Journal of Statistical Software

August 2018, Volume 86, Book Review 2. *doi: 10.18637/jss.v086.b02*

Reviewer: Hakan Demirtas
University of Illinois at Chicago

Handbook of Fitting Statistical Distributions with R

Zaven A. Karian, Edward J. Dudewicz
Chapman & Hall/CRC, Boca Raton, 2010.
ISBN 978-1584887119. xlv + 1672 pp. USD 210.00 (HC).
https://www.crcpress.com/9781584887119

Handbook of Fitting Statistical Distributions with R is a useful reference book that should be in the bookshelf of anybody who deals with the theory and applications of statistical distributions. The book has enormous breadth and depth, an immense amount of material is covered in about 1700 pages; its scale, scope, and the information it carries are equivalent to those of several books combined. The book is designed around the generalized lambda distribution (GLD). The GLD is one of the few important generalized classes of distributions; its introduction goes back to nearly half a century ago. It has four parameters (location, scale, skewness, and kurtosis). It covers a wide region in the skewness-kurtosis plane and takes many well-known distributions as a special case. This level of generality and flexibility makes the GLD a versatile option in a broad spectrum of research fields including the physical, medical, biobehavioral, social, and managerial sciences as well as in engineering. The downside is that the parameter estimation process can be challenging for such an inclusive distribution.

The GLD's more general form EGLD (extended GLD) is established with the addition of the generalized beta distribution (GBD) into the formulation to cover more space in the symmetry-elongation plane. The Johnson system and the kappa distribution are also delineated. Major estimation techniques such as method of moments (MoM), method of percentiles (MoP), and method of L-moments (MoL) are described for most of the above distributional setups. The book also encompasses other pivotal topics such as inference based on quantile functions, assessment of quality of fit, random variate generation (RVG) procedures, the generalized bootstrap (GB) and Monte Carlo (MC) methods, and some peripheral topics such as response modeling methodology (RMM) and design of experiments (DoE). The book comes with an accompanying CD-ROM, which includes computer programs that implement the methodologies appeared in the book.

The main authors (Karian and Dudewicz) are well-recognized authorities in statistical communities. The book is divided into eight parts (seven textual ones and appendices) and 34 chapters within these parts that are written by 46 authors in total. Each part has 3–10 chapters with the exception of Part I (a single overview chapter) and Part VIII (tabular appendices). The two main authors wrote 10 chapters, one or both appear in three more

chapters. The range of authors is wide, including people working in academia and industry, from all continents. Six chapters have end-of-chapter exercises.

In general, it is a decent book. Positives are summarized below. The book

- presents a comprehensive set of methods, algorithms, and computations for fitting distributions to data;
- provides some in-depth coverage of applications;
- explains how to model/fit a continuous probability distribution to data well;
- includes chapters written by experts who offer deep insight on a range of topics;
- uses real data sets in applications that span many areas, including agriculture, reliability estimation, weather phenomena, water systems, insurance and inventory management, and materials science;
- discusses the roles of DoE, RVG, and assessment of fit quality in many applications;
- contains proofs of key results throughout as well as necessary tables in the appendices;
- provides programs on an accompanying CD-ROM.

On the other side of the coin, negatives include:

- The organization could have been better. There are some unnecessary repetitions. Each chapter is supposed to be self-contained, so it is tolerable. When many authors write around a unified theme, it is what usually happens.
- Some chapters seem redundant, out of place, or somewhat irrelevant. The book occasionally gives an impression that a garden variety of papers are put together without paying much attention to integrity.
- Typographical and grammatical errors are palpable but for a book of this size, it may be natural.
- The title is not really accurate in the sense that (a) the crux of the book is built around GLD, GBD, EGLD, the Johnson family, and the kappa distribution. It is comprehensive but it does not include all statistical distributions in a way the title may suggest, and (b) R is not the only software throughout the book, Maple, MATLAB, C, and SAS codes are also utilized. The title is probably chosen for better marketability, which is mildly annoying. It may not necessarily be a bad thing though considering the book's potential to improve the awareness among practitioners.
- This book is too thick and dense; it could have been leaner, with a more efficient, orderly, and consistent structure.
- Author names do not appear in the table of contents.

We are now ready to delve into the specifics and coverage in detail. Part I has only one chapter, which is an overview. It gives historical background of the GLD and summarizes the organization and covered topics for the rest of the book. Part II starts with the definition of the GLD, the parameter space and associated regions for validity and shapes of the GLD density functions for a set of parameter values, RVG for this family, and the fitting process. Most chapters in Part II follow the same format, leading to an easy-to-read framework. This structure involves appropriate moments, symmetry-peakedness plane, fitting continuous distributions to data, approximations to well-known distributions, examples, handling data that come in the form of a histogram, and RVG. This part proceeds with two chapters that cover MoM estimation for the GLD and EGLD systems. The EGLD connects the GBD and GLD to accommodate a larger spectrum of parameter combinations, making it a strong contender in statistical modeling. Two key chapters are dedicated to MoP and MoL estimation. In situations where some moments do not exist, or they are difficult to estimate, or they have a large variability, percentile-based approaches and L-moments could be particularly operational. One chapter is concerned with how to deal with multiple solutions to GLD that ensue via moments and percentiles. Subsequent chapters cover mixture distributions, the bivariate GLD, and fitting GLD with location- and scale-free functionals. The last chapter is pertinent to DoE, which is useful in general, but seems a little out-of-context for the purposes of this book.

Part III augments the discussion into the world of quantile functions. Statistical modeling and fitting based on quantile distribution functions (Chapter 12 is extremely informative), RMM, and fitting GLDs and mixture of GLDs to data using quantile matching method, along with information about the relevant software tools (Chapters 14–15) are included in this part. Part IV is an interesting chunk of the book; it enriches the content markedly by introducing two other generalized families, the Johnson system and the kappa distribution. The Johnson system consists of distributions that, through specified transformations, can be reduced to the standard normal variable. MoM estimation for this system and approximations to well-known distributions are considered. The four parameter kappa distribution fit through MoP and MoL are also covered. The GLD and kappa distributions jointly span a large region in the L-moment space, which makes this duo broadly applicable in many contexts and disciplines. Inclusion of the Johnson system and the kappa distribution significantly enhances the scope of the book. This part continues with an intriguing, must-read chapter about non-normal error distributions (skewed generalized t, the GBD of the first and second kind), probability- and quantile-based distributional regression, quantile-based partially adaptive estimation, and the expanded version of the GLD with five parameters (Chapter 18). This part ends with a multivariate gamma distribution for linearly related proportional outcomes.

Part V is concerned with the GB and MC methods. The regular bootstrap method has serious limitations and the generalized form is designed to handle these complications. The GB essentially fits an ELGD to the available data, and takes samples from the fitted distribution. This parametric bootstrap procedure via a comprehensive distributional family (EGLD) is shown to perform better than the standard nonparametric bootstrap in this part in terms of model fit in general and confidence intervals for high quantiles. Part VI is allocated to assessment of quality of fits, which is an important issue whenever statistical modeling is involved, but this part is more geared towards theoreticians rather than practitioners, if one has to be economical on reading time, this part can be skipped. Part VII explores real-world applications in agriculture, reliability estimation, hurricanes, hail storms, water

systems, insurance and inventory management, and materials science. The applications in these chapters complement others in the book that deal with competitive bidding, medicine, biology, meteorology, bioassays, economics, quality management, engineering, control, and planning. Five chapters are involved with the GLD, the other four are about more general discussions on practical elements of distributions in applied settings.

In a nutshell, it is a rich reference book on fitting a few important classes of continuous distributions to data, it is neither a textbook nor a self-study book. It has substantial intellectual value as a source for theoretical, application-related, and computational aspects of fitting several statistical distributions. However, a reader should know how to get the best out of it. It is not a kind of book to read from beginning to end, it would be suboptimal for most people, selective reading is probably a better approach. The recommended style would be reading the overview (Chapter 1), digesting the coverage throughout the book roughly, taking notes, and getting back to it on an as-needed basis. Beginners are advised to start with a more basic/accessible book.

Reviewer:

Hakan Demirtas
Division of Epidemiology and Biostatistics (MC 923)
University of Illinois at Chicago, School of Public Health
1603 West Taylor Street, Room 950
Chicago, IL, 60612-4336, United States of America
E-mail: demirtas@uic.edu
URL: http://www.healthstats.org/members/hdemirtas.html

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
August 2018, Volume 86, Book Review 2 *Published:* 2018-09-04
doi:10.18637/jss.v086.b02