\eta}}{\sqrt{n}} + E\,\eta + F\,\eta^{3/2}\,, \tag{28}$$

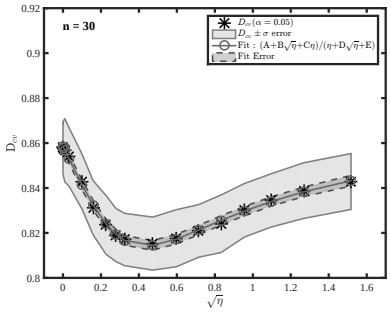
and the fit results are given in Table 10.

Table 10: The fit parameters in Equation (28) are presented here for Cases II and IIIb.

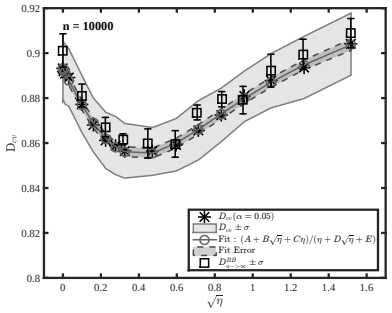
Case II					
A	B	C	D	E	F
0.894±0.001	-0.196±0.01	-0.178±0.007	-0.096±0.02	0.263±0.012	-0.092±0.006
Case IIIb					
A	B	C	D	E	F
1.311± 0.003	-0.164± 0.028	0.187±0.17	-0.036±0.041	-0.495±0.028	0.198±0.013

6.5. Exploring CV's for the dependence of Weibull parameter ranges

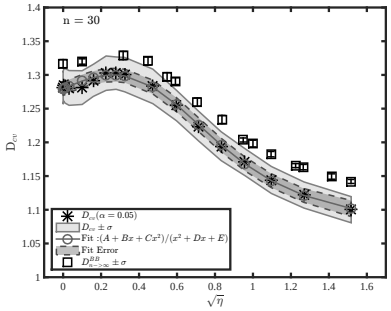
This section numerically explores the effects of the range of the Weibull parameters on the critical values as discussed in section 3. For this purpose, we consider various combinations of the scale parameter $\alpha = 1, 400, 1000, 2000$ and shape parameter, $\beta = 0.2, 0.35, 0.58, 0.8, 1$. The results are displayed in Figure 5, where the critical values plotted as a function η for sample sizes $n = 30(left)$ for Case II, Case IIIb. All the curves for different parameter combinations overlap with each other to show the insensitivity to different parameter values. In Case I and IIIa the CV's are independent of parameter, as is well known.



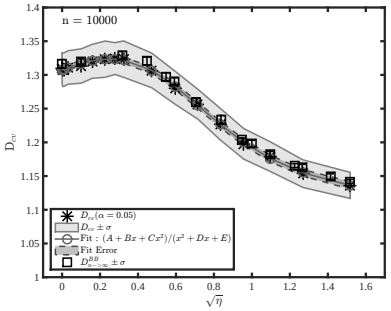
(a) Both α and β are unknown – Case II.



(b) Both α and β are unknown – Case II.



(c) β is unknown and α is known – Case IIIb.

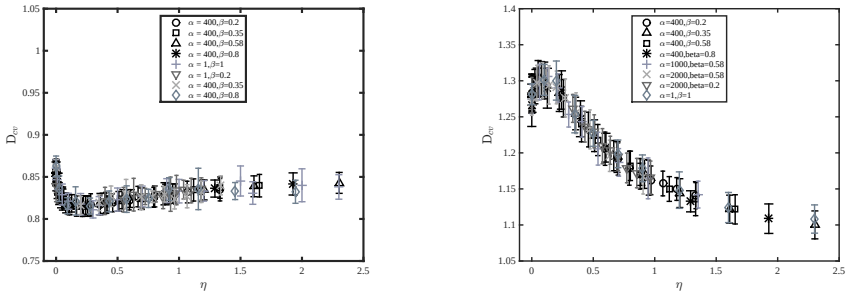


(d) β is unknown and α is known – Case IIIb

Figure 4: Critical values obtained from the quantile analysis and their fits are plotted as a function of $\sqrt{\eta}$ for a sample sizes $n = 30$ (left), $n = 10,000$ (right).

6.6. Comparison of the results with literature

Comparison of our CV's with those already published are shown in Tables 11-14. We can see that there is excellent agreement. All of the previous studies in the literature only considered complete (untruncated) data, whereas our study considers a range of truncations, including the untruncated case. Therefore, we can only compare the complete case results with the literature. Also, we wish to remind the reader that the Weibull distribution is a special case of the generalised extreme value distribution.



(a) n=30 – Case II. (b) n=30 – Case IIIb.

Figure 5: Critical values versus η for various combinations of α and β. $\alpha = 1, 400, 1000, 2000$, $\beta = 0.2, 0.35, 0.58, 0.8, 1$ ranges for $n = 30$.

Table 11: Comparison of our results with the available literature for Case I (α and β are known). To the best of our knowledge, no data is available for the CV's of left-truncated Weibull distribution. For *complete* data sets, ($\tau_L = 0 = \eta = 0 = p = 0$, our error is ± 0.025).

Authors	Estimation	Distribution	n	$D_{cv}(95\%)$	Our Results
Smirnov 1948 Smirnov (1948)	-	all	∞	1.36	1.356
Massey 1951 Massey (1951)	-	all	30	1.32	1.323
	-	all	∞	1.36	1.356
Birnbaum 1952 Birnbaum (1952)	-	all	30	1.3238	1.323
	-	all	50	1.3322	1.329
	-	all	100	1.3400	1.337
	-	all	∞	1.3581	1.356
Miller 1956 Miller (1956)	-	all	30	1.324	1.323
	-	all	50	1.332	1.329
	-	all	100	1.340	1.337
	-	all	∞	1.358	1.356

Table 12: Comparison of our results with the available literature for Case II (α and β are unknown). To the best of our knowledge, no CV's of left-truncated Weibull distributions are available. For *complete* data sets, ($\tau_L = 0 = \eta = 0 = p = 0$, our error is ± 0.015).

Authors	Estimation	Distribution	n	$D_{cv}(95\%)$	Our Results
Littell et al. 1979 Littell *et al.* (1979)	MLE	Weibull	30	0.854	0.858
Parsons & Wirsching 1982 Parsons and Wirsching (1982)	MLE	Weibull	30	0.854	0.858
	MLE	Weibull	∞	0.865	0.896
Chandra et al. 1981 Chandra *et al.* (1981)	MLE	extreme value	50	0.856	0.865
	MLE	extreme value	∞	0.874	0.896
D'Agostino & Stephens 1986Agostino and Stephens (1986)	MLE	extreme value	50	0.856	0.865
	MLE	extreme value	∞	0.874	0.896
Evans et al. 1989 Evans, Johnson, and Green (1989)	MLE	Weibull	30	0.8599	0.858
	MLE	Weibull	50	0.8697	0.865
	MLE	Weibull	100	0.8740	0.874
	MLE	Weibull	200	0.8796	0.881
	MLE	Weibull	∞	0.8982	0.896

Table 13: Comparison of our results with the available literature for Case IIIa (α unknown and β known). To the best of our knowledge, no CV's of left-truncated Weibull distributions are available. For *untruncated (complete)* data sets, ($\tau_L = 0 = \eta = 0 = p = 0$, our error is ± 0.020).

Authors	Estimation	Distribution	n	$D_{cv}(95\%)$	Our Results
Lilliefors 1969 Lilliefors (1969)	MLE	Exponential	30	1.052	1.055
	MLE	Exponential	∞	1.060	1.093
Durbin 1975 Durbin (1975)	MLE	Exponential	30	1.0580	1.055
	MLE	Exponential	50	1.0668	1.065
	MLE	Exponential	100	1.0753	1.073
Chandra et al. 1981 Chandra *et al.* (1981)	MLE	extreme value	50	1.067	1.064
	MLE	extreme value	∞	1.094	1.093
Woodruff et al 1983 Woodruff, Moore, Dunne, and Cortes (1983)	MLE	Weibull	30	1.057	1.055
D'Agostino & Stephens 1986 Agostino and Stephens (1986)	MLE	Exponential	50	1.061	1.065
	MLE	Exponential	100	1.072	1.073
	MLE	Exponential	∞	1.094	1.093
Shorack & Wellner p 239 Shorack and Wellner (2009)	MLE	Exponential	∞	1.094	1.093

Table 14: Comparison of our results with the available literature for Case IIIb (α known and β unknown). To the best of our knowledge, no CV's of left-truncated Weibull distributions are available. For *untruncated (complete)* data sets, ($\tau_L = 0 = \eta = 0 = p = 0$, our error is ± 0.025).

Authors	Estimation	Distribution	n	$D_{cv}(95\%)$	Our Results
D'Agostino & Stephens 1986 Agostino and Stephens (1986)	MLE	extreme value	50	1.29	1.289
	MLE	extreme value	∞	1.29	1.317

7. Interpretation and evaluation of results

7.1. The eta-parameter in practical applications

If α and β are unknown then p and hence η are estimated from the sample so that $\hat{\eta} = \left(\frac{\tau_L}{\hat{\alpha}}\right)^{\hat{\beta}}$ and $\hat{p} = 1 - e^{-\hat{\eta}}$. As $\hat{\eta}$ is a non-linear function of $\hat{\alpha}$ and $\hat{\beta}$ then $\hat{\eta}$ will be a biased estimate of η. As discussed in Appendix B, for a sample size n the bias in η is defined as

$$\mathbb{E}\left[\Delta\hat{\eta}\right] = \mathbb{E}\left[\hat{\eta} - \eta\right] = \mathbb{E}\left[\hat{\eta}\right] - \eta\,, \tag{29}$$

so that an unbiased estimate of η is given in Appendix B by Equation (32) in conjunction with Equation (36). Estimated $\hat{\eta}$ and corrected (unbiased) $\hat{\eta}$ values for various sample sizes and truncation levels are given in the Tables 22 and 23 for Case II and Case IIIb respectively. Making use of these tables, we demonstrate the passing rates with and without the bias-correction of $\hat{\eta}$ in Tables 15 and 16 for Case II and Case IIIb, respectively. For large sample size the bias vanishes in accordance with Theorem 1 in Kreer *et al.* (2015). Furthermore, for small truncation parameters η the bias is of no relevance. Only for small sample sizes ($n = 30, 50, 100$) and truncation levels, p above 0.7 does the correction (unbiasing) formula need to be applied.

Table 16: Percentage pass rates in KS-test with and without $\hat{\eta}$-correction for 10000 simulations in Case IIIb (error is less than ±0.5%).

p	η	n = 30		n=50		n=100		n=1000	
		$\hat{\eta}$ uncorrected	$\hat{\eta}$ corrected	$\hat{\eta}$ uncorrected	$\hat{\eta}$ corrected	$\hat{\eta}$ uncorrected	$\hat{\eta}$ corrected	$\hat{\eta}$ uncorrected	$\hat{\eta}$ corrected
0	0	95.6	95.6	95.2	95.2	95.0	95.0	95.2	95.2
0.1	0.1	94.7	94.7	95.0	95.0	95.0	95.0	95.2	95.2
0.2	0.2	94.8	94.8	94.6	94.6	94.9	94.9	94.5	94.5
0.3	0.35	95.1	95.0	95.0	94.9	95.1	95.0	94.9	94.9
0.4	0.5	95.7	95.6	95.0	95.0	95.4	95.4	95.0	95.0
0.5	0.7	95.2	95.2	95.4	95.4	95.5	95.5	95.4	95.4
0.6	0.9	95.3	95.3	95.5	95.5	95.2	95.2	95.4	95.4
0.7	1.2	95.2	95.2	95.1	95.1	95.3	95.3	94.9	94.9
0.8	1.6	94.6	94.6	95.0	95.0	94.2	94.2	94.8	94.8
0.9	2.3	95.4	95.4	94.9	94.9	95.5	95.5	95.2	95.2

Table 15: Percentage pass rates in KS-test with and without $\hat{\eta}$-correction for 10000 simulations in Case II (error is less than ±0.5%).

p	η	n = 30		n=50		n=100		n=1000	
		$\hat{\eta}$ uncorrected	$\hat{\eta}$ corrected	$\hat{\eta}$ uncorrected	$\hat{\eta}$ corrected	$\hat{\eta}$ uncorrected	$\hat{\eta}$ corrected	$\hat{\eta}$ uncorrected	$\hat{\eta}$ corrected
0	0	95.3	95.3	95.0	95.0	95.2	95.2	95.4	95.4
0.1	0.1	95.1	95.1	94.9	94.9	94.8	94.8	95.5	95.5
0.2	0.2	95.2	95.2	95.2	95.2	94.9	94.9	95.1	95.1
0.3	0.35	94.8	94.8	94.8	94.8	94.8	94.8	94.6	94.6
0.4	0.5	94.2	94.3	94.7	94.6	94.8	94.8	94.6	94.6
0.5	0.7	94.3	94.7	94.4	94.3	95.1	95.0	94.6	94.6
0.6	0.9	93.6	94.5	94.3	94.2	94.9	94.8	95.0	95.0
0.7	1.2	90.5	94.1	93.8	94.7	94.6	94.6	95.2	95.2
0.8	1.6	86.7	94.3	90.7	94.3	93.9	94.6	95.1	95.1
0.9	2.3	76.7	94.6	81.6	94.3	88.6	94.4	94.8	94.9

7.2. Power studies: comparison with other distributions in Case II

In order to answer the question " What is the chance that data drawn from some alternative distribution will pass the hypothesis test for a Weibull distribution?", the power test is employed.

We compare the power of our out-sample (Case I) and in-sample (Case II) tests by drawing the random numbers of our samples from alternative distributions commonly used in the literature for making goodness-of-fit comparisons. We follow Aho, Bain, and Engelhardt (1985) and consider as possible alternatives to the 2-parameter Weibull distribution, those distributions defined on the positive range. In particular, we consider the log-normal, log-Cauchy, Pareto (power law), log-double exponential, log-logistic and chi-square distributions with 1, 3 and 4 degrees of freedom (note that the chi-square distribution with 2 deegrees of freedom is the exponential and thus not in the scope here). We consider the chi-square distributions with 1, 3 and 4 degrees of freedom as academic only, as they only permit one to fit one single parameter, i.e. the degree of freedom k. As noted earlier by Aho *et al.* (1985), for the complete data set our test performs well for log-Cauchy, Pareto, log-double-exponential and log-logistic, namely one can rule out these distributions as candidates explaining the data set. On the other hand, we found that the power-testing does have problems ruling out χ^2-distributions with 1, 3 and 4 degrees of freedom and log-normal distributions. The latter can be ruled out by a likelihood ratio test in the spirit of Dumonceaux and Antle (1973). The results are summarized in Table 17 for the complete and the truncated Case I and Case II.

Table 17: Summary of in-sample KS-test, truncation rate $p = 0, 0.5$ for Case I and Case II, number of simulations, $N = 1000$.

		Case I		Case II	
		$p = 0$	$p = 0.5$	$p = 0$	$p = 0.5$
distribution	sample size n	pass rate %	pass rate %	pass rate %	pass rate %
Weibull2d	30	96	96	93	93
Weibull2d	100	95	96	93	93
Weibull2d	500	-	-	95	97
log-Cauchy	30	50	4	6	3
log-Cauchy	100	2	1	0	0
log-Cauchy	500	-	-	0	0
log-double exp.	30	57	42	39	62
log-double exp.	100	6	1	3	52
log-double exp.	500	-	-	0	43
log-logistic	30	46	1	63	85
log-logistic	100	1	0	16	85
log-logistic	500	-	-	0	56
log-normal	30	55	65	73	93
log-normal	100	4	17	30	93
log-normal	500	-	-	0	89
Pareto	30	0	1	1	42
Pareto	100	0	0	0	52
Pareto	500	-	-	0	44
chi-square(k=1)	30	56	78	92	95
chi-square(k=1)	100	8	40	81	96
chi-square(k=1)	500	-	-	43	94
chi-square(k=3)	30	0	3	93	89
chi-square(k=3)	100	0	0	92	95
chi-square(k=3)	500	-	-	81	95
chi-square(k=4)	30	0	0	92	90
chi-square(k=4)	100	0	0	87	88
chi-square(k=4)	500	-	-	63	87

8. Application of our modified KS test

8.1. US data on duration of ethnically mixed marriages

Data on the duration of marriages that end in divorce in the US is publicly available at (http://data.princeton.edu/wws509/datasets/#divorce). Most states in the United States require a minimum legal separation time prior to divorce, although not all do. The duration of marriages that ultimately end in a divorce in the database will therefore contain a mixture of those with a minimum duration (from 0 to 12 months). In order to determine the distribution that describes the duration of failed marriages in the US, it is therefore necessary to left-truncate the data.

We have taken a subset of 230 divorced couples where husband and wife belong to different ethnic groups. We then analyze the duration of the marriages for a range of left-truncation values, specifically $\tau_L = 0.25$, 1, 5 and 10 years in Table 18. We observe from the data also that the smallest life time is bigger than 0.25 years. This is further evidence that the data is left-truncated. Before starting our Weibull analysis, we firstly generate a Q-Q plot for the most commonly used alternatives: Weibull, Pareto and log-normal distribution. In our case the Pareto distribution can clearly be singled out by purely looking at its curved graph in the Q-Q plot. To decide for either Weibull or log-normal is more delicate as both graphs in the Q-Q plot are more or less straight lines. Here we use a likelihood ratio test as proposed firstly by Dumonceaux and Antle (1973) for the discrimination between (un-truncated) log-normal and (un-truncated) Weibull distributions. As their table covers only sample sizes of $n = 20, 30, 40, 50$ we had to extend it to sample sizes $n = 100, 200, 300$. The likelihood ratio test gives a clear verdict in favor of the Weibull distribution.[3]

In the following Weibull analysis, truncation rates p are given as percentage of data which have been eliminated by the truncation procedure. From the estimated parameters $\hat{\alpha}$ and

[3] The results of Dumonceaux and Antle (1973) have been modified by the authors to account also for left-truncation.

$\hat{\beta}$ we got $\hat{\eta}$ as estimator for our critical value using Equation (28) and the KS distance D_n is calculated from the data using Equation (12). Due to moderate truncation levels we do not need to un-bias the value of $\hat{\eta}$. Hence, we can not reject the hypothesis, that the data come from a left-truncated Weibull distribution for a wide range of truncation levels with $\beta = 1.25 \pm 0.07$ and $\alpha = 11.4 \pm 0.06$ years. The details of this analysis can be seen in the Table 18.

Table 18: Duration of ethnically mixed marriages ending in divorce in the US. y indicates year as a unit.

τ_L [y]	n	$\hat{\alpha}$ [y]	$\hat{\beta}$	$\hat{\eta}$	p[%]	D_n	$D_{cv}(n, \hat{\eta}, 0.05)/\sqrt{n}$	H_0
-	230	11.5 ± 0.6	1.29 ± 0.07	0.00 ± 0.00	0.0	0.0502	0.0581	Accept
0.25	230	11.3 ± 0.6	1.25 ± 0.07	0.01 ± 0.00	0.0	0.0437	0.0570	Accept
1	222	11.2 ± 0.7	1.24 ± 0.07	0.05 ± 0.01	3.5	0.0425	0.0572	Accept
5	157	11.4 ± 0.8	1.24 ± 0.08	0.36 ± 0.02	31.7	0.0551	0.0672	Accept
10	96	14.0 ± 0.9	1.50 ± 0.11	0.60 ± 0.02	58.3	0.0626	0.0861	Accept

8.2. Time between major terrorist attacks with minimum 10 casualties

The worldwide probability distribution of terrorist attacks has been investigated by Clauset and Woodard (2013). We utilize the RAND-MIPT database (available at http://www.rand.org/nsrd/projects/ terrorism-incidents/download.html) containing 13,274 terrorist events worldwide from 1968 to 2007. Like Clauset and Woodard (2013) we are interested in "major attacks", defined as terrorist events with at least 10 casualties. We investigate the times between these major attacks and find that a large proportion of their tail can be described as left-truncated Weibull. From the estimated parameters $\hat{\alpha}$ and $\hat{\beta}$ we get $\hat{\eta}$ as estimator for our critical value using Equation (28) and the KS distance D_n is calculated from the data using Equation (12). Results are given in Table 19. We note that the tail of the distribution can be described by a Weibull distribution with shape parameter $\beta \simeq 0.50$ whereas the short-end is described by something else and does not pass the Weibull hypothesis.

Table 19: Time between major terrorist attacks with minimum 10 casualties. d indicates day as a unit.

τ_L [d]	n	$\hat{\alpha}$ [d]	$\hat{\beta}$	$\hat{\eta}$	p[%]	D_n	$D_{cv}(n, \hat{\eta}, 0.05)/\sqrt{n}$	H_0
-	926	9.1 ± 0.5	0.61 ± 0.02	0.00 ± 0.00	0.0	0.2292	0.0290	Decline
10	204	12.7 ± 2.0	0.48 ± 0.03	0.89 ± 0.07	78.0	0.0491	0.0604	Accept
12	187	12.6 ± 2.0	0.48 ± 0.03	0.98 ± 0.07	79.8	0.0426	0.0632	Accept
14	173	12.5 ± 2.1	0.48 ± 0.03	1.06 ± 0.08	81.3	0.0447	0.0659	Accept
16	161	12.2 ± 2.1	0.47 ± 0.03	1.14 ± 0.08	82.6	0.0465	0.0684	Accept
18	148	15.6 ± 2.7	0.50 ± 0.03	1.08 ± 0.08	84.0	0.0526	0.0711	Accept
20	140	14.4 ± 2.6	0.49 ± 0.03	1.17 ± 0.08	84.9	0.0539	0.0733	Accept
22	132	14.2 ± 2.7	0.49 ± 0.03	1.24 ± 0.09	85.7	0.0557	0.0756	Accept
24	124	15.9 ± 3.0	0.51 ± 0.04	1.23 ± 0.09	86.6	0.0588	0.0779	Accept

8.3. Stock market data

We investigate the difference in arrival times between consecutive orders at the New York Stock Exchange (NYSE) for a given stock. The free data provided by www.tickdata.com comprises the entire trading day of shares of ITT Corp. on 11 January 2011, from 9:30 to 16:00 EST. For this example we only look at a snapshot from 12:00:00 to 12:00:21 EST, i.e. 21 seconds of data. The resolution of the arrival times is milliseconds.

Truncation of arrival time differences is the process of taking the differences between consecutive arrival times and keeping only those with differences greater than $\tau_L = 1, 2, 5, 10$ milliseconds. As we did in the previous examples, having singled out the alternatives of Pareto and log-normal distribution, we estimate the Weibull parameters and perform the hypothesis test; the results are given in Table 20.

Table 20: Arrival times of for ITT Corp. orders on NYSE on 11 Jan. 2011, 12:00:00-12:00:21.

τ_L [ms]	n	$\hat{\alpha}$ [ms]	$\hat{\beta}$	$\hat{\eta}$	p[%]	D_n	$D_{cv}(n,\hat{\eta},0.05)/\sqrt{n}$	H_0
-	100	-	-	0	0	-	0.0874	Decline
1	61	128 ± 38	0.46 ± 0.05	0.1073	39%	0.0684	0.1064	Accept
2	57	169 ± 46	0.51 ± 0.05	0.1041	43%	0.0734	0.1100	Accept
5	54	190 ± 49	0.55 ± 0.06	0.1352	46%	0.0797	0.1127	Accept
10	51	179 ± 50	0.53 ± 0.06	0.2168	49%	0.0839	0.1155	Accept

From Table 20 we see that we can not reject the hypothesis that our truncated samples come from a Weibull distribution. However when we analyse the complete (untruncated) sample we see by a similar computation that it leads to the rejection of the Weibull hypothesis as the zero-inflated data with arrival time differences below 1 millisecond prevent the MLE converging onto a solution. One millisecond truncation seems to corrupt the estimation of the Weibull parameters due to the error in time measurement of ± 1 millisecond. From 2 millisecond truncation onwards one finds consistent parameter estimation. Taking the weighted means and errors from the truncated data sets with truncations of 2, 5 and 10 milliseconds we find for the parameters $\hat{\alpha} = 179 \pm 37$ milliseconds and $\hat{\beta} = 0.53 \pm 0.04$.

8.4. Time intervals for radioactive decay of Americium-241

Since the pioneering work of Geiger and Rutherford (1910) the counting process of the particles arising from radioactive decay have been found to be described by a Poisson process. Due to the so-called "dead time" of the detection device, certain decay events might not be measured because the detector is still busy with "detecting" the previous event. Thus, the data set will be incomplete due to "truncation". This has given rise to certain corrections for the Poisson process. Only 60 years later it was possible to measure waiting times between radioactive decay events with acceptable accuracy using multichannel analyzers. Garfinkel and Mann (1968) did one of the first measurement using a probe of 0.2 μCi Americium-231 as a nearly pure α-source Their entire data set, comprising some 300'000 time intervals, was evaluated later by Berkson (1975) albeit under the assumption of a Poisson process and performing a χ^2-test on the bin-ed data. Here, we want to demonstrate our analysis of a smaller sample which is displayed in Garfinkel and Mann (1968) on page 709. We use the second, third and fourth block only because the first block contains some control measurements for calibration. Our data sample comprises 300 measurement points describing the time between subsequent α-particles. The dead time was estimated by the authors to be 2.54 T.U.(1 T.U. denotes a time unit and corresponds to the pulse frequency of 370 kHz). Our results are displayed in Table 21. We recover as expected a shape parameter $\beta = 1$ indicating that the waiting times are exponentially distributed giving rise to the Poisson process discovered in Geiger and Rutherford (1910).

Table 21: Time intervals for radioactive decay of Americium-241. *T.U.* indicates time unit.

τ_L [T.U.]	n	$\hat{\alpha}$ [T.U.]	$\hat{\beta}$	$\hat{\eta}$	p[%]	D_n	$D_{cv}(n,\hat{\eta},0.05)/\sqrt{n}$	H_0
-	300	15605 ± 947	1.00 ± 0.05	0.00 ± 0.00	0.0	0.0491	0.0510	Accept
3	300	15596 ± 947	1.00 ± 0.05	0.00 ± 0.00	0.0	0.0493	0.0508	Accept
10	300	15576 ± 948	1.00 ± 0.05	0.00 ± 0.00	0.0	0.0498	0.0507	Accept
100	298	15597 ± 951	1.00 ± 0.05	0.01 ± 0.01	0.6	0.0500	0.0504	Accept

9. Conclusion

The Weibull distributions with a shape parameter less than one is known as "heavy-tailed" because it has significant probabilities quite far from its mean. In insurance and other industries the cost of rare events due to "heavy tails" can be very high, so it is important to determine exactly how rare they actually are. This can only be done by taking the available data and testing it against hypothesized distributions.

Data obtained from real life examples are often left-truncated. To test the hypothesis that the data are sampled from a left-truncated Weibull distribution, one can perform a Kolmogorov-Smirnov goodness-of-fit test. If the shape and scale parameters are not known they must be estimated from the data itself. The commonly used maximum likelihood estimator does not always give a non-trivial solution to estimating the shape and scale parameters, especially for small sample sizes. For a small sample size there is a chance that the solution of the maximum likelihood estimate lie on the trivial boundaries where either one or both of the parameters vanish. A criterion for determining when non-vanishing solutions for the parameters exist was given in this paper. We demonstrated also that with increasing sample size non-trivial estimates exist with probability tending to one and these estimates are consistent, asymptotically normal, and efficient. Having obtained non-trivial estimates, a goodness-of-fit can be judged using a Kolmogorov-Smirnov test. If either the shape and/or scale parameters are unknown the critical values differ significantly from those when the parameters are known. If both the parameters or only the shape parameter are unknown the critical values depend on the truncation value as well the number of data.

The modified critical values presented here should be used to test if a set of data is sampled from a left-truncated Weibull distribution with a known truncation point but unknown shape and/or scale parameter. When both the parameters or only the shape parameter are unknown and the truncation level is greater than 10%, then the dependence of the critical value on the truncation level must be included, otherwise incorrect conclusions from the hypothesis tests will be drawn. We provided the modified CVs in Tables (3) - (6) for various sample sizes and truncation ranges and also formulas Equation (27) and Equation (28) where one can calculate them for any desired p (or η) for given n and for combination of (p (or η), n), respectively.

Although the results presented here on the left truncated Weibull distribution can be applied to a wide range of applications in many disciplines we are not aware of any other comprehensive studies that discuss the effects of truncation dependence on the critical values and parameter estimation. We are in the process of applying our techniques to investigate financial, insurance, and real estate data using our tables and models for the critical values which include the dependence on truncation and sample size.

10. Acknowledgement(s)

The authors are grateful to Ross Frick from University of South Australia for a careful reading of the manuscript which led to significant improvements. This work was supported by the University of Adelaide and by the Australian Research Council (AWT, FL0992247).

A. Left-truncated Weibull random variates and their representation by exponential variates

Let $u_i \in (0, 1)$ denote the standard uniform random variable. Then from the cdf in equation (1) we obtain for the left-truncated Weibull random variable τ_i

$$\tau_i = \alpha \cdot \left[\left(\frac{\tau_L}{\alpha} \right)^{\beta} + \log \frac{1}{u_i} \right]^{1/\beta} = \alpha \cdot [\eta + y_i]^{1/\beta} \tag{30}$$

where y_i is a standard exponentially distributed random variable and $\eta \equiv (\tau_L/\alpha)^{\beta}$.

B. Bias in estimates of eta

The estimated value of η (i.e. $\hat{\eta} = (\tau_L/\hat{\alpha})^{\hat{\beta}}$), will have an estimation error $\Delta\hat{\eta}$. In this sense,

both $\hat{\eta}$ and $\Delta\hat{\eta}$ are random variables whereas η is a fixed real number:

$$\eta = \hat{\eta} - \Delta\hat{\eta} \quad \Rightarrow \quad \eta = \mathbb{E}[\hat{\eta}] - \mathbb{E}[\Delta\hat{\eta}]\,. \tag{31}$$

By definition $\eta \geq 0$ hence we use an un-biasing formula motivated by Equation (31)

$$\eta = \max\{0, \hat{\eta} - \mathbb{E}[\Delta\hat{\eta}]\} \tag{32}$$

where the individual $\hat{\eta}$ is unbiased by a correction term $\mathbb{E}[\Delta\hat{\eta}]$ subject to $\eta \geq 0$.

Defining the parameter estimation vector (suppressing the index n) as $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta})$ and the true parameter vector as $\boldsymbol{\theta^0} = (\alpha^0, \beta^0)$, from Section 2 for large sample size n the difference $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta^0})$ is asymptotically normal with vector mean zero and covariance matrix $Z^{-1}(\boldsymbol{\theta^0})$, the inverse of the Fisher matrix Equation (11). Thus

$$\Delta\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta^0} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{n}\boldsymbol{Z}^{-1}\right) \quad . \tag{33}$$

To estimate the effect of errors in η due to errors in $\hat{\boldsymbol{\theta}}$ we write similarly

$$\hat{\alpha} = \alpha^0 + \Delta\hat{\alpha} \qquad \hat{\beta} = \beta^0 + \Delta\hat{\beta} \tag{34}$$

$$\text{where} \qquad \mathbb{E}\left[\begin{bmatrix}\Delta\hat{\alpha}\\ \hat{\Delta}\beta\end{bmatrix}[\Delta\hat{\alpha} \quad \hat{\Delta}\beta]\right] = \begin{bmatrix}\sigma^2_{\hat{\alpha}} & \sigma_{\hat{\alpha}\hat{\beta}}\\ \sigma_{\hat{\alpha}\hat{\beta}} & \sigma^2_{\hat{\beta}}\end{bmatrix} = \boldsymbol{Z}^{-1}\,. \tag{35}$$

Here, we have used Equation (33) to calculate the expectation. Then the Taylor expansion of $\Delta\hat{\eta}$ gives

$$\begin{aligned}
\Delta\hat{\eta} &= \frac{\partial\hat{\eta}}{\partial\hat{\alpha}}\,\Delta\hat{\alpha} + \frac{\partial\hat{\eta}}{\partial\hat{\beta}}\,\Delta\hat{\beta} + \frac{1}{2}\,\frac{\partial^2\hat{\eta}}{\partial\hat{\alpha}^2}\,\Delta\hat{\alpha}^2 + \frac{1}{2}\,\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}^2}\,\Delta\hat{\beta}^2 + \frac{\partial^2\hat{\eta}}{\partial\hat{\alpha}\,\partial\hat{\beta}}\,\Delta\hat{\alpha}\,\Delta\hat{\beta} + \cdots\\
\mathbb{E}[\Delta\hat{\eta}] &= \frac{1}{2}\,\frac{\partial^2\hat{\eta}}{\partial\hat{\alpha}^2}\,\sigma^2_{\hat{\alpha}} + \frac{1}{2}\,\frac{\partial^2\hat{\eta}}{\partial\hat{\beta}^2}\,\sigma^2_{\hat{\beta}} + \frac{\partial^2\hat{\eta}}{\partial\hat{\alpha}\,\partial\hat{\beta}}\,\sigma_{\hat{\alpha}\,\hat{\beta}} + \cdots\\
&= \hat{\eta}\left\{\left[\frac{1}{2\hat{\alpha}^2}\,\hat{\beta}(1+\hat{\beta})\right]\sigma^2_{\hat{\alpha}} + \left[\frac{1}{2\hat{\beta}^2}\,\log\hat{\eta}^2\right]\sigma^2_{\hat{\beta}} - \frac{1}{\hat{\alpha}}\,[1+\log\hat{\eta}]\,\sigma_{\hat{\alpha}\,\hat{\beta}} + \cdots\right\}\,.
\end{aligned} \tag{36}$$

Note that in Equation (36) we take the expectations only over the $\Delta\hat{\alpha}$ and $\Delta\hat{\beta}$ but not over the estimates $\hat{\alpha}$ or $\hat{\beta}$. Estimated $\hat{\eta}$ (uncorrected) values and corrected (unbias) $\hat{\eta}$ using Equation (36) for various sample sizes and truncation levels are given in Table 22 for Case II and in Table 23 for Case IIIb .

Table 22: Estimated $\hat{\eta}$ and unbiased $\hat{\eta}$ for 10,000 simulations for sample sizes $n = 30, 50, 100, 1000$ in Case II.

p	η	n = 30		n=50		n=100		n=1000	
		$\hat{\eta}$	$\hat{\eta} - unbias$	$\hat{\eta}$	$\hat{\eta} - unbias$	$\hat{\eta}$	$\hat{\eta} - unbias$	$\hat{\eta}$	$\hat{\eta} - unbias$
0	0	0	0	0	0	0	0	0	0
0.1	0.1	0.106	0.106	0.103	0.104	0.102	0.102	0.100	0.100
0.2	0.2	0.212	0.203	0.205	0.203	0.202	0.201	0.200	0.200
0.3	0.35	0.382	0.327	0.360	0.343	0.355	0.349	0.351	0.350
0.4	0.5	0.583	0.433	0.528	0.476	0.512	0.498	0.502	0.501
0.5	0.7	0.895	0.548	0.767	0.629	0.719	0.677	0.701	0.698
0.6	0.9	1.291	0.647	1.040	0.744	0.946	0.844	0.902	0.897
0.7	1.2	2.084	0.763	1.496	0.894	1.300	1.059	1.202	1.192
0.8	1.6	3.149	0.900	2.307	1.063	1.827	1.252	1.611	1.587
0.9	2.3	5.629	1.092	4.405	1.245	2.968	1.528	2.319	2.254

Table 23: Estimated $\hat{\eta}$ and unbiased $\hat{\eta}$ for 10,000 simulations for sample sizes $n = 30, 50, 100, 1000$ in Case IIIb.

p	η	n = 30		n=50		n=100		n=1000	
		$\hat{\eta}$	$\hat{\eta} - unbias$	$\hat{\eta}$	$\hat{\eta} - unbias$	$\hat{\eta}$	$\hat{\eta} - unbias$	$\hat{\eta}$	$\hat{\eta} - unbias$
0	0	0	0	0	0	0	0	0	0
0.1	0.1	0.097	0.104	0.097	0.101	0.099	0.101	0.100	0.100
0.2	0.2	0.191	0.201	0.195	0.201	0.197	0.200	0.200	0.200
0.3	0.35	0.034	0.347	0.342	0.348	0.346	0.349	0.350	0.350
0.4	0.5	0.488	0.497	0.493	0.497	0.497	0.499	0.500	0.500
0.5	0.7	0.693	0.697	0.696	0.698	0.698	0.699	0.700	0.700
0.6	0.9	0.898	0.899	0.898	0.899	0.899	0.900	0.900	0.900
0.7	1.2	1.205	1.202	1.203	1.201	1.201	1.201	1.200	1.200
0.7	1.6	1.608	1.603	1.608	1.604	1.604	1.602	1.600	1.600
0.9	2.3	2.288	2.281	2.297	2.292	2.302	2.298	2.301	2.301

References

Agostino R, Stephens MA (1986). *Goodness-of-fit Techniques.* M. Dekker, New York. ISBN 0-8247-7487-6.

Aho M, Bain LJ, Engelhardt M (1985). "Goodness-of-fit Tests for the Weibull Distribution with Unknown Parameters and Heavy Censoring." *Journal of Statistical Computation and Simulation.*, **21(3-4)**, 213–225.

Anderson T, Stephens M (1997). "The Continuous and Discrete Brownian Bridges: Representations and applications." *Linear Algebra and its Applications*, **264 (6)**, 145–171.

Balakrishnan N, Cohen A (1991). *Order Statistics and Inference: Estimation Methods.* Academic Press, Boston.

Balakrishnan N, Mitra D (2012). "Left Truncated and Right Censored Weibull Data and Likelihood Inference with an Illustration." *Computational Statistics & Data Analysis*, **56**, 4011 – 4025.

Barr DR, Davidson T (1973). "A Kolmogorov-Smirnov Test for Censored Samples." *Technometrics*, **15(4)**, 739-757.

Berkson J (1975). "Do Radioactive Decay Events Follow a Random Poisson-Exponential?" *International Journal of Applied Radiation and Isotopes,*, **26(9)**, 543–549.

Birnbaum FJ (1952). "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size." *Journal of the American Statistical Association*, **47**(259), 425–441.

Chandra M, Singapurwalla N, Stephens M (1981). "Kolmogorov Statistics for Tests of Fit for the Extreme Value and Weibull Distributions." *Journal of the American Statistical Association*, **76 (375)**, 729–731.

Clauset A, Woodard R (2013). "Estimating the Historical and Future Probabilities of Large Terrorist Events." *Annals of Applied Statistics*, **7 (4)**, 1838–1865.

Cohen CA (1965). "Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and on Censored Samples." *Technometrics*, **7 (4)**, 579–588.

David F, Johnson N (1948). "The Probability Integral Transformation When Parameters are Estimated from the Sample." *Biometrika*, **35 (1/2)**, 182–190.

Dufour R, Maag UR (1978). "Distribution Results for Modified Kolmogorov-Smirnov Statistics for Truncated or Censored Samples." *Technometrics*, **20(1)**, 29–32.

Dumonceaux, R, Antle CE J (1973). "Discrimination between the Log-normal and the Weibull Distributions." *Techometrics*, **15 (4)**, 923–926.

Durbin J (1973). "Weak Convergence of the Sample Distribution Function when Parameters Are Estimated." *The Annals of Statistics*, **1 (2)**, 279–290.

Durbin J (1975). "Kolmogorov-Smirnov Tests when Parameters are Estimated with Applications to Tests of Exponentiality and Tests on Spacings." *Biometrica*, **62(1)**, 5–22.

Evans JW, Johnson RA, Green DW (1989). *Two-and Three-parameter Weibull Goodness-of-fit Tests*, volume 493. US Department of Agriculture, Forest Service, Forest Products Laboratory.

Garfinkel S, Mann W (1968). "A Method for Obtaining Large Numbers of Measured Time Intervals in Radioactive Decay." *International Journal of Applied Radiation and Isotopes,*, **19(9)**, 707–709.

Geiger H, Rutherford E (1910). "The Number of α Particles Emitted by Uranium and Thorium and by Uranium Minerals." *Philolophical Magazine*, **20(118)**, 691–698.

Kendall MG, Stuart A (1979). *The Advanced Theory of Statistics - Inference and Relationship*, volume 2. 4th revised edition edition. C. Griffin, London. ISBN 978—0-8-52-64-2 -255- 2.

Koziol JA , Byar DP (1975). "Percentage Points of the Asymptotic Distributions of One and Two Sample K-S Statistics for Truncated or Censored Data." *Technometrics,* **17(4)**,507–510.

Kreer M, Kizilersu A, Thomas AW, dos Reis AE (2015). "Goodness-of-fit Tests and Applications for Left-truncated Weibull Distributions to Non-life Insurance." *European Actuarial Journal*, **5 (1)**. doi:10.1007/s13385-015-0105-8.

Lehmann E, Casella G (1998). *Theory of Point Estimation* -. 2nd ed. 1998. corr. 4th printing 2003 edition. Springer, Berlin, Heidelberg. ISBN 978-0-387-98502-2.

Lilliefors H (1969). "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown." *Journal of the American Statistical Association*, **64 (325)**, 387–389.

Littell RC, McClave JT, Offen WW (1979). "Goodness-of-fit Tests for the Two Parameter Weibull Distribution." *Communications in Statististics - Simulation and Computation*, **8 (3)**, 257–269.

Malevergne Y, Pisarenko V, Sornette D (2005). "Empirical Distributions of Stock Returns: between the Stretched Exponential and the Power Law?" *Quantitative Finance*, **5 (4)**, 379–401.

Massey FJ (1951). "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association*, **46**(253), 68–78.

Miller LH (1956). "Table of Percentage Points of Kolmogorov Statistics." *Journal of the American Statistical Association*, **51**(273), pp. 111–121.

Nadarajah S, Kotz S (2006). "R Programs for Truncated Distributions" *Journal of Statistical Software, Code Snippets*, **16**(2), pp. 1–8.

Parsons F, Wirsching PH (1982). "A Kolmogorov-Smirnov Goodness-of-fit Test for the Two-parameter Weibull Distribution when the Parameters Are Estimated from the Data." *Microelectronics Reliability*, **22(2)**, 163–167.

Rinne H (2009). *The Weibull Distribution: A Handbook*. 1st edition. CRC Press, Taylor & Francis Group, Boca Raton New York. ISBN 978-1-4200-8743-7.

Shorack GR, Wellner JA (2009). *Empirical Processes with Applications to Statistics*. SIAM, Philadelphia. ISBN 978-0-898-71901-7.

Smirnov N (1948). "Table for Estimating the Goodness of Fit of Empirical Distributions." *The Annals of Mathematical Statistics*, **19(2)**, 279–281.

Stephens M (1977). "Goodness of Fit for the Extreme Value Distribution." *Biometrika*, **64 (3)**, 583–588.

Thoman DR, Bain LJ, Antle CE (1969). "Inferences on the Parameters of the Weibull Distribution." *Technometrics*, **11 (3)**, 445–460.

Weibull W (1951). "A Statistical Distribution Function of Wide Applicability." *J. Appl. Mech.-Trans. ASME*, **73**, 293.

Wingo DR (1989). "The Left-truncated Weibull Distribution: Theory and Computation." *Statistical Papers*, **30**, 39–48.

Woodruff BW, Moore AH, Dunne EJ, Cortes R (1983). "A Modified Kolmogorov-Smirnov Test for Weibull Distributions with Unknown Location and Scale Parameters." *IEEE Transactions on Reliability*, **R-32 (2)**, 209–213.

Affiliation:

Ayşe Kızılersü
Special Research Centre for the Subatomic Structure of Matter (CSSM)
Department of Physics
University of Adelaide
5005, Adelaide,Australia
E-mail: ayse.kizilersu@adelaide.edu.au
URL: http://www.physics.adelaide.edu.au/cssm/index.html

Austrian Journal of Statistics
September 2016, Volume 45, 43–69.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i3.107

The Log-logistic Weibull Distribution with Applications to Lifetime Data

Broderick O. Oluyede, Susan Foya, Gayan Warahena-Liyanage, Shujiao Huang
Georgia Southern University, Botswana International University of Science and Technology, Central Michigan University, University of Houston

Abstract

In this paper, a new generalized distribution called the log-logistic Weibull (LLoGW) distribution is developed and presented. This distribution contain the log-logistic Rayleigh (LLoGR), log-logistic exponential (LLoGE) and log-logistic (LLoG) distributions as special cases. The structural properties of the distribution including the hazard function, reverse hazard function, quantile function, probability weighted moments, moments, conditional moments, mean deviations, Bonferroni and Lorenz curves, distribution of order statistics, L-moments and Rényi entropy are derived. Method of maximum likelihood is used to estimate the parameters of this new distribution. A simulation study to examine the bias, mean square error of the maximum likelihood estimators and width of the confidence intervals for each parameter is presented. Finally, real data examples are presented to illustrate the usefulness and applicability of the model.

Keywords: generalized distribution, log-logistic distribution, Weibull distribution, log-logistic Weibull distribution, probability weighted moments, L-moments, maximum likelihood estimation.

1. Introduction

There are several generalizations of univariate distributions including those of (Eugene, Lee, and Famoye 2002) dealing with the beta-normal distribution, as well general family of univariate distributions generated from the Weibull distribution that was introduced by Gurvich, Dibenedetto, and Ranade (1997). The cumulative distribution function (cdf) given by (Gurvich *et al.* 1997) is

$$G(x;\alpha,\boldsymbol{\Theta}) = 1 - \exp[-\alpha H(x;\boldsymbol{\Theta})], \quad x \in \mathcal{C},\ \alpha > 0, \tag{1}$$

where $\mathcal{C}$ is a subset of $\mathbf{R}$, and $H(x;\boldsymbol{\Theta})$ is a non-negative monotonically increasing function that depends on the vector of parameters $\boldsymbol{\Theta}$. The corresponding probability density function (pdf) is given by

$$g(x;\alpha,\boldsymbol{\Theta}) = \alpha \exp[-\alpha H(x;\boldsymbol{\Theta})] h(x;\boldsymbol{\Theta}), \tag{2}$$

where $h(x;\boldsymbol{\Theta})$ is the derivative of $H(x;\boldsymbol{\Theta})$. The choice of the function $H(x;\boldsymbol{\Theta})$ can lead to

different models including for example, exponential distribution with $H(x;\boldsymbol{\Theta}) = x$, Rayleigh distribution is obtained from $H(x;\boldsymbol{\Theta}) = x^2$ and Pareto distribution from setting $H(x;\boldsymbol{\Theta}) = \log(x/k)$.

There are several ways of generating new probability distributions from classic ones to relative new distributions in the literature. Nelson mentioned in (Nelson 1982) that distributions with bathtub-shaped failure rate are sufficiently complex and, therefore, difficult to model. The distribution proposed by (Hjorth 1980) is such an example. Later on, (Rajarshi and Rajarshi 1988) presented a revision of these distributions, and (Haupt and Schäbe 1992) put forward a new lifetime model with bathtub-shaped failure rates. Unfortunately, these models are not sufficient to address various practical situations, so new classes of distributions were presented based on the modifications of the Weibull distribution to satisfy non-monotonic failure rate. For a review of these models, the reader can refer to (Mudholkar and Srivastava 1993), and (Pham and Lai 2007), where the authors summarized some generalizations of Weibull distribution in their papers. Other generalizations include the exponentiated Weibull (EW) (Gupta and Kundu 2001), the modified Weibull (MW) (Lai, Xie, and Murthy 2003), and the beta exponential (BE) (Nadarajah and Kotz 2006). Some more recent extensions are the generalized modified Weibull (GMW) (Carrasco, Ortega, and Cordeiro 2008), the beta modified Weibull (BMW) (Silva, Ortega, and Cordeiro 2010), (Nadarajah, Cordeiro, and Ortega 2011), the Weibull-G family (Bourguignon, Silva, and Cordeiro 2014) and the Gamma-exponentiated Weibull distributions (GEW) (Pinho, Cordeiro, and Nobre 2012). (Gurvich *et al.* 1997) developed a new statistical distribution for characterizing the random strength of brittle materials.

To motivate the model under study, consider a series system and assume that the lifetime of the components follow the log-logistic and Weibull distributions with with reliability functions $R_1(t) = (1 + (\frac{t}{s})^c)^{-1}$ and $R_2(t) = e^{-\alpha t^\beta}$, respectively. The reliability $R(t) = P(T > t)$ of the system is given by

$$R(t) = \prod_{i=1}^{2} R_i(t). \tag{3}$$

In some context, a series model is referred to as a competing risk model.

Also, a primary motivation for developing this model is the advantages presented by this generalized distribution with respect to having a hazard function that exhibits increasing, decreasing and bathtub shapes, as well as the versatility and flexibility of the log-logistic and Weibull distributions in modeling lifetime data. We propose and study this new distribution called the log-logistic Weibull distribution which inherits these desirable properties and also covers quite a variety of shapes.

There is an added advantage to this model, in that it has an additional dispersion parameter, depending on the overall form that accounts for the scale of the underlying random variable. The distribution also has exponential dumping in the upper tail making the distribution suitable for modeling samples that display power behavior for intermediate observations and decrease in tail probability for large observations or beyond a certain threshold or specified value.

The proposed new distribution generalizes the log-logistic and Weibull distributions. Some structural properties of this distribution are obtained and estimation the parameters via the method of maximum likelihood presented.

This paper is organized as follows. In section 2, we present the generalized distribution including the corresponding probability density functions (pdf), hazard and reverse hazard functions, quantile function and various sub-models. In section 3, the probability weighted moments, moments and conditional moments are presented. Section 4 contain the derivation of the mean deviations, Bonferroni and Lorenz curves. Section 5 is concerned with Rényi entropy, distribution of order statistics and L-moments. Estimation of model parameters is presented in section 6. Monte Carlo simulation study is conducted in section 7 to examine the bias and mean square

error of the maximum likelihood estimators and the width of the confidence intervals for each parameter. Applications of the proposed model to real data are given in section 8, followed by concluding remarks.

2. The log-logistic Weibull distribution

In this section, we present some statistical properties of the new log-logistic Weibull (LLoGW) distribution, including pdf, cdf, quantile function, hazard and reverse hazard functions. Plots of the hazard rate function for selected values of the model parameters are also given. We first of all present the Burr-XII, log-logistic and Weibull distributions. The very popular Burr Type III and Type XII distributions attract special attention because they include several families of distributions (e.g., the gamma distribution) with varying degrees of skewness and kurtosis. Further, these distributions have applications in a wide variety of areas in statistics and applied mathematics including modeling events associated with fracture roughness, life testing, operational risk, option market price distributions, forestry, meteorology, modeling crop prices, software reliability growth, and reliability analysis. See (Burr 1942), (Burr 1973) for additional details. The cdf and pdf of Burr XII distribution are given by

$$F_{_B}(x) = 1 - \left(1 + \left(\frac{x}{s}\right)^c\right)^{-k}, \tag{4}$$

and

$$f_{_B}(x) = \frac{kc}{s}\left(\frac{x}{s}\right)^{c-1}\left(1 + \left(\frac{x}{s}\right)^c\right)^{-k-1}, \quad \text{for } s,\, c,\, k, \text{ and } x \geq 0,$$

respectively. The reliability and hazard rate functions are given by

$$\overline{F}_{_B}(x) = \left(1 + \left(\frac{x}{s}\right)^c\right)^{-k}, \quad \text{and} \quad h_{_{F_B}}(x) = \frac{kc}{s}\left(\frac{x}{s}\right)^{c-1}\left(1 + \left(\frac{x}{s}\right)^c\right)^{-1},$$

respectively. Note that the pdf is unimodal with mode at $x_0 = ((c-1)/(ck+1))^{1/c}$ when $c > 1$, and $L-$shaped when $c = 1$. The r^{th} non-central moment is given by

$$E(X^r) = ks^r B(k - rc^{-1}, 1 + rc^{-1}), \quad \text{for } ck > r.$$

Note that k and c are shape parameters and s is a scale parameter. When $k = 1$ we obtain the log-logistic distribution. The cdf of the well known Weibull (W) distribution is given by

$$F_{_W}(x; \alpha, \beta) = 1 - \exp(-\alpha x^\beta), \quad x \geq 0,\, \alpha > 0,\, \beta > 0,$$

where α and β are scale and shape parameter, respectively.

Now, consider the log-logistic Weibull (LLoGW) distribution obtained by taking $R_1(x) = \overline{F}_1(x) = (1 + (\frac{x}{s})^c)^{-1}$ and $R_2(x) = \overline{F}_2(x) = exp(-\alpha x^\beta)$ in equation (1) to obtain the new LLoGW cdf $G(x) = G(x; s, c, \alpha, \beta)$ given by

$$G(x) = 1 - \left(1 + \left(\frac{x}{s}\right)^c\right)^{-1} \exp(-\alpha x^\beta), \tag{5}$$

for s, c, α, $\beta > 0$ and $x \geq 0$. If a random variable X has the LLoGW cdf, we write $X \sim LLoGW(s, c, \alpha, \beta)$. The corresponding LLoGW pdf is given by

$$g(x) = e^{-\alpha x^\beta}\left[1 + \left(\frac{x}{s}\right)^c\right]^{-1}\left\{\alpha\beta x^{\beta-1} + \frac{cx^{c-1}}{(s^c + x^c)}\right\}, \tag{6}$$

s, c, α, $\beta > 0$, and $x \geq 0$. Plots of the pdf for selected values of the model parameters are given in Figure 1. The plots suggests that the LLoGW pdf can be right skewed or decreasing for the selected values of the parameters.

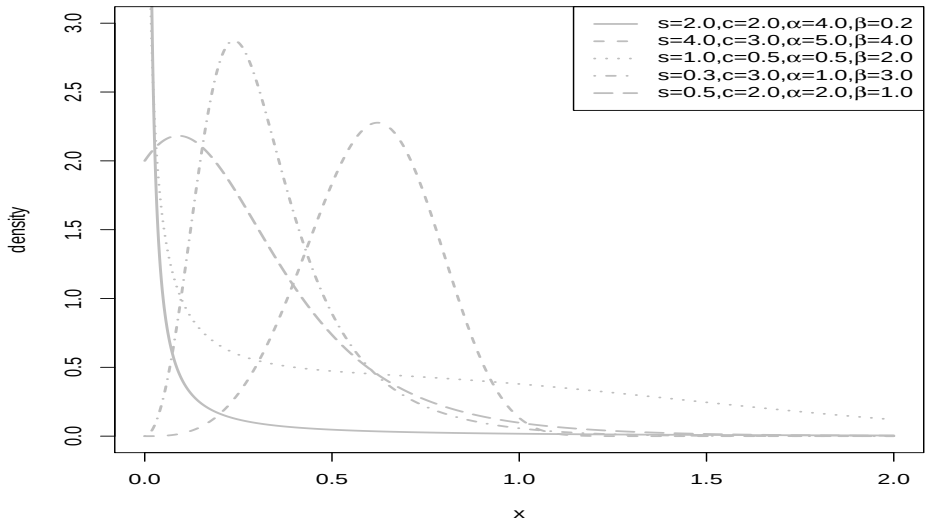


Figure 1: Plots of LLoGW pdf

2.1. Quantile function

The LLoGW quantile function can be obtained by inverting $\overline{G}(x) = 1 - u$, where $G(x) = u$, $0 \leq u \leq 1$, and

$$\overline{G}(x) = \left(1 + \left(\frac{x}{s}\right)^c\right)^{-1} e^{-\alpha x^\beta}. \tag{7}$$

The quantile function of the LLoGW distribution is obtained by solving the equation

$$\alpha x^\beta + \log\left(1 + \left(\frac{x}{s}\right)^c\right) + \log(1-u) = 0, \tag{8}$$

using numerical methods. Consequently, random number can be generated based on equation (8). Table 1 lists the quantile for selected values of the parameters of the LLoGW distribution.

Table 1: LLoGW quantile for selected values

	(s, c, α, β)				
u	(1.5,1.5,0.5,0.5)	(1.5,0.5,1.5,0.5)	(0.5,0.5,1.5,1.5)	(0.3,1.0,0.3,0.8)	(1.0,1.0,2.0,2.0)
0.1	0.1388	0.0021	0.0061	0.0277	0.0923
0.2	0.2296	0.0095	0.0290	0.0629	0.1754
0.3	0.3148	0.0247	0.0744	0.1075	0.2547
0.4	0.4012	0.0516	0.1436	0.1656	0.3338
0.5	0.4936	0.0968	0.2351	0.2444	0.4157
0.6	0.5972	0.1730	0.3489	0.3574	0.5040
0.7	0.7202	0.3068	0.4907	0.5339	0.6046
0.8	0.8800	0.5675	0.6779	0.8520	0.7287
0.9	1.1292	1.2234	0.9688	1.6268	0.9098

2.2. Some new and known sub-models

There are several new as well as well known distributions that can be obtained from the LLoGW distribution. Note that when $s = m^{-1}$, we have the log-logistic Weibull (LLoGW) distribution with the survival function $S(x; m, c, \alpha, \beta) = [1 + (xm)^c]^{-1} e^{-\alpha x^\beta}$. When $c = 1$ it reduces to the generalized Pareto type II Weibull (GP-II-W) distribution.

- If $\beta = 1$, we obtain the log-logistic exponential (LLoGE) distribution.
- If $\beta = 2$, we have the log-logistic Rayleigh (LLoGR) distribution.
- When $\alpha \to 0$, we have the log-logistic (LLoG) distribution.
- If $c = 1$ and $s \to \infty$, we obtain Weibull (W) distribution.
- If $c = 1$, $s \to \infty$, and $\beta = 2$ we have Rayleigh (R) distribution.
- If $c = 1$, $s \to \infty$, and $\beta = 1$, we have the exponential (E) distribution.
- If $c = 1$, then the LLoGW cdf reduces to the three parameter distribution with cdf given by

$$G(x) = 1 - \left(1 + \left(\frac{x}{s}\right)\right)^{-1} \exp(-\alpha x^\beta), \tag{9}$$

  for s, α, $\beta > 0$, and $x \geq 0$.
- If $c = \beta = 1$ then the LLoGW cdf reduces to to the two parameter distribution given by

$$G(x) = 1 - \left(1 + \left(\frac{x}{s}\right)\right)^{-1} \exp(-\alpha x), \tag{10}$$

  for s, $\alpha > 0$, and $x \geq 0$.
- If $c = 1$ and $\beta = 2$, then the LLoGW cdf reduces to the two parameter model

$$G(x) = 1 - \left(1 + \left(\frac{x}{s}\right)\right)^{-1} \exp(-\alpha x^2), \tag{11}$$

  for s, $\alpha > 0$, and $x \geq 0$.

2.3. Hazard and reverse hazard functions

In general, if X is a continuous random variable with cdf F, and pdf f, then the hazard function, reverse hazard function and mean residual life function are given by

$$\lambda_F(x) = \frac{f(x)}{\overline{F}(x)}, \quad \tau_F(x) = \frac{f(x)}{F(x)}, \quad \text{and} \quad \delta_F(x) = \frac{\int_x^\infty \overline{F}(u)du}{\overline{F}(x)},$$

respectively. The functions $\lambda_F(x)$, $\delta_F(x)$, and $\overline{F}(x)$ are equivalent. See (Shaked and Shanthikumar 1994) and references therein. In this subsection, the hazard and reverse hazard functions of the LLoGW distribution are presented. The hazard and reverse hazard functions of the LLoGW distribution are

$$h_G(x) = \left\{\alpha\beta x^{\beta-1} + \frac{cx^{c-1}}{(s^c + x^c)}\right\},$$

and

$$\tau_G(x) = \left[1 - \left(1 + \left(\frac{x}{s}\right)^c\right)^{-1} e^{-\alpha x^\beta}\right]^{-1} e^{-\alpha x^\beta} \left[1 + \left(\frac{x}{s}\right)^c\right]^{-1} \left\{\alpha\beta x^{\beta-1} + \frac{cx^{c-1}}{(s^c + x^c)}\right\} \tag{12}$$

for $x \geq 0$, s, c, α, $\beta > 0$, respectively. The limiting behavior of the hazard function of the LLoGW distribution, which can be readily established is as follows:

- For $\beta < 1$ and $c = 1$, $\lim_{x\to 0} h_{_G}(x) = \infty$ and $\lim_{x\to\infty} h_{_G}(x) = 0$.
- For $\beta = 1$,
$$\lim_{x\to 0} h_{_G}(x) = \begin{cases} \infty & 0 < c < 1, \\ \alpha & c > 1. \end{cases}$$
- For $\beta = 1$, and for each $c > 0$, $\lim_{x\to\infty} h_{_G}(x) = \alpha$.
- For $\beta = 1$ and $c = 1$, $\lim_{x\to 0} h_{_G}(x) = \alpha + \frac{1}{s}$ and $\lim_{x\to\infty} h_{_G}(x) = \alpha$.
- For $\beta < 1$ and $c < 1$, $\lim_{x\to 0} h_{_G}(x) = \infty$ and $\lim_{x\to\infty} h_{_G}(x) = 0$.
- For $\beta > 1$ and $c > 1$, $\lim_{x\to 0} h_{_G}(x) = 0$ and $\lim_{x\to\infty} h_{_G}(x) = \infty$.
- For $\beta > 1$ and $c = 1$, $\lim_{x\to 0} h_{_G}(x) = \frac{1}{s}$ and $\lim_{x\to\infty} h_{_G}(x) = \infty$.

Plots of the hazard function are given in Figure 2. The graphs exhibit increasing, decreasing, bathtub, upside down bathtub, and upside down bathtub followed by bathtub shapes for the selected values of the model parameters. This very attractive flexibility makes the LLoGW hazard function useful and suitable for non-monotonic empirical hazard behaviors which are more likely to be encountered in practice or real life situations.

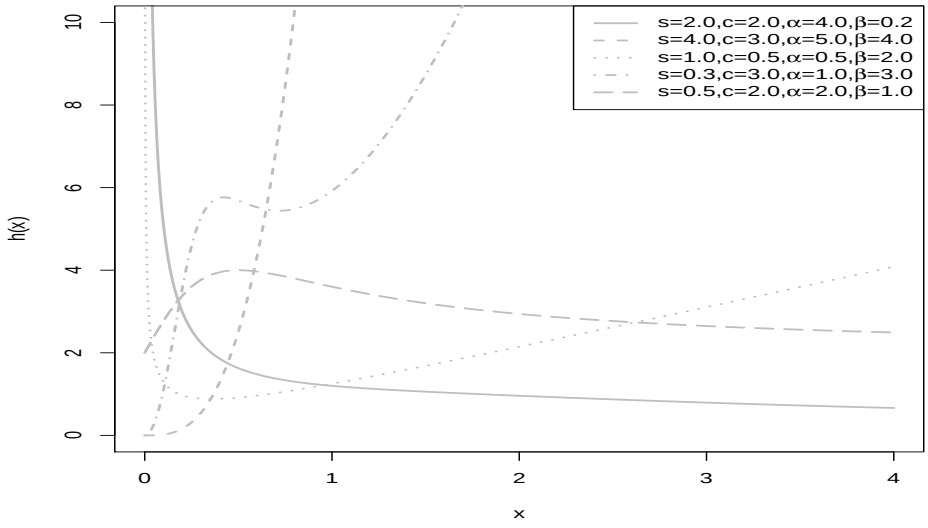


Figure 2: LLoGW hazard functions

3. Probability weighted moments, moments and conditional moments

In this section, we obtain probability weighted moments (PWMs) (Greenwood, Landwehr, Matalas, and Wallis 1979), moments and conditional moments for the LLoGW distribution. The probability weighted moments (PWMs) of the LLoGW distribution is given by

$$\begin{aligned} E(X^r G^l(X)\overline{G}^m(X)) &= \int_0^\infty x^r G^l(x)\overline{G}^m(x)g(x)dx \\ &= \sum_{j=0}^{\infty} \frac{(-1)^j \Gamma(l+1)E(X^r(\overline{G}(X))^{j+m})}{\Gamma(l+1-j)\Gamma(j+1)}. \end{aligned}$$

Now, by setting $y = (1 + (x/s)^c)^{-1}$, the PWMs of the LLoGW distribution can be written as:

$$\begin{aligned} E(X^r G^l(X)\overline{G}^m(X)) &= \sum_{j,k=0}^{\infty} \frac{(-1)^{j+k}\Gamma(l+1)[\alpha(j+m+1)]^k}{\Gamma(l+1-j)\Gamma(j+1)k!} \\ &\times \left[\int_0^{\infty} x^{r+k\beta}\left(1+\left(\frac{x}{s}\right)^c\right)^{-(j+m+2)} \frac{c}{s}\left(\frac{x}{s}\right)^{c-1} dx \right. \\ &+ \left. \alpha\beta\int_0^{\infty} x^{r+k\beta+\beta-1}\left(1+\left(\frac{x}{s}\right)^c\right)^{-(j+m+1)} dx\right] \\ &= \sum_{j,k=0}^{\infty} \frac{(-1)^{j+k}\Gamma(l+1)[\alpha(j+m+1)]^k}{\Gamma(l+1-j)\Gamma(j+1)k!} \\ &\times \left[s^{r+k\beta}\int_0^1 y^{j+m-(\frac{r+k\beta}{c})}(1-y)^{\frac{r+k\beta}{c}} dy \right. \\ &+ \left. \frac{\alpha\beta s^{r+k\beta+\beta}}{c}\int_0^1 y^{j+m-(\frac{r+k\beta+\beta}{c})}(1-y)^{\frac{r+k\beta+\beta}{c}-1} dy\right]. \end{aligned}$$

Consequently, the PWMs of the LLoGW distribution is given by

$$\begin{aligned} E(X^r G^l(X)\overline{G}^m(X)) &= \sum_{j,k=0}^{\infty} \frac{(-1)^{j+k}\Gamma(l+1)[\alpha(j+m+1)]^k s^{r+k\beta}}{\Gamma(l+1-j)\Gamma(j+1)k!} \\ &\times \left[B\left(j+m+1-\left(\frac{r+k\beta}{c}\right), \frac{r+k\beta+c}{c}\right)\right. \\ &+ \left. \frac{\alpha\beta s^{\beta}}{c} B\left(j+m+1-\left(\frac{r+k\beta+\beta}{c}\right), \frac{r+k\beta+\beta}{c}\right)\right]. \end{aligned}$$

Remarks: special cases

- When $l = m = 0$, we obtain the r^{th} non-central moment μ'_r given by

$$\begin{aligned} \mu'_r = E(X^r) &= \sum_{k=0}^{\infty} \frac{(-1)^k \alpha^k s^{k\beta+r}}{k!}\left(\frac{\alpha\beta s^{\beta}}{c} B\left(1-\frac{k\beta+\beta+r}{c}, \frac{k\beta+\beta+r}{c}\right)\right. \\ &+ \left. B\left(1-\frac{k\beta+r}{c}, 1+\frac{k\beta+r}{c}\right)\right), \end{aligned}$$

  where $B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}dy$ is the beta function.

- When $r = l = 0$, we have

$$\begin{aligned} E(\overline{G}^m(X)) &= \sum_{k=0}^{\infty} \frac{(-1)^k[\alpha(m+1)]^k s^{k\beta}}{k!} \\ &\times \left[B\left(m+1-\left(\frac{k\beta}{c}\right), \frac{k\beta+c}{c}\right)\right. \\ &+ \left. \frac{\alpha\beta s^{\beta}}{c} B\left(m+1-\left(\frac{k\beta+\beta}{c}\right), \frac{k\beta+\beta}{c}\right)\right]. \end{aligned}$$

- When $l = 0$, the LLoGW PWMs reduces to

$$\begin{aligned} E(X^r\overline{G}^m(X)) &= \sum_{k=0}^{\infty} \frac{(-1)^k[\alpha(m+1)]^k s^{r+k\beta}}{k!} \\ &\times \left[B\left(m+1-\left(\frac{r+k\beta}{c}\right), \frac{r+k\beta+c}{c}\right)\right. \\ &+ \left. \frac{\alpha\beta s^{\beta}}{c} B\left(m+1-\left(\frac{r+k\beta+\beta}{c}\right), \frac{r+k\beta+\beta}{c}\right)\right]. \end{aligned}$$

- When $m = 0$, the LLoGW PWMs reduces to

$$\begin{aligned} E(X^r G^l(X)) &= \sum_{j,k=0}^{\infty} \frac{(-1)^{j+k}\Gamma(l+1)[\alpha(j+1)]^k s^{r+k\beta}}{\Gamma(l+1-j)\Gamma(j+1)k!} \\ &\times \left[B\left(j+1-\left(\frac{r+k\beta}{c}\right), \frac{r+k\beta+c}{c}\right)\right. \\ &+ \left.\frac{\alpha\beta s^{\beta}}{c} B\left(j+1-\left(\frac{r+k\beta+\beta}{c}\right), \frac{r+k\beta+\beta}{c}\right)\right]. \end{aligned}$$

- When $r = m = 0$, we have

$$\begin{aligned} E(G^l(X)) &= \sum_{j,k=0}^{\infty} \frac{(-1)^{j+k}\Gamma(l+1)[\alpha(j+1)]^k s^{k\beta}}{\Gamma(l+1-j)\Gamma(j+1)k!} \\ &\times \left[B\left(j+1-\left(\frac{k\beta}{c}\right), \frac{k\beta+c}{c}\right)\right. \\ &+ \left.\frac{\alpha\beta s^{\beta}}{c} B\left(j+1-\left(\frac{k\beta+\beta}{c}\right), \frac{k\beta+\beta}{c}\right)\right]. \end{aligned}$$

- When $r = 0$, we have

$$\begin{aligned} E(G^l(X)\overline{G}^m(X)) &= \sum_{j,k=0}^{\infty} \frac{(-1)^{j+k}\Gamma(l+1)[\alpha(j+m+1)]^k s^{k\beta}}{\Gamma(l+1-j)\Gamma(j+1)k!} \\ &\times \left[B\left(j+m+1-\left(\frac{k\beta}{c}\right), \frac{k\beta+c}{c}\right)\right. \\ &+ \left.\frac{\alpha\beta s^{\beta}}{c} B\left(j+m+1-\left(\frac{k\beta+\beta}{c}\right), \frac{k\beta+\beta}{c}\right)\right]. \end{aligned}$$

Note that the r^{th} raw moment μ'_r can also be obtained as follows:

$$\begin{aligned} \mu'_r = E(X^r) &= \int_0^{\infty} x^r g(x; s, c, \alpha, \beta) dx \\ &= \int_0^{\infty} \alpha\beta x^{r+\beta-1} \left(1+\left(\frac{x}{s}\right)^c\right)^{-1} e^{-\alpha x^{\beta}} dx \\ &+ \int_0^{\infty} \frac{c}{s^c} x^{r+c-1} \left(1+\left(\frac{x}{s}\right)^c\right)^{-2} e^{-\alpha x^{\beta}} dx. \end{aligned}$$

Let $y = (x/s)^c$, and $u = y^{\beta/c}$ then

$$\begin{aligned} E(X^r) &= \frac{\alpha\beta s^{r+\beta}}{c} \int_0^{\infty} y^{\frac{r+\beta}{c}-1}[1+y]^{-1} e^{\alpha s^{\beta} y^{\beta/c}} dy + s^r \int_0^{\infty} y^{\frac{r}{c}}[1+y]^{-2} e^{\alpha s^{\beta} y^{\beta/c}} dy \\ &= \alpha s^{r+\beta} \int_0^{\infty} u^{\frac{r+\beta}{\beta}-1}[1+u^{c/\beta}]^{-1} e^{-\alpha s^{\beta} u} du \\ &+ \frac{cs^r}{\beta} \int_0^{\infty} u^{\frac{r+c}{\beta}-1}[1+u^{c/\beta}]^{-2} e^{-\alpha s^{\beta} u} du. \end{aligned}$$

We apply the following results which holds for m and k positive integers, $w > -1$ and $p > 0$ ((Prudnikov, Brychkov, and Marichev 1992), page 21):

$$\begin{aligned} I(p, w, \frac{m}{k}, \nu) &= \int_0^{\infty} x^w [1+x^{\frac{m}{k}}]^{\nu} e^{-px} dx \\ &= \frac{k^{-\nu} m^{w+1/2}}{(2\pi)^{\frac{m-1}{2}}\Gamma(-\nu)p^{w+1}} G^{k,k+m}_{k+m,k}\left(\frac{m^m}{p^m} \,\middle|\, \begin{matrix} \Delta(m,-w), & \Delta(k,\nu+1) \\ \Delta(k,0) & \end{matrix}\right), \end{aligned}$$

where $\Delta(k,a) = \frac{a}{k}, \frac{a+1}{k}, ..., \frac{a+k}{k}$, and the Meijer G function is defined by

$$G_{p,q}^{m,n}\left(x \left| \begin{matrix} a_1, ... & ,a_p \\ b_1, ... & ,b_q \end{matrix} \right.\right) = \frac{1}{2\pi i}\int_L \frac{\prod_{j=1}^{m}\Gamma(b_j+t)\prod_{j=1}^{n}\Gamma(1-a_j-t)}{\prod_{j=n+1}^{p}\Gamma(a_j+t)\prod_{j=m+1}^{p}\Gamma(1-b_j-t)}x^{-t}dt,$$

where $i = \sqrt{-1}$ is the complex unit and L is the integration path (see (Gradshtein and Ryzhik 2000), sec. 9.3 for a description of this path).
Consequently, the r^{th} moment of the LLoGW distribution is given by

$$\begin{aligned} E(X^r) &= \frac{\beta c^{\frac{r}{\beta}+\frac{1}{2}}}{(2\pi)^{\frac{c-1}{2}}\alpha^{\frac{r}{\beta}}}G_{\beta+c,\beta}^{\beta,\beta+c}\left(\frac{c^c}{(\alpha s^\beta)^c}\left| \begin{matrix} \Delta(c,-\frac{r}{\beta}), & \Delta(\beta,0) \\ \Delta(\beta,0) \end{matrix} \right.\right) \\ &+ \frac{\beta c^{\frac{r+c}{\beta}+\frac{1}{2}}}{(2\pi)^{\frac{c-1}{2}}\alpha^{\frac{r+c}{\beta}}s^c}G_{\beta+c,\beta}^{\beta,\beta+c}\left(\frac{c^c}{(\alpha s^\beta)^c}\left| \begin{matrix} \Delta(c,1-\frac{r+c}{\beta}), & \Delta(\beta,-1) \\ \Delta(\beta,0) \end{matrix} \right.\right). \end{aligned}$$

Alternatively, the following direct computation also gives the r^{th} moment of the LLoGW distribution when we use the fact that $e^{-\alpha x^\beta} = \sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k x^{k\beta}}{k!}$.

Theorem 3.1. *The r^{th} raw moment $\mu_r' = E(X^r)$ of the LLoGW distribution is*

$$\begin{aligned} E(X^r) &= \sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k s^{k\beta+r}}{k!}\left(\frac{\alpha\beta s^\beta}{c}B\left(1-\frac{k\beta+\beta+r}{c}, \frac{k\beta+\beta+r}{c}\right)\right. \\ &+ \left. B\left(1-\frac{k\beta+r}{c}, 1+\frac{k\beta+r}{c}\right)\right), \end{aligned}$$

where $B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1}dy$ is the beta function.

Proof: Note that

$$\begin{aligned} E(X^r) &= \int_0^\infty x^r g(x;s,c,\alpha,\beta)dx \\ &= \int_0^\infty \alpha\beta x^{r+\beta-1}\left(1+\left(\frac{x}{s}\right)^c\right)^{-1}e^{-\alpha x^\beta}dx \\ &+ \int_0^\infty \frac{c}{s^c}x^{r+c-1}\left(1+\left(\frac{x}{s}\right)^c\right)^{-2}e^{-\alpha x^\beta}dx \end{aligned}$$

Let $y = (1+(\frac{x}{s})^c)^{-1}$, and apply $e^{-\alpha x^\beta} = \sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k x^{k\beta}}{k!}$, to obtain

$$\begin{aligned} E(X^r) &= \sum_{k=0}^{\infty}\frac{(-1)^k\alpha^{k+1}\beta s^{k\beta+\beta+r}}{k!c}\int_0^1 y^{1-\frac{k\beta+\beta+r}{c}-1}(1-y)^{\frac{k\beta+\beta+r}{c}-1}dy \\ &+ \sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k s^{k\beta+r}}{k!}\int_0^1 y^{1-\frac{k\beta+r}{c}-1}(1-y)^{\frac{k\beta+r}{c}+1-1}dy \\ &= \sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k s^{k\beta+r}}{k!}\left(\frac{\alpha\beta s^\beta}{c}B\left(1-\frac{k\beta+\beta+r}{c}, \frac{k\beta+\beta+r}{c}\right)\right. \\ &+ \left. B\left(1-\frac{k\beta+r}{c}, 1+\frac{k\beta+r}{c}\right)\right), \end{aligned} \tag{13}$$

for $c > k\beta+\beta+r$. To obtain the moment generating function (MGF) of the LLoGW distribution, recall the Taylor's series expansion of the function $e^{tx} = \sum_{j=0}^{\infty}\frac{(tx)^j}{j!}$, so that, we have $M_X(t) = E(e^{tX}) = \sum_{n=0}^{\infty}\frac{t^n}{n!}E(X^n)$, where $E(X^n)$ is given above.

Table 2 lists the first six moments together with the standard deviation (SD or σ), coefficient of variation (CV), coefficient of skewness (CS) and coefficient of kurtosis (CK) of the LLoGW distribution for selected values of the parameters, by fixing $\alpha = 1.5$ and $\beta = 1.5$. And Table 3 lists the first six moments, SD, CV, CS and CK of the LLoGW distribution for selected values of the parameters, by fixing $s = 1.5$ and $c = 1.0$. These values can be determined numerically using R and MATLAB. The SD, CV, CS and CK are given by $\sigma = \sqrt{\mu_2' - \mu^2}$,

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\mu_2' - \mu^2}}{\mu} = \sqrt{\frac{\mu_2'}{\mu^2} - 1},$$

$$CS = \frac{E\left[(X-\mu)^3\right]}{\left[E(X-\mu)^2\right]^{3/2}} = \frac{\mu_3' - 3\mu\mu_2' + 2\mu^3}{(\mu_2' - \mu^2)^{3/2}},$$

and

$$CK = \frac{E\left[(X-\mu)^4\right]}{\left[E(X-\mu)^2\right]^2} = \frac{\mu_4' - 4\mu\mu_3' + 6\mu^2\mu_2' - 3\mu^4}{(\mu_2' - \mu^2)^2},$$

respectively.

Table 2: Moments of the LLoGW distribution for some parameter values; $\alpha = 1.5$ and $\beta = 1.5$.

μ_s'	$s = 0.5, c = 0.5$	$s = 0.5, c = 1.5$	$s = 1.5, c = 0.5$	$s = 1.5, c = 1.5$
μ_1'	0.3784	0.4060	0.4627	0.5777
μ_2'	0.3181	0.2696	0.4099	0.4973
μ_3'	0.3680	0.2485	0.4871	0.5543
μ_4'	0.5244	0.2916	0.7062	0.7482
μ_5'	0.8740	0.4124	1.1918	1.1759
μ_6'	1.6522	0.6780	2.2751	2.0961
SD	0.4181	0.3237	0.4425	0.4044
CV	1.1049	0.7972	0.9565	0.7001
CS	1.5772	1.5918	1.3412	1.1800
CK	5.8595	6.6587	5.0516	4.8231

Table 3: Moments of the LLoGW distribution for some parameter values; $s = 1.5$ and $c = 1.0$.

μ_s'	$\alpha = 0.2, \beta = 1.8$	$\alpha = 2.5, \beta = 0.8$	$\alpha = 1.0, \beta = 1.0$	$\alpha = 0.1, \beta = 3.5$
μ_1'	1.2593	0.2890	0.6724	1.1314
μ_2'	2.7456	0.2143	0.9828	1.8173
μ_3'	7.9772	0.2744	2.2886	3.3807
μ_4'	28.0396	0.5190	7.4228	6.8809
μ_5'	113.5654	1.3309	31.0823	14.9509
μ_6'	514.8466	4.3789	160.0520	34.2132
SD	1.0770	0.3617	0.7285	0.7330
CV	0.8552	1.2517	1.0835	0.6479
CS	1.2797	2.8920	2.3639	0.2764
CK	4.7851	16.8416	11.7875	2.1587

3.1. Conditional moments

For lifetime models, it is also of interest to find the conditional moments and the mean residual

lifetime function. The r^{th} conditional moments for LLoGW distribution is given by

$$\begin{aligned} E(X^r|X>t) &= \frac{1}{\overline{G}(t)}\int_t^\infty x^r g_{LLoGW}(x)dx \\ &= \frac{1}{\overline{G}(t)}\left[\sum_{k=0}^\infty \frac{(-1)^k\alpha^k}{k!}\alpha\beta\int_t^\infty x^{k\beta+r+\beta-1}\left[1+\left(\frac{x}{s}\right)^c\right]^{-1}dx\right. \\ &+ \left.\sum_{k=0}^\infty \frac{(-1)^k c\alpha^k}{s^c k!}\int_t^\infty x^{k\beta+r+c-1}\left[1+\left(\frac{x}{s}\right)^c\right]^{-2}dx\right]. \end{aligned} \tag{14}$$

Let $y=(1+(x/s)^c)^{-1}$, then

$$\begin{aligned} E(X^r|X>t) &= \frac{1}{\overline{G}(t)}\left[\sum_{k=0}^\infty \frac{(-1)^{k+1}\alpha^{k+1}\beta}{k!}s^{k\beta+\beta+r}\right. \\ &\times \int_{(1+(t/s)^c)^{-1}}^1 y^{1-\frac{k\beta+\beta+r}{c}-1}(1-y)^{\frac{k\beta+\beta+r}{c}-1}dy \\ &+ \left.\sum_{k=0}^\infty \frac{(-1)^{k+1}\alpha^k\beta}{k!}s^{k\beta+r}\int_{(1+(t/s)^c)^{-1}}^1 y^{1-\frac{k\beta+r}{c}-1}(1-y)^{1+\frac{k\beta+r}{c}-1}dy\right] \\ &= \frac{1}{\overline{G}(t)}\left[\sum_{k=0}^\infty \frac{(-1)^{k+1}\alpha^k\beta s^{k\beta+r}}{k!}\right. \\ &\times \left[\alpha\beta s^\beta\left(\left(B\left(1-\frac{k\beta+\beta+r}{c},\frac{k\beta+\beta+r}{c}\right)\right.\right.\right. \\ &- \left.B_{(1+(t/s)^c)^{-1}}\left(1-\frac{k\beta+\beta+r}{c},\frac{k\beta+\beta+r}{c}\right)\right) \\ &+ \left(B\left(1-\frac{k\beta+r}{c},\frac{k\beta+r}{c}\right)\right. \\ &- \left.\left.\left.\left.B_{(1+(t/s)^c)^{-1}}\left(1-\frac{k\beta+r}{c},1+\frac{k\beta+r}{c}\right)\right)\right)\right]\Bigg], \end{aligned}$$

where $B_x(a,b)=\int_0^x y^{a-1}(1-y)^{b-1}dy$ is the incomplete beta function, and $c>k\beta+\beta+r$. Alternatively, consider the following integral for $q>0$ and $b>0$, (Paranaíba, Ortega, Cordeiro, and Pescim 2011),

$$\begin{aligned} J(q,a,b) &= \int_q^\infty y^a[1+y]^{-b}dy \\ &= -\left[\frac{{}_2F_1(b,a+1;a+2;-q)q^{a+1}}{a+1}+\frac{\Gamma(b-a-1)\pi(\pi a)}{\Gamma(b)\Gamma(-a)}\right], \end{aligned} \tag{15}$$

where ${}_2F_1$ is the hypergeometric function given by

$${}_2F_1(a,b;c;x)=\sum_{k=0}^\infty \frac{(a)_k(b)_k}{(c)_k}\frac{x^k}{k!}, \tag{16}$$

and $(a)_k=a(a+1)....(a+k-1)$ is the ascending factorial. Now, consider the integral $\int_t^\infty x^{k\beta+r+\beta-1}[1+(x/s)^c]^{-1}dx$ and let $y=(x/s)^c$, then

$$\begin{aligned} \int_t^\infty x^{k\beta+r+\beta-1}[1+(x/s)^c]^{-1}dx &= \frac{s^{k\beta+r+\beta}}{c}\int_{(t/s)^c}^\infty y^{\frac{k\beta+r+\beta}{c}-1}[1+y]^{-1}dy \\ &= \frac{s^{k\beta+r+\beta}}{c}J\left(\left(\frac{t}{s}\right)^c,\frac{k\beta+r+\beta-c}{c},1\right). \end{aligned}$$

Consequently, the r^{th} conditional moment of the LLoGW distribution is given by

$$\begin{aligned} E(X^r|X>t) &= \frac{1}{\overline{G}(t)}\Bigg[\sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k}{k!}\frac{s^{k\beta+r}}{c}\Bigg(\alpha\beta s^{\beta}J\Bigg(\left(\frac{t}{s}\right)^c,\frac{k\beta+r+\beta-c}{c},1\Bigg) \\ &+ s^cJ\Bigg(\left(\frac{t}{s}\right)^c,\frac{k\beta+r}{c},2\Bigg)\Bigg)\Bigg]. \end{aligned}$$

The mean residual lifetime function of the LLoGW distribution is $E(X|X>t)-t$.

4. Mean deviations

The amount of scatter in a population is evidently measured to some extent by the totality of deviations from the mean and median. These are known as the mean deviation about the mean and the mean deviation about the median, and defined by

$$\delta_1(x)=\int_0^{\infty}|x-\mu|g(x)dx \quad \text{and} \quad \delta_2(x)=\int_0^{\infty}|x-M|g(x)dx, \tag{17}$$

respectively, where $\mu=E(X)$ is the mean and M =Median (X) is the median. The measures $\delta_1(x)$ and $\delta_2(x)$ can be calculated using the relationships

$$\delta_1(x)=2\mu G(\mu)-2\mu+2\int_{\mu}^{\infty}xg(x)dx, \tag{18}$$

and

$$\delta_2(x)=-\mu+2\int_{M}^{\infty}xg(x)dx, \tag{19}$$

respectively. When $r=1$, we get the mean $\mu=E(X)$. Note that $T(\mu)=\int_{\mu}^{\infty}xg(x)dx$ and $T(M)=\int_{M}^{\infty}xg(x)dx$ are given by

$$\begin{aligned} T(\mu) &= \Bigg[\sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k}{k!}\frac{s^{k\beta+r}}{c}\Bigg(\alpha\beta s^{\beta}J\Bigg(\left(\frac{\mu}{s}\right)^c,\frac{k\beta+r+\beta-c}{c},1\Bigg) \\ &+ s^cJ\Bigg(\left(\frac{\mu}{s}\right)^c,\frac{k\beta+r}{c},2\Bigg)\Bigg)\Bigg] \end{aligned}$$

and

$$\begin{aligned} T(M) &= \Bigg[\sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k}{k!}\frac{s^{k\beta+r}}{c}\Bigg(\alpha\beta s^{\beta}J\Bigg(\left(\frac{M}{s}\right)^c,\frac{k\beta+r+\beta-c}{c},1\Bigg) \\ &+ s^cJ\Bigg(\left(\frac{M}{s}\right)^c,\frac{k\beta+r}{c},2\Bigg)\Bigg)\Bigg], \end{aligned}$$

respectively. Alternatively,

$$\begin{aligned} T(\mu) &= \int_{\mu}^{\infty} xg(x)dx \\ &= \Bigg[\sum_{k=0}^{\infty} \frac{(-1)^{k+1}\alpha^k\beta s^{k\beta+r}}{k!} \\ &\times \Bigg[\alpha\beta s^{\beta}\Bigg(\Bigg(B\Bigg(1-\frac{k\beta+\beta+r}{c}, \frac{k\beta+\beta+r}{c}\Bigg) \\ &- B_{(1+(\mu/s)^c)^{-1}}\Bigg(1-\frac{k\beta+\beta+r}{c}, \frac{k\beta+\beta+r}{c}\Bigg)\Bigg) \\ &+ \Bigg(B\Bigg(1-\frac{k\beta+r}{c}, \frac{k\beta+r}{c}\Bigg) \\ &- B_{(1+(\mu/s)^c)^{-1}}\Bigg(1-\frac{k\beta+r}{c}, 1+\frac{k\beta+r}{c}\Bigg)\Bigg)\Bigg)\Bigg]\Bigg]. \end{aligned}$$

Consequently, the mean deviation about the mean is

$$\delta_1(x) = 2\mu G(\mu) - 2\mu + 2T(\mu)$$

and the mean deviation about the median is

$$\delta_2(x) = -\mu + 2T(M).$$

4.1. Bonferroni and Lorenz curves

In this subsection, we present Bonferroni and Lorenz Curves. Bonferroni and Lorenz curves have applications not only in economics for the study income and poverty, but also in other fields such as reliability, demography, insurance and medicine. Bonferroni and Lorenz curves for the LLoGW distribution are given by

$$B(p) = \frac{1}{p\mu}\int_0^q xg(x)dx = \frac{1}{p\mu}[\mu - T(q)],$$

and

$$L(p) = \frac{1}{\mu}\int_0^q xg(x)dx = \frac{1}{\mu}[\mu - T(q)],$$

respectively, where $T(q) = \int_q^{\infty} xg(x)dx$, and $q = G^{-1}(p)$, $0 \leq p \leq 1$.

5. Order statistics, L-moments and Rényi entropy

The concept of entropy plays a vital role in information theory. The entropy of a random variable is defined in terms of its probability distribution and can be shown to be a good measure of randomness or uncertainty. In this section, we present the distribution of the order statistic, L-moments and Rényi entropy for the LLoGW distribution.

5.1. Order statistics

Order statistics play an important role in probability and statistics. In this subsection, we present the distribution of the i^{th} order statistic from the LLoGW distribution. The pdf of the

i^{th} order statistic from the LLoGW pdf $g(x)$ is given by

$$\begin{aligned} g_{i:n}(x) &= \frac{n!g(x)}{(i-1)!(n-i)!}[G(x)]^{i-1}[1-G(x)]^{n-i} \\ &= \frac{n!g(x)}{(i-1)!(n-i)!}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}[G(x)]^{j+i-1} \end{aligned}$$

by using the binomial expansion $[1-G(x)]^{n-i} = \sum_{m=0}^{n-i}\binom{n-i}{m}(-1)^m[G(x)]^m$. Consequently,

$$\begin{aligned} g_{i:n}(x) &= \frac{1}{B(i,n-i+1)}\sum_{m=0}^{n-i}\binom{n-i}{m}\frac{(-1)^m}{m+i}(m+i)[G(x)]^{m+i-1}g(x) \\ &= \sum_{m=0}^{n-i}w_{i,m}g_{m+i}(x), \end{aligned}$$

where $g_{m+i}(x)$ is the pdf of the exponentiated LLoGW distribution with parameters s, c, α, β and $(m+i)$, $B(.,.)$ is the beta function and the weights $w_{i,m}$ are given by

$$w_{i,m} = \frac{1}{B(i,n-i+1)}\frac{(-1)^m}{m+i}\binom{n-i}{m} = (-1)^m\binom{m+i-1}{m}\binom{n}{m+i}.$$

The t^{th} moment of the i^{th} order statistics from the LLoGW distribution can be derived via a result of (Barakat and Abdelkader 2004) as follows:

$$E(X_{i:n}^t) = t\sum_{p=n-i+1}^{n}(-1)^{p-n+i-1}\binom{p-1}{n-i}\binom{n}{p}\int_0^\infty x^{t-1}[1-G(x)]^p dx. \tag{20}$$

Note that

$$\int_0^\infty x^{t-1}[1-G(x)]^p dx = \sum_{k=0}^{\infty}\frac{(-1)^k(p\alpha)^k}{k!}\int_0^\infty x^{k\beta+t-1}[1+(x/s)^c]^{-p}dx.$$

Let $y = [1+(x/s)^c]^{-1}$, then

$$\int_0^\infty x^{k\beta+t-1}[1+(x/s)^c]^{-p}dx = \frac{s^{k\beta+t}}{c}\int_0^1 y^{p-\frac{k\beta}{c}-\frac{t}{c}-1}(1-y)^{\frac{k\beta+t}{c}-1}dy.$$

Now,

$$\begin{aligned} E(X_{i:n}^t) &= t\sum_{p=n-i+1}^{n}\sum_{k=0}^{\infty}(-1)^{p-n+i+k}\frac{(p\alpha)^k}{k!}\binom{p-1}{n-i}\binom{n}{p}\frac{s^{k\beta+t}}{c} \\ &\times B\left(p-\frac{k\beta+t}{c},\frac{k\beta+t}{c}\right), \end{aligned} \tag{21}$$

where $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$ is the beta function.

5.2. L-moments

The $L-$moments (Hoskings 1990) are expectations of some linear combinations of order statistics and they exist whenever the mean of the distribution exits, even when some higher moments may not exist. They are relatively robust to the effects of outliers and are given by

$$\lambda_{k+1} = \frac{1}{k+1}\sum_{j=0}^{k}(-1)^j\binom{k}{j}E(X_{k+1-j:k+1}), \quad k = 0,1,2,..... \tag{22}$$

The $L-$moments of the LLoGW distribution can be readily obtained from equation (21). The first four $L-$moments are given by $\lambda_1 = E(X_{1:1})$, $\lambda_2 = \frac{1}{2}E(X_{2:2} - X_{1:2})$, $\lambda_3 = \frac{1}{3}E(X_{3:3} - 2X_{2:3} + X_{1:3})$ and $\lambda_4 = \frac{1}{4}E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4})$, respectively.

5.3. Rényi entropy

In this subsection, Rényi entropy of the LLoGW distribution is derived. An entropy is a measure of uncertainty or variation of a random variable. Rényi entropy is an extension of Shannon entropy and is defined to be

$$I_R(v) = \frac{1}{1-v}\log\left(\int_0^\infty [g(x;s,c,\alpha,\beta)]^v dx\right), v \neq 1, v > 0. \tag{23}$$

Rényi entropy tends to Shannon entropy as $v \to 1$. Note that $[g(x;s,c,\alpha,\beta)]^v = g^v(x)$ can be written as

$$\begin{aligned} g^v(x) &= \left[\frac{cx^{c-1}}{s^c}\left(1+\left(\frac{x}{s}\right)^c\right)^{-2} e^{-\alpha x^\beta} + \alpha\beta x^{-\beta-1}\left(1+\left(\frac{x}{s}\right)^c\right)^{-1} e^{-\alpha x^\beta}\right]^v \\ &= \sum_{k=0}^{v}\binom{v}{k}\left(\frac{cx^{c-1}}{s^c}\left[1+\left(\frac{x}{s}\right)^c\right]^{-2k} e^{-k\alpha x^\beta}\right. \\ &\times \left.\left(\alpha\beta x^{\beta-1}\left[1+\left(\frac{x}{s}\right)^c\right]^{-1} e^{-\alpha x^\beta}\right)^{v-k}. \end{aligned}$$

Now,

$$\begin{aligned} \int_0^\infty g^v(x)dx &= \sum_{k=0}^{v}\sum_{m=0}^{\infty}\binom{v}{k}\frac{(-1)^m(v\alpha)^m c^k(\alpha\beta)^{v-k}}{m!s^{ck}} \\ &\times \int_0^\infty x^{m\beta+v\beta-k\beta+ck-v}\left[1+\left(\frac{x}{s}\right)^c\right]^{-v-k} dx \\ &= \sum_{k=0}^{v}\sum_{m=0}^{\infty}\binom{v}{k}\frac{(-1)^m(v\alpha)^m c^{k-1}(\alpha\beta)^{v-k}}{m!} s^{m\beta+v\beta-k\beta-v+1} \\ &\times \int_0^1 y^{v+\frac{v+k\beta-v\beta-m\beta-1}{c}-1}(1-y)^{\frac{m\beta+v\beta-k\beta-v+1}{c}} dy \\ &= \sum_{k=0}^{v}\sum_{m=0}^{\infty}\binom{v}{k}\frac{(-1)^m(v\alpha)^m c^{k-1}(\alpha\beta)^{v-k}}{m!} s^{m\beta+v\beta-k\beta-v+1} \\ &\times B\left(v+\frac{v+k\beta-v\beta-m\beta-1}{c}, 1+\frac{m\beta+v\beta-k\beta-v+1}{c}\right). \end{aligned}$$

Consequently, Rényi entropy is given by

$$\begin{aligned} I_R(v) &= \left(\frac{1}{1-v}\right)\log\left[\sum_{k=0}^{v}\sum_{m=0}^{\infty}\binom{v}{k}\frac{(-1)^m(v\alpha)^m c^{k-1}(\alpha\beta)^{v-k}}{m!} s^{m\beta+v\beta-k\beta-v+1}\right. \\ &\times \left. B\left(v+\frac{v+k\beta-v\beta-m\beta-1}{c}, 1+\frac{m\beta+v\beta-k\beta-v+1}{c}\right)\right], \end{aligned} \tag{24}$$

for $v \neq 1$, and $v > 0$.

6. Maximum likelihood estimation

Let $X \sim LLoGW(s,c,\alpha,\beta)$ and $\boldsymbol{\Delta} = (s,c,\alpha,\beta)^T$ be the parameter vector. The log-likelihood

$\ell = \ell(\boldsymbol{\Delta})$ for a single observation x of X is given by

$$\ell(\boldsymbol{\Delta}) = -\alpha x^{\beta} - \log(1+(x/s)^c) + \log\left(\alpha\beta x^{\beta-1} + \frac{cx^{c-1}}{s^c+x^c}\right). \tag{25}$$

The first derivative of the log-likelihood function with respect to $\boldsymbol{\Delta} = (s, c, \alpha, \beta)^T$ are given by

$$\frac{\partial \ell}{\partial \alpha} = -x^{\beta} + \frac{\beta x^{\beta-1}}{\alpha\beta x^{\beta-1} + \frac{cx^{c-1}}{[s^c+x^c]}},$$

$$\frac{\partial \ell}{\partial \beta} = -\alpha x^{\beta}\log(x) + \frac{\alpha x^{\beta} + \alpha\beta x^{\beta-1}\log(x)}{\alpha\beta x^{\beta-1} + cx^{c-1}[s^c+x^c]^{-1}},$$

$$\frac{\partial \ell}{\partial s} = -(1+(x/s)^c)^{-1}(c/s)(x/s)^c - \frac{c(x/s)^{c-1}(s^c+x^c)^{-2}}{\alpha\beta x^{\beta-1} + cx^{c-1}(s^c+x^c)^{-1}},$$

and

$$\frac{\partial \ell}{\partial c} = -\frac{(1+(x/s)^c)^{-2}(x/s)^c\log(x/s)}{(1+(x/s)^c)^{-1}} + \frac{(s^c+x^c)^{-1}(x^{c-1}+cx^{c-1}\log(x)+cx^{c-1}-[s^c+x^c]^{-2}(s^c\log(s)+x^c\log(x)))}{\alpha\beta x^{\beta-1}+cx^{c-1}(s^c+x^c)^{-1}}.$$

The total log-likelihood function based on a random sample of n observations: $x_1, x_2, ..., x_n$ drawn from the LLoGW distribution is given by $\ell_n^* = L(\boldsymbol{\Delta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\Delta})$, where $\ell_i(\boldsymbol{\Delta})$, $i = 1, 2, ..., n$ is given by equation (25). The equations obtained by setting the above partial derivatives to zero are not in closed form and the values of the parameters s, c, α, β must be found by using iterative methods. The maximum likelihood estimates of the parameters, denoted by $\hat{\boldsymbol{\Delta}}$ is obtained by solving the nonlinear equation $(\frac{\partial \ell}{\partial s}, \frac{\partial \ell}{\partial c}, \frac{\partial \ell}{\partial \alpha}, \frac{\partial \ell}{\partial \beta})^T = \mathbf{0}$, using a numerical method such as Newton-Raphson procedure. The Fisher information matrix is given by $\mathbf{I}(\boldsymbol{\Delta}) = [\mathbf{I}_{\theta_i,\theta_j}]_{4X4} = E(-\frac{\partial^2 \ell}{\partial\theta_i\partial\theta_j})$, $i, j = 1, 2, 3, 4$, can be numerically obtained by MATLAB, SAS or R software, where $(\theta_1, \theta_2, \theta_3, \theta_4) = (s, c, \alpha, \beta)$. The total Fisher information matrix $n\mathbf{I}(\boldsymbol{\Delta})$ can be approximated by

$$\mathbf{J}_n(\hat{\boldsymbol{\Delta}}) \approx \left[-\frac{\partial^2 \ell}{\partial\theta_i\partial\theta_j}\bigg|_{\boldsymbol{\Delta}=\hat{\boldsymbol{\Delta}}} \right]_{4X4}, \quad i, j = 1, 2, 3, 4. \tag{26}$$

For a given set of observations, the matrix given in equation (26) is obtained after the convergence of the Newton-Raphson procedure.

6.1. Asymptotic confidence intervals

In this subsection, we present the asymptotic confidence intervals for the parameters of the LLoGW distribution. The expectations in the Fisher Information Matrix (FIM) can be obtained numerically. Let $\hat{\Delta} = (\hat{s}, \hat{c}, \hat{\alpha}, \hat{\beta})$ be the maximum likelihood estimate of $\boldsymbol{\Delta} = (s, c, \alpha, \beta)$. Under the usual regularity conditions and that the parameters are in the interior of the parameter space, but not on the boundary, we have: $\sqrt{n}(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \xrightarrow{d} N_4(\underline{0}, I^{-1}(\boldsymbol{\Delta}))$, where $I(\boldsymbol{\Delta})$ is the expected Fisher information matrix. The asymptotic behavior is still valid if $I(\boldsymbol{\Delta})$ is replaced by the observed information matrix evaluated at $\hat{\boldsymbol{\Delta}}$, that is $J(\hat{\boldsymbol{\Delta}})$. The multivariate normal distribution $N_4(\underline{0}, J(\hat{\boldsymbol{\Delta}})^{-1})$, where the mean vector $\underline{\mathbf{0}} = (0, 0, 0, 0)^T$, can be used to construct confidence intervals and confidence regions for the individual model parameters and for the survival and hazard rate functions. That is, the approximate $100(1-\eta)\%$ two-sided confidence intervals for s, c, α, and β are given by:

$$\widehat{s} \pm Z_{\frac{\eta}{2}}\sqrt{\mathbf{I}_{ss}^{-1}(\widehat{\boldsymbol{\Delta}})}, \quad \widehat{c} \pm Z_{\frac{\eta}{2}}\sqrt{\mathbf{I}_{cc}^{-1}(\widehat{\boldsymbol{\Delta}})}, \quad \widehat{\alpha} \pm Z_{\frac{\eta}{2}}\sqrt{\mathbf{I}_{\alpha\alpha}^{-1}(\widehat{\boldsymbol{\Delta}})} \quad \text{and} \quad \widehat{\beta} \pm Z_{\frac{\eta}{2}}\sqrt{\mathbf{I}_{\beta\beta}^{-1}(\widehat{\boldsymbol{\Delta}})},$$

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

7. Simulation study

In this section, we study the performance and accuracy of maximum likelihood estimates of the LLoGW model parameters by conducting various simulations for different sample sizes and different parameter values. Equation (8) is used to generate random data from the LLoGW distribution. The simulation study is repeated $N = 5,000$ times, each with sample size $n = 25, 50, 75, 100, 200, 400$ and parameter values $I : s = 5.5, c = 2.5, \alpha = 0.8, \beta = 0.2$ and $II : s = 5.5, c = 8.5, \alpha = 0.5, \beta = 0.5$. Four quantities are computed in this simulation study.

(b) Average bias of the MLE $\hat{\vartheta}$ of the parameter $\vartheta = s, c, \alpha, \beta$:

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\vartheta} - \vartheta).$$

(b) Root mean squared error (RMSE) of the MLE $\hat{\vartheta}$ of the parameter $\vartheta = s, c, \alpha, \beta$:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\vartheta} - \vartheta)^2}.$$

(c) Coverage probability (CP) of 95% confidence intervals of the parameter $\vartheta = s, c, \alpha, \beta$, i.e., the percentage of intervals that contain the true value of parameter ϑ.

(d) Average width (AW) of 95% confidence intervals of the parameter $\vartheta = s, c, \alpha, \beta$.

Table 4 presents the Average Bias, RMSE, CP and AW values of the parameters s, c, α and β for different sample sizes. From the results, we can verify that as the sample size n increases, the RMSEs decay toward zero. We also observe that for all the parametric values, the biases decrease as the sample size n increases. Also, the table shows that the coverage probabilities of the confidence intervals are quite close to the nominal level of 95% and that the average confidence widths decrease as the sample size increases. Consequently, the MLE's and their asymptotic results can be used for estimating and constructing confidence intervals even for reasonably small sample sizes.

Table 4: Monte Carlo simulation results: average bias, RMSE, CP and AW

		I				II			
Parameter	n	Average Bias	RMSE	CP	AW	Average Bias	RMSE	CP	AW
s	25	0.52811	2.64774	0.80320	7.31782	0.03521	0.63589	0.81620	1.84526
	50	0.32822	1.61445	0.88920	5.31108	0.01472	0.41423	0.89860	1.41529
	75	0.25455	1.25934	0.91700	4.27320	0.01039	0.32745	0.91660	1.17925
	100	0.20528	1.05246	0.92200	3.72238	0.00437	0.27415	0.92720	1.00656
	200	0.08874	0.66811	0.94000	2.55273	0.00214	0.18193	0.94120	0.70739
	400	0.07761	0.47016	0.94740	1.79814	0.00097	0.12919	0.94140	0.49884
c	25	2.42437	11.61202	0.93080	8.93786	5.95590	17.41663	0.93140	24.84591
	50	0.68223	5.33933	0.94160	4.00937	1.67338	5.23907	0.94320	12.11561
	75	0.36400	1.06252	0.94480	2.81549	0.90145	4.07377	0.94500	8.94710
	100	0.36400	1.06252	0.94480	2.81549	0.61040	2.21012	0.94460	7.36365
	200	0.09824	0.40474	0.95220	1.47864	0.26576	1.30460	0.95380	4.89920
	400	0.04759	0.26770	0.95320	1.01592	0.13375	0.88148	0.95400	3.38478
α	25	0.00953	0.23419	0.93000	0.87833	-0.00224	0.14260	0.92620	0.53624
	50	0.00756	0.16517	0.93520	0.62066	-0.00165	0.09929	0.93560	0.37911
	75	0.00320	0.13366	0.93240	0.50368	-0.00094	0.08211	0.93400	0.30948
	100	0.00320	0.11391	0.94540	0.43755	-0.00048	0.06811	0.94740	0.26802
	200	0.00146	0.07917	0.94460	0.30843	-0.00017	0.04886	0.94600	0.18899
	400	0.00130	0.05606	0.95060	0.21773	-0.00002	0.03443	0.95100	0.13387
β	25	0.02114	0.08255	0.96400	0.22165	0.02681	0.13760	0.94160	0.48931
	50	0.01420	0.03978	0.96100	0.14848	0.01408	0.08995	0.94700	0.33926
	75	0.01397	0.03350	0.95480	0.12064	0.01086	0.07166	0.95500	0.27499
	100	0.01181	0.02751	0.95900	0.10348	0.00608	0.06133	0.94760	0.23555
	200	0.00590	0.01832	0.95920	0.07108	0.00382	0.04249	0.94620	0.16553
	400	0.00503	0.01325	0.95080	0.04994	0.00159	0.02938	0.95140	0.11628

8. Applications

In this section, we present examples to illustrate the flexibility of the LLoGW distribution and its sub-models for data modeling. Estimates of the parameters of LLoGW distribution (standard error in parentheses), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Cramer von Mises (W^*), Andersen-Darling (A^*), and sum of squares (SS) from the probability plots are presented for each data set. We also compare the LLoGW distribution with the gamma-Dagum (GD) (Oluyede, Huang, and Pararai 2014) and beta Weibull (BW) (Lee, Famoye, and Olumolade 2007), (Famoye, Lee, and Olumolade 2005) distributions. The GD and BW pdfs are given by

$$g_{GD}(x) = \frac{\lambda\beta\delta x^{-\delta-1}}{\Gamma(\alpha)}(1+\lambda x^{-\delta})^{-\beta-1}(-\log[1-(1+\lambda x^{-\delta})^{-\beta}])^{\alpha-1}, \quad x>0,$$

and

$$g_{BW}(x) = \frac{k\lambda^k}{B(a,b)}x^{k-1}e^{-b(\lambda x)^k}(1-e^{-(\lambda x)^k})^{a-1}, \quad x>0,$$

respectively.

The maximum likelihood estimates (MLEs) of the LLoGW parameters $\Delta = (s, c, \alpha, \beta)$ are computed by maximizing the objective function via the subroutine NLMIXED in SAS as well as the function nlm in R (R Development Core Team 2011). The estimated values of the parameters (standard error in parenthesis), -2log-likelihood statistic, Akaike Information Criterion, $AIC = 2p - 2\ln(L)$, and Bayesian Information Criterion, $BIC = p\ln(n) - 2\ln(L)$, where $L = L(\hat{\Theta})$ is the value of the likelihood function evaluated at the parameter estimates, n is the number of observations, and p is the number of estimated parameters are presented in Tables 5 and 6. The LLoGW distribution is fitted to the data sets and these fits are compared to the fits using LLoGR, LLoGE and LLoG distributions.

As stated earlier, we maximize the likelihood function using NLmixed in SAS as well as the function nlm in R (R Development Core Team 2011). These functions were applied and executed for wide range of initial values. This process often results or lead to more than one maximum, however, in these cases, we take the MLEs corresponding to the largest value of the maxima. In a few cases, no maximum was identified for the selected initial values. In these cases, a new initial value was tried in order to obtain a maximum. The issues of existence and uniqueness of the MLEs are theoretical interest and has been studied by several authors for different distributions including (Seregin 2010), (Silva and Tenreyro 2010), (Zhou 2009), and (Xia, Mi, and Zhou 2009).

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

Plots of the fitted densities, the histogram of the data and probability plots (Chambers, Cleveland, Kleiner, and Tukey 1983) are given in Figures 3 and 4 for the first data set and Figures 5 and 6 for the second dataset. For the probability plot, we plotted $G(x_{(j)}; \hat{s}, \hat{c}, \hat{\alpha}, \hat{\beta})$ against $\frac{j - 0.375}{n + 0.25}$, $j = 1, 2, \cdots, n$, where $x_{(j)}$ are the ordered values of the observed data. The measure of closeness given by the sum of squares $SS = \sum_{j=1}^{n} \left[G(x_{(j)}) - \left(\frac{j - 0.375}{n + 0.25}\right)\right]^2$, was computed for each fitted model.

The goodness-of-fit statistics W^* and A^*, described by (Chen and Balakrishnan 1995) are also presented in the tables. These statistics can be used to verify which distribution fits better to the data. In general, the smaller the values of W^* and A^*, the better the fit. Let $G(x; \boldsymbol{\Delta})$ be the cdf, where the form of G is known but the k-dimensional parameter vector, say $\boldsymbol{\Delta}$ is unknown. We can obtain the statistics W^* and A^* as follows: (i) Compute $u_i = G(x_i; \hat{\boldsymbol{\Delta}})$, where the x_i's are in ascending order; (ii) Compute $y_i = \Phi^{-1}(u_i)$, where $\Phi(.)$ is the standard normal cdf and $\Phi^{-1}(.)$ its inverse; (iii) Compute $v_i = \Phi((y_i - \overline{y})/s_y)$, where $\overline{y} = n^{-1}\sum_{i=1}^{n} y_i$ and $s_y^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$; (iv) Calculate $W^2 = \sum_{i=1}^{n}\{v_i - (2i-1)/(2n)\}^2 + 1/(12n)$ and $A^2 = -n - n^{-1}\sum_{i=1}^{n}\{(2i-1)\log(v_i) + (2n+1-2i)\log(1-v_i)\}$; (v) Modify W^2 into $W^* = W^2(1 + 0.5/n)$ and A^2 into $A^* = A^2(1 + 0.75/n + 2.25/n^2)$.

8.1. Fracture toughness of alumina (Al_2O_3)(in the units of MPa $m^{1/2}$)

This data set consists of 119 observations on fracture toughness of Alumina (Al_2O_3)(in the units of MPa $m^{1/2}$. The data are taken from the web-site:
http://www.ceramics.nist.gov/ srd/summary/ftmain.htm. The same data set has also been studied by (Nadarajah and Kotz 2007). Initial values for the LLoGW model in R code are $s = 2.34, c = 3.42, \alpha = 0.155, \beta = 1$. Estimates of the parameters of LLoGW distribution and its related sub-models (standard error in parentheses), AIC, BIC, W*, A* and SS are give in Table 5. Plots of the fitted densities and the histogram are given in Figure 3, and plots of the observed probability vs predicted probability are given in Figure 4. The estimated variance-covariance matrix for the LLoGW distribution is

$$\begin{pmatrix} 0.09121 & -0.09084 & -0.00039 & 0.1929 \\ -0.09084 & 2.9051 & 0.00136 & -0.5062 \\ -0.00039 & 0.00136 & 4.21E-06 & -0.00166 \\ 0.1929 & -0.5062 & -0.00166 & 0.6950 \end{pmatrix},$$

and the 95% confidence intervals for the model parameters are given by $s \in (4.9735 \pm 1.96 \times 0.3020), c \in (10.4933 \pm 1.96 \times 1.7044), \alpha \in (0.00196 \pm 1.96 \times 0.0021)$ and $\beta \in (3.6271 \pm 1.96 \times$

0.8337), respectively.

Table 5: Estimates of models for fracture toughness of alumina data

	Estimates				Statistics					
Model	$\hat{s}$	$\hat{c}$	$\hat{\alpha}$	$\hat{\beta}$	$-2\log L$	AIC	BIC	W^*	A^*	SS
LLoGW	4.9735	10.4933	0.00196	3.6271	334.7	342.7	353.8	0.0416	0.2545	0.0408
	(0.3020)	(1.7044)	(0.0021)	(0.8337)						
LLoGE	4.5223	9.5011	0.02847	1	350.0	356.0	364.3	0.0659	0.5475	0.0395
	(0.0988)	(1.1499)	(0.0122)	-						
LLoG	4.3234	7.0649	0	1	356.7	360.7	366.2	0.3549	2.2314	0.1803
	(0.0959)	(0.5535)	-	-						
LLoGR	4.6670	10.6049	0.01245	2	339.6	345.6	354.0	0.0603	0.4009	0.0386
	(0.1060)	(1.2344)	(0.0037)	-						
	$\hat{\lambda}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\alpha}$						
GD	0.2687	0.5433	17.2156	26.4212	387.6	395.6	406.7	0.9288	5.3449	1.0675
	(0.8648)	(0.0.1388)	(2.4638)	(0.5059)						
	$\hat{\lambda}$	$\hat{k}$	$\hat{a}$	$\hat{b}$						
BW	0.09969	5.6903	0.8010	55.2052	337.1	345.1	356.2	0.0829	0.5021	0.0785
	(0.0102)	(1.4654)	(0.3214)	(4.7545)						

Note. Standard errors are in parentheses.

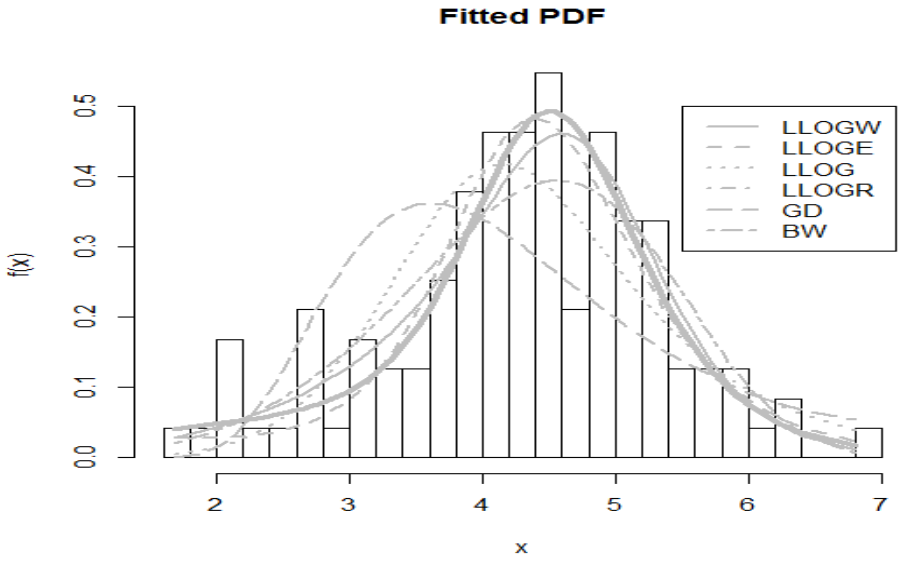


Figure 3: Fitted densities for fracture toughness of alumina data

The LR test statistic of the hypothesis H_0: LLoGE against H_a: LLoGW, H_0:LLoG against H_a: LLoGW, and H_0:LLoGR against H_a: LLoGW are 15.3 (p-value < 0.0001), 22.0 (p-value < 0.0001), and 4.9 (p-value=0.02686 < 0.05). We can conclude that there are significant differences between LLoGW and LLoGE, LLoG, LLoGR distributions. The values of the statistics: AIC and BIC also shows that the LLoGW distribution is a better fit than the non-nested GD and BW distributions for the fracture toughness of alumina data. There is also clear evidence based on the goodness-of-fit statistics W^* and A^* that the LLoGW distribution is by far the better fit for the fracture toughness of alumina data.

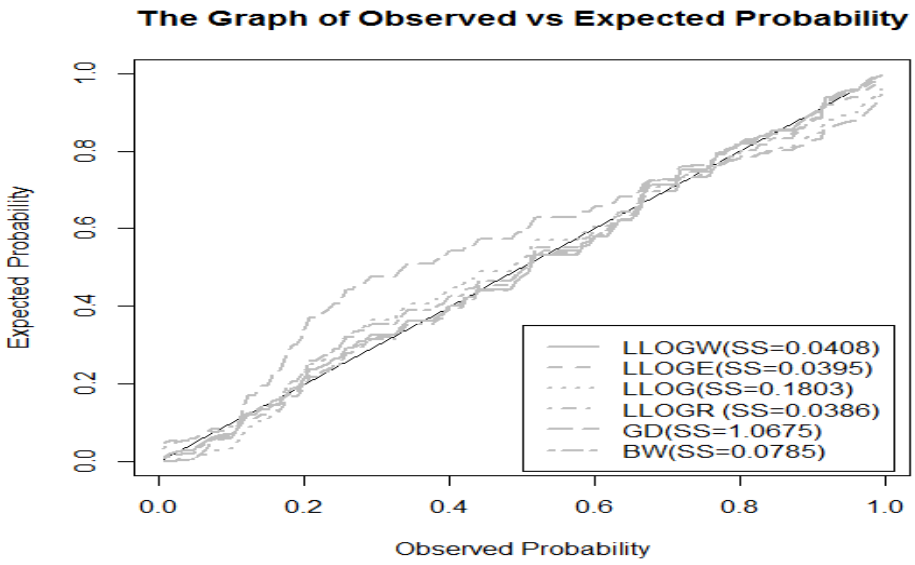


Figure 4: Probability pots for fracture toughness of alumina data

8.2. Breaking stress of carbon fibres (in Gba)

This data set consists of 100 uncensored data on breaking stress of carbon fibres (in Gba), (Nichols and Padgett 2006). Initial values for the LLoGW model in R code are $s = 1, c = 1, \alpha = 1, \beta = 3$. Estimates of the parameters of LLoGW distribution and its related sub-models (standard error in parentheses), AIC, BIC, W*, A* and SS are give in Table 6. Plots of the fitted densities and the histogram, as well as observed probability vs predicted probability are given in Figures 5 and 6, respectively. The estimated variance-covariance matrix for the LLoGW distribution is

$$\begin{pmatrix} 0.5771 & -0.5628 & 0.0014 & 0.2189 \\ -0.5628 & 4.1583 & 0.0415 & -0.9957 \\ 0.0014 & 0.0415 & 0.0008 & -0.0128 \\ 0.2189 & -0.9957 & -0.0128 & 0.3387 \end{pmatrix},$$

and the 95% confidence intervals for the model parameters are given by $s \in (3.4839 \pm 1.96 \times 0.7597), c \in (4.3649 \pm 1.96 \times 2.0392), \alpha \in (0.0418 \pm 1.96 \times 0.0274)$ and $\beta \in (2.5196 \pm 1.96 \times 0.5820)$, respectively.

The LR test statistic of the hypothesis H_0: LLoGE against H_a: LLoGW and H_0:LLoG against H_a: LLoGW are 7.5791 (p-value = 0.0059) and 10.1860 (p-value = 0.0014). We can conclude that there are significant difference between LLoGW and LLoGE distributions as well between LLoGW and LLoG distributions. There is also very clear and convincing evidence based on the goodness-of-fit statistics W^* and A^* that the LLoGW distribution is by far the better fit than the sub-models. The SS value of 0.0584 for the LLoGW distribution is smaller than the values for the non-nested GD and BW distributions. The values of AIC and BIC also shows that the LLoGW distribution is a better fit than the non-nested GD and BW distributions for the breaking stress of carbon fibres data.

9. Concluding remarks

We have presented a new distribution called the log-logistic Weibull (LLoGW) distribution that is suitable for applications in various areas including reliability, survival analysis and actuarial

Table 6: Estimates of models for breaking stress of carbon fibres data

	Estimates				Statistics					
Model	$\hat{s}$	$\hat{c}$	$\hat{\alpha}$	$\hat{\beta}$	$-2\log L$	AIC	BIC	W^*	A^*	SS
LLoGW	3.4839	4.3649	0.0418	2.5196	282.37	290.37	300.79	22.2883	133.4936	0.0584
	(0.7597)	(2.0392)	(0.0274)	(0.5820)						
LLoGE	2.6224	4.5962	0.0311	1	289.95	295.95	303.77	24.5245	137.8047	0.1268
	(0.1482)	(0.5469)	(0.0298)	-						
LLoG	2.4984	4.1179	0	1	292.56	298.56	306.37	23.8320	139.0570	0.1664
	(0.1054)	(0.3441)	-	-						
LLoGR	3.1737	5.3261	0.0595	2	283.48	289.48	297.30	23.6256	133.9022	0.0507
	(0.3043)	(1.1471)	(0.0213)	-						
	$\hat{\lambda}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\alpha}$						
GD	36.8189	4.1911	3.7238	0.2278	289.06	297.06	307.48	23.7881	138.2885	0.1716
	(16.3529)	(1.5857)	(0.3264)	(0.1076)						
	$\hat{\lambda}$	$\hat{k}$	$\hat{a}$	$\hat{b}$						
BW	0.2002	2.3259	1.3707	4.8215	282.69	290.69	301.11	22.9775	133.4390	0.0680
	(0.0179)	(0.7563)	(0.7529)	(0.0031)						

Note. Standard errors are in parentheses.

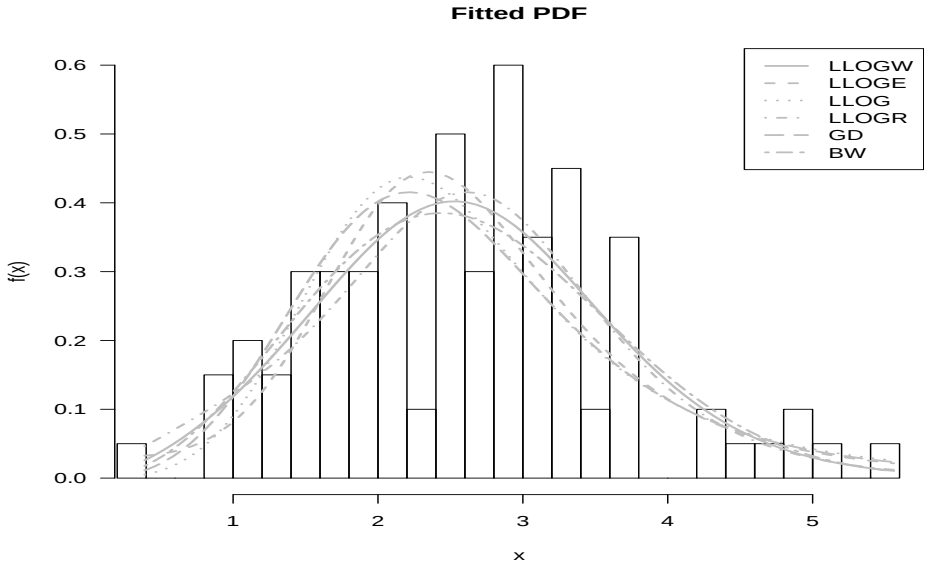


Figure 5: Fitted densities for breaking stress of carbon fibres data

sciences just to mention a few areas. The structural properties including hazard and reverse hazard functions, quantile function, probability weighted moments (PWMs), moments, conditional moments, mean deviations, Bonferroni and Lorenz curves, Rényi entropy, distribution of order statistics, maximum likelihood estimates, asymptotic confidence intervals are presented. Applications of the model to real data sets are given in order to illustrate the applicability and usefulness of the proposed distribution.

Acknowledgements
The authors are very grateful to the referees for some useful comments on an earlier version of this manuscript which led to this improved version.

References

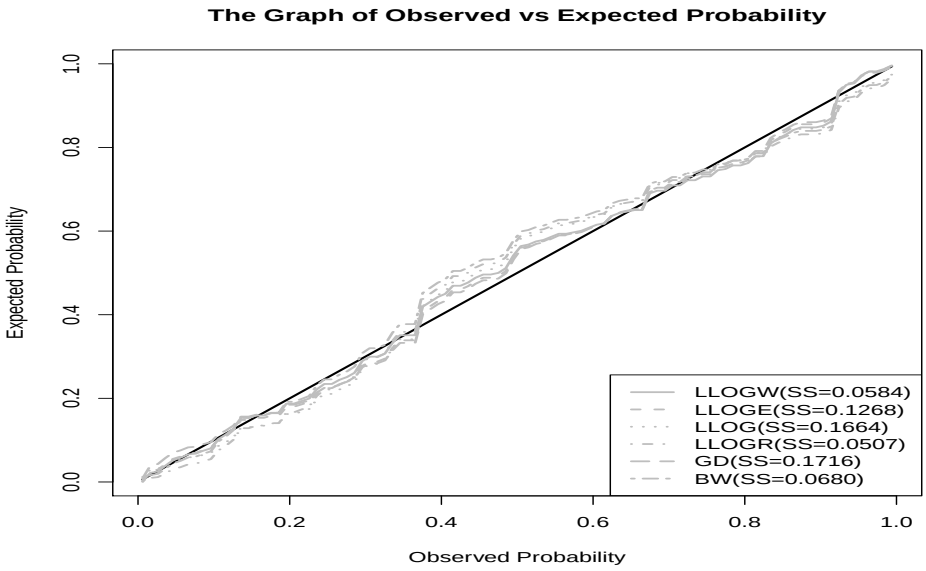


Figure 6: Probability plots for breaking stress of carbon fibres data

Barakat HM, Abdelkader YH (2004). "Computing the Moments of Order Statistics from Nonidentical Random Variables." *Statistical Methods and Applications*, **13**(1), 15–26.

Bourguignon M, Silva RB, Cordeiro GM (2014). "The Weibull–G Family of Probability Distributions." *Journal of Data Science*, **12**(1), 53–68.

Burr IW (1942). "Cumulative Frequency Functions." *The Annals of Mathematical Statistics*, **13**(2), 215–232.

Burr IW (1973). "Pameters for a General System of Distributions to Match a Grid of $\alpha 3$ and $\alpha 4$." *Communications in Statistics-Theory and Methods*, **2**(1), 1–21.

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

Chambers JM, Cleveland WS, Kleiner B, Tukey PA (1983). *Graphical Methods of Data Analysis*. Chapman and Hall.

Chen G, Balakrishnan N (1995). "A General Purpose Approximate Goodness-of-fit Test." *Journal of Quality Technology*, **27**(2), 154–161.

Durbin J, Koopman SJ (2012). *Time Series Analysis by State Space Methods*. 38. Oxford University Press.

Eugene N, Lee C, Famoye F (2002). "Beta-normal Distribution and Its Applications." *Communications in Statistics-Theory and Methods*, **31**(4), 497–512.

Famoye F, Lee C, Olumolade O (2005). "The Beta-Weibull Distribution." *Journal of Statistical Theory and Applications*, **4**(2), 121–136.

Gradshtein IS, Ryzhik IM (2000). *Tables of Integrals, Series and Products*. Academic Press, San Diego.

Greenwood J, Landwehr JM, Matalas NC, Wallis JR (1979). "Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Expressable in Inverse Form." *Water Resources Research*, **15**(5), 1049–1054.

Gupta RD, Kundu D (2001). "Exponentiated Exponential Family: An Alternative to Gamma and Weibull Distributions." *Biometrical Journal*, **43**(1), 117–130.

Gurvich MR, Dibenedetto AT, Ranade SV (1997). "A New Statistical Distribution for Characterizing the Random Strength of Brittle Materials." *Journal of Materials Science*, **32**(10), 2559–2564.

Haupt E, Schäbe H (1992). "A New Model for a Lifetime Distribution with Bathtub Shaped Failure Rate." *Microelectronics Reliability*, **32**(5), 633–639.

Hjorth U (1980). "A Reliability Distribution with Increasing, Decreasing, Constant and Bathtub-Shaped Failure Rates." *Technometrics*, **22**(1), 99–107.

Hoskings JRM (1990). "L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics." *Journal of the Royal Statistical Society*, **B52**, 105–124.

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

Lai CD, Xie M, Murthy DNP (2003). "A Modified Weibull Distribution." *IEEE Transactions on Reliability*, **52**(1), 33–37.

Lee C, Famoye F, Olumolade O (2007). "Beta-Weibull Distribution: Some Properties and Applications to Censored Data." *Journal of Modern Applied Statistical Methods*, **6**(1), 173–186.

Mudholkar GS, Srivastava DK (1993). "Exponentiated Weibull Family for Analyzing Bathtub Failure-Rate Data." *IEEE Transactions on Reliability*, **42**(2), 299–302.

Murthy DNP, Xie M, Jiang R (2004). *Weibull Models*, volume 505. John Wiley & Sons.

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

Nadarajah S, Kotz S (2006). "The Beta Exponential Distribution." *Reliability Engineering & System Safety*, **91**(6), 689–697.

Nadarajah S, Kotz S (2007). "On the Alternative to the Weibull Function." *Engineering Fracture Mechanics*, **74**(3), 451–456.

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

Nelson W (1982). *Lifetime Data Analysis*. Wiley, New York.

Nichols MD, Padgett WJ (2006). "A Bootstrap Control Chart for Weibull Percentiles." *Quality and Reliability Engineering International*, **22**(2), 141–151.

Oluyede BO, Huang S, Pararai M (2014). "A New Class of Generalized Dagum Distribution with Applications to Income and Lifetime Data." *Journal of Statistical and Econometric Methods*, **3**(2), 125–151.

Paranaíba PF, Ortega EMM, Cordeiro GM, Pescim RR (2011). "The Beta Burr XII Distribution with Application to Lifetime Data." *Computational Statistics & Data Analysis*, **55**(2), 1118–1136.

Pham H, Lai CD (2007). "On Recent Generalizations of the Weibull Distribution." *IEEE Transactions on Reliability*, **56**(3), 454–458.

Pinho LGB, Cordeiro GM, Nobre JS (2012). "The Gamma–Exponentiated Weibull Distribution." *Journal of Statistical Theory and Applications*, **11**(4), 379–395.

Prudnikov AP, Brychkov YA, Marichev O I (1992). *Integrals and Series*, volume 4. Gordon and Breach Science Publishers, Amsterdam.

R Development Core Team (2011). "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2011."

Rajarshi S, Rajarshi MB (1988). "Bathtub Distributions: A Review." *Communications in Statistics-Theory and Methods*, **17**(8), 2597–2621.

Renyi A (1960). "On Measures of Entropy and Information." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561.

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

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

Silva GO, Ortega EMM, Cordeiro GM (2010). "The Beta Modified Weibull Distribution." *Lifetime Data Analysis*, **16**(3), 409–430.

Silva JMCS, Tenreyro S (2010). "On the Existence of the Maximum Likelihood Estimates in Poisson Regression." *Economics Letters*, **107**(2), 310–312.

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

Zhou C (2009). "Existence and Consistency of the Maximum Likelihood Estimator for the Extreme Value Index." *Journal of Multivariate Analysis*, **100**(4), 794–815.

R code

In this section, the R codes to compute cdf, pdf, moments, Rényi entropy, mean deviations, maximum likelihood estimates and variance-covariance matrix for the LLoGW distribution are presented.

```
#Define the pdf of LLoGW distribution
f1=function(x,s,c,alpha,beta){
  y=(((1+(x/s)**c)**(-1))*exp(-alpha*(x**beta)))
  *(((1+(x/s)**c)**(-1))*(c/s)*(x/s)**(c-1)
  +alpha*beta*(x**(beta-1)))
  return(y)
}

#Define the cdf of LLoGW distribution
F1=function(x,s,c,alpha,beta){
  y=1-((1+(x/s)**c)
  **(-1))*exp(-alpha*x**beta)
  return(y)
}
```

```
#Define the moments of LLoGW distribution
moment=function(s,c,alpha,beta,r){
f=function(x,s,c,alpha,beta,r)
  {(x^r)*(f1(x,s,c,alpha,beta))}
  y=integrate(f,lower=0,upper=Inf,
  subdivisions=100,s=s,c=c,alpha=alpha,beta=beta,r=r)
  return(y)
}

#Define the quantile of LLoGW distribution
quantile=function(s,c,alpha,beta,u){
f=function(x){alpha*x^beta+log(1+(x/s)^c)+log(1-u)}
  rc <- uniroot(f, lower=0, upper=100, tol = 1e-9)
  result=rc$root
  #check
  error=F1(result,s,c,alpha,beta)-u
  return(list("result"=result,"error"=error))
}

#Define Mean Deviation about the mean of LLoGW distribution
DU=function(s,c,alpha,beta){
  mu=moment(s,c,alpha,beta,1)$ value
  f=function(x,s,c,alpha,beta){(abs(x-mu)*f1(x,s,c,alpha,betaa)}
  y=integrate(f,lower=0,upper=Inf,subdivisions=100
  ,s=s,c=c,alpha=alpha,beta=beta)
  return(y)
}

#Define Mean Deviation about the median of LLoGW distribution
DM=function(s,c,alpha,beta){
  M=median(c(X)) #X is the data set
  f=function(x,s,c,alpha,beta){(abs(x-M)*f1(x,s,c,alpha,beta)}
  y=integrate(f,lower=0,upper=Inf,subdivisions=100
  ,s=s,c=c,alpha=alpha,beta=beta)
  return(y)
}

Define the Renyi entropy of LLoGW distribution
t=function(s,c,alpha,beta,v){
  f=function(x,s,c,alpha,beta,v)
    {(f1(x,s,c,alpha,beta))^(v)}
    y=integrate(f,lower=0,upper=Inf,subdivisions=100
    ,s=s,c=c,alpha=alpha,beta=beta,v=v)$ value
    return(y)
  }
  Renyi=function(s,c,alpha,beta,v){
  y=log(t(s,c,alpha,beta,v))/(1-v)
  return(y)
}

#Calculate the maximum likelihood estimators
#of LLoGW distribution
library('bbmle')
xvec<-c(X) #X is the data set
ln<-function(s,c,alpha,beta){
  -sum(log((((1+(x/s)**c)**(-1))*exp(-alpha*(x**beta)))
  *(((1+(x/s)**c)**(-1))*(c/s)*(x/s)**(c-1)
```

```
  +alpha*beta*(x**(beta-1)))))
}
mle.results1<-mle2(ln, start=list(s=s, c=c, alpha=alpha
, beta=beta), hessian.opt=TRUE)
summary(mle.results1)

# Variance-covariance matrx of LLoGW distribution
vcov(mle.results1)
```


Affiliation:

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

Susan Foya
Department of Mathematics and Computational Sciences
Botswana International University of Science and Technology
Palapye, BW

Gayan Warahena-Liyanage
Department of Mathematics
Central Michigan University
Mount Pleasant, MI, 48859, USA

Shujiao Huang
Department of Mathematics
University of Houston
Houston, TX, 77004, USA

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 45
September 2016

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

Submitted: 2015-04-20
Accepted: 2015-06-15

Austrian Journal of Statistics
September 2016, Volume 45, 71–93.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i3.109

Another Generalized Transmuted Family of Distributions: Properties and Applications

Faton Merovci
University of Mitrovica

Morad Alizadeh
Persian Gulf University

G. G. Hamedani
Marquette University

Abstract

We introduce and study general mathematical properties of a new generator of continuous distributions with two extra parameters called the *Another generalized transmuted family of distributions*. We present some special models. We investigate the asymptotes and shapes. The new density function can be expressed as a linear combination of exponentiated densities based on the same baseline distribution. We obtain explicit expressions for the ordinary and incomplete moments and generating functions, Bonferroni and Lorenz curves, asymptotic distribution of the extreme values, Shannon and Rényi entropies and order statistics, which hold for any baseline model, certain characterisations are presented. Further, we introduce a bivariate extensions of the new family. We discuss the different method of estimation of the model parameters and illustrate the potentiality of the family by means of two applications to real data. A brief simulation for evaluating Maximum likelihood estimator is done.

Keywords: transmuted distribution, generated family, maximum likelihood, moment, order statistic, quantile function, Rényi entropy, characterizations..

1. Introduction

Numerous classical distributions have been extensively used over the past decades for modeling data in several areas such as engineering, actuarial, environmental and medical sciences, biological studies, demography, economics, finance and insurance. However, in many applied areas such as lifetime analysis, finance and insurance, there is a clear need for extended forms of these distributions. For that reason, several methods for generating new families of distributions have been studied. Some attempts have been made to define new families of probability distributions that extend well-known families of distributions and at the same time provide great flexibility in modeling data in practice.

In many practical situations, classical distributions do not provide adequate fits to real data. For example, if the data are asymmetric, the normal distribution will not be a good choice. So, several generators employing one or more parameters to generate new distributions have been proposed in the statistical literature. Some well-known generators are Marshal-Olkin generated family (MO-G) Marshall and Olkin (1997), the beta-G by Eugene, Lee, and Famoye (2002) and Jones (2004), Kumaraswamy-G (Kw-G for short) Cordeiro and de Castro (2011),

McDonald-G (Mc-G) by Alexander, Cordeiro, Ortega, and Sarabia (2012), gamma-G (type 1) by Zografos and Balakrishnan (2009), gamma-G (type 2) by Ristić and Balakrishnan (2012), gamma-G (type 3) by Torabi and Hedesh (2012), log-gamma-G by Amini, MirMostafaee, and Ahmadi (2012), logistic-G by Tahir, Cordeiro, Alzaatreh, Mansoor, and Zubair (2015a), exponentiated generalized-G by Cordeiro, Ortega, and da Cunha (2013), Transformed-Transformer (T-X) by Alzaatreh, Lee, and Famoye (2013), exponentiated (T-X) by Alzaghal, Famoye, and Lee (2013), Weibull-G by Bourguignon, Silva, and Cordeiro (2014), Exponentiated half logistic generated family by Cordeiro, Alizadeh, and Ortega (2014a), Lomax-G by Cordeiro, Ortega, Popović, and Pescim (2014b), Kumaraswamy Odd log-logistic-G by Alizadeh, Emadi, Doostparast, Cordeiro, Ortega, and Pescim (2015b), Kumaraswamy Marshall-Olkin by Alizadeh, Tahir, Cordeiro, Mansoor, Zubair, and Hamedani (2015c), Beta Marshall-Olkin by Alizadeh, Cordeiro, De Brito, and Demétrio (2015a), Type 1 Half-Logistic family of distributions by Cordeiro, Alizadeh, and Diniz Marinho (2015) and Odd generalized exponential-G by Tahir, Cordeiro, Alizadeh, Mansoor, Zubair, and Hamedani (2015b).

Let $r(t)$ be the probability density function (pdf) of a random variable $T \in [a, b]$ for $-\infty < a < b < \infty$ and let $W[G(x)]$ be a function of the cumulative distribution function (cdf) of a random variable X satisfying the following conditions:

$$\begin{cases} (i) & W[G(x)] \in [a, b], \\ (ii) & W[G(x)] \text{ is differentiable and monotonically non-decreasing, and} \\ (iii) & W[G(x)] \to a \text{ as } x \to -\infty \text{ and } W[G(x)] \to b \text{ as } x \to \infty. \end{cases} \tag{1}$$

Recently, Alzaatreh *et al.* (2013) defined the *T-X family* of distributions by

$$F(x) = \int_a^{W[G(x)]} r(t)\, dt, \tag{2}$$

where $W[G(x)]$ satisfies the conditions (1). The pdf corresponding to (2) is given by

$$f(x) = \left\{ \frac{d}{dx} W[G(x)] \right\} r\{W[G(x)]\}. \tag{3}$$

Taking $W[G(x)] = 1 - \left(\bar{G}(x)\right)^{\alpha}$ and $r(t) = 1 + \lambda - 2\lambda t \;\; , 0 < t < 1$, we define the cumulative distribution function (cdf) of the *Another Generalized Transmuted Class (AGT-G for short)* of distributions by

$$F(x; \lambda, \alpha, \boldsymbol{\xi}) = (1 + \lambda)\left[1 - \left(\bar{G}(x; \boldsymbol{\xi})\right)^{\alpha}\right] - \lambda\left[1 - \left(\bar{G}(x; \boldsymbol{\xi})\right)^{\alpha}\right]^2 \quad , \alpha > 0 \; , |\lambda| \leq 1 \tag{4}$$

where $G(x; \boldsymbol{\xi})$ is the baseline cdf depending on a parameter vector $\boldsymbol{\xi}$ and $\alpha > 0$ and $|\lambda| \leq 1$ are two additional shape parameters. For each baseline cdf G, the AGT-G family of distributions is defined by the cdf (4). It includes the *Transmuted family* of distributions and the proportional reversed hazard rate models. Some special models are given in Table 1.

This paper is organized as follows. In Section 2, we define the AGT-G family. Three special cases of this family are defined in Section 3. In Section 4, the asymptotic and shape of the density and hazard rate functions are expressed analytically. Some useful expansions are derived in Section 5. In Section 6, we provide explicit expressions for the moments, incomplete moments, generating function and mean deviation. Extreme values are discussed in Section 7. General expressions for the Rényi and Shannon entropies are presented in Section 8. General results for order statistics are obtained in Section 9. Certain characterisations are given in Section 10. In Section 11, we introduce a bivariate extension of the new family. Estimation procedures of the model parameters are presented in Section 12. Applications to two real data sets illustrate the performance of the new family in Section 13. The paper is concluded in Section 14.

Table 1: Some known special cases of the AGT-G model.

α	λ	$G(x)$	Reduced distribution
1	-	$G(x)$	Transmuted G family of distributions Shaw and Buckley (2009)
-	0	$G(x)$	Proportioanl hazard rate family Gupta and Gupta (2007)
1	0	$G(x)$	$G(x)$
1	-	exponentiated exponential	Transmuted exponentiated exponential distribution Merovci (2013a)
1	-	Pareto	Transmuted Pareto distribution Merovci and Puka (2014)
1	-	Gumbel	Transmuted Gumbel distribution Aryal and Tsokos (2009)
1	-	Weibull	Transmuted Weibull distribution Aryal and Tsokos (2011)
1	-	inverse Weibull	Transmuted inverse Weibull distribution Merovci, Elbatal, and Ahmed (2014)
1	-	Lindley	Transmuted Lindley distribution Merovci (2013b)
1	-	Lindley-geometric	Transmuted Lindley-geometric Merovci and Elbatal (2014a)
1	-	Weibull-geometric	Transmuted Weibull-geometric Merovci and Elbatal (2014b)
1	-	Rayligh	Transmuted Rayligh distribution Merovci (2013c)
-	-	Generalized Rayligh	Transmuted Generalized Rayligh distribution Merovci (2014)
1	-	extreme value	Transmuted extreme value distribution Aryal and Tsokos (2009)
1	-	log-logistic	Transmuted log-logistic distribution Aryal (2013)

2. The new family

The corresponding density function to (4) is given by

$$f(x;\lambda,\alpha,\boldsymbol{\xi}) = \alpha g(x,\boldsymbol{\xi})\left(\bar{G}(x,\boldsymbol{\xi})\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}(x;\boldsymbol{\xi})\right)^{\alpha}\right]\right\} \tag{5}$$

where $g(x;\boldsymbol{\xi})$ is the baseline pdf. Equation (5) will be most tractable when the cdf $G(x)$ and the pdf $g(x)$ have simple analytic expressions. Hereafter, a random variable X with density function (5) is denoted by $X \sim$ AGT-G$(\alpha,\lambda,\boldsymbol{\xi})$. Further, we can omit (sometimes) the dependence on the vector $\boldsymbol{\xi}$ of the parameters and simply write $G(x) = G(x;\boldsymbol{\xi})$.

The hazard rate function (hrf) of X becomes

$$h(x;\lambda,\alpha,\boldsymbol{\xi}) = \frac{\alpha g(x,\boldsymbol{\xi})\left(\bar{G}(x,\boldsymbol{\xi})\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}(x;\boldsymbol{\xi})\right)^{\alpha}\right]\right\}}{1-(1+\lambda)\left[1-\left(\bar{G}(x;\boldsymbol{\xi})\right)^{\alpha}\right]+\lambda\left[1-\left(\bar{G}(x;\boldsymbol{\xi})\right)^{\alpha}\right]^2} \tag{6}$$

To motivate the new family, let Z_1, Z_2 be $i.i.d$ random variables from $1-\left(\bar{G}(x;\boldsymbol{\xi})\right)^{\alpha}$ and $Z_{1:2} = \min\{Z_1, Z_2\}$ and $Z_{2:2} = \max\{Z_1, Z_2\}$, and let

$$V = \begin{cases} Z_{1:2}, & \text{with probability } \frac{1+\lambda}{2}; \\ Z_{2:2}, & \text{with probability } \frac{1-\lambda}{2}. \end{cases}$$

Then $F_V(x;\lambda,\alpha,\boldsymbol{\xi}) = (1+\lambda)\left[1-\left(\bar{G}(x;\boldsymbol{\xi})\right)^{\alpha}\right]-\lambda\left[1-\left(\bar{G}(x;\boldsymbol{\xi})\right)^{\alpha}\right]^2$, which is the proposed family. The AGT-G family of distributions is easily simulated by inverting (4) as follows: if U has a uniform $U(0,1)$ distribution,then

$$X_U = G^{-1}\left\{1-\left[\frac{\lambda-1+\sqrt{(1+\lambda)^2-4\lambda U}}{2\lambda}\right]^{\frac{1}{\alpha}}\right\} \quad for\ \lambda \neq 0 \tag{7}$$

has the density function (5).

3. Special AGT-G distributions

In the following sections, we study some mathematical properties of AGT-G distribution since it extends several widely-known distributions in the literature. First, we discuss some special AGT-G distributions.

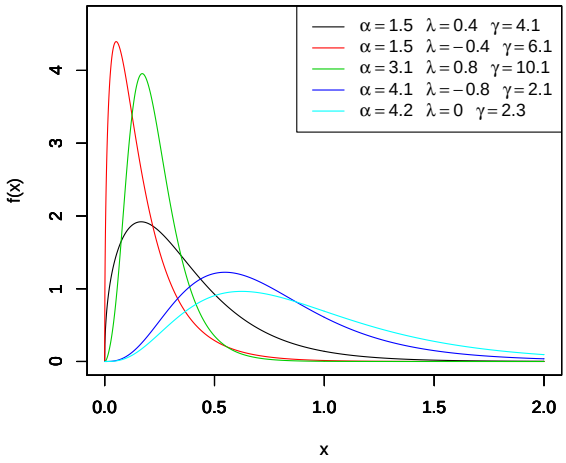


Figure 1: The pdf's of various AGT-E distributions .

3.1. AGT-Exponential(AGT-E) distribution

The parent exponential distribution has pdf and cdf given, respectively, by

$$g(x,\gamma) = \gamma \exp(-\gamma\ x) \tag{8}$$

and

$$G(x,\gamma) = 1 - \exp(-\gamma\ x) \tag{9}$$

The cdf and pdf of AGT-Exponential distribution are given by $(x > 0)$

$$F(x;\lambda,\alpha,\gamma) = (1+\lambda)\left[1-\exp(-\alpha\ \gamma\ x)\right] - \lambda\left[1-\exp(-\alpha\ \gamma\ x)\right]^2 \quad ,\alpha > 0\ ,|\lambda| \leq 1 \tag{10}$$

$$f(x;\lambda,\alpha,\gamma) = \alpha\gamma \exp(-\alpha\ \gamma\ x)\left\{1+\lambda-2\lambda\left[1-\exp(-\alpha\ \gamma\ x)\right]\right\} \tag{11}$$

Figure 1 illustrates some of the possible shapes of the pdf of the AGT-E distribution.
The expectation and variance of AGT-E are:

$$E(X) = \frac{2-\lambda}{2\alpha\gamma} \text{ and } var(X) = \frac{4-3\lambda}{4\alpha^2\gamma^2.}$$

3.2. AGT-Fréchet (AGT-F) distribution

The parent Fréchet distribution has cdf and pdf given, respectively, by

$$G(x;a,b) = \exp\left(-\left(\frac{b}{x}\right)^a\right), \quad a>0, b>0, x>0, \tag{12}$$

and

$$g(x;a,b) = ab^a x^{-(a+1)} \exp\left(-\left(\frac{b}{x}\right)^a\right) \tag{13}$$

The cdf and pdf of AGT-Fréchet distribution are given by $(x > 0)$:

$$\begin{aligned} F(x;\lambda,\alpha,a,b) = (1+\lambda)&\left[1-\left[1-\exp\left(-\left(\frac{b}{x}\right)^a\right)\right]^\alpha\right] \\ &-\lambda\left[1-\left[1-\exp\left(-\left(\frac{b}{x}\right)^a\right)\right]^\alpha\right]^2 ,\alpha>0\ ,|\lambda|\leq 1, a>0, b>0, \end{aligned} \tag{14}$$

and

$$f(x;\lambda,\alpha,a,b) = \alpha a b^a x^{-(a+1)} \exp\left(-\left(\tfrac{b}{x}\right)^a\right)\left[1-\exp\left(-\left(\tfrac{b}{x}\right)^a\right)\right]^{\alpha-1} \times\left\{1+\lambda-2\lambda\left[1-\left[1-\exp\left(-\left(\tfrac{b}{x}\right)^a\right)\right]^{\alpha}\right]\right\} \tag{15}$$

3.3. AGT-Normal(AGT-N) distribution

The cdf and pdf of AGT-Normal distribution are given by:

$$F(x;\lambda,\alpha,\mu,\sigma) = (1+\lambda)\left[1-\left(1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right)^{\alpha}\right] - \lambda\left[1-\left(1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right)^{\alpha}\right]^2, \alpha>0\ , |\lambda|\le 1, \tag{16}$$

and

$$f(x;\lambda,\alpha,\mu,\sigma) = \alpha\phi\left(\frac{x-\mu}{\sigma}\right)\left(1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right)^{\alpha-1} \times\left\{1+\lambda-2\lambda\left[1-\left(1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right)^{\alpha}\right]\right\} \tag{17}$$

3.4. The AGT-Uniform (AGT-U) distribution

The parent uniform distribution in the interval $(0,\theta),\quad \theta>0$ has cdf and pdf given, respectively, by

$$G(x;\theta) = \frac{x}{\theta} \tag{18}$$

and

$$g(x;\theta) = \frac{1}{\theta} \tag{19}$$

The cdf and pdf of AGT-Uniform distribution are given by:

$$F(x;\lambda,\alpha,\theta) = (1+\lambda)\left[1-\left(1-\frac{x}{\theta}\right)^{\alpha}\right] - \lambda\left[1-\left(1-\frac{x}{\theta}\right)^{\alpha}\right]^2, \alpha>0\ , |\lambda|\le 1 \tag{20}$$

and

$$f(x;\lambda,\alpha,\theta) = \frac{\alpha}{\theta}\left(1-\frac{x}{\theta}\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left[1-\left(1-\frac{x}{\theta}\right)^{\alpha}\right]\right\} \tag{21}$$

3.5. The AGT-Weibull (AGT-W) distribution

The parent Weibull distribution has cdf and pdf given by, respectively:

$$G(x;a,b) = 1-\exp\left(-bx^a\right) \tag{22}$$

and

$$g(x;a,b) = abx^{a-1}\exp\left(-bx^a\right) \tag{23}$$

The cdf and pdf of AGT-Weibull distribution are given by

$$F(x;\lambda,\alpha,a,b) = (1+\lambda)\left[1-\exp\left(-\alpha bx^a\right)\right] - \lambda\left[1-\exp\left(-\alpha bx^a\right)\right]^2, \alpha>0\ , |\lambda|\le 1 \tag{24}$$

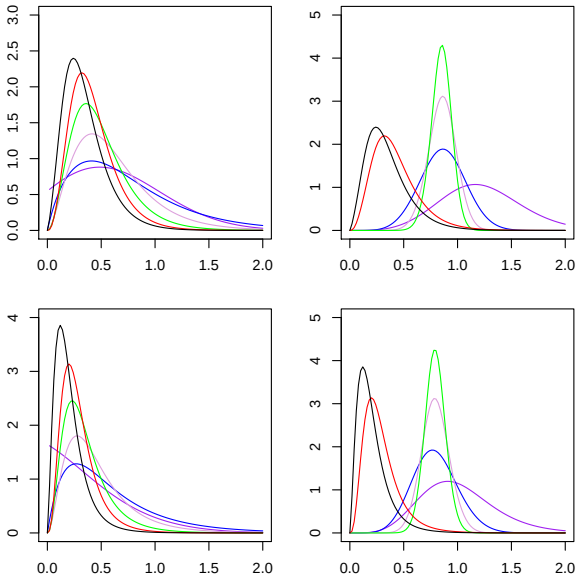


Figure 2: The pdf's of various AGT-Weibull distributions .

$$f(x;\lambda,\alpha,a,b) = \alpha a b x^{a-1}\exp\left(-\alpha b x^{a}\right) \\ \left\{1+\lambda-2\lambda\left[1-\exp\left(-\alpha b x^{a}\right)\right]\right\} \tag{25}$$

Figure 2 illustrates possible shapes of the density functions for some AGT-Weibull distributions.

4. Asymptotics and shapes

Proposition 1 *The asymptotics of equations (4), (5) and (6) as* $G(x)\to 0$ *are given by*

$$F(x)\sim \alpha\left(1+\lambda\right)G(x) \quad as \;\; G(x)\to 0,$$
$$f(x)\sim \alpha\left(1+\lambda\right)g(x) \quad as \;\; G(x)\to 0,$$
$$h(x)\sim \alpha\left(1+\lambda\right)g(x) \quad as \;\; G(x)\to 0.$$

Proposition 2 *The asymptotics of equations (4), (5) and (6) as* $x\to\infty$ *are given by*

$$1-F(x)\sim \left(\bar{G}(x)\right)^{\alpha} \quad \text{as } \mathrm{x}\to\infty,$$
$$f(x)\sim \alpha\, g(x)\,\left(\bar{G}(x)\right)^{\alpha-1} \quad \text{as } \mathrm{x}\to\infty,$$
$$h(x)\sim \frac{\alpha\, g(x)}{\bar{G}(x)} \quad \text{as } \mathrm{x}\to\infty.$$

The shapes of the density and hazard rate functions can be described analytically. The critical points of the AGT-G density function are the roots of the equation

$$\frac{g'(x)}{g(x)}+(1-\alpha)\frac{g(x)}{\bar{G}(x)}-\frac{2\alpha\lambda g(x)\left(\bar{G}(x)\right)^{\alpha-1}}{1+\lambda-2\lambda\left[1-\left(\bar{G}(x)\right)^{\alpha}\right]}=0. \tag{26}$$

The critical points of $h(x)$ are obtained from the equation

$$\frac{g'(x)}{g(x)}+(1-\alpha)\frac{g(x)}{\bar{G}(x)}-\frac{2\alpha\lambda g(x)\left(\bar{G}(x)\right)^{\alpha-1}}{1+\lambda-2\lambda\left[1-\left(\bar{G}(x)\right)^{\alpha}\right]} \\ +\frac{\alpha g(x)\left(\bar{G}(x)\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}(x)\right)^{\alpha}\right]\right\}}{1-(1+\lambda)\left[1-\left(\bar{G}(x)\right)^{\alpha}\right]+\lambda\left[1-\left(\bar{G}(x)\right)^{\alpha}\right]^{2}}=0. \tag{27}$$

5. Useful expansions

By using generalized binomial expansion we can show that the cdf (4) of X has the expansion

$$F(x;\alpha,\lambda,\boldsymbol{\xi}) = \sum_{k=0}^{\infty} c_k \left(G(x)\right)^k = \sum_{k=0}^{\infty} c_k H_k(x) \tag{28}$$

where $c_0 = 0$ and for $k \geq 1$,

$$c_k = (-1)^k \left[(\lambda - 1)\binom{\alpha}{k} - \lambda\binom{2\alpha}{k}\right] \tag{29}$$

and $H_a(x) = (G(x))^a$ denotes the exponentiated-G ("exp-G" for short) cumulative distribution. We can prove $\sum_{k=0}^{\infty} c_k = 1$. Some structural properties of the exp-G distributions are studied by Mudholkar *et al.* (1996), Gupta and Kundu (2001) and Nadarajah and Kotz (2006), among others.

The density function of X can be expressed as an infinite linear combination of exp-G density functions

$$f(x;\alpha,\lambda,\boldsymbol{\xi}) = \sum_{k=0}^{\infty} c_{k+1} h_{k+1}(x), \tag{30}$$

where $h_{k+1} = (k+1)G(x)^k g(x)$(for $k \geq 0$) is the exp-G density with power parameter $k+1$. Equation (30) reveals that the AGT-G density function is a linear combination of exp-G density functions. Thus, some mathematical properties of the new model can be derived from those properties of the exp-G distribution. For example, the ordinary and incomplete moments and moment generating function (mgf) of X can be obtained from those quantities of the exp-G distribution.

The formulae derived throughout the paper can be easily handled in most symbolic computation software plataforms such as Maple, Mathematica and Matlab. These plataforms have currently the ability to deal with analytic expressions of formidable size and complexity. Established explicit expressions to calculate statistical measures can be more efficient than computing them directly by numerical integration. The infinity limit in these sums can be substituted by a large positive integer such as 20 or 30 for most practical purposes.

6. Some measures

6.1. Moments

Let Y_k be a random variable with exp-G distribution with power parameter $k+1$, i.e., with density $h_{k+1}(x)$. A first formula for the nth moment of $X \sim$AGT-G follows from (30) as

$$E(X^n) = \sum_{k=0}^{\infty} c_{k+1}\, E(Y_k^n), \tag{31}$$

where $\sum_{k=0}^{\infty} c_k = 1$. Expressions for moments of several exp-G distributions are given in Nadarajah and Kotz (2006b), which can be used to obtained $E(X^n)$.

A second formula for $E(X^n)$ can be written from (31) in terms of the G quantile function as

$$E(X^n) = \sum_{k=0}^{\infty} (k+1)\, c_{k+1}\, \tau(n,k), \tag{32}$$

where $\tau(n,k) = \int_{-\infty}^{\infty} x^n \left(G(x)\right)^k g(x)dx = \int_0^1 \left(Q_G(u)\right)^n u^k du$. Cordeiro, Nadarajah *et al.* (2011) obtained $\tau(n,k)$ for some well known distribution such as Normal, Beta, Gamma and Weibull distributions, which can be used to find moments of AGT-G.

For empirical purposes, the shape of many distributions can be usefully described by what we call the incomplete moments. These types of moments play an important role in measuring inequality, for example, income quantiles and Lorenz and Bonferroni curves, which depend on the incomplete moments of a distribution. The nth incomplete moment of X is

$$m_n(y) = E(X^n|X < y) = \sum_{k=0}^{\infty}(k+1)\, c_{k+1} \int_0^{G(y)} (Q_G(u))^n\, u^k du. \tag{33}$$

The last integral can be computed for most G distributions.

6.2. Generating function

Let $M_X(t) = E(\mathrm{e}^{\mathrm{t\,X}})$ be mgf of $X \sim$ AGT-G,then, the first form of $M_X(t)$ comes from (30) as

$$M_X(t) = \sum_{k=0}^{\infty} c_{k+1}\, M_k(t), \tag{34}$$

where $M_k(t)$ is the mgf of Y_k. Hence, $M_X(t)$ can be determined from the exp-G generating function.

A second formula for $M_X(t)$ can be derived from (30) as

$$M_X(t) = \sum_{i=0}^{\infty}(k+1)\, c_{k+1}\, \rho(t,k), \tag{35}$$

where $\rho(t,k) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{t\,x}}\, (\mathrm{G(x)})^{\mathrm{k}}\, \mathrm{g(x)dx} = \int_0^1 \exp[\mathrm{t\,Q_G(u)}]\, \mathrm{u^k du}$.

We can obtain the mgfs of several distributions directly from equation (35).

6.3. Mean deviation

The mean deviation about the mean ($\delta_1 = E(|X - \mu_1'|)$) and about the median ($\delta_2 = E(|X - M|)$) of X can be expressed as

$$\delta_1(X) = 2\mu_1' F\left(\mu_1'\right) - 2m_1\left(\mu_1'\right) \qquad \text{and} \qquad \delta_2(X) = \mu_1' - 2m_1(M), \tag{36}$$

respectively, where $\mu_1' = E(X)$, $M = Median(X)$ is the median defined by $M = Q(0.5)$, $F(\mu_1')$ is easily calculated from the cdf (4) and $m_1(z) = \int_{-\infty}^{z} xf(x)dx$ is the first incomplete moment obtained from (33) with $n = 1$.

Now, we provide two alternative ways to compute δ_1 and δ_2. A general equation for $m_1(z)$ can be derived from (30) as

$$m_1(z) = \sum_{k=0}^{\infty} c_{k+1}\, J_k(z), \tag{37}$$

where $J_k(z) = \int_{-\infty}^{z} x\, h_{k+1}(x)dx$ is the basic quantity to compute the mean deviation for the exp-G distributions. Hence, the mean deviation in (36) depend only on the mean deviation of the exp-G distribution. So, alternative representations for δ_1 and δ_2 are

$$\delta_1(X) = 2\mu_1' F\left(\mu_1'\right) - 2\sum_{k=0}^{\infty} c_{k+1}\, J_k(\mu_1') \qquad \text{and} \qquad \delta_2(X) = \mu_1' - 2\sum_{k=0}^{\infty} c_{k+1}\, J_k(M).$$

A simple application of $J_k(z)$ refers to the the AGT-G distribution discussed in Section 3.5. The exponentiated Weibull with parameter $k+1$ has pdf (for $x > 0$) given by

$$h_{k+1}(x) = \frac{(k+1)\eta}{\sigma^{\eta}}\, x^{\eta-1}\, \exp\left[-(\frac{x}{\sigma})^{\eta}\right] \left\{1 - \exp\left[-(\frac{x}{\sigma})^{\eta}\right]\right\}^k$$

and then

$$\begin{aligned} J_k(z) &= \frac{(k+1)\eta}{\sigma^\eta}\int_0^z x^\eta \exp\left[-(\frac{x}{\sigma})^\eta\right]\left\{1-\exp\left[-(\frac{x}{\sigma})^\eta\right]\right\}^k dx \\ &= \frac{(k+1)\eta}{\sigma^\eta}\sum_{r=0}^{k}(-1)^r\binom{k}{r}\int_0^z x^\eta \exp\left[-(r+1)(\frac{x}{\sigma})^\eta\right] \end{aligned}$$

The last integral is just the incomplete gamma function and then the mean deviation for the AGT-G distribution can be determined from

$$m_1(z) = \sum_{k=0}^{\infty}\sum_{r=0}^{k}\frac{(k+1)b_{k+1}\,(-1)^r\binom{k}{r}}{(r+1)^{1+\eta^{-1}}\,\sigma^{2\eta+1}}\,\gamma(1+\eta^{-1},(r+1)(\frac{z}{\sigma})^\eta)$$

A second general formula for $m_1(z)$ can be derived by setting $u = G(x)$ in (30)

$$m_1(z) = \sum_{k=0}^{\infty}(k+1)\,c_{k+1}\,T_k(z), \tag{38}$$

where $T_k(z) = \int_0^{G(z)} Q_G(u)\,u^k du$ is a simple integral defined from the baseline quantile function and $Q_G(u) = G^{-1}(u)$.

Remark: Applications of these equations employed to obtain Bonferroni and Lorenz curves defined for a given probability π by

$$B(\pi) = \frac{T(q)}{\pi\mu_1'} \qquad \text{and} \qquad L(\pi) = \frac{T(q)}{\mu_1'},$$

respectively, where $\mu_1' = E(X)$ and $q = Q(\pi)$ is the qf of X at π.

7. Extreme values

Let $\overline{X} = (X_1 + \cdots + X_n)/n$ denote the mean of a random sample from (5), then by the usual central limit theorem $\sqrt{n}(\overline{X} - E(X))/\sqrt{Var(X)}$ approaches the standard normal distribution as $n \to \infty$ under suitable conditions. Sometimes one would be interested in the asymptotics of the extreme values $M_n = \max\{X_1, \ldots, X_n\}$ and $m_n = \min\{X_1, \ldots, X_n\}$.

First, suppose G belongs to the max domain of attraction of Gumbel extreme value distribution. Then by Leadbetter, Lindgren, and Rootzén (2012) (chapter 1), there must exist a strictly positive function, say $h(t)$, such that

$$\lim_{t\to\infty}\frac{1-G(t+xh(t))}{1-G(t)} = \mathrm{e}^{-x},$$

for every $x \in (-\infty, \infty)$. But

$$\lim_{t\to\infty}\frac{1-F(t+xh(t))}{1-F(t)} = \lim_{x\to\infty}\frac{xf(tx)}{f(t)} = \mathrm{e}^{-\alpha x},$$

for every $x \in (-\infty, \infty)$. So, it follows from Leadbetter *et al.* (2012) (chapter 1) that F belongs to the max domain of attraction of the Gumbel extreme value distribution with

$$\lim_{n\to\infty} P[a_n(M_n - b_n \le x)] = \exp\left[-\exp(-\alpha x)\right]$$

for some suitable norming constants $a_n > 0$ and b_n. Second, suppose G belongs to the max domain of attraction of the Fréchet extreme value distribution. Then from Leadbetter *et al.* (2012) (Chapter 1), there must exist a $\beta > 0$ such that

$$\lim_{t\to\infty}\frac{1-G(t+xh(t))}{1-G(t)} = x^c$$

for every $x \in (-\infty, \infty)$. But

$$\begin{aligned}\lim_{t\to\infty} \frac{1-F(t+xh(t))}{1-F(t)} &= \lim_{t\to\infty} \frac{xf(tx)}{f(t)} \\ &= x^{\alpha\, c},\end{aligned}$$

for every $x > 0$. So, it follows from Leadbetter *et al.* (2012) (chapter 1) that F belongs to the max domain of attraction of the Gumbel extreme value distribution with

$$\lim_{n\to\infty} P[a_n(M_n - b_n \leq x)] = \exp(-x^{\alpha\, c})$$

for some suitable norming constants $a_n > 0$ and b_n. Third, suppose G belongs to the max domain of attraction of the Weibull extreme value distribution. Then, Leadbetter *et al.* (2012) (chapter 1), there must exist a $\beta > 0$ such that

$$\lim_{t\to 0} \frac{G(tx)}{G(t)} = x^{\beta}$$

for every $x < 0$. But

$$\lim_{t\to 0} \frac{F(tx)}{F(t)} = \lim_{t\to 0} \frac{xf(tx)}{f(t)} = x^{\beta}$$

for every $x < 0$. So, it follows from Leadbetter *et al.* (2012) (chapter 1) that F belongs to the max domain of attraction of the Weibull extreme value distribution with

$$\lim_{n\to\infty} P[a_n(M_n - b_n \leq x)] = \exp\left[-(-x)^{\beta}\right]$$

for some suitable norming constants $a_n > 0$ and b_n. We conclude that F belongs to the same max domain of attraction as that of G. The same argument applies to min domain of attraction.That is, F belongs to the same max domain of attraction as that of G.

8. Entropies

An entropy is a measure of variation or uncertainty of a random variable X. Two popular entropy measures are the Rényi and Shannon entropies Renyi (1961) , Shannon (2001). The Rényi entropy of a random variable with pdf $f(x)$ is defined as

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

for $\gamma > 0$ and $\gamma \neq 1$. The Shannon entropy of a random variable X is defined by $E\{-\log[f(X)]\}$. It is the special case of the Rényi entropy when $\gamma \uparrow 1$. Direct calculation yields

$$\begin{aligned}\mathrm{E}\{-\log[f(X)]\} &= -\log(\alpha) - \mathrm{E}\{\log[g(X;\boldsymbol{\xi})]\} + (1-\alpha)\,\mathrm{E}\left\{\log\left[\bar{G}(X;\boldsymbol{\xi})\right]\right\} \\ &- \mathrm{E}\left\{\log\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}(X;\boldsymbol{\xi})\right)^{\alpha}\right]\right\}\right\}\end{aligned}$$

First we define and compute

$$A(a_1, a_2; \lambda, \alpha) = \int_0^1 x^{a_1}(1 - \frac{2\lambda}{1+\lambda}(1-(1-x)^{\alpha}))^{a_2}dx. \tag{39}$$

Using generalized binomial expansion and then after some algebraic manipulations, we obtain

$$A(a_1, a_2; \lambda, \alpha) = \sum_{i=0}^{\infty}\sum_{j=0}^{i}(-1)^{i+j}\binom{a_2}{i}\binom{i}{j}(\frac{2\lambda}{1+\lambda})^i\, Beta(a_1+1, \alpha\, j+1)$$

Proposition 3 *Let X be a random variable with pdf (5). Then,*

$$\mathrm{E}\left\{\log\left[G(X)\right]\right\} = \frac{\alpha}{1+\lambda}\,\frac{\partial}{\partial t} A(\alpha+t-1,1;\lambda,\alpha)\big|_{t=0}$$

$$\mathrm{E}\left\{\log\left\{1+\lambda-2\lambda\left[1-\bar{G}(X;\boldsymbol{\xi})^{\alpha}\right]\right\}\right\} = \frac{\alpha}{1+\lambda}\,\frac{\partial}{\partial t}\frac{1}{(1+\lambda)^t}\,A(\alpha-1,t+1;\lambda,\alpha)\big|_{t=0}$$

The simplest formula for the entropy of X is given by

$$\begin{aligned}\mathrm{E}\left\{-\log[f(X)]\right\} &= -\log(\alpha) - \mathrm{E}\left\{\log[g(X;\boldsymbol{\xi})]\right\}\\ &+ (1-\alpha)\frac{\alpha}{1+\lambda}\,\frac{\partial}{\partial t}A(\alpha+t-1,1;\lambda,\alpha)\big|_{t=0}\\ &- \frac{\alpha}{1+\lambda}\,\frac{\partial}{\partial t}\frac{1}{(1+\lambda)^t}\,A(\alpha-1,t+1;\lambda,\alpha)\big|_{t=0}\end{aligned}$$

After some algebraic manipulations, we obtain an alternative expression for $I_R(\gamma)$

$$I_R(\gamma) = \frac{\gamma}{1-\gamma}\log(\frac{\alpha}{1+\lambda}) + \frac{1}{1-\gamma}\log\left\{\sum_{i=0}^{\infty}\sum_{j=0}^{i} w^*_{i,j}\,\mathrm{E}_{Y_j}[g^{\gamma-1}[G^{-1}(Y)]]\right\} \tag{40}$$

where $Y_i \sim Beta(\gamma(\alpha-1)+1, \gamma j+1)$ and

$$w^*_{i,j} = \frac{(-1)^{i+j}\binom{\gamma}{i}\binom{i}{j}}{Beta(\gamma(\alpha-1)+1,\gamma j+1)}\,(\frac{2\lambda}{1+\lambda})^i$$

9. Order statistics

Order statistics make their appearance in many areas of statistical theory and practice. Suppose $X_1,\ldots,X_n$ is a random sample from the AGT-G family of distributions. We can write the density of the ith order statistic, say $X_{i:n}$, as

$$f_{i:n}(x) = K\,f(x)\,(F(x))^{i-1}\,\{1-F(x)\}^{n-i} = K\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j} f(x)\,(F(x))^{j+i-1},$$

where $K = n!/[(i-1)!\,(n-i)!]$.

Following similar algebraic manipulations, we can write the density function of the i^{th} order statistic, $X_{i:n}$, as

$$f_{i:n}(x) = \sum_{r,k=0}^{\infty} m_{r,k}\,h_{r+k+1}(x), \tag{41}$$

where $h_{r+k+1}(x)$ denotes the exp-G density function with power parameter $r+k+1$,

$$m_{r,k} = \frac{n!\,(r+1)\,(i-1)!\,c_{r+1}}{(r+k+1)}\sum_{j=0}^{n-i}\frac{(-1)^j\,f_{j+i-1,k}}{(n-i-j)!\,j!},$$

and c_k is defined in equation (29). Here, the quantities $f_{j+i-1,k}$ are obtained recursively by $f_{j+i-1,0} = c_0^{j+i-1}$ and (for $k \geq 1$)

$$f_{j+i-1,k} = (k\,c_0)^{-1}\sum_{m=1}^{k}[m(j+i)-k]\,c_m\,f_{j+i-1,k-m}.$$

Equation (41) is the main result of this section. It reveals that the pdf of the AGT-G order statistic is a linear combination of exp-G density functions. So, several mathematical quantities of the AGT-G order statistics such as ordinary, incomplete and factorial moments, mgf, mean deviation and several others can be obtained from those quantities of the exp-G distribution.

10. Characterization results

In designing a stochastic model for a particular modeling problem, an investigator will be vitally interested to know if their model fits the requirements of a specific underlying probability distribution. To this end, the investigator will rely on the characterizations of the selected distribution. Generally speaking, the problem of characterizing a distribution is an important problem in various fields and has recently attracted the attention of many researchers. Consequently, various characterization results have been reported in the literature. These characterizations have been established in many different directions. In this Section, we present characterizations of AGT-G distribution. These characterizations are based on: (i) a simple relationship between two truncated moments; (ii) conditional expectations of a function of the random variable. We like to mention that the characterization (i) which is expressed in terms of the ratio of truncated moments is stable in the sense of weak convergence. It also serves as a bridge between a first order differential equation and probability and it does not require a closed form of the cdf.

10.1. Characterizations based on two truncated moments

In this subsection we present characterizations of ATG-G distribution in terms of a simple relationship between two truncated moments. Our characterization results presented here will employ an interesting result due to Glänzel (1987) (Theorem 4 below). The advantage of the characterizations given here is that, cdf F need not have a closed form and is given in terms of an integral whose integrand depends on the solution of a first order differential equation, which can serve as a bridge between probability and differential equation.

Theorem 4 *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a given probability space and let $H = [a, b]$ be an interval for some $a < b$ ($a = -\infty$, $b = \infty$ might as well be allowed). Let $X : \Omega \to H$ be a continuous random variable with the distribution function F and let q_1 and q_2 be two real functions defined on H such that*

$$\mathbf{E}\left[q_1(X) \mid X \geq x\right] = \mathbf{E}\left[q_2(X) \mid X \geq x\right] \eta(x), \qquad x \in H,$$

is defined with some real function η. Assume that $q_1, q_2 \in C^1(H)$, $\eta \in C^2(H)$ and F is twice continuously differentiable and strictly monotone function on the set H. Finally, assume that the equation $q_2\eta = q_1$ has no real solution in the interior of H. Then F is uniquely determined by the functions q_1, q_2 and η, particularly

$$F(x) = \int_a^x C \left| \frac{\eta'(u)}{\eta(u) q_2(u) - q_1(u)} \right| \exp\left(-s(u)\right) du,$$

where the function s is a solution of the differential equation $s' = \frac{\eta' q_2}{\eta q_2 - q_1}$ and C is a constant, chosen to make $\int_H dF = 1$.

Remarks 5 (*a*) *In Theorem 4, the interval* H *need not be closed since the condition is only on the interior of* H. (*b*) *Clearly, Theorem 4 can be stated in terms of two functions* q_1 *and* η *by taking* $q_2(x) \equiv 1$, *which will reduce the condition given in Theorem 4 to* $E[q_1(X) \mid X \geq x] = \eta(x)$. *However, adding an extra function will give a lot more flexibility, as far as its application is concerned.*

Proposition 6 *Let* $X : \Omega \to (0, \infty)$ *be a continuous random variable and let* $q_2(x) = \left\{1 - \lambda + 2\lambda\left(\overline{G}(x)\right)^{\alpha}\right\}^{-1}$ *and* $q_1(x) = q_2(x)\left(\overline{G}(x)\right)$ *for* $x > 0$. *The pdf of* X *is* (5) *if and only if the function* η *defined in Theorem 4 has the form*

$$\eta(x) = \frac{\alpha}{\alpha+1}\overline{G}(x) \ , \quad x > 0.$$

Proof. Let X have pdf (5), then

$$(1 - F(x)) \ \mathbf{E}[q_2(X) \mid X \geq x] = \left(\overline{G}(x)\right)^{\alpha}, \quad x > 0,$$

and

$$(1 - F(x)) \ \mathbf{E}[q_1(X) \mid X \geq x] = \frac{\alpha}{\alpha+1}\left(\overline{G}(x)\right)^{\alpha+1}, \quad x > 0,$$

and finally

$$\eta(x)\, q_2(x) - q_1(x) = -\frac{1}{\alpha+1} q_2(x)\, \overline{G}(x) < 0 \ , \quad x > 0.$$

Conversely, if η is given as above, then

$$s'(x) = \frac{\eta'(x) \ q_2(x)}{\eta(x) \ q_2(x) - q_1(x)} = \alpha g(x)\left(\overline{G}(x)\right)^{-1} \ , \quad x > 0 \ ,$$

and hence

$$s(x) = -\log(\left(\overline{G}(x)\right)^{\alpha}) \ , \quad x > 0.$$

Now, in view of Theorem 4, X has pdf (5).

Corollary 7 *Let* $X : \Omega \to (0, \infty)$ *be a continuous random variable and let* $q_2(x)$ *be as in Proposition 6. The pdf of* X *is* (5) *if and only if there exist functions* q_1 *and* η *defined in Theorem 4 satisfying the differential equation*

$$\frac{\eta'(x) \ q_2(x)}{\eta(x) \ q_2(x) - q_1(x)} = \alpha g(x)\left(\overline{G}(x)\right)^{-1}, \quad x > 0.$$

Proof. Is straightforward and hence omitted.

Remarks 8 (*a*) *The general solution of the differential equation in Corollary* **??** *is*

$$\eta(x) = \left(\overline{G}(x)\right)^{-\alpha}\left[-\int \alpha g(x)\left(\overline{G}(x)\right)^{\alpha-1}\left(q_2(x)\right)^{-1} q_1(x)\, dx + D\right],$$

for $x > 0$ *, where* D *is a constant. One set of appropriate functions is given in Proposition 8 with* $D = \frac{1}{2}$.

(*b*) *Clearly there are other triplets of functions* (q_1, q_2, η) *satisfying the conditions of Theorem 4. We presented one such triplet in Proposition .*

10.2. Characterizations based on conditional expectation of a function of the variable

In this subsection we employ a single function ψ of X and state characterization results in terms of $\psi(X)$.

Proposition 9 *Let* $X : \Omega \to (a, b)$ *be a continuous random variable with cdf* F *. Let* $\psi(x)$ *be a differentiable function on* (a, b) *with* $\lim_{x \to a^+} \psi(x) = \delta > 1$ *and* $\lim_{x \to b^-} \psi(x) = \infty$ *. Then*

$$E\left[(\psi(X))^{\delta} \mid X \leq x\right] = \delta(\psi(x))^{\delta - 1}, \quad x \in (a, b), \tag{42}$$

implies

$$\psi(x) = \delta\left[1 - (F(x))^{\frac{1}{\delta - 1}}\right]^{-1}, \quad x \in (a, b). \tag{43}$$

Proof. From (42), we have

$$\int_a^x (\psi(u))^{\delta} f(u)\, du = \delta(\psi(x))^{\delta - 1} F(x).$$

Taking derivatives from both sides of the above equation, we arrive at

$$(\psi(x))^{\delta} f(x) = \delta\left\{(\delta - 1)\psi'(x)(\psi(x))^{\delta - 2} F(x) + (\psi(x))^{\delta - 1} f(x)\right\},$$

from which we have

$$\frac{f(x)}{F(x)} = (\delta - 1)\left\{-\frac{\psi'(x)}{\psi(x)} + \frac{\psi'(x)}{\psi(x) - \delta}\right\}.$$

Integrating bothe sides of this equation from x to b and using the condition $\lim_{x \to b^-} \psi(x) = \infty$, we obtain (43).

It is easy to see that for $\delta = 2$, "implies" in Proposition 8 will be replaced by "if and only if".

Proposition 10 *Let* $X : \Omega \to (a, b)$ *be a continuous random variable with cdf* F*. Let* $\psi_1(x)$ *be a differentiable function on* (a, b) *with* $\lim_{x \to a^+} \psi_1(x) = \delta/2 > 1/2$ *and* $\lim_{x \to b^-} \psi_1(x) = \delta$*. Then*

$$E\left[(\psi_1(X))^{\delta} \mid X \geq x\right] = \delta(\psi_1(x))^{\delta - 1}, \quad x \in (a, b), \tag{44}$$

if and only if

$$\psi_1(x) = \delta\left[1 + (1 - F(x))^{\frac{1}{\delta - 1}}\right]^{-1}, \quad x \in (a, b). \tag{45}$$

Proof. Is similar to that of Proposition 9.

Remarks 11 (a) *Taking, e.g.,* $(a,b)=(0,\infty)$ *and*

$$\psi(x)=\delta\left[1+\left(\left(1-\left(\overline{G}(x)\right)^{\alpha}\right)\left(1+\lambda\left(\overline{G}(x)\right)^{\alpha}\right)\right)^{\frac{1}{\delta-1}}\right]^{-1},$$

Proposition 9 gives a characterization of ATG-G distribution.

(b) *Taking, e.g.,* $(a,b)=(0,\infty)$ *and* $\psi_1(x)=\delta\left[1+\left(\left(\overline{G}(x)\right)^{\alpha}\left(1+\lambda-\lambda\left(\overline{G}(x)\right)^{\alpha}\right)\right)^{\frac{1}{\delta-1}}\right]^{-1}$, *Proposition 10 gives a characterization of ATG-G distribution.*

11. Bivariate extention

In this section we introduce a bivariate version of the proposed model. The joint cdf is given by

$$F_{X,Y}(x,y)=(1+\lambda)\left\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\right\}-\lambda\left\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\right\}^2 \tag{46}$$

where $G(x,y;\boldsymbol{\xi})$ is a bivariate continuous distribution with mariginal cdf's $G_1(x;\boldsymbol{\xi})$ and $G_2(y;\boldsymbol{\xi})$. We denote this distribution by *another bivariate Generalized Transmuted G* (ABGT-G) distribution. The marginal cdf's are given by

$$F_X(x)=(1+\lambda)\left[1-\left(\bar{G}_1(x;\boldsymbol{\xi})\right)^{\alpha}\right]-\lambda\left[1-\left(\bar{G}_1(x;\boldsymbol{\xi})\right)^{\alpha}\right]^2 \quad \text{and}$$
$$F_Y(y)=(1+\lambda)\left[1-\left(\bar{G}_2(y;\boldsymbol{\xi})\right)^{\alpha}\right]-\lambda\left[1-\left(\bar{G}_2(y;\boldsymbol{\xi})\right)^{\alpha}\right]^2$$

The joint pdf of (X,Y) is easily obtained by $f_{X,Y}(x,y)=\dfrac{\partial^2 F_{X,Y}(x,y)}{\partial x\partial y}$

$$f_{X,Y}(x,y) = \alpha A(x,y;\alpha,\lambda\boldsymbol{\xi})\left(1-G(x,y;\boldsymbol{\xi})\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\right\}\right\}$$

where

$$\begin{aligned}A(x,y;\alpha,\lambda,\boldsymbol{\xi}) &= g(x,y;\boldsymbol{\xi})+\frac{1-\alpha}{1-G(x,y;\boldsymbol{\xi})}\frac{\partial G(x,y,\boldsymbol{\xi})}{\partial x}\frac{\partial G(x,y,\boldsymbol{\xi})}{\partial y}\\ &- \frac{2\alpha\lambda\left(1-G(x,y;\boldsymbol{\xi})\right)^{\alpha-1}}{1+\lambda-2\lambda\left\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\right\}}\frac{\partial G(x,y,\boldsymbol{\xi})}{\partial x}\frac{\partial G(x,y,\boldsymbol{\xi})}{\partial y}.\end{aligned}$$

The marginal pdf's are

$$f_X(x)=\alpha g_1(x,\boldsymbol{\xi})\left(\bar{G}_1(x,\boldsymbol{\xi})\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}_1(x;\boldsymbol{\xi})\right)^{\alpha}\right]\right\}$$

and

$$f_Y(y)=\alpha g_2(y,\boldsymbol{\xi})\left(\bar{G}_2(y,\boldsymbol{\xi})\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}_2(y;\boldsymbol{\xi})\right)^{\alpha}\right]\right\}.$$

The conditional cdf's are

$$F_{X|Y}(x|y) = \frac{(1+\lambda)\left\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\right\}-\lambda\left\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\right\}^2}{(1+\lambda)\left[1-\left(\bar{G}_2(y;\boldsymbol{\xi})\right)^{\alpha}\right]-\lambda\left[1-\left(\bar{G}_2(y;\boldsymbol{\xi})\right)^{\alpha}\right]^2}$$

and

$$F_{Y|X}(y|x) = \frac{(1+\lambda)\left\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\right\}-\lambda\left\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\right\}^2}{(1+\lambda)\left[1-\left(\bar{G}_1(x;\boldsymbol{\xi})\right)^{\alpha}\right]-\lambda\left[1-\left(\bar{G}_1(x;\boldsymbol{\xi})\right)^{\alpha}\right]^2}.$$

The conditional density functions are

$$f_{X|Y}(x|y) = \frac{A(x,y;\alpha,\lambda\boldsymbol{\xi})\,(1-G(x,y;\boldsymbol{\xi}))^{\alpha-1}\,\{1+\lambda-2\lambda\,\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\}\}}{g_2(y,\boldsymbol{\xi})\,\left(\bar{G}_2(y,\boldsymbol{\xi})\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}_2(y;\boldsymbol{\xi})\right)^{\alpha}\right]\right\}}$$

and

$$f_{Y|X}(y|x) = \frac{A(x,y;\alpha,\lambda\boldsymbol{\xi})\,(1-G(x,y;\boldsymbol{\xi}))^{\alpha-1}\,\{1+\lambda-2\lambda\,\{1-(1-G(x,y;\boldsymbol{\xi}))^{\alpha}\}\}}{g_1(x,\boldsymbol{\xi})\,\left(\bar{G}_1(x,\boldsymbol{\xi})\right)^{\alpha-1}\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}_1(x;\boldsymbol{\xi})\right)^{\alpha}\right]\right\}}$$

12. Estimation

Here, we determine the maximum likelihood estimates (MLEs) of the model parameters of AGT-G from complete samples only. Let $x_1,\ldots,x_n$ be observed values from the AGT-G distribution with parameters α,λ and $\boldsymbol{\xi}$. Let $\Theta=(\alpha,\lambda,\boldsymbol{\xi})^{\top}$ be the $r\times 1$ parameter vector. The total log-likelihood function for Θ is given by

$$\begin{aligned}\ell_n &= \ell_n(\Theta) = n\,\log(\alpha)+\sum_{i=1}^{n}\log\left[g(x_i;\boldsymbol{\xi})\right]+(\alpha-1)\sum_{i=1}^{n}\log\left[\bar{G}(x_i;\boldsymbol{\xi})\right]\\ &+ \sum_{i=1}^{n}\log\left\{1+\lambda-2\lambda\left[1-\left(\bar{G}(x_i;\boldsymbol{\xi})\right)^{\alpha}\right]\right\} \qquad (47)\end{aligned}$$

The log-likelihood function can be maximized either directly by using the SAS (PROC NLMIXED) or the Ox program (sub-routine MaxBFGS) (see Doornik, 2007) or by solving the nonlinear likelihood equations obtained by differentiating (47). The components of the score function $U_n(\Theta)=(\partial\ell_n/\partial\alpha,\partial\ell_n/\partial\lambda,\partial\ell_n/\partial\boldsymbol{\xi})^{\top}$ are

$$\begin{aligned}\frac{\partial\ell_n}{\partial\alpha} &= \frac{n}{\alpha}+\sum_{i=1}^{n}\log\left[\bar{G}(x_i;\boldsymbol{\xi})\right]+2\lambda\sum_{i=1}^{n}\frac{\left(\bar{G}(x_i;\boldsymbol{\xi})\right)^{\alpha}\log\left[\bar{G}(x_i;\boldsymbol{\xi})\right]}{1+\lambda-2\lambda\left[1-\left(\bar{G}(x_i;\boldsymbol{\xi})\right)^{\alpha}\right]},\\ \frac{\partial\ell_n}{\partial\lambda} &= \sum_{i=1}^{n}\frac{1-2\left[1-\left(\bar{G}(x_i;\boldsymbol{\xi})\right)^{\alpha}\right]}{1+\lambda-2\lambda\left[1-\left(\bar{G}(x_i;\boldsymbol{\xi})\right)^{\alpha}\right]},\\ \frac{\partial\ell_n}{\partial\boldsymbol{\xi}} &= \sum_{i=1}^{n}\frac{g^{(\xi)}(x_i,\boldsymbol{\xi})}{g(x_i,\boldsymbol{\xi})}+(1-\alpha)\sum_{i=1}^{n}\frac{G^{(\xi)}(x_i,\boldsymbol{\xi})}{\bar{G}(x_i,\boldsymbol{\xi})}\\ &\quad+2\alpha\lambda\sum_{i=1}^{n}\frac{G^{(\xi)}(x_i,\boldsymbol{\xi})\left(\bar{G}(x_i;\boldsymbol{\xi})\right)^{\alpha-1}}{1+\lambda-2\lambda\left[1-\left(\bar{G}(x_i;\boldsymbol{\xi})\right)^{\alpha}\right]}\end{aligned}$$

where $h^{(\xi)}(\cdot)$ denotes the derivative of the function h with respect to ξ.

12.1. Maximum product spacing estimates

The maximum product spacing (MPS) method has been proposed by Cheng and Amin (1983). This method is based on an idea that the differences (spacings) of the consecutive points should be identically distributed. The geometric mean of the differences is given as

$$GM = \sqrt[n+1]{\prod_{i=1}^{n+1}D_i} \qquad (48)$$

where the difference D_i is defined by

$$D_i = \int_{x_{(i-1)}}^{x_{(i)}} f\,(x,\lambda,\alpha,\boldsymbol{\xi})\,dx;\quad i=1,2,\ldots,n+1. \qquad (49)$$

Here, $F(x_{(0)}, \lambda, \alpha, \xi) = 0$ and $F(x_{(n+1)}, \lambda, \alpha, \boldsymbol{\xi}) = 1$. The MPS estimators $\hat{\lambda}_{PS}, \hat{\alpha}_{PS}$ and $\hat{\boldsymbol{\xi}}_{PS}$ of λ, α and $\boldsymbol{\xi}$ are obtained by maximizing the geometric mean (GM) of the differences. Substituting pdf of AGT-G in (49) and taking logarithm of the above expression, we will have

$$LogGM = \frac{1}{n+1}\sum_{i=1}^{n+1}\log\left[F(x_{(i)}, \lambda, \alpha, \boldsymbol{\xi}) - F(x_{(i-1)}, \lambda, \alpha, \boldsymbol{\xi})\right]. \tag{50}$$

The MPS estimators $\hat{\lambda}_{PS}, \hat{\alpha}_{PS}$ and $\hat{\boldsymbol{\xi}}_{PS}$ of λ, α and $\boldsymbol{\xi}$ can be obtained as the simultaneous solution of the following non-linear equations:

$$\frac{\partial LogGM}{\partial \lambda} = \frac{1}{n+1}\sum_{i=1}^{n+1}\left[\frac{F'_{\lambda}(x_{(i)}, \lambda, \alpha, \boldsymbol{\xi}) - F'_{\lambda}(x_{(i-1)}, \lambda, \alpha, \boldsymbol{\xi})}{F(x_{(i)}, \lambda, \alpha, \boldsymbol{\xi}) - F(x_{(i-1)}, \lambda, \alpha, \boldsymbol{\xi})}\right] = 0$$

$$\frac{\partial LogGM}{\partial \alpha} = \frac{1}{n+1}\sum_{i=1}^{n+1}\left[\frac{F'_{\alpha}(x_{(i)}, \lambda, \alpha, \boldsymbol{\xi}) - F'_{\alpha}(x_{(i-1)}, \lambda, \alpha, \boldsymbol{\xi})}{F(x_{(i)}, \lambda, \alpha, \boldsymbol{\xi}) - F(x_{(i-1)}, \lambda, \alpha, \boldsymbol{\xi})}\right] = 0$$

$$\frac{\partial LogGM}{\partial \boldsymbol{\xi}} = \frac{1}{n+1}\sum_{i=1}^{n+1}\left[\frac{F'_{\boldsymbol{\xi}}(x_{(i)}, \lambda, \alpha, \boldsymbol{\xi}) - F'_{\boldsymbol{\xi}}(x_{(i-1)}, \lambda, \alpha, \boldsymbol{\xi})}{F(x_{(i)}, \lambda, \alpha, \boldsymbol{\xi}) - F(x_{(i-1)}, \lambda, \alpha, \boldsymbol{\xi})}\right] = 0$$

12.2. Least square estimates

Let $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$ be the ordered sample of size n drawn the AGT-G population pdf. Then, the expectation of the empirical cumulative distribution function is defined as

$$E\left[F\left(X_{i:n}\right)\right] = \frac{i}{n+1}; i = 1, 2, \ldots, n \tag{51}$$

The least square estimates (LSEs) $\hat{\lambda}_{LS}, \hat{\alpha}_{LS}$ and $\hat{\boldsymbol{\xi}}_{LS}$ of λ, α and $\boldsymbol{\xi}$ are obtained by minimizing

$$Z\left(\lambda, \alpha, \boldsymbol{\xi}\right) = \sum_{i=1}^{n}\left(F\left(x_{i:n}, \lambda, \alpha, \boldsymbol{\xi}\right) - \frac{i}{n+1}\right)^2$$

Therefore, $\hat{\lambda}_{LS}, \hat{\alpha}_{LS}$ and $\hat{\boldsymbol{\xi}}_{LS}$ of λ, α and $\boldsymbol{\xi}$ can be obtained as the solution of the following system of equations:

$$\frac{\partial Z\left(\lambda, \alpha, \boldsymbol{\xi}\right)}{\partial \lambda} = \sum_{i=1}^{n} F'_{\lambda}(x_{i:n}, \lambda, \alpha, \boldsymbol{\xi})\left(F\left(x_{i:n}, \lambda, \alpha, \boldsymbol{\xi}\right) - \frac{i}{n+1}\right) = 0$$

$$\frac{\partial Z\left(\lambda, \alpha, \boldsymbol{\xi}\right)}{\partial \alpha} = \sum_{i=1}^{n} F'_{\alpha}(x_{i:n}, \lambda, \alpha, \boldsymbol{\xi})\left(F\left(x_{i:n}, \lambda, \alpha, \boldsymbol{\xi}\right) - \frac{i}{n+1}\right) = 0$$

$$\frac{\partial Z\left(\lambda, \alpha, \boldsymbol{\xi}\right)}{\partial \boldsymbol{\xi}} = \sum_{i=1}^{n} F'_{\boldsymbol{\xi}}(x_{i:n}, \lambda, \alpha, \boldsymbol{\xi})\left(F\left(x_{i:n}, \alpha, \theta\right) - \frac{i}{n+1}\right) = 0$$

13. Applications

Now we use a real data set to show that the AGT-E can be a better model than the beta-exponential (Nadarajah and Kotz (2006a)), Kumaraswamy-exponential distribution and exponential distribtuion.

We consider a data set of the life of fatigue fracture of Kevlar 373/epoxy that are subject to constant pressure at the 90% stress level until all had failed, so we have complete data with the exact times of failure. For previous studies with the data sets see Andrews & Herzberg (1985) and Barlow, Toland & Freeman (1984).
These data are:

0.0251, 0.0886, 0.0891, 0.2501, 0.3113, 0.3451, 0.4763, 0.5650, 0.5671, 0.6566, 0.6748, 0.6751, 0.6753, 0.7696, 0.8375, 0.8391, 0.8425, 0.8645, 0.8851, 0.9113, 0.9120, 0.9836, 1.0483, 1.0596, 1.0773, 1.1733, 1.2570, 1.2766, 1.2985, 1.3211, 1.3503, 1.3551, 1.4595, 1.4880, 1.5728, 1.5733, 1.7083, 1.7263, 1.7460,

Table 2: Estimated parameters of the AGT-E, BE and KwE distribution for data set.

Model	ML Estimate	Standard Error	$-\ell(\cdot;x)$	LSE	PS Estimator
AGTE	$\hat{\lambda}=0.733$	0.274	121.3219	-0.636	-0.760
	$\hat{\alpha}=1.197$	0.344		1.631	1.038
	$\hat{\gamma}=0.769$	0.101		0.907	0.704
Exponential	$\hat{\lambda}=0.510$	0.058	127.114	0.981	0.926
Beta	$\hat{a}=1.679$	0.374	122.227	2.235	1.520
Exponential	$\hat{b}=1.508$	6.760		1.558	1.082
	$\hat{\lambda}=0.484$	1.981		0.586	0.598
Kumaraswamy	$\hat{a}=1.556$	0.401	122.0942	1.987	1.426
Exponential	$\hat{b}=2.448$	6.065		2.228	2.243
	$\hat{\lambda}=0.328$	0.691		0.453	0.316

1.7630, 1.7746, 1.8275, 1.8375, 1.8503, 1.8808, 1.8878, 1.8881, 1.9316, 1.9558, 2.0048, 2.0408, 2.0903, 2.1093, 2.1330, 2.2100, 2.2460, 2.2878, 2.3203, 2.3470, 2.3513, 2.4951, 2.5260, 2.9911, 3.0256, 3.2678, 3.4045, 3.4846, 3.7433, 3.7455, 3.9143, 4.8073, 5.4005, 5.4435, 5.5295, 6.5541, 9.0960

The variance covariance matrix $I(\hat{\theta})^{-1}$ of the MLEs under the AGT-E distribution is computed as

$$\begin{pmatrix} 0.075308090 & -0.07199734 & -0.005229209 \\ -0.071997344 & 0.11877571 & 0.021537278 \\ -0.005229209 & 0.02153728 & 0.010247905 \end{pmatrix}.$$

Thus, the variances of the MLE of λ, α and γ is $var(\hat{\lambda}) = 0.075472979, var(\hat{\alpha}) = 0.11889578$ and $var(\hat{\gamma}) = 0.010248121$.

The LR test statistic to test the hypotheses $H_0 : \lambda = 0 \,\&\alpha = 1$ versus $H_1 : \lambda \neq 0 \,\&\alpha \neq 1$ for data set is $\omega = 11.584 > 5.991 = \chi^2_{2;0.05}$, so we reject the null hypothesis.

Table 3: Criteria for comparison.

Model	K-S	-2ℓ	AIC	AICC	BIC
AGTE	0.0954	242.643	248.643	249.143	255.636
Beta-E	0.0962	244.455	250.455	250.621	257.447
Kw-E	0.0988	244.188	250.188	250.521	257.180
Expoential	0.512	254.228	256.228	256.282	258.559

In order to compare the two distribution models, we consider criteria like Kolmogorov-Smirnov (K-S) statistics, -2ℓ, AIC (Akaike information criterion), and AICC (corrected Akaike information criterion) for the data set. The better distribution corresponds to smaller KS, -2ℓ, AIC,AICC and BIC values:

- K-S distance $D_n = \sup_x |F(x) - F_n(x)|$, where, $F_n(x)$ is the empirical distribution function,
- $AIC = -2\log \ell\left(\underset{\sim}{x}, \alpha, \lambda, \boldsymbol{\xi}\right) + 2p$,
- $AICC = AIC + \frac{2p(p+1)}{n-p-1}$,
- $BIC = -2\log \ell\left(\underset{\sim}{x}, \alpha, \lambda, \boldsymbol{\xi}\right) + p\log(n)$,
  where, p is the number of parameters are to be estimated from the data and n the sample size.

Also, here for calculating the values of KS we use the sample estimates of α, λ and γ. Table 3 shows the MLEs under both distributions, Table 3 shows the values of KS, -2ℓ, AIC, AICC, and BIC values. The values in Table 3 indicate that the AGT-E leads to a better fit than the exponential distribution.

The P-P plots, fitted distribution function and density functions of the considered models are plotted in Figures 3 and 4, respectively, for the data set.

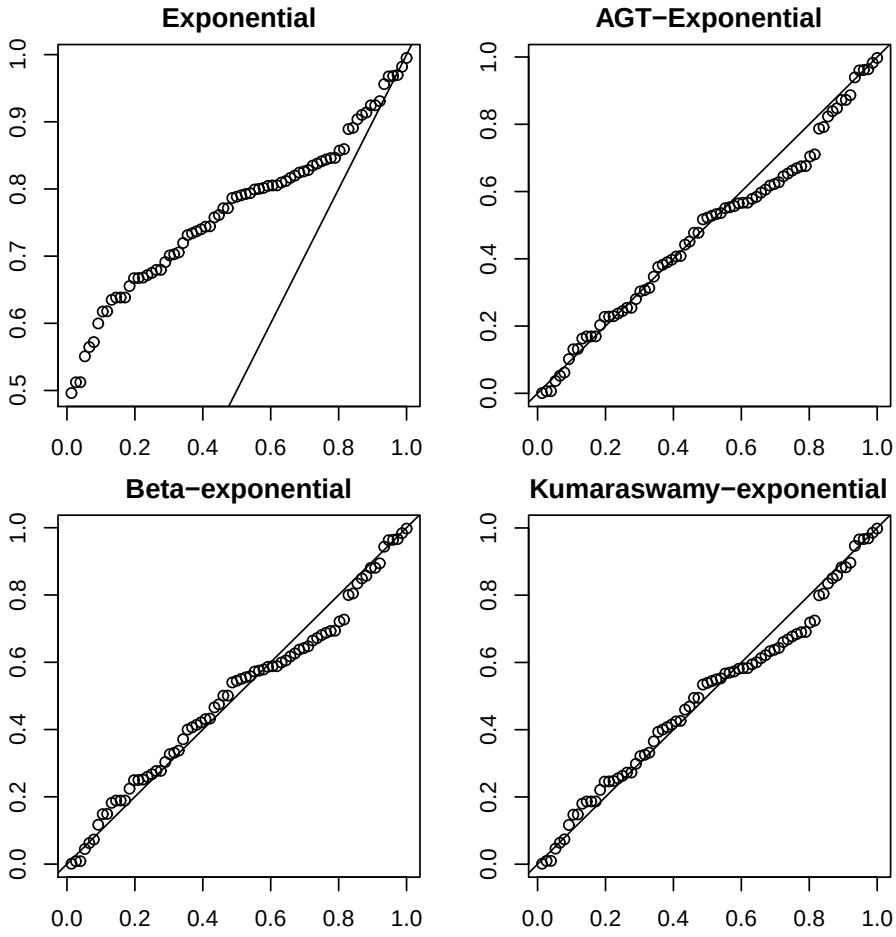


Figure 3: The P-P plots for the real data set

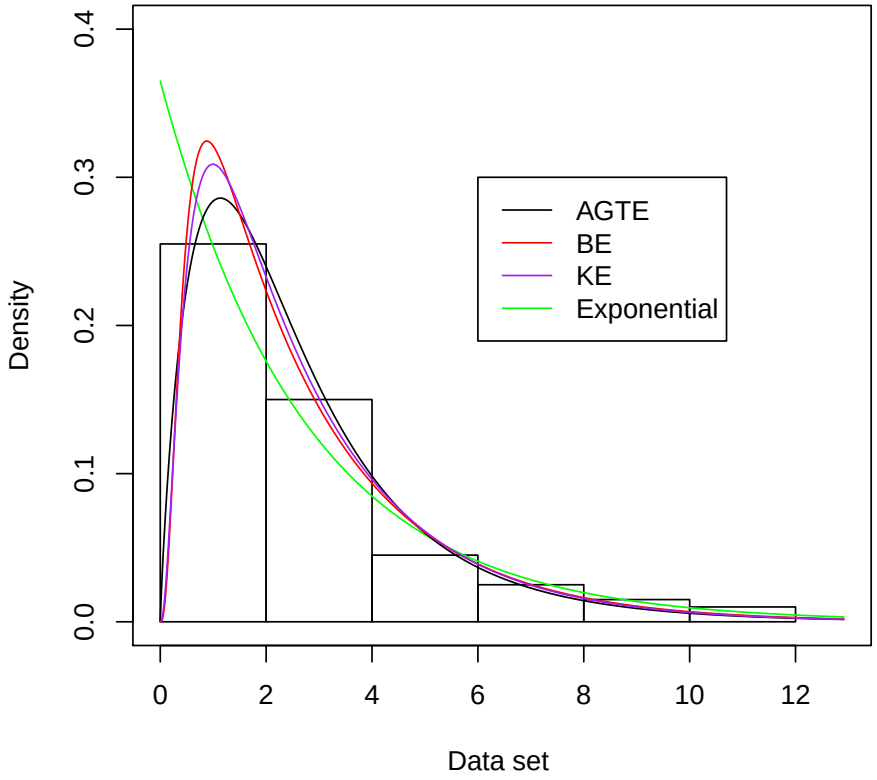


Figure 4: Fitted pdfs plots of the considered distributions for the real data set

References

Alexander C, Cordeiro GM, Ortega EM, Sarabia JM (2012). "Generalized Beta-generated Distributions." *Computational Statistics & Data Analysis*, **56**(6), 1880–1897.

Alizadeh M, Cordeiro GM, De Brito E, Demétrio CGB (2015a). "The Beta Marshall-Olkin Family of Distributions." *Journal of Statistical Distributions and Applications*, **2**(1), 1–18.

Alizadeh M, Emadi M, Doostparast M, Cordeiro GM, Ortega EM, Pescim RR (2015b). "A New Family of Distributions: the Kumaraswamy Odd Log-logistic, Properties and Applications." *Hacettepa Journal of Mathematics and Statistics (to appear)*.

Alizadeh M, Tahir M, Cordeiro GM, Mansoor M, Zubair M, Hamedani G (2015c). "The Kumaraswamy Marshal-Olkin Family of Distributions." *Journal of the Egyptian Mathematical Society*.

Alzaatreh A, Lee C, Famoye F (2013). "A New Method for Generating Families of Continuous Distributions." *Metron*, **71**(1), 63–79.

Alzaghal A, Famoye F, Lee C (2013). "Exponentiated T-X Family of Distributions with Some Applications." *International Journal of Statistics and Probability*, **2**(3), p31.

Amini M, MirMostafaee S, Ahmadi J (2012). "Log-gamma-generated Families of Distributions." *Statistics*, (ahead-of-print), 1–20.

Aryal GR (2013). "Transmuted Log-logistic Distribution." *Journal of Statistics Applications & Probability*, **2**(1), 11–20.

Aryal GR, Tsokos CP (2009). "On the Transmuted Extreme Value Distribution with Application." *Nonlinear Analysis: Theory, Methods & Applications*, **71**(12), e1401–e1407.

Aryal GR, Tsokos CP (2011). "Transmuted Weibull Distribution: A Generalization of the Weibull Probability Distribution." *European Journal of Pure and Applied Mathematics*, **4**(2), 89–102.

Bourguignon M, Silva RB, Cordeiro GM (2014). "The Weibull-G Family of Probability Distributions." *Journal of Data Science*, **12**(1), 53–68.

Cheng R, Amin N (1983). "Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 394–403.

Cordeiro GM, Alizadeh M, Diniz Marinho PR (2015). "The Type I Half-logistic Family of Distributions." *Journal of Statistical Computation and Simulation*, (ahead-of-print), 1–22.

Cordeiro GM, Alizadeh M, Ortega EM (2014a). "The Exponentiated Half-Logistic Family of Distributions: Properties and Applications." *Journal of Probability and Statistics*, **2014**.

Cordeiro GM, de Castro M (2011). "A New Family of Generalized Distributions." *Journal of Statistical Computation and Simulation*, **81**(7), 883–898.

Cordeiro GM, Nadarajah S, *et al.* (2011). "Closed-form Expressions for Moments of a Class of Beta Generalized Distributions." *Brazilian journal of probability and statistics*, **25**(1), 14–33.

Cordeiro GM, Ortega EM, da Cunha DC (2013). "The Exponentiated Generalized Class of Distributions." *Journal of Data Science*, **11**(1), 1–27.

Cordeiro GM, Ortega EM, Popović BV, Pescim RR (2014b). "The Lomax Generator of Distributions: Properties, Minification Process and Regression Model." *Applied Mathematics and Computation*, **247**, 465–486.

Eugene N, Lee C, Famoye F (2002). "Beta-normal Distribution and Its Applications." *Communications in Statistics-Theory and methods*, **31**(4), 497–512.

Glänzel W (1987). "A Characterization Theorem Based on Truncated Moments and Its Application to Some Distribution Families." In *Mathematical statistics and probability theory*, pp. 75–84. Springer.

Gupta RC, Gupta RD (2007). "Proportional Reversed Hazard Rate Model and Its Applications." *Journal of Statistical Planning and Inference*, **137**(11), 3525–3536.

Jones M (2004). "Families of Distributions Arising from Distributions of Order Statistics." *Test*, **13**(1), 1–43.

Leadbetter MR, Lindgren G, Rootzén H (2012). *Extremes and Related Properties of Random Sequences and Processes*. Springer Science & Business Media.

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

Merovci F (2013a). "Transmuted Exponentiated Exponential Distribution." *Mathematical Sciences And Applications E-Notes*, **1**(2), 112–122.

Merovci F (2013b). "Transmuted Lindley Distribution." *International Journal of Open Problems in Computer Science & Mathematics*, **6**(2), 63–72.

Merovci F (2013c). "Transmuted Rayleigh Distribution." *Austrian Journal of Statistics*, **42**(1), 21–31.

Merovci F (2014). "Transmuted Generalized Rayleigh Distribution." *Journal of Statistics Applications and Probability*, **3**(1), 9–20.

Merovci F, Elbatal I (2014a). "Transmuted Lindley-geometric Distribution and Its Applications." *Journal of Statistics Applications & Probability*, **3**(1), 77–91.

Merovci F, Elbatal I (2014b). "Transmuted Weibull-geometric Distribution and Its Applications." *School of Mathematics Northwest University Xian, Shaanxi, PR China*, **10**(1), 68–82.

Merovci F, Elbatal I, Ahmed A (2014). "The Transmuted Generalized Inverse Weibull Distribution." *Austrian Journal of Statistics*, **43**(2), 119–131.

Merovci F, Puka L (2014). "Transmuted Pareto Distribution." *Probstat*, **7**, 1–11.

Nadarajah S, Kotz S (2006a). "The Beta Exponential Distribution." *Reliability engineering & system safety*, **91**(6), 689–697.

Nadarajah S, Kotz S (2006b). "The Exponentiated Type Distributions." *Acta Applicandae Mathematica*, **92**(2), 97–111.

Renyi A (1961). "On Measures of Entropy and Information." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 547–561. University of California Press, Berkeley, Calif. URL http://projecteuclid.org/euclid.bsmsp/1200512181.

Ristić MM, Balakrishnan N (2012). "The Gamma-exponentiated Exponential Distribution." *Journal of Statistical Computation and Simulation*, **82**(8), 1191–1206.

Shannon CE (2001). "A mathematical theory of communication." *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**(1), 3–55.

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 arXiv:0901.0434*.

Tahir M, Cordeiro GM, Alzaatreh A, Mansoor M, Zubair M (2015a). "The Logistic-X Family of Distributions and Its Applications." *Commun. Stat. Theory Methods (2015a). forthcoming.*

Tahir MH, Cordeiro GM, Alizadeh M, Mansoor M, Zubair M, Hamedani GG (2015b). "The Odd Generalized Exponential Family of Distributions with Applications." *Journal of Statistical Distributions and Applications*, **2**(1), 1–28.

Torabi H, Hedesh NM (2012). "The Gamma-uniform Distribution and Its Applications." *Kybernetika*, (1), 16–30.

Zografos K, Balakrishnan N (2009). "On Families of Beta-and Generalized Gamma-generated Distributions and Associated Inference." *Statistical Methodology*, **6**(4), 344–362.

Affiliation:

Faton Merovci
University of Mitrovica "Isa Boletini"
PIM Trepça 40000, Mitrovicë, Kosovo
E-mail: fmerovci@yahoo.com

Morad Alizadeh
Department of Statistics, Faculty of Sciences,
Persian Gulf University, Bushehr, 75169, Iran.
E-mail: moradalizadeh78@gmail.com

G. G. Hamedani
Department of Mathematics, Statistics and Computer Science,
Marquette University, USA.
E-mail: gholamhoss.hamedani@marquette.edu

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 45 *Submitted:* 2015-05-08
September 2016 *Accepted:* 2015-09-14

Austrian Journal of Statistics
September 2016, Volume 45, 95–102.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i3.596

Some Statistics Concerning the Austrian Presidential Election 2016

Erich Neuwirth
University of Vienna
Faculty of Computer Science

Walter Schachermayer
University of Vienna
Faculty of Mathematics

Abstract

The 2016 Austrian presidential runoff election have been repealed by the Austrian constitutional court. The results of the counted votes had yielded a victory of Alexander van der Bellen by a margin of 30.863 votes as compared to the votes for Norbert Hofer. However, the constitutional court found that 77.769 votes were "contaminated" as there have been - at least on a formal level - violations of the legal procedure when counting those votes. For example, the envelopes were opened prematurely, or not all the members of the electoral board were present during the counting etc. Hence the court considered the scenario that the irregular counting of these votes might have caused a reversal of the result as *possible*. The constitutional court sentenced that this *possibility* presents a sufficient irregularity in order to order a repetition of the entire election.

While it is, of course, *possible* that the irregular counting of those 77.769 votes reversed the result, we shall show that the probability, that this indeed has happened, is ridiculously low.

Keywords: election analysis, election fraud detection, election forensics.

1. Introduction

On May 22, 2016 the Austrians voted in a run-off election between Norbert Hofer (candidate 1) and Alexander van der Bellen (candidate 2). The result after counting the votes was 49.7 : 50.3 in favor of van der Bellen. For the precise data we refer to Bundesministerium für Inneres (2016).

The party supporting the candidate Norbert Hofer subsequently appealed to the Austrian constitutional court, claiming irregularities in the procedure of counting the mail votes.

Here are the details. In Austria there are two ways of casting one's vote. Either by showing up personally at the poll site and delivering the vote into the ballot box (ballot voting), or by sending the vote by mail during a well-defined period preceding the voting day (mail voting).

The allegation of Hofer's party was that in some districts the counting of the mail votes violated the procedure stipulated by the law. For example, the letters of the outer envelopes containing these votes (in an inner envelope) should only not be opened before 9:00 a.m. of the subsequent Monday, May 23, the reason being that the electoral board for the mail votes

for districts only is called to duty for this time. By opening these letters prematurely these votes became invalid, as argued by the alleging party. Several other accusations were made, involving different degrees of severity Verfassungsgerichtshof (2016).

The constitutional court carefully investigated these accusations and concluded that in 11 of the 117 voting districts there have indeed happened violations of the law during the procedure of counting the mail votes. The court sentenced that in total there were 77.769 mail votes counted in an irregular way. The central argument of the court in favor of ordering a repetition of the election was that there was the *possibility* that manipulations on such a number of votes might have led to a reversal of the result. After all, the margin was only 30.863 votes.

The constitutional court states explicitly in its finding that there was no evidence that there actually have been manipulations of the votes. What has been proven were several violations of the legally prescribed procedure of counting the votes.

2. The analysis

Our goal is to obtain a quantitative analysis of the probability that there was in reality a victory of Norbert Hofer which was only turned afterwards – in whatever way – into a victory of Alexander van der Bellen because of wrong-counting the mail votes in the incriminated 11 districts.

To do so, we first compare the results in the $N = 106 = 117 - 11$ "green" or "uncontaminated" districts where the court did not find violations of the legal procedures, with the $M = 11$ "red" or "contaminated" districts districts where the court found violations of these procedures.

Each "green" district corresponds to a green dot: on the x-axis we plot the percentage of votes for candidate 1 (Norbert Hofer) among the ballot votes, and on the y-axis the percentage of the votes for candidate 1 among the votes by mail. The picture clearly indicates a linear relation between these two ratios. One also sees that the slope of the regression line is smaller than 1 which does not come as a surprise. Among the voters of Norbert Hofer the propensity to use the possibility of voting by mail is smaller than among the voters of Alexander van der Bellen. We can also observe that the intercept of the regression line essentially vanishes, i.e. the regression line essentially passes through the origin.

While the green dots correspond to the "uncontaminated" districts, the red dots in Figure 1 correspond to the 11 "contaminated" ones. Glancing at Figure 1 one cannot see any alarming behavior of the red dots.

The above Figure 1 corresponds to the *counted* votes. In the subsequent analysis we shall only accept the green dots as valid data. As regards the red dots, we only take their x-coordinate as granted: recall that the x-coordinate corresponds to the percentage of ballot votes in favor of candidate 1. As regards the y-coordinates of the red dots, it is precisely our point to analyze whether the *true* votes gave different results than the *counted* votes in a degree which could have resulted in a reversal of the election result.

As an illustration, Figure 2 below indicates a scenario for the *true* votes which would have yielded a victory for candidate 1 (by a margin of 1 vote). To obtain Figure 2, we have assigned – hypothetically – 15.432 (half the missing 30.863 votes rounded up) proportionally to the 11 "contaminated" districts, and subsequently recalculated the corresponding percentages. This procedure implements a scenario where that many votes have wrongly been counted for candidate 2 instead of candidate 1. For more detailed information we refer to the web site of the first named author Neuwirth (2016) (`http://www.wahlanalyse.com/WahlkartenDifferenzenVfGh.html`).

It is evident that a scenario as in diagram 2 does not look very likely to have happened in reality. This diagram only has an illustrative character for our purposes in order to visualize the absurdity of such a scenario. It will not play any role in the subsequent analysis. In particular, we shall not assume a certain given assignment of the missing 30.863 votes to the

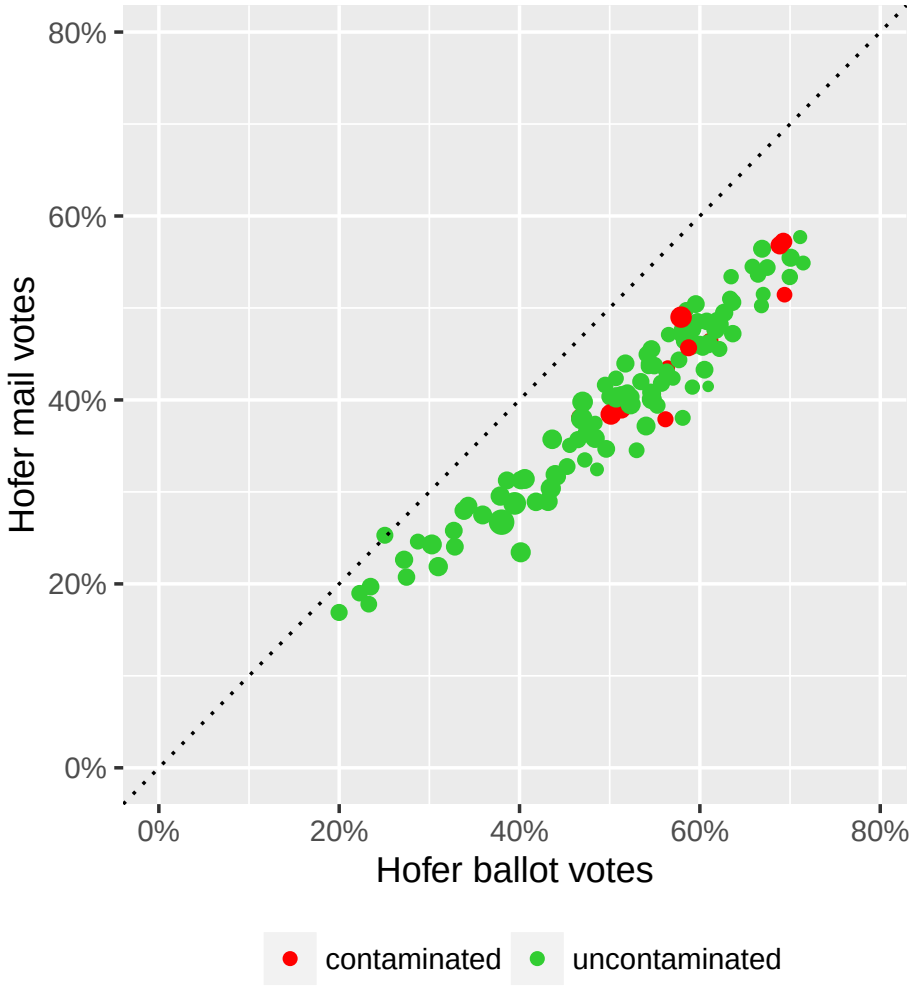


Figure 1: Mail and ballot vote percentages - official results

11 red districts. We shall only be interested in their *total sum*. Speaking mathematically, we shall eventually calculate the probability distribution of the total number of *true* mail votes in the incriminated districts, conditionally on the given results of the "green" districts, and calculate the probability of the event that they would have resulted in a victory of candidate 1.

To analyze the probability that the *true* votes by mail in the red districts would have yielded a victory for candidate 1, we apply a weighted linear regression model to the green dots in diagram 1.

In fact, since the calculations become easier to write and to program, we use a regression model on the number of votes instead of the percentages, which is mathematically equivalent. Since the expected variation of the votes depends on the total number of votes, the model exhibits heteroskedascity, and we have to use a weighted regression.

Regression models include prediction intervals for observations with known values for the independent variable(s) and known standard deviations (compared to standard deviations of completely known cases). Using these procedures and assuming that the mail vote results in the contaminated districts follow the model of the uncontaminated districts, we can then compute the distribution of the sum of the expected votes in the contaminated 11 districts (which, under the present model assumptions, follows a rescaled t-distribution).

Using the distribution of this expected value, we can then calculate the probability that the *true votes* would have resulted in an election of candidate 1. The numerical value equals

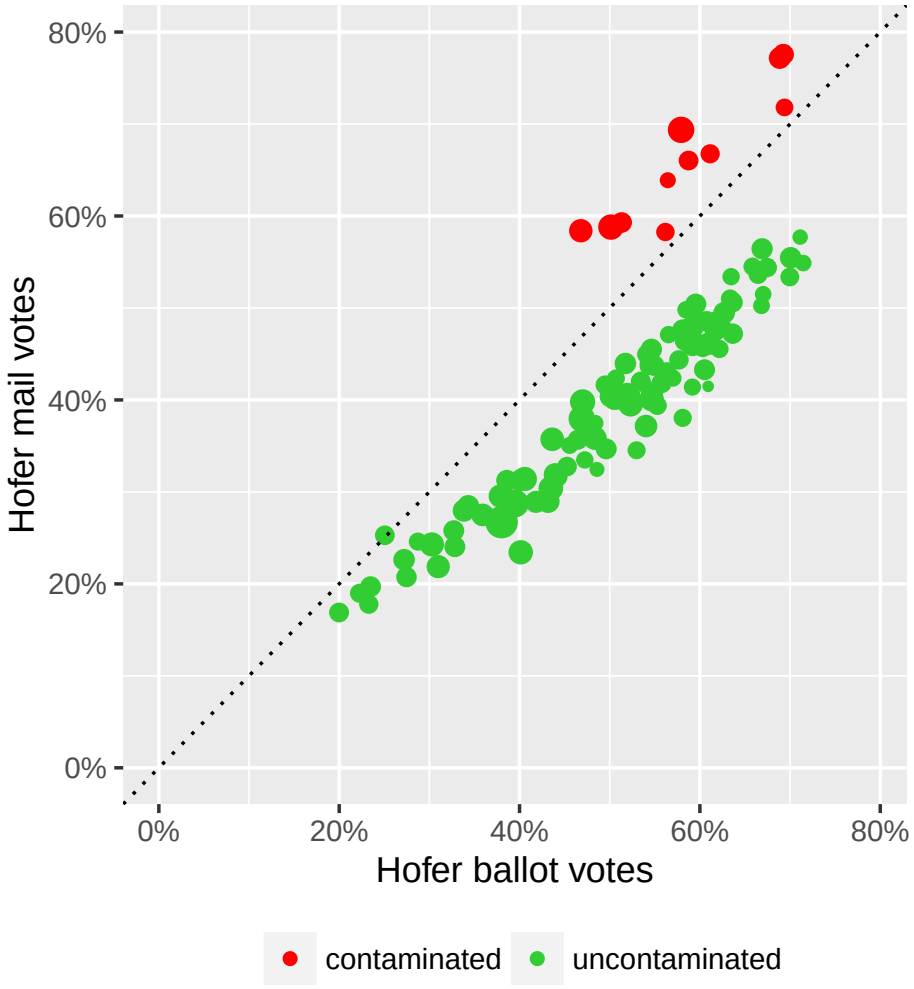


Figure 2: Mail and ballot vote percentages - modified results

$p = 1.322065 \cdot 10^{-10}$.

Actually, it turns out that the sentence of the constitutional court may also lead to a slightly different calculation. Apart from the 11 "contaminated" districts it identified 3 more "dubious" districts where things are not so clear. We refer to Verfassungsgerichtshof (2016) for the details. Although the sentence of the court did not take into account these districts, one might argue that – possibly – also in these districts the counting of the mail votes was not reliable. Mathematically speaking, this leads to the consideration of M = 14 "red" and N = 117 -14 = 103 "green" districts. The calculations are identical to the above considered case and lead to a numerical value of $p = 5.151422 \cdot 10^{-8}$.

This modified analysis still gives an extremely low probability which for practical purposes rules out the possibility of manipulations having taken place. It also shows that the results of the analysis are quite robust under slightly different model assumptions.

3. The model

We build our model following the ideas of model based survey approach, namely we interpret the results of the 106 districts without irregularities as a sample of all mail results and use for prediction of the overall results an estimate for the 11 districts with possible irregularities. This can be seen as as an application of the ratio estimator (see e.g. Valliant, Dorfman, and Royall (2000)).

We consider $N = 117-11 = 106$ voting districts with a total of t_n valid votes for $n = 1, \ldots, N$. In district n v_n votes were counted for candidate 1 and $\bar{v}_n$ votes for candidate 2 so that $v_n + \bar{v}_n = t_n$.

The votes t_n split into b_n many ballot votes and m_n many mail votes which again are divided into $v_{b,n} (resp.\ v_{m,n})$ many votes for candidate 1 and $\bar{v}_{b,n} = b_n - v_{b,n}$ $(resp.\ \bar{v}_{m,n} = m_n - v_{m,n})$ many votes for candidate 2.

Our objects of interest are the vote numbers

$$v_{b,n} \quad \text{and} \quad v_{m,n}, \quad n = 1, \ldots, N.$$

These numbers denote the counted votes for candidate 1 among the ballot and mail votes respectively. As indicated by the diagrams above, a linear relation between these quantities is justified as model assumption. To make the plots easier to understand, we used percentages instead of votes there.

While the numbers $(v_{b,n})_{n=1}^N$ are considered as given data the numbers $(v_{m,n})_{n=1}^N$ are considered as realizations of the following random variables:

$$V_{m,n} = k\, v_{b,n} + \epsilon_n, \quad n = 1, \ldots, N \tag{1}$$

Here k is an unknown deterministic number while $(\epsilon_n)_{n=1}^N$ are independent centered Gaussian random variables. Variances of votes for parties being proportional to the number of total votes is a standard model assumption in statistical voting analysis procedures (see Bruckmann (1966), Neuwirth (1984), Neuwirth (1994), Neuwirth (2012), and Ledl (2007)).

Assuming independence of vote counts in different districts seems very natural. Independence of the decision of single voters is more difficult to justify. In fact, if this were the case, one could model the sum of votes as a binomial (or multinomial) random variable. We can, however, assume that for small groups of voters (e.g. families) there is a fixed covariance structure of the voting decisions for the members of such a group. Then, the sum of votes for each group of a fixed size has fixed variance. Assuming the the vote sums of different groups are independent, we see that the variance of the vote sums of a district is proportional to the number of groups, and, since we assume the groups to be of essentially equal sizes, also proportional to the number of voters.

Therefore, the variances of our random variables $V_{m,n}$, have values

$$\operatorname{var}(V_{m,n}) = \operatorname{var}(\epsilon_n) = \sigma^2 m_n \tag{2}$$

for some (unknown) deterministic number $\sigma > 0$.

We are thus facing a heteroskedastic, linear regression model.

Applying standard regression theory, we obtain the estimators $\hat{k}, \hat{\sigma}$ which we consider as random variables. In particular, the estimator $\hat{k}$ follows a (rescaled) t-distribution whose parameters can be explicitly calculated for the given data.

We next consider $M = 11$ many "contaminated" districts, disjoint from the N " uncontaminated" districts.

Assuming that the results $V_{m,j} \quad j = 1, \ldots, M.$ also follow model 1, they can be considered random variables

$$V_{m,j} = k\, v_{b,j} + \epsilon_j, \quad j = 1, \ldots, M.$$

As we do not know the true value of k we have to consider the estimated variable

$$\hat{V}_{m,j} = \hat{k} v_{b,j} + \epsilon_j, \quad j = 1, \ldots, M.$$

The new noise variables $(\epsilon_j)_{j=1}^M$ are such that $((\epsilon_n)_{n=1}^N, (\epsilon_j)_{j=1}^M)$ are independent. The variance of the ϵ_j again is given by $\sigma^2 m_j$.

Finally we consider the sum

$$\hat{V} = \sum_{j=1}^{M} \hat{V}_{m,j},$$

the total ballot result of candidate 1 in the contaminated districts, $v_b = \sum_{j=1}^{M} v_{b,j}$, the total number of votes there, $m = \sum_{j=1}^{M} m_j$.

$\hat{V}$ is the random variable modeling the total mail votes for candidate 1 in the M "contaminated" districts.

The random variable $\hat{V}$ follows the model

$$\hat{V} = \hat{k}\, v_b + \epsilon$$

where ϵ is a centered Gaussian variable with variance $\sigma^2 m$.

A standard tool of regression theory allow us to compute a prediction interval for $\hat{V}$.

If σ^2 were known, $\hat{V} = \hat{k}v_b + \epsilon$ were distributed with mean $\hat{k}v$ and variance of $\hat{V}$ equal to $\sigma^2 \left(\frac{v_b^2}{\sum_{n=1}^{N} \frac{v_{b,n}^2}{m_n}} + m \right)$

Using this fact we use the regression model estimate for $\hat{\sigma}$ as substitute for the unknown constant σ. Using this, the random variable

$$\frac{\hat{V} - \hat{k}v_b}{\hat{\sigma}\sqrt{\frac{v_b^2}{\sum_{n=1}^{N} \frac{v_{b,n}^2}{m_n}} + m}}$$

follows a t-distribution with 105 degrees of freedom.

We compare this random variable with the critical number $\tilde{V}$ which would be necessary for candidate 1 reversing the result. Finally we compute

$$\mathbb{P}[V \geq \tilde{V}] \tag{3}$$

which can be computed explicitly.

4. Confidence intervals for prediction

We use results for the standard heteroskedastic linear model

$$y = X\beta + \epsilon$$

with covariance matrix $(cov)(\epsilon) = \sigma^2 W$.

In this model, the best linear unbiased estimator for β is

$$\hat{\beta} = (X'W^{-1}X)^{-1}X'W^{-1}y$$

The covariance of this estimator is

$$\begin{aligned} \operatorname{cov}(\hat{\beta}) &= (X'W^{-1}X)^{-1}X'W^{-1}\operatorname{cov}(y)((X'W^{-1}X)^{-1}X'W^{-1})' \\ &= (X'W^{-1}X)^{-1}X'W^{-1}\sigma^2 WW^{-1}X(X'W^{-1}X)^{-1} \\ &= \sigma^2(X'W^{-1}X)^{-1} \end{aligned}$$

In our case, X is the Nx1-matrix $(v_{b,n})_{n=1}^{N}$ and W is the diagonal matrix $\operatorname{diag}((m_n)_{n=1}^{N})$. The parameter β in our case is the scalar k and its estimator is $\hat{k}$, and $\operatorname{cov}(\hat{\beta})$ becomes $\operatorname{var}(\hat{k})$

Therefore $X'W^{-1}X = \sum_{n=1}^{N} v_{b,n}\frac{1}{m_n}v_{b,n} = \sum_{n=1}^{N}\frac{v_{b,n}^2}{m_n}$ and

$$\mathrm{var}(\hat{k}) = \sigma^2\frac{1}{\sum_{n=1}^{N}\frac{v_{b,n}^2}{m_n}}$$

We want to compute the distribution of $\hat{V} = \hat{k}v_b + \epsilon$ with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{var}(\epsilon) = \sigma^2 m$.
We have

$$\begin{aligned}\mathrm{E}(\hat{k}v_b + \epsilon) &= \mathrm{E}(\hat{k}v_b) + \mathrm{E}(\epsilon) = kv_b \\ \mathrm{var}(\hat{k}v_b + \epsilon) &= \mathrm{var}(\hat{k}v_b) + \mathrm{var}(\epsilon) = \sigma^2\left(\frac{v_b^2}{\sum_{n=1}^{N}\frac{v_{b,n}^2}{m_n}} + m\right)\end{aligned}$$

Since σ^2 is unknown, we have to replace it by the estimator $\hat{\sigma}^2$ and then $\frac{\hat{V}-\hat{k}v_b}{\hat{\sigma}\sqrt{\frac{v_b^2}{\sum_{n=1}^{N}\frac{v_{b,n}^2}{m_n}}+m}}$ follows a t-distribution with $N-1$ degrees of freedom, an from that confidence intervals for $\hat{V}$ are easily derived.

5. The results

All the data and the code for performing our analysis can be found at https://github.com/neuwirthe/AustrianPresidentialElection.

Hofer had 34479 votes in the contaminated districts, and he would need additional 15432 votes, so that in total he needs $\tilde{V} = 34479 + 15432 = 49911$ votes to overturn the result.

Using the R code from the URL above to compute the probability of a result overturning the result in favor of Hofer, we get the value

$$1.322065 \cdot 10^{-10}.$$

References

Bruckmann G (1966). *Schätzung von Wahlresultaten aus Teilergebnissen.* Physica-Verlag, Wien.

Bundesministerium für Inneres (2016). http://www.bmi.gv.at/cms/BMI_wahlen/bundespraes/bpw_2016/. [Online, accessed 20-September-2016].

Ledl T (2007). *Modellierung von Wechselwählerverhalten als Multinomialexperiment.* Ph.D. thesis, Fakultät für Wirtschaftswissenschaften und Informatik, Universität Wien.

Neuwirth E (1984). "Schätzung von Wählerübergangswahrscheinlichkeiten." In M Holler (ed.), *Wahlanalyse – Hypothesen, Methoden und Ergebnisse.* tuduv-Buch.

Neuwirth E (1994). "Prognoserechnung am Beispiel der Wahlhochrechnung." In P Mertens (ed.), *Prognoserechnung.* physica-Verlag.

Neuwirth E (2012). "Wahlhochrechnung: ein kurzer Überblick über den Einsatz bei bundesweiten Wahlen in Österreich." In W Lutz, H Strasser (eds.), *Österreich 2032 (Festschrift zum 80. Geburtstag von Gerhart Bruckmann).* Verlag der österreichischen Akademie der Wissenschaften.

Neuwirth E (2016). http://www.wahlanalyse.com/WahlkartenDifferenzenVfGh.html. [Online, accessed 20-September-2016].

Valliant R, Dorfman AH, Royall RM (2000). *Finite Population Sampling and Inference*. Wiley.

Verfassungsgerichtshof (2016). "Erkenntnis *W I 6/2016-125*." https://www.vfgh.gv.at/cms/vfgh-site/attachments/5/7/8/CH0003/CMS1468412977051/w_i_6_2016.pdf. [Online, accessed 20-September-2016].

Affiliation:

Erich Neuwirth
Faculty of Computer Science
University of Vienna
Währinger Straße 29 A-1090 Vienna, Austria
E-mail: erich.neuwirth@univie.ac.at
URL: http://homepage.univie.ac.at/erich.neuwirth

Walter Schachermayer
Faculty of Mathematics
Universitity of Vienna
Oskar-Morgenstern-Platz 1
A-1090 Vienna, Austria, and
Institute for Theoretical Studies, ETH Zurich
E-mail: walter.schachermayer@univie.ac.at
Partially supported by the Austrian Science Fund (FWF) under grant P25815, the Vienna Science and Technology Fund (WWTF) under grant MA09-003 and Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zurich Foundation.

Austrian Journal of Statistics
September 2016, Volume 45, 103–105.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i3.573

Laudatio verfasst von Johann Bacher (JKU Linz)
Wien, am 4.4.2016

Laudatio zur Verleihung des Bruckmannpreises 2016 der Österreichischen Statistischen Gesellschaft an A. Univ.-Prof. Mag. Dr. Andreas Quatember

Sehr geehrte Damen und Herren,

sehr geehrte Festgäste,

lieber Andreas Quatember,

es ist mir eine große Freude und Ehre anlässlich der Verleihung des Gerhart Bruckmann-Preises für dich, Andreas, eine kurze Laudatio zu halten.

Freude und Ehre aus folgenden drei Gründen:

1.) Es wird ein Mitglied unserer Fakultät geehrt und diese Ehre strahlt auch auf unsere Fakultät aus. Auch sie wird mitausgezeichnet - zumindest beansprucht sie das. Lieber Andreas, dagegen kannst du dich nicht wehren!

2.) Geehrt wird ein Kollege, dem Statistik und öffentliche Aufklärung über Statistik schon immer ein zentrales Anliegen waren und sind. Für ihn war der aktuelle „Modetrend“ der Dissemination, dass sich Universitäten und WissenschaftlerInnen öffentlich äußern, nicht erforderlich. Seit langem weist du, Andreas, in der Öffentlichkeit auf falsche Anwendungen und Interpretationen von Statistiken hin.

3.) Freude schließlich, da ein Kollege geehrt wird, der bereit ist sich zu engagieren, der immer ein offenes Ohr für die unterschiedlichen Anliegen hat, für den Kollegialität und nicht Eigennutz an erster Stelle steht.

Geboren wurde Andreas Quatember Anfang der 1960er in Linz. Nach der Volksschule besucht er die Fadingerschule. Die Fadingerschule ist - zu Ihrer Information, sehr geehrte Damen und Herren - eine Linzer Institution, ein Bundesrealgymnasium mit über 500 SchülerInnen im Zentrum von Linz mit einem naturwissenschaftlichen Schwerpunkt. Eine - wie ich mich in einem Sparkling-Science-Projekt überzeugen konnte - faszinierende Schule, die SchülerInnen viele Freiräume lässt, aber als Gegenleistung auch Selbstorganisation und Selbständigkeit einfordert.

Andreas besuchte in den 1970er Jahren die Fadingerschule zur Zeit der Bildungsexpansion, in der sich die Gymnasien gegenüber breiten Bevölkerungsgruppen quantitativ und auch kognitiv, mentalitätsmäßig öffneten. Auch wenn wir aus der empirischen Bildungsforschung wissen, dass der Einfluss der Schule auf den Bildungserfolg und die Persönlichkeitsentwicklung nicht überschätzt werden darf, so hat das Gymnasium Andreas Quatember geprägt. Sein Interesse an Mathematik wurde in der Unterstufe durch einen engagierten Professor gefördert, in der Pubertät gewährte die Schule ausreichende Freiheiten und hat ihn letztlich nicht zu stark geprägt, sodass er nach der Matura noch nicht genau wusste, was er studieren möchte. Seine

Präferenzen schwankten zwischen Mathematik und Soziologie und er entschied sich schließlich 1983 für das Studium der Sozial- und Wirtschaftsstatistik an der Johannes Kepler Universität Linz, das er 1989 mit einer Arbeit zum sequentiellen Quotiententest in der statistischen Qualitätskontrolle erfolgreich abschloss.

Bereits während des Studiums war er als Studienassistent am Institut für Angewandte Statistik tätig, über eine Karenzvertretung erhielt er schließlich einen Assistentenvertrag. 1996 schloss er erfolgreich seine Dissertation ab, 2014 habilitierte er sich in Statistik.

Bei seiner Dissertation kam ihm die während seiner Schulzeit erworbene Selbständigkeit zugute. Seine Begeisterung für wissenschaftliches Arbeiten und die Statistik kamen hinzu, sodass er als einer von wenigen in diesem Zeitraum seine Dissertation zum Quotenverfahren erfolgreich abschloss.

Fragen der Stichprobentheorie bzw. allgemein der Gewinnung von statistisch validen Daten ließen ihn seitdem nicht mehr los. In seiner Habilitationsschrift entwickelt er mit dem Konzept der Pseudo-Population einen einheitlichen Rahmen zur Analyse und Modellierung unterschiedlicher Fehlerquellen, die bei der Gewinnung und Aufbereitung von Daten mittels Stichproben auftreten können. Eine anregende und spannend zu lesende Arbeit.

Ein Merkmal der Arbeiten von Andreas Quatember ist, dass er - wie bereits erwähnt - Modetrends ignoriert, ihnen nicht folgt. Es geht ihm um die statistische Analyse und Bewertung von realen Anwendungsproblemen der Umfrageforschung, unabhängig davon, ob sich der Mainstream der Statistik gerade damit beschäftigt oder nicht. Andreas Quatember hat als Folge dieser Eigenständigkeit - Eigenständigkeit im positiven Sinn - für meinen Arbeitsbereich, die empirische Sozialforschung, aber auch für andere angewandte Forschungsfelder, wie z.B. die Markt-, Innovations- oder Unternehmensforschung, wichtige Grundlagenarbeit geleistet. Manche seiner gewonnenen Erkenntnisse sind für uns enttäuschend", da wir in der Praxis oft nach einfachen Rezepten und Regeln suchen. Aber Andreas Quatember hält auch diesen Verführungen stand. Obwohl ich es gelegentlich immer noch versuche, ihm statistische Rezepte zu entlocken, bleibt er seinem wissenschaftlichen Ethos treu, dass zunächst eine sorgfältige statistische Analyse eines Problems erforderlich ist, bevor eine Antwort auf eine Anwendungsfrage gegeben werden kann, und dass simple Verallgemeinerung oft nicht möglich sind.

Teil dieser Eigenständigkeit ist auch seine Wertschätzung der Lehre, die heute - wo universitäre Leistungen vielfach nur mehr an Journalartikeln gemessen werden - unterbewertet wird. Andreas Quatember lehrt gerne und gut, wie die positiven Rückmeldungen Studierender zu seinen Lehrveranstaltungen zeigen.

Identifikation mit seinem Fach, wissenschaftliches Ethos und Eigenständigkeit erklären auch, warum der Preisträger nicht einfach Pressemeldungen mit statistischem Unsinn hinnimmt. Kontinuierlich wächst seine Homepage, in der er über Grafiken mit vertauschten X-Y-Achsen, über Diagramme mit verzerrenden Skalierungen, über falsche Interpretationen von Mittelwerten, über die Verwechslung von Stichprobengröße und Repräsentativität sowie über vermeintlichen Zusammenhängen informiert. So z.B. verlautet eine Presseaussendung von ORF-Science „Die Lebenserwartung verbessert sich mit zunehmendem Pensionistenalter" (**?**).

Diese falschen bzw. problematischen Darstellungen und Interpretationen werden vielfach überlesen, bestimmen aber leider den öffentlichen Diskurs, wenn gesellschaftlich relevante Themenbereiche angesprochen werden, wie etwa PISA oder das oben genannte Pensionsbeispiel, wo nicht ausgeschlossen werden kann, dass die Ergebnisse in der öffentlichen Diskussion als Argument für eine Anhebung des Pensionsalters verwendet werden, da dies die Gesundheit fördere und die Lebenserwartung erhöhe.

Daher ist Aufklärung darüber, welche Aussagen aufgrund statistischer Erhebungen und Analysen zulässig sind und welche nicht, wichtig. Zum einen für die Reputation der Fachdisziplin, zum anderen ist sie ein wichtiger allgemeiner Beitrag zur Förderung eines mündigen Bürgertums. Ich möchte daher der Jury danken, dass Sie als diesjährigen Preisträger Andreas Quatember ausgewählt hat. Als Laudator erlaube ich mir, dir Andreas zu diesem Preis sehr herzlich zu gratulieren. Bedanken möchte ich mich abschließend nochmals für deine eingangs

erwähnte Kollegialität. Sie ist nicht selbstverständlich. Danke und alles Gute!

Literatur

Quatember A (2016). "Universum in den Medien - Mittelwerte." http://www.jku.at/ifas/content/e101235/e101336 , (20.06.2016).

Address of author:

Johann Bacher
Dekan der Sozial- und Wirtschaftswissenschaftlichen Fakultät
Johannes Kepler Universität Linz
E-mail: johann.bacher@jku.at

ISSN 1026–597X

Contents

Page