Generalizations of Tukey distributions

Czesław Domański
University of Lodz, Lodz, Poland czedoman@uni.lodz.pl

Katarzyna Bolonek-Lasoń
University of Lodz, Lodz, Poland kbolonek1@wp.pl

Abstract

The Generalized Lambda Distribution (GλD) is a four-parameter generalization of Tukey's Lambda family. Several methods for estimating the parameters of the GλD have been reported in the literature, but the most popular one is the moment-method matching proposed by Ramberg and Schmeiser (1974). One of the drawbacks of the GλD is that the shape parameter also determines skewness. It seems reasonable that there should be three linear parameters determining position, scale, and skewness and two parameters determining the shapes of the two tails. This suggests a natural generalization of the GλD to give a five-parameter lambda distribution (FPLD).

Keywords: Tukey distribution, moment method, lambda distribution.

1. Foreword

Varied empirical distributions have been reported in statistical literature. The Pearson system (Pearson (1948), the Johnson system (Hahn and Shapiro (1967)), the Burr distribution (Burr (1973)) and Tukey distribution (1960) belong to the most important empirical distributions.

In order to derive probability distribution function (PDF) one can follow two approaches.

Pearson (1895) specified differential equation as the basis to derive the family of PDF

$$\frac{df(x)}{dx} = \frac{(x-c)f(x)}{a_0 + a_1 x + a_2 x^2} \tag{1}$$

where $f(x)$ is a density function. The constant coefficients entering the distribution can be described in terms of the first four moments (when they exist).

Solutions of eq. (1) are classified according to the character of roots of equation $a_0 + a_1 x + a_2 x^2 = 0$, where

$$a_0 + a_1 x + a_2 x^2 = a_2\left(x - \frac{\left(a_1 - \sqrt{a_1^2 - 4a_0 a_2}\right)}{2a_2}\right)\left(x - \frac{\left(a_1 + \sqrt{a_1^2 - 4a_0 a_2}\right)}{2a_2}\right). \tag{2}$$

Edgeworth (1898) presented another approach, the so called method of translation which is based on two functions. D'Addaro (1949) called these functions the probability generating function (PGF) and transformation function (TF).

The Johnson system (1949) provides wide variety of distribution shapes as wide as that of Pearson system. In Edgeworth system, the normal distribution plays a role of PGF and the logarithm of income – the role of TF.

2. Generalizations of Tukey-lambda distributions

In most cases a continuous probability distribution is defined by means the cumulative distribution function (CDF) or probability density function (PDF); alternatively, it can be defined in terms of percentile functions. This is the case for

Tukey-lambda distribution we look for the distribution parameters on the basis of the percentile function which is the inverse of the cumulative distribution function.
A particular example of the percentile function is Tukey's λ function

$$R(p) = \frac{p^{\lambda} - (1-p)^{\lambda}}{\lambda}, \qquad 0 \le p \le 1, \quad \lambda \ne 0. \tag{3}$$

Ramberg and Schmeiser (1974) generalized Tukey's λ distribution to a four-parameter distribution defined by the percentile function

$$R(p) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, \qquad 0 \le p \le 1, \tag{4}$$

where λ_1 is a location parameter, λ_2 is a scale parameter, λ_3 is a skewness parameter and λ_4 is a kurtosis one.
The density function is defined by

$$f(x) = \frac{1}{\frac{dR(p)}{dp}} = \frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1}}, \qquad 0 \le p \le 1. \tag{5}$$

Some suggestions concerning GλD appeared in literature, namely that GλD is not suitable for description of the asymmetrical distributions. This is due to the fact that the shape parameter determines also skewness, so there should be rather three linear parameters determining position, scale, and skewness together with two parameters determining the shapes of the two tails. This suggests a natural generalization of the GλD to give a five-parameter lambda distribution (FPLD).

The five–parameter version of generalized lambda distribution was proposed by Gilchrist (2000). The quantile function is given by

$$Q(\lambda) = \lambda_1 + \frac{\lambda_2}{2}\left((1-\lambda_3)\frac{p^{\lambda_4} - 1}{\lambda_4} - (1+\lambda_3)\frac{(1-p)^{\lambda_5} - 1}{\lambda_5} \right)$$

(6)

where $0 \le p \le 1$, $\lambda_2 \ge 0$, $\lambda_3 \in \langle -1, 1 \rangle$.

3. Comparison of GλD and FPLD approximations for asymmetric distribution

Several methods for estimating the parameters of the GλD exist.

1) The oldest method, the so called moment-matching method was proposed by Ramberg and Schmeiser (1979). In this method we equate the mean, the variance, and the third and fourth moments of GλD to the corresponding mean μ^*, variance $(\sigma^*)^2$, skewness β_3^*, and kurtosis β_4^* of the sample and then compute the parameters λ from these four equations.
2) Öztürk and Dale (1985) proposed the least squares method. In this method one finds the values of λ for which the differences between the observed and predicted order statistics are as small as possible.
3) The starship method was proposed by King and MacGillivray (1999). First, we select a region in four-dimensional space that covers the range of the four parameters λ and cover it by a four-dimensional rectangular grid. Next we evaluate the grid points by performing a goodness-of-fit test on the corresponding distributions. If the test is satisfied, we stop procedure; otherwise we continue with the next point in the gird.

Lakhany and Maussner (2000) compared the results of the above mentioned estimation methods. It follows from their analysis that the method proposed by

Öztürk and Dale is one of the best. For that reason we use only this method. In more detail, the procedure based on the least squares method can be described as follows. Let x_i, $i=1,...,n$ denote the *i*th order statistic of data which is to be represented by the quantile function $R(u)$ and let U_i, $i=1,...,n$ denote the order statistic of the corresponding uniformly distributed random variable. The method consists in determining the values of λ for which the differences between the observed and predicted order statistics are as small as possible. So, in the case of GλD we must minimize the function

$$G(\lambda)=\sum_{i=1}^{n}\left(x_{(i)}-\lambda_1-\frac{Z_i}{\lambda_2}\right)^2 \tag{7}$$

where

$$\begin{aligned} Z_i &= \frac{1}{\lambda_3}\left(EU_{(i)}^{\lambda_3}-1\right)-\frac{1}{\lambda_4}\left(E\left(1-U_{(i)}\right)^{\lambda_4}-1\right)\\ EU_{(i)}^{\lambda_3} &= \frac{\Gamma(n+1)\Gamma(i+\lambda_3)}{\Gamma(i)\Gamma(n+\lambda_3+1)}\\ E\left(1-U_{(i)}\right)^{\lambda_4} &= \frac{\Gamma(n+1)\Gamma(n-i+\lambda_4+1)}{\Gamma(n-i+1)\Gamma(n+\lambda_4+1)} \end{aligned} \tag{8}$$

On the other hand, in the case of FPLD we minimize the function

$$G(\lambda)=\sum_{i=1}^{n}\left(x_{(i)}-Z_{(i)}\right)^2 \tag{9}$$

where

$$Z_{(i)}=\lambda_1+\frac{(1-\lambda_3)\lambda_2}{2\lambda_4}\left(\frac{\Gamma(n+1)\Gamma(i+\lambda_4)}{\Gamma(i)\Gamma(n+1+\lambda_4)}-1\right)+\frac{(1+\lambda_3)\lambda_2}{2\lambda_5}\left(1-\frac{\Gamma(n+1)\Gamma(n+1-i+\lambda_5)}{\Gamma(n+1-i)\Gamma(n+1+\lambda_5)}\right) \tag{10}$$

In both cases we minimalize functions with the help of means of Nelder-Mead method.

We have drawn random samples of size $n=400$ for the chosen asymmetric continuous distributions. We calculated the parameters λ_i of GλD and FPLD (Table 1) for every distribution. As a measure of quantitative closeness of two statistical distribution we used, following Taristano (2010), the Tchebycheff metric, which determine the maximum absolute value of the difference between the theoretical quantiles $R(p)$ and fitted quantiles $R(p,\lambda)$:

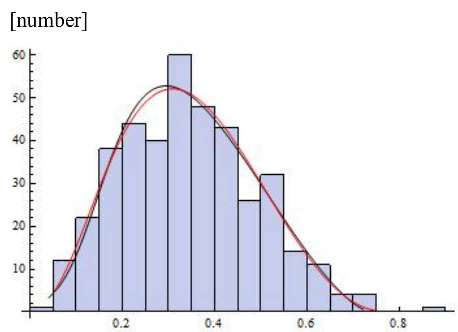
$$\max_{1\le i\le 39}\left|R(p_i)-R(p_i,\lambda)\right| \qquad \text{where } p_i=\frac{i}{40},\ i=1,2,...,39. \tag{11}$$

Comparing distances for four- and five-parameters functions, calculated on the basis of eq. (11), we always obtained smaller distance for FPLD. This means that the FPLD provides a better fit than GλD.

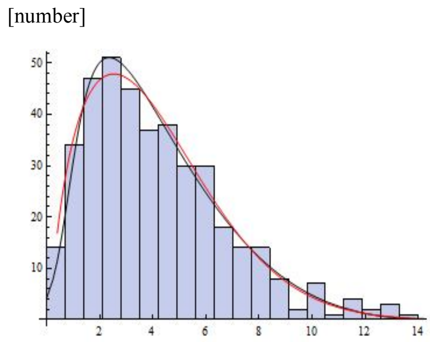
Table 1. Comparison of fit GλD i FPLD for asymmetric probabilistic distribution.

Model		λ_1	λ_2	λ_3	λ_4	λ_5	$\max_{1\le i\le 39}\lvert R(p_i)-R(p_i,\lambda)\rvert$
Beta[3,6]	GλD	0.25263	2.05295	0.10853	0.38998		0.0154
	FPLD	0.29944	0.23902	0.23451	0.29279	0.30638	0.0120
Chi-square[4]	GλD	1.16817	0.05299	0.01038	0.18126		0.0553
	FPLD	2.37635	3.99807	0.27658	0.71823	0.09292	0.0389
Gamma[2,2]	GλD	1.79032	0.07079	0.02577	0.23227		0.7068
	FPLD	2.9177	3.81181	0.26153	0.4982	0.13424	0.6090
Lognormal[0,1]	GλD	0.429	-0.21452	-0.02024	-0.1977		0.1349
	FPLD	2.62411	4.74602	-0.91609	1.89853	-0.5331	0.0517

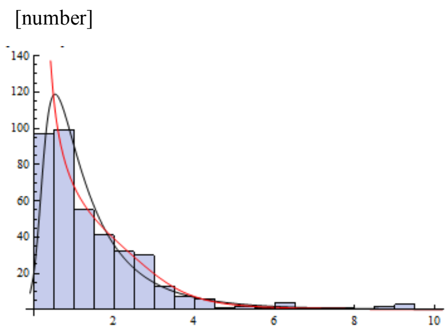
Source: own calculations.



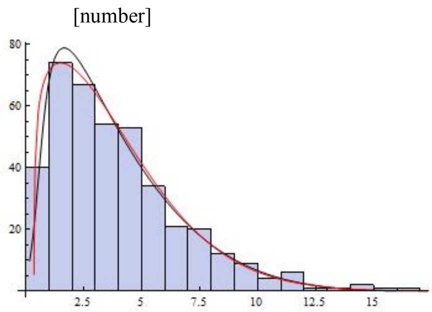
a) Beta[3,6]



b) Gamma[2,2]



c) Lognormal[0,1]



d) Chi-square[4]

Figure. 1. Histograms of random samples and estimated functions GλD (black curve) and FPLD (red curve).

4. Conclusion

The generalized Tukey distribution permits to obtain a wide variety of curves, the simplest examples being presented on Fig. 1. Due to the large flexibility of this distribution, the GλD finds many applications in the cases when the real distribution is not known.

References

Burr, I. W. (1973) " Parameters for a general system of distributions to match a grid of α_3 and α_4 ",. *Comm. Statist.*, 2,1-21

Chalabi Y., Scott D.J., Wuertz D. (2012) "An asymmetry-steepness parameterization of the generalized lambda distribution", http://mpra.ub.uni-muenchen.de/37814

Domański Cz., Pruska K. (2000) *Nieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa

Edgeworth F.Y. (1898) "On the Representation of Statistics by Mathematical Formule", *Journal of the Royal Statistical Society* 1, 670-700

D'Addaro R. (1949), Ricerche sulla curva dei redditi. Giornale degli Economisti e Annali di Economia, 8, 91–114

Gilchrist W.G. (2000) "Statistical Modelling with Quantile Functions", Chapman & Hall/CRC Press

Hahn G.J. and Shapiro, S. S. (1967). *Statistical Models in Engineering*. New York: John Wiley & Sons

Johnson N.L. (1949) "Systems of Frequency Curves Generated by Methods of Translation", *Biometrika* 44, s. 147-176

King R.A.R., MacGillivray H.L. (1999) "A starship estimation method for the generalized λ distributions", *Austral. & New Zealand J. Statist*. 41(3),353-374

Lakhany A., Mausser H. (2000) "Estimating the parameters of generalized lambda distribution", *Algo Research Quarterly*, vol. 3, No.3

Ozturk A., Dale R. (1985), "Least squares estimation of the parameters of the generalized lambda distribution". *Technometrics* 27(1), 81-84

Pearson K. (1895) "Contributions to the Mathematical Theory of Evolution", *II Skew Variation in Homogeneous Material Philosophical*, 186, w K. Pearson (1948), 41-112

Pearson K. (1948), *Karl Pearson's Early Statistical Papers*, Cambridge University Press

Ramberg J. S. and Schmeister B. W. (1974) „An approximate method for generating asymmetric random variables", *Canon. ACM*, 17, 78-82

Tarsitano A. (2010) "Comparing estimation methods for the FPLD", *Journal of Probability and Statistics*, Volume 2010, Article ID 295042

Tukey J. W. (1960), "The Practical Relationship Between the Common Transformations of Percentages of Counts and of Amounts" Technical Report 36, Statistical Techniques Research Group, Princeton University.