Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 231-240

Estimating the Area under the ROC Curve in the Framework of Lindley Centered Distributions

Balaswamy, S.[1] and Vishnu Vardhan, R.[2]
[1]*Department of Statistics, Indira Gandhi National Tribal University, Madhya Pradesh*
[2]*Department of Statistics, Pondicherry University, Puducherry*

Received: 23 May 2020; Revised: 22 December 2020; Accepted: 27 December 2020

Abstract

In the context of ROC curve analysis, the most widely used ROC form is the Binormal ROC curve. But due to the theoretical structures and distributional assumptions, many more bi-distributional ROC curve models have been proposed over the years. In this paper, an attempt has been made to overcome few limitations of ROC curve that emanated from exponential distribution. To address this, we have considered different forms of Lindley distributions and taking its mathematical advantages and mathematical flexibility, three new ROC curves are proposed. The proposed methodology is supported by simulation studies.

Key words: ROC curves; Lindley distribution; Power Lindley; Extended Lindley distributions and AUC.

1. Introduction

In statistical theory and practice, classification problems have gained lot of attention by many researchers in solving problems that are trivial as well as complex. Identification of class label is one of the major objectives in classification, for which several statistical techniques have been developed and proposed. Basing on the prominence and demand to handle such problems, those varieties of statistical tools have emerged and were brought under the hub of Statistical Decision Theory (SDT). The common problem of interest in classification is in allocating an individual or object to one of the predefined groups (or populations) by using a threshold. These problems were addressed by using a performance tool namely, Receiver Operating Characteristic (ROC) Curve, which evolved during World War II.

ROC Curve analysis was first presented to Psychologists by Tanner and Swets (1954), who brought out the concept from the Theory of Signal Detectability (TSD), which was introduced by Peterson *et al.* (1954) during World War II for analyzing the radar signals to detect enemy objects in battlefield *i.e*., identifying the signal as signal and noise as noise. Its expansion to other fields was prompt, for instance, in Psychology it was used to study the perceptual detection of stimuli (Swets, 1996). In medicine, one of the earliest applications was proposed by Lusted (1971), in which he postulated that to measure the worth of a diagnostic test, one must measure the performance of observers with the test and argued that ROC Curve provides an ideal means of studying observer performance.

Corresponding author: Vishnu Vardhan R
Email: vrstatsguru@gmail.com

Swets and Pickett (1982) noted two other key features of ROC Curves that make them ideal for studying diagnostic tests. First, the curves display all possible cut points and thus supply estimates of the frequency of various outcomes (*i.e.*, true positives, true negatives, false positives, and false negatives) at each cut point. Second, the curve allows the use of previous probabilities of condition, as well as calculations for the benefits of correct and incorrect decisions, to determine the best cut point for a given test in a given set up. Suppose the outcome S of a medical test is a measurement on a continuous scale (score), then there exists a threshold t of the test score, which can be used to classify subjects. For instance, a person with $S \leq t$ may be classified as healthy (normal or benign), otherwise as diseased (abnormal or malignant). Basing on the above classification, a 2 × 2 contingency table, namely the "*confusion matrix*" can be generated with four possible states, *viz*., True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Relatively few diagnostic tests correctly classify all subjects tested as diseased "D" (abnormal) or healthy "H" (normal). Sometimes, the threshold considered for classification, classifies few healthy ones as diseased and vice versa. This wrong classification leads to the terminology of False Positive Rate (FPR) and False Negative Rate (FNR). The probabilistic definitions pertaining to four possible states are given below:

i. The probability that an individual from D is correctly classified.
   True Positive Rate, TPR = $P(S > t \mid D)$ (*Sensitivity*)
ii. The probability that an individual from H is misclassified.
   False Positive Rate, FPR = $P(S > t \mid H)$ (1-*Specificity*)
iii. The probability that an individual from H is correctly classified.
   True Negative Rate, TNR = $P(S \leq t \mid H)$ (*Specificity*)
iv. The probability that an individual from D is misclassified.
   False Negative Rate, FNR = $P(S \leq t \mid D)$. (1-*Sensitivity*)

These four probabilities describe the performance of the test at this cutoff. It is important to note that all the intrinsic measures correspond to a given value t. As t changes these measures change. One of the problems of interest is to determine such t which optimizes one or more intrinsic measures, usually referred as *"optimal cutoff"*. With these probabilistic definitions of intrinsic measures, the ROC Curve can be defined as $ROC(t) = f(FPR(t), TPR(t))$. This means that the ROC Curve is generated by a set of pairs of FPR and TPR, which are obtained at every threshold point, that are actually observed test scores. So, each test score will act as a cutoff, which in turn generates the co-ordinates (FPR, TPR). ROC Curve is a tradeoff between FPR and TPR at every t. A test is said to be a *better* one, if it has maximum TPR and a *reasonably low* FPR.

An assumption in ROC curve is that the test scores of diseased populations will be greater than that of the healthy populations. For instance, if the Creatinine (one of the indicators of severe kidney impairment) levels are 5.0 or more in adults, then those adults are classified as risk group and the rest are non-risk group. Similar examples are HbA1c, LDL, Cholesterol, *etc.*, The bi-distributional ROC curves namely Bi-Normal, Bi-Exponential, Bi-Gamma *etc*., will fit to the above situation. But, these ROC forms do not fit to deal with situation where lower values of variables indicate risk and higher indicate non-risk group. For example, if the HDL is less than 40mg/dL (for men aged more than 20), then such individuals are considered as at risk and may be prone to cardiac issues and higher values relates to non-risk group. Few more variables that take similar phenomena as that of HDL are copper, iron, *etc*. Thus, it is important to address the problem of defining an ROC curve and also the corresponding intrinsic measures that can fit for the situation of lower test scores indicating risk group and higher scores indicate non-risk group.

With this background, we have considered Lindley (L) (Lindley, 1958; Ghitany *et al.*, 2008), Power Lindley (PL) (Ghitany *et al.*, 2013) and Extended Power Lindley (EPL) distributions (Said, 2015) to propose new ROC forms. The parameter combinations of these distributions have some interesting points that help us in constructing the ROC forms of the required nature. Another main reason to consider variant forms of Lindley is that it has a better fit than the exponential distribution (Ghitany *et al.*, 2008). It is also known that the power transformation and having additional parameter for basic Lindley form provides a lot of mathematical flexibility in explaining the shape and dispersion of heavy tail.

2. Family of Some Lindley Distributions

In this section, we start with the probability density functions and cumulative distribution function of the three distributions *i.e.*, Lindley, Power Lindley and Extended Power Lindley respectively.

2.1. Lindley distribution

$$f(x,\theta) = \frac{\theta^2}{\theta+1}(1+x)e^{-\theta x} \quad ;\ \theta, x > 0 \tag{1}$$

$$F(x) = 1 - \left(1 + \frac{\theta x}{\theta+1}\right)e^{-\theta x} \quad ;\ \theta, x > 0 \tag{2}$$

where $\theta \in (0,1)$, is a scale parameter.

2.2. Power Lindley distribution

$$f(x;\theta,\alpha) = \frac{\alpha\theta^2}{\theta+1}(1+x^\alpha)x^{\alpha-1}e^{-\theta x^\alpha} \quad ;\theta,\alpha,x > 0 \tag{3}$$

$$F(x;\theta,\alpha) = 1 - \left(1 + \frac{\theta}{\theta+1}x^\alpha\right)e^{-\theta x^\alpha} \quad ;\theta,\alpha,x > 0 \tag{4}$$

where θ is a scale parameter and α is a shape parameter. The purpose and reason to work on Power Lindley distribution is to overcome the theoretical and practical limitations of Lindley distribution. PL distribution is more flexible and this can be viewed as mixture of Weibull distribution due to the power transformation (shape α and scale θ), and a generalized gamma distribution (with shape parameters 2, α and scale θ), with mixing proportion $p = \theta/(\theta + 1)$ (Ghitany *et al*. 2013). For the values of α between 0 and 1, and with $\theta > 0$, we can have the increasing and decreasing nature of the density function.

2.3. Extended Power Lindley distribution

$$f(x;\theta,\beta,\alpha) = \frac{\alpha\theta^2}{\theta+\beta}(1+\beta x^\alpha)x^{\alpha-1}e^{-\theta x^\alpha} \quad ;\theta,\beta,\alpha,x > 0 \tag{5}$$

$$F(x;\theta,\beta,\alpha) = 1 - \left(1 + \frac{\theta\beta}{\theta+\beta}x^\alpha\right)e^{-\theta x^\alpha} \quad ;\theta,\beta,\alpha,x > 0 \tag{6}$$

EPL distribution can be shown as the mixture of Weibull distribution (with shape α and scale θ), and a generalized gamma distribution (with shape parameters 2, α and scale θ), with mixing proportion $p = \theta/(\theta + \beta)$.

3. Family of Three Lindley ROC Curves

In this section, we have developed a family of Lindley ROC Curves based on the considered Lindley distributions.

3.1. Bi-Lindley (L) ROC curve

It is assumed that the test scores (*S*, which is attributed as random variable) of normal/population I (denoted with "0") and abnormal/population II (denoted with "1") follow Lindley Distribution and the expression for the *FPR (1- specificity)* is defined as

$$FPR = x(t) = \left(1 + \frac{\theta_0}{\theta_0+1} t\right) e^{-\theta_0 t} \tag{7}$$

The threshold values can be obtained using the following formula

$$t = \left[\left(\frac{\theta_0+1}{\theta_0^2}\right) \log\left(\frac{1}{x(t)}\right)\right] \tag{8}$$

Here, an approximation of the type $\log(1+x) \approx x$ is used in driving the expression of "*t*", since our interest is only involved in the first order term in ROC form and the *TPR* (*sensitivity*) is obtained as

$$TPR = y(t) = \left(1 + \frac{\theta_1}{\theta_1+1} t\right) e^{-\theta_1 t} \tag{9}$$

on substituting the "*t*" value in above expression, the Lindley ROC Curve can be estimated as

$$y(t) = \left(1 + \frac{\theta_1}{\theta_1+1}\left[\left(\frac{\theta_0+1}{\theta_0^2}\right) \log\left(\frac{1}{x(t)}\right)\right]\right) e^{-\theta_1\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]} \tag{10}$$

Further, the Area under the Lindley ROC Curve can be estimated as

$$AUC = \int_0^1 \left(1 + \frac{\theta_1}{\theta_1+1}\left[\left(\frac{\theta_0+1}{\theta_0^2}\right) \log\left(\frac{1}{x(t)}\right)\right]\right) e^{-\theta_1\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]} dx(t) \tag{11}$$

on further simplification, the expression for AUC is

$$AUC = \frac{\theta_0^2}{\theta_0^2+\theta_1(\theta_0+1)}\left[\frac{\theta_1(\theta_0+1)(\theta_1+1)+\theta_0^2(\theta_1+1)+\theta_1(\theta_0+1)}{(\theta_1+1)\left(\theta_0^2+\theta_1(\theta_0+1)\right)}\right] \tag{12}$$

3.2. Bi-Power Lindley (PL) ROC curve

The *FPR* for the Power Lindley distribution can be derived as follows

$$FPR = x(t) = \left(1 + \frac{\theta_0}{\theta_0+1} t^{\alpha_0}\right) e^{-\theta_0 t^{\alpha_0}} \tag{13}$$

From the above expression, the threshold value can be found at each and every test score as

$$t = \left[\left(\frac{\theta_0+1}{\theta_0^2}\right) \log\left(\frac{1}{x(t)}\right)\right]^{\frac{1}{\alpha_0}} \tag{14}$$

Further, the expression for the *TPR* under Power Lindley distribution is derived as

$$TPR = y(t) = \left(1 + \frac{\theta_1}{\theta_1+1} t^{\alpha_1}\right) e^{-\theta_1 t^{\alpha_1}} \tag{15}$$

on substituting the expression for "*t*" in the above equation, the Power Lindley ROC Curve can be obtained as

$$y(t) = \left(1 + \frac{\theta_1}{\theta_1+1}\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right) e^{-\theta_1\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}} \tag{16}$$

Further, the Area under the Power Lindley ROC Curve can be estimated as follows

$$AUC = \int_0^1 \left(1 + \frac{\theta_1}{\theta_1+1}\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right) e^{-\theta_1\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}} dx(t) \tag{17}$$

The above expression does not have a closed form solution and has to be evaluated by numerical integration method.

3.3. Bi- Extended Power Lindley (EPL) ROC curve

ROC Curve based on the EPL distribution is constructed as follows. The False Positive Rate is given by

$$FPR = x(t) = \left(1 + \frac{\theta_0\beta_0}{\theta_0+\beta_0}t^{\alpha_0}\right)e^{-\theta_0 t^{\alpha_0}} \tag{18}$$

on further simplification, the expression for the threshold 't' is given by

$$t = \left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{1}{\alpha_0}} \tag{19}$$

The True Positive Rate is given by

$$TPR = y(t) = \left(1 + \frac{\theta_1\beta_1}{\theta_1+\beta_1}t^{\alpha_1}\right)e^{-\theta_1 t^{\alpha_1}} \tag{20}$$

The Extended Power Lindley ROC (EPLROC) Curve can be defined on substituting the expression for "t" in the above equation as follows.

$$y(t) = \left(1 + \frac{\theta_1\beta_1}{\theta_1+\beta_1}\left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right) e^{-\theta_1\left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}} \tag{21}$$

Further, the Area under the EPLROC Curve can be derived as follows,

$$AUC = \int_0^1 \left(1 + \frac{\theta_1\beta_1}{\theta_1+\beta_1}\left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right) e^{-\theta_1\left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}} dx(t) \tag{22}$$

The above expression does not have a closed form solution and has to be evaluated by numerical integration method.

4. Simulation Studies

The application of the proposed three new ROC forms is demonstrated using simulated data. For each ROC type, that is L, PL and EPL, the random numbers (RNs) are generated according to their distribution functionalities. With respect to Lindley distribution, the RNs are generated using quantile function. For Power Lindley and Extended Power Lindley forms, the RNs are generated using mixture of Weibull and Generalized Gamma distributions (Ghitany *et al.* (2013) and Said (2015)).

To demonstrate different forms of ROC curves (worst, moderate and best), simulations are carried out with different parameter combinations and the optimal threshold is deduced using Youden's index (J) for each of the combinations. In table 1, the results pertaining to all the three ROC curves are reported and the figures (ROC Curves) are depicted in Figure 1, which shows the comparison between the proposed ROC curves. The ROC curves are in the order of the parameter combination which is displayed as in Table 1.

Table 1: AUC and J values of L, PL and EPL ROC curves

θ_0	θ_1	AUC_L	J_L
0.6	0.5	0.4530	0.095
0.9	0.5	0.6265	0.2958
1	0.5	0.6675	0.3442
1.3	0.5	0.7567	0.4564
1.8	0.5	0.8396	0.5765

θ_0	θ_1	α_0	α_1	AUC_{PL}	J_{PL}
0.6	0.5	2	0.5	0.7486	0.6438
0.9	0.5	2	0.5	0.786	0.6784
1	0.5	2	0.5	0.795	0.6870
1.3	0.5	2	0.5	0.8148	0.7077
1.8	0.5	2	0.5	0.8357	0.7318

θ_0	θ_1	β_0	β_1	α_0	α_1	AUC_{EPL}	J_{EPL}
0.6	0.5	3	2.5	2	0.5	0.7487	0.7006
0.9	0.5	3.2	2.5	2	0.5	0.7945	0.7335
1	0.5	3.5	2.5	2	0.5	0.8016	0.7413
1.3	0.5	3.8	2.5	2	0.5	0.8243	0.7604
1.8	0.5	4	2.5	2	0.5	0.8504	0.7826

In Figure 1, the advantage of power transformation to Lindley and its extension by having additional parameter can be seen clearly. That is, the ROC curve of Lindley is very close to the chance line, which is not a preferable form for a better classification, whereas with additional shape parameter, the ROC forms of PL and EPL have shifted towards the top left corner of the unit square plot. In the context of ROC methodology, any test's or procedure's ROC curve should be far away from the chance line indicating that the test/procedure can classify the subjects with greater accuracy. Here, with these simulations, the advantage of having additional shape parameter has boosted the performance of a classifier witnessing a better ROC curve. However, with increase in the scale parameter values of population I, the ROC curve of Lindley distribution gradually shifted towards the top left corner of unit square plot. The gradual improvement in ROC curve of each L, PL and EPL ROC curves can be seen in Figure 2.

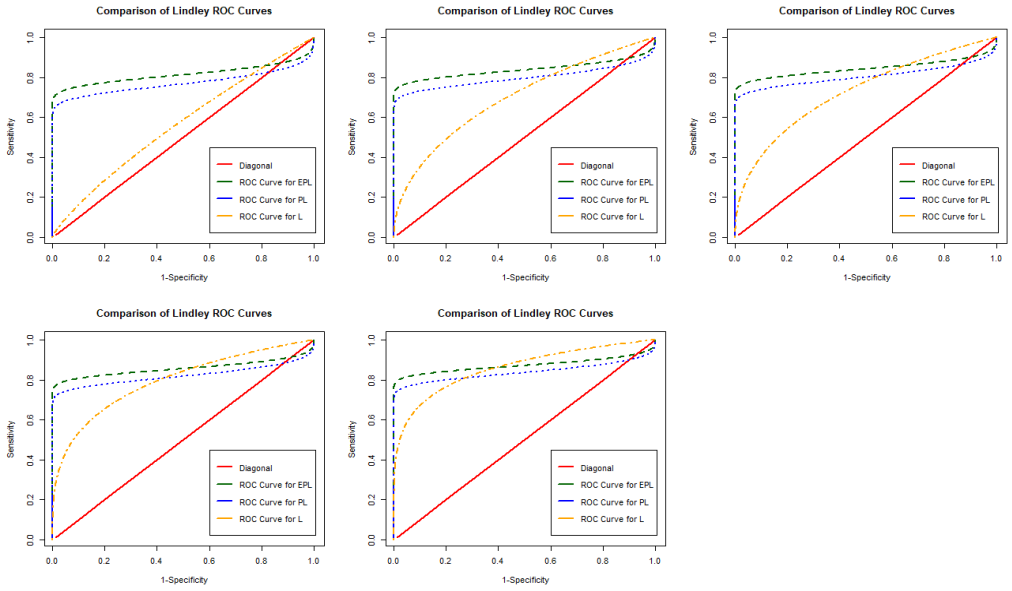

Figure 1: Graphical Comparison between L, PL and EPL ROC curves

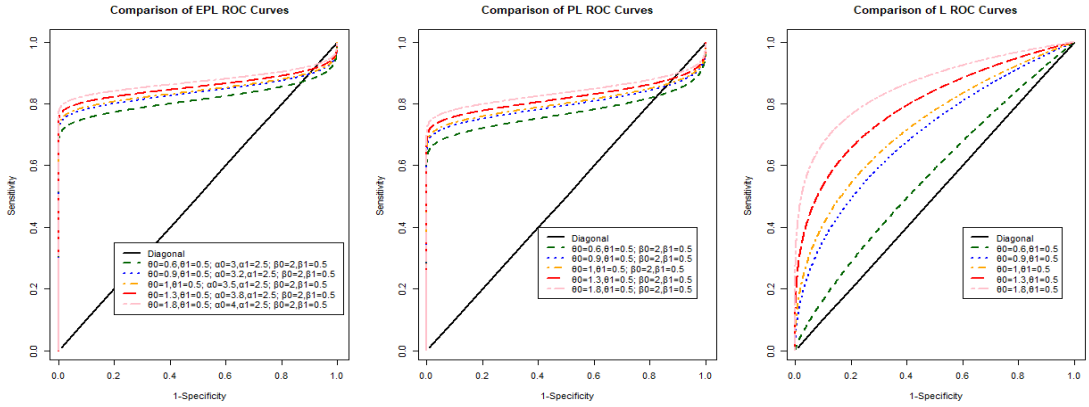

Figure 2: L, PL and EPL ROC curves with different parameter combinations

One more illustration is also carried out to address the question; "what happens to the nature of the ROC curve if the parameter combinations of population I are kept constant and varying parameter values in population II?". In the previous illustration, the simulations and ROC curves are exhibited for the case where the parameter values of population II are fixed. With the second illustration, it is observed that it affects the performance of the classifier and will not have impact in having a better accuracy (Table 2). With a large difference of scale and shape values between the populations I and II, some sort of improvement in ROC curves can be witnessed (Figures 3 and 4). This is due to the basic nature of the distribution forms that the values of population I should be at the higher side than that of population II, which is a very rare phenomenon in the general context of ROC methodology. Hence, the distributional forms of L, PL and EPL distributions has a very rare functionality of having higher values on population I (Normal or Healthy) than that of population II (Abnormal or Diseased), and these proposed ROC forms can be applied to such situations to explain the accuracy and other measures.

Table 2: AUC and J values of L, PL and EPL ROC curves

θ_0	θ_1	AUC_L	J_L
1.8	1.8	0.4766	0.0000
1.8	1.5	0.5342	0.0871
1.8	1.2	0.6052	0.1942
1.8	0.9	0.6929	0.3295
1.8	0.6	0.8001	0.5057

θ_0	θ_1	α_0	α_1	AUC_{PL}	J_{PL}
1.8	1.8	2	2	0.4766	0.0000
1.8	1.5	2	1.7	0.5201	0.0916
1.8	1.2	2	1.4	0.5858	0.2183
1.8	0.9	2	1.1	0.6788	0.3927
1.8	0.6	2	0.8	0.7974	0.6201

θ_0	θ_1	β_0	β_1	α_0	α_1	AUC_{EPL}	J_{EPL}
1.8	1.8	4	4	2	2	0.3615	0.0000
1.8	1.5	4	3.7	2	1.7	0.4167	0.1094
1.8	1.2	4	3.4	2	1.4	0.5052	0.2558
1.8	0.9	4	3.1	2	1.1	0.6350	0.4470
1.8	0.6	4	2.8	2	0.8	0.7924	0.6769

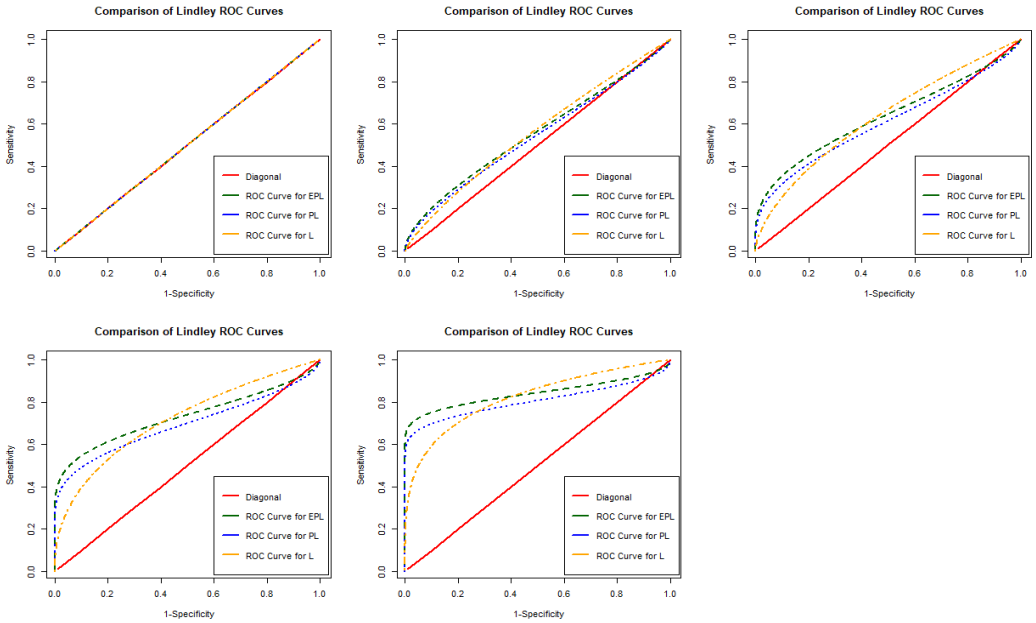

Figure 3: Graphical Comparison between L, PL and EPL ROC curves

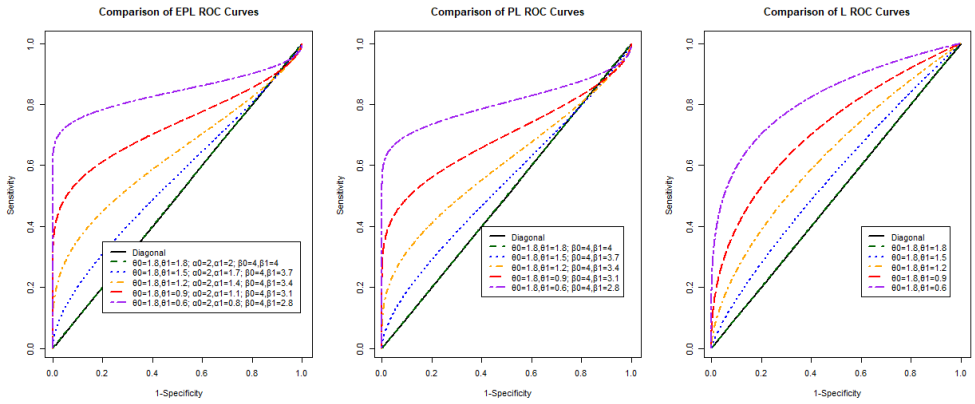

Figure 4: L, PL and EPL ROC curves with different parameter combinations

5. Conclusion

Different ROC curves have been studied in this paper by taking into consideration the three variant forms of Lindley distribution. The reason being certain mathematical and functional advantages such as superiority of Lindley over Exponential, the ease of Power transformation to the basic Lindley form and having additional shape parameter to the Power Lindley distribution. These considerations have been the support and motivation to propose three ROC curves namely, L, PL and EPL ROC forms. The advantages and flexibility of having power transformation and additional parameter is well demonstrated through simulation studies and also using graphical comparisons. Further, from the simulations and parameter combinations, an interesting fact that the ROC pattern and assumption of scores in population I and population II are in reverse pattern than that of the usual assumption made in several bi-distributional ROC forms such as Bi-Normal, Bi-Gamma and Bi-Exponential ROC curves *etc.*, was revealed. At most attention to the type of data is needed before fitting the proposed ROC Curves. The three L, PL and EPL ROC curves are quite applicable and apt to the practical contexts where the above said situation is witnessed.

References

Balaswamy. S., and Vishnu Vardhan. R. (2016). An anthology of parametric ROC models. *Research and Reviews: Journal of Statistics*, **5(2)**, 32-46.

Ghitany, M., Atieh, B., and Nadadrajah, S. (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation*, **78**, 493-506.

Ghitany, M., Al-Mutairi, D., Balakrishnan, N., and Al-Enezi, I. (2013). Power Lindley distribution and associated inference. *Computational Statistics and Data Analysis*, **64**, 20-33.

Krzanowski, W. J., and Hand, D. J. (2009). *ROC Curves for Continuous Data*. Monographs on Statistics and Applied Probability. New York, NY: CRC Press (ISBN: 978-1-4398-0021-8).

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society*, **B20**,102-107.

Lusted, L. B. (1971). Signal detectability and medical decision making, *Science*, **171**(**3977**), 1217-1219.

Said Hofan Alkarni (2015). Extended power Lindley distribution: A new statistical model for non-monotone survival data. *European Journal of Statistics and Probability*, **3**(**3**), 19-34.

Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Lawrence Erlbaum Associates, New Jersey (ISBN 978-1-1389-8191-1)

Swets, J. A., and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press (ISBN 978-0-12-679080-1).

Tanner Jr, W. P., and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, **61**(**6**), 401-409.