Some recent developments in probability distributions

Felix Famoye[*1], Carl Lee[1], and Ayman Alzaatreh[2]
[1]Dept. of Mathematics, Central Michigan University, Mount Pleasant, Michigan 48859
[2]Dept. of Mathematics & Statistics, Austin Peay State University, Clarksville, TN 37044
[*]Corresponding author: Felix Famoye, e-mail: felix.famoye@cmich.edu

Abstract

The paper by Eugene, Lee and Famoye in 2002 pioneered the class of beta generated probability distributions. Since then, many members of the class have appeared in the literature along with their various properties and applications. The class was extended to the Kumaraswamy generated class. A more general class, the class of *T*-*X* distributions, was recently developed. This talk will address some of the important properties and applications of these different classes. The present challenges with the estimation of distribution parameters and some open problems will be mentioned.

Key Words: beta-generated class, estimation, hazard function, simulation, *T*-*X* class

1. Introduction

The books by Johnson, et al. (1994, 1995) contain a lot of probability distributions and in particular, volume 1 of these books detailed the early development of probability distributions. The development started with the Pearson system of continuous distributions. Some specific cases were considered in the literature. Another development was the Burr system (Burr, 1942), which led to the twelve different Burr type distributions.

In order to include skewness in the normal distribution, Azzalini (1985) introduced the skew normal family of distributions. A lot of developments in this area and others were recently reviewed by Lee et al. (2013). There are other notable developments in distribution theory. This talk will now focus on the beta-generated class and its subsequent extensions.

Eugene et al. (2002) used the beta distribution with shape parameters α and β to develop the beta-generated distributions. The cumulative distribution function (cdf) of a beta-generated random variable *X* is defined as

$$G(x) = \int_0^{F(x)} b(t)dt\,, \tag{1}$$

where $b(t)$ is the probability density function (pdf) of the beta random variable and $F(x)$ is the cdf of any random variable *X*. The pdf, when *X* is continuous, for to the beta-generated distributions in (1) is given by

$$g(x) = \frac{1}{B(\alpha,\beta)} f(x)F^{\alpha-1}(x)(1-F(x))^{\beta-1}. \tag{2}$$

If *X* is discrete, the probability mass function is given by $g(x) = G(x) - G(x-1)$.

One important property of this class is that it is a generalization of the distribution of order statistic for the random variable *X* with cdf $F(x)$ as pointed out by Eugene et al. (2002) and Jones (2004). Since the paper by Eugene et al. (2002) in which the normal distribution was defined and studied, many papers have appeared in this class. These include beta-Fréchet (Nadarajah and Gupta, 2004), beta-Weibull (Famoye et al., 2005), beta-Pareto (Akinsete et al., 2008), β-Birnbaum-Saunders (Cordeiro and Lemonte, 2011), and beta-Cauchy (Alshawarbeh, et al., 2012).

Jones (2009) and Cordeiro and de Castro (2011) extended the beta-generated class by replacing the beta pdf in (1) with Kumaraswamy distribution, $b(t) = \alpha\beta x^{\alpha-1}(1-x^{\alpha})^{\beta-1}$,

$x \in (0, 1)$ (Kumaraswamy, 1980). The pdf of the Kumaraswamy-generated distributions (*KW-G*) is given by

$$g(x) = \alpha\beta f(x) F^{\alpha-1}(x)(1 - F^{\alpha}(x))^{\beta-1}. \tag{3}$$

Many generalized distributions from (3) have been studied in the literature including the Kumaraswamy Weibull distribution (Cordeiro et al., 2010) and the Kumaraswamy generalized gamma distribution (de Castro et al., 2011).

The basic idea for the beta-generated and the Kumaraswamy-generated distributions is that the generator beta distribution or the Kumaraswamy distribution has a support of (0, 1), which is the same for any cdf $F(x)$. Thus, the $b(t)$ in (1) needs to be a distribution with support (0, 1). Alzaatreh et al. (2013a) generalized both the beta-generated and the Kumaraswamy-generated distributions by considering any distribution as a generator.

In section 2, we define the *T-X* families and in section 3, we discuss an example of a *T-X* distribution. In section 4, we mention some applications and conclude with section 5.

2. The *T-X* families of distributions

Suppose $r(t)$ is the pdf of a random variable $T \in [a, b]$, for $-\infty \le a < b \le \infty$. Let $W(F(x))$ be a function of the cdf $F(x)$ of any random variable X so that $W(F(x))$ satisfies the following conditions:

$$\left.\begin{array}{l} W\left(F(x)\right) \in [a, b] \\ W\left(F(x)\right) \text{ is differentiable and monotonically non-decreasing} \\ W\left(F(x)\right) \to a \text{ as } x \to -\infty \text{ and } W\left(F(x)\right) \to b \text{ as } x \to \infty. \end{array}\right\} \tag{4}$$

The cdf of a new family of distributions is defined as

$$G(x) = \int_a^{W(F(x))} r(t)dt = R(W(F(x))), \tag{5}$$

where $W(F(x))$ satisfies the conditions in (4) and $R(t)$ is the cdf of the random variable T. The corresponding pdf associated with (5) is

$$g(x) = \left\{\frac{d}{dx} W\left(F(x)\right)\right\} r\left\{W\left(F(x)\right)\right\}. \tag{6}$$

Note that:

- The pdf $r(t)$ in (6) is "transformed" into a new cdf $G(x)$ through the function, $W(F(x))$, which acts as a "transformer". Hence, Alzaatreh et al. (2013a) referred to the distribution $g(x)$ in (6) as transformed from random variable T through the transformer random variable X and called it "Transformed-Transformer" or "*T-X*" distribution.
- The random variable X may be discrete and in such a case, $G(x)$ is the cdf of a family of discrete distributions.

Different $W(F(x))$ will give a new family of distributions. The definition of $W(F(x))$ depends on the support of the variable T. The following are some examples of $W(.)$.

1. When the support of T is bounded: Without loss of generality, we assume the support of T is (0, 1). Distributions for such T include uniform (0, 1), beta, Kumaraswamy and other types of generalized beta distributions. The distributions belong to the beta--generated class and they have been well studied. The function $W(F(x))$ can be defined as $F(x)$ or $F^{\alpha}(x)$.
2. When the support of T is $[a, \infty)$, $a \ge 0$: Without loss of generality, we assume $a = 0$. The function $W(F(x))$ can be defined as $-\ln(1 - F(x))$, $F(x)/(1 - F(x))$, $-\ln(1 - F^{\alpha}(x))$, or $F^{\alpha}(x)/(1 - F^{\alpha}(x))$, where $\alpha > 0$.

3. When the support of T is $(-\infty, \infty)$: Examples of $W(F(x))$ are $\ln[-\ln(1-F(x))]$, $\ln[F(x)/(1-F(x))]$, $\ln[-\ln(1-F^{\alpha}(x))]$, and $\ln[F^{\alpha}(x)/(1-F^{\alpha}(x))]$.

Alzaatreh et al. (2013a) used $W\left(F(x)\right)=-\log\left(1-F(x)\right)$ to study a T-X family defined as

$$G(x)=\int_0^{-\log(1-F(x))} r(t)dt = R\left\{-\log\left(1-F(x)\right)\right\}, \tag{8}$$

where $R(t)$ is the cdf of the random variable T. The corresponding pdf from (8) is

$$g(x)=\frac{f(x)}{1-F(x)}r\left(-\log\left(1-F(x)\right)\right)=h(x)\,r\left(-\log\left(1-F(x)\right)\right)=h(x)r\left(H(x)\right), \tag{9}$$

where $h(x)$ is the hazard function for the random variable X with the cdf $F(x)$ and $H(x)$ is the cumulative hazard function for X. Hence, the family of distributions can be considered as a family arising from a weighted hazard function.

Since $G(x)=R\left(-\log(1-F(x))\right)$, we have the following relationship between the random variables X and T:

$$X = F^{-1}(1-e^{-T}). \tag{10}$$

The result in (10) provides an easy method to simulate the random variable X and it can also be used to determine the moments of X. For example, $E(X)=E[F^{-1}(1-e^{-T})]$.

From (8), the quantile function $Q(\lambda)$, $0 < \lambda < 1$, for the T-X family of distributions is given by

$$Q(\lambda)=F^{-1}\left(1-e^{-R^{-1}(\lambda)}\right). \tag{11}$$

Theorem: If a random variable X follows the T-X pdf

$$g(x)=\frac{f(x)}{1-F(x)}r\left(-\log\left(1-F(x)\right)\right),$$

given in (9), then the Shannon entropy of X is given by

$$\eta_X = -E\left\{\log f\left(F^{-1}(1-e^{-T})\right)\right\}-\mu_T+\eta_T, \tag{12}$$

where μ_T and η_T are the mean and the Shannon entropy for the variable T with pdf $r(t)$.

Proof: See Alzaatreh et al. (2013a)

Alzaatreh et al. (2013a) defined the sub-families gamma-*X*, beta-exponentiated-*X* and Weibull-*X* distributions. Among the special cases of the gamma-*X* sub-family are the pdf of the n^{th} upper record value and the generalized gamma distribution defined and studied by Amoroso (1925). Among the special cases of beta-exponentiated-*X* sub-family are the beta-generated distributions, Kumaraswamy-generated distributions, the exponentiated-Weibull distribution defined by Mudholkar et al. (1995), the exponentiated-exponential distribution defined by Gupta and Kundu (2001), and the type I generalized logistic distribution in Johnson et al. (1995, p. 140). A member of the Weibull-*X* sub-family, the Weibull-Pareto distribution, is discussed in the next section.

3. Weibull-Pareto distribution

Alzaatreh et al. (2013b) defined and studied the Weibull-Pareto distribution. If a random variable T follows the Weibull distribution with parameters c and γ, then $r(t)=(c/\gamma)(t/\gamma)^{c-1}\exp[-(t/\gamma)^c],\ t\geq 0$. The Weibull-*X* family is given by

$$g(x)=\frac{c}{\gamma}\frac{f(x)}{1-F(x)}\left\{-[\log\left(1-F(x)\right)]/\gamma\right\}^{c-1}\exp\left\{-\left(-[\log\left(1-F(x)\right)]/\gamma\right)^{c}\right\}. \qquad (13)$$

If X is a Pareto random variable with pdf $f(x)=k\theta^{k}x^{-k-1},\ x>\theta$, then the Weibull-Pareto distribution is defined as

$$g(x)=(\beta c/x)\left\{\beta\log\left(x/\theta\right)\right\}^{c-1}\exp\left\{-\left(\beta\log\left(x/\theta\right)\right)^{c}\right\},\quad x>\theta,\ c,\beta,\theta>0, \qquad (14)$$

where $\beta=k/\gamma$. The corresponding cdf to the pdf in (14) is given by

$$G(x)=1-\exp\left\{-\left(\beta\log\left(x/\theta\right)\right)^{c}\right\}.$$

When θ = 1, the Weibull-Pareto distribution (WPD) reduces to the log-Weibull distribution defined by Sayana and Sekine (2004). When c = 1, the WPD reduces to the Pareto distribution with parameters β and θ.

Alzaatreh et al. (2013b) gave some properties of the WPD. The distribution is unimodal and it can be left or right skewed. The authors addressed two problems when using the maximum likelihood estimation (MLE) method for the WPD parameters. The first problem occurs when the parameter c < 1. The WPD likelihood function tends to infinity as θ goes to the sample minimum $x_{(1)}$ and hence, when c < 1 and θ is estimated by $x_{(1)}$, no MLE for c or β exists. The authors used an alternative MLE (AMLE) method proposed by Smith (1985). The second problem is when $c\gg 1$. For this situation, the WPD is left skewed with a long tail which makes $x_{(1)}$ a poor estimate for θ and this produces an unusually large bias in the AMLE for c and β. A modification of the regular MLE is proposed to deal with this large bias problem.

For AMLE, the log-likelihood function to maximize is $L_{*}=\Sigma_{x_i\neq x_{(1)}}\ln g(x_i;x_{(1)},\beta,c)$. Observe that θ has been replaced with $x_{(1)}$. The initial estimates for c and β are the moment estimates of Weibull distribution (Johnson et al., 1994, pp. 642-643). The sample minimum $x_{(1)}$ is not a good estimate of θ when the WPD shifts from being a right skewed distribution to being a left skewed distribution. In summary, the simulation results by Alzaatreh et al. (2013b) showed that the AMLE method did not provide good estimates when c > 1.

In view of the problem with AMLE, Alzaatreh et al. (2013b) considered a modified MLE (MMLE) proposed by Smith (1985) for c > 1. The log-likelihood function to be maximized in this case is $L_n(c,\beta,\theta)=\Sigma_{i=1}^{n}\ln g(x_i;\theta,\beta,c)$, which is defined only for $x_{(1)}$ > θ. On differentiating $L_n(c,\beta,\theta)$ with respect to the parameter θ, it is not difficult to show that the derivative is continuous when 0 < θ < $x_{(1)}$ and so the derivative exits. The MMLE method showed considerable improvement over the AMLE in terms of bias. In order to compare the performance of AMLE and MMLE when c > 1, Alzaatreh et al (2013b) computed and compared the mean square error (MSE). The results indicated that MMLE consistently had smaller MSE than AMLE when c > 1.

In practice, one should first obtain a graphical display of the data. If the data has a reversed J-shape, we recommend the use of AMLE for estimation. Otherwise, we suggest the use of MMLE method, since biases are reduced dramatically when compared with AMLE. Further research is needed for developing better estimation method for the *T-X* family of distributions.

4. Some applications

The beta-Pareto distribution (Akinsete et al. 2008) was applied to fit flood data. It outperformed the three-parameter Weibull distribution and the generalized Pareto distribution. The beta-Weibull distribution (Famoye et al., 2005) has a constant, a decreasing, an increasing, a bathtub, or a unimodal failure rate. The distribution was applied to fit various survival (complete or censored) data sets.

Alzaatreh et al. (2013b) fitted the WPD to three biological data sets and compared the fits with other known distributions. The results showed that the WPD gave a good fit to each data set and provided the best fit to the right and left tails. Alzaatreh et al. (2012) defined and studied the gamma-Pareto distribution. The distribution is unimodal and it is positively skewed. The distribution provided good fits to flood, reliability and biological data sets.

Cordeiro and Lemonte (2011) applied the beta-Birnbaum-Saunders distribution to three failure data sets: breaking stress of carbon fibres, number of successive failures for the air conditioning system in jet planes, and strengths of 1.5 cm glass fibres. The beta generalized Rayleigh distribution (Cordeiro et al. 2013) was found to provide good fits to life time data sets. de Castro et al. (2011) applied the Kumaraswamy-generated gamma distribution to right-censored life time data on 148 children contaminated with HIV.

Razzaghi (2009) used the beta-normal distribution in dose-response modeling and risk assessment. Razzaghi noted that the most important features of the beta-normal distribution is its generality and it encompasses a wide range of distributional shapes. Ortega et al. (2012) developed the negative binomial regression model to study the recurrence of prostate cancer and to predict the cure rate for patients.

5. Conclusion

Is the distribution theory still relevant? The answer to the question is emphatic YES based on the following reasons:

Quite a large number of researchers continue to work on distribution theory and many are finding out that the new distributions are providing better fit than the existing distributions. In parametric modeling, like Poisson regression or negative binomial regression, these models are based on probability distributions and inferences on them are based on distribution theory. Recent developments in the beta-generated distributions have been successfully applied in statistical modeling. Other distributions in the *T-X* families are yet to be explored.

References

Akinsete, A., Famoye, F. & Lee, C. (2008). The beta-Pareto distribution, *Statistics*, 42, 547-563.

Alshawarbeh, E., Lee, C. & Famoye, F. (2012). Beta-Cauchy distribution, *Journal of Probability and Statistical Science*, 10, 41-57.

Alzaatreh, A, Famoye, F. & Lee, C. (2013b). Weibull-Pareto distribution and its applications, To appear in *Communications in Statistics: Theory & Methods*.

Alzaatreh, A., Famoye, F. & Lee, C. (2012). Gamma-Pareto distribution and its applications, *Journal of Modern Applied Statistical Methods*, 11(1), 78-94.

Alzaatreh, A., Lee, C. & Famoye, F. (2013a). A new method for generating families of continuous distributions, To appear in *Metron*.

Amoroso, L. (1925). Ricerche intorno alla curva dei redditi, *Annali de Mathematica*, series 4, 2, 123-159.

Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, 12, 171-178.

Burr, I.W. (1942). Cumulative frequency functions, *Annals of Mathematical Statistics* 13, 215-232.

Cordeiro, G.M., Cristino, C.T., Hashimoto, E.M. & Ortega, E.M.M. (2013). The beta generalized Rayleigh distribution with applications to lifetime data, *Statistical Papers*, 54, 133-161.

Cordeiro, G.M. & de Castro, M. (2011). A new family of generalized distributions, *Journal of Statistical Computation and Simulation*, 81, 883-898.

Cordeiro, G.M. & Lemonte, A.J. (2011). The β-Birnbaum-Saunders distribution: An improved distribution for fatigue life modeling, *Computational Statistics and Data Analysis*, 55, 1445-1461.

Cordeiro, G.M., Ortega, E.M.M. & Nadarajah, S. (2010). The Kumaraswamy Weibull distribution with application to failure data, *Journal of the Franklin Institute*, 347, 1399-1429.

de Castro, M.A.R., Ortega, E.M.M. & Cordeiro, G.M. (2011). The Kumaraswamy generalized gamma distribution with application in survival analysis, *Statistical Methodology*, 8(5), 411-433.

Eugene, N., Lee, C. & Famoye, F. (2002). The 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.

Gupta, R.D. & Kundu, D. (2001). Exponentiated-exponential family: An alternative to gamma and Weibull distributions, *Biometrical Journal*, 43, 117-130.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd edition, John Wiley and Sons, Inc., New York.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Vol. 2, 2nd edition, John Wiley and Sons, Inc., New York.

Jones, M.C. (2004). Families of distributions arising from distributions of order statistics, *Test*, 13(1), 1-43.

Jones, M.C. (2009). Kumaraswamy’s distribution: A beta-type distribution with tractability advantages, *Statistical Methodology*, 6, 70-81.

Kumaraswamy, P. (1980). A generalized probability density functions for double-bounded random processes, *Journal of Hydrology*, 46, 79-88.

Lee, C., Famoye, F. and Alzaatreh, A. (2013). Methods for generating families of univariate continuous distributions in the recent decades, To appear in *WIREs Computational Statistics*.

Mudholkar, G.S., Srivastava, D.K. & Freimer, M. (1995). The exponentiated Weibull family: A reanalysis of the bus-motor-failure data, *Technometrics*, 37(4), 436-445.

Nadarajah, S. & Gupta, A.K. (2004). The beta Fréchet distribution, *Far East Journal of Theoretical Statistics*, 14, 15-24.

Ortega, E.M.M., Cordeiro, G.M. & Kattan, M.W. (2012). The negative binomial-beta Weibull regression model to predict the cure of prostate cancer, *Journal of Applied Statistics*, 39, 1191-1210.

Razzaghi, M. (2009). Beta-normal distribution in dose-response modeling and risk assessment for quantitative responses, *Environmental and Ecological Statistics*, 16, 25-36.

Sayama, S. & Sekine, M. (2004). Suppression of sea-ice clutter observed by a millimeter wave radar using a new log-Weibull/CFAR system, *International Journal of Infrared and Millimeter Waves*, 25, 1481-1494.

Smith, L.R. (1985). Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, 72(1), 67-90.