A Study of Exponential Mixture ROC Model

Sudesh Pundir[1], Azharuddin[2]

[1,2]*Department of Statistics, Pondicherry University, Puducherry, India*

Abstract: **In the case of heterogeneity in the data and heavily skewed distribution, we proposed Exponential-Mixture Receiver Operating Characteristic (ROC) model and discussed the properties of this ROC Curve. Area Under the Exponential-mixture ROC Curve (AUC) and its variance are also found. Estimates of the parameters of Exponential-mixture ROC model are derived by using the maximum Likelihood method and method of moments. The proposed model is validated by using the simulation studies**

Keywords: **Exponential-mixture distribution, ROC Curve, Area Under the ROC Curve, Estimates of the Area Under the ROC Curve, Variance of AUC and Monte Carlo simulation.**

1. INTRODUCTION

Mixture distribution arises where a statistical population contains two or more sub-populations. A random variable (X) is said to follow a finite mixture distribution if it has a probability density function of the form

$$f(x) = \sum_{i=1}^{k} p_i f_i(x/\theta_i),\ p_i > 0,\ i = 1,2,...k \quad and \quad \sum_{i=1}^{k} p_i = 1. \tag{1.1}$$

The parameters, $p_i's$ are mixing weights, and $f_i(x/\theta_i)$ is the i^{th} component density of the mixture.

The finite mixture distribution is first introduced by Newcomb in (1886) for outliers. Pearson (1894) also used the mixture distribution for estimating the parameters of two component Normal-mixture distribution by the method of moments. The research work on mixture distribution is growing very fast due to the advent of computers. The first monograph on finite mixture distribution is introduced by Everitt and Hand (1981). Titterington et al. (1985) discussed the finite mixture model in his book. McLachlan and Peel (2000) also discussed the finite mixture model. They discussed the estimation method of finite mixture model.

The mixture distribution have many applications in a large number of areas e.g. Econometrics, Sociology, Engineering, Reliability estimation, Remote sensing, Medical diagnosis and Discriminant analysis.

Schlatmann (2009) discussed the application of mixture model in medical diagnosis. In this paper, we discussed the mixture distribution in medical diagnosis.

Exponential distribution is the only continuous life distribution which has a constant failure rate. The Exponential distribution is also very famous for its memory-less property. Dass and Seong (2011) discussed the multivariate Normal-mixture ROC model. Gonen (2013) also discussed the p-component Normal-mixture ROC model.

He also discussed the Area under the Curve of p-component normal-mixture ROC curve. In this paper we discuss the mixture of two exponential density functions. Exponential distribution is a highly skewed distribution. It is also used in those cases where certain events occur with a constant probability per unit length. Everitt and Hand (1981) discussed the exponential mixture distribution in their comprehensive monograph. Titterington et al. (1985) also discussed the exponential mixture distribution in their book. The probability density function of the Exponential mixture distribution is as follows

$$f(x) = p\frac{1}{\lambda_1}e^{-\left(\frac{x}{\lambda_1}\right)} + (1-p)\frac{1}{\lambda_2}e^{-\left(\frac{x}{\lambda_2}\right)},\ 0 < \lambda_1, \lambda_2 < \infty,\ 0 \le x < \infty. \tag{1.2}$$

The cumulative distribution function of Exponential mixture distribution is as follows

$$F(x) = p\left[1 - e^{\left(-\frac{x}{\lambda_1}\right)}\right] + (1-p)\left[1 - e^{\left(-\frac{x}{\lambda_2}\right)}\right],\ 0 < \lambda_1, \lambda_2 < \infty,\ 0 \le x < \infty\ . \tag{1.3}$$

where p is the weight of the mixture distribution. The sum of the weights of distribution should be equal to 1. The Exponential mixture distribution is very useful in life testing problems.

Receiver Operating Characteristic (ROC) curve is used for the classification of the objects by using the cut-off values. Here Receiver Operating Characteristic curve is defined only for the continuous random variables. The ROC Curve is first introduced by Radar engineer and Electrical engineer 1950's during World War II in Signal detection theory. It is also used in the Psychophysics to classify human detection of weak signals. ROC curve have many applications in Medical as well as Non-Medical fields. ROC curve is also used in the diagnostic accuracy. It classifies the healthy and diseased populations.

The classification of the healthy and diseased population depends on the threshold or cut-off values. Suppose X is a random variable and t is the threshold value in a classification rule, so that an individual belongs to healthy population if classification scores (s) are greater than t otherwise an individual belong to disease population. Suppose that we take healthy population (H) and diseased population (D). There are four classification rates which gives the probability for understanding the discrimination of the patients. True Positive Rate (TPR) is defined as the probability that an individual from disease cases is correctly classified. True Negative Rate (TNR) is defined as the probability that an individual from healthy control is correctly classified. False Positive Rate (FPR) is defined as the probability that an individual from healthy control is misclassified and False Negative Rate (FNR) is defined as the probability that an individual from disease cases is misclassified. The mathematical definitions of TPR, FPR, TNR and FNR is as follows

TPR= $P\left(s > t / D\right)$,

FPR = $P\left(s > t / H\right)$,

TNR = $P\left(s \le t / H\right)$,

FNR = $P\left(s \le t / D\right)$.

Sensitivity and Specificity are statistical terms which are used to evaluate a diagnostic test. These are used to measure the performance of a binary classification test. They are independent of the population of interest subject to the test. Sensitivity is also known as True Positive Rate and Specificity is also known as True Negative Rate. In medical terminology, Sensitivity is defined as the proportion of patients who are suffering from disease and also the test is positive. Specificity is defined as the proportion of patients who are healthy and also the test is negative. Mathematically, Sensitivity and Specificity are defined as follows,

$$Sensitivity = y(t) = \frac{True\ Positive}{True\ Positive + False\ Negative},$$

$$Specificity = x(t) = \frac{True\ Negative}{True\ Negative + false\ Positive}.$$

For getting 100% classification of a test, we need high sensitivity and low 1-Specificity or high True Positive Rate (TPR) and low False Positive Rate (FPR). For the 100% classification, the coordinates of FPR and TPR should be 0 and 1.

ROC Curve is simply a curve between the False Positive Rate (FPR) and True Positive Rate (TPR). The discrimination of the ROC curve varies with the threshold or cut-off values. FPR is also known as 1-Specificity and TPR is also known as Sensitivity. The ROC curve is only a single Curve which tells the information in the cumulative distribution functions of the scores of the two classes. There are three approaches for estimating the ROC curve - Parametric, Non-Parametric and semi-parametric method. In this paper, we discuss only parametric method for estimating the ROC curve. Mathematical definition of ROC curve is given as

$$y\left(t\right) = 1 - G\left[\, F^{-1}\left(1 - x\left(t\right)\right)\right],\ 0 \le x(t) \le 1. \qquad (1.4)$$

where, F(.) and G(.) denote the cumulative distribution function of healthy observations and the cumulative distribution function of diseased observations.

Properties of ROC Curve

1. The test values of X are smaller than Y.
2. ROC curve is a monotonically increasing function.
3. ROC curve does not change if the classification scores undergo a strictly increasing transformation.
4. The slope of the ROC curve at a threshold value t is given as
$$slope = \frac{p(t)}{q(t)} \qquad 1.5$$
5. For the symmetric properties we use the Kullback-Leibler divergence that gives a critical values of the TPR and FPR. For checking the symmetric property, first we draw the negative chance diagonal about False positive Rate. The ROC curve is called as TPR asymmetric if the curve may adhere to the left edge of the ROC space longer than it does to the top. Similarly, the ROC curve is called as TNR asymmetric if the curve may adhere to the top edge of the ROC space longer than it does to the left edge of the ROC curve (Ref. Hughes and Bhattacharya (2013)). Suppose that X is a continuous random variable, we denote pdfs $f_1(x)$ for diseased cases and $f_2(x)$ for healthy controls. Then the Kullback-Leibler divergence is defined as

$$KL\left(f_1, f_2\right) = \int_D f_1(x) \ln\left[\frac{f_1(x)}{f_2(x)}\right] dx \qquad (1.6)$$

where $f_1(x)$ is the comparison distribution and $f_2(x)$ is the reference distribution. Similarly,

$$KL\left(f_2, f_1\right) = \int_D f_2(x) \ln\left[\frac{f_2(x)}{f_1(x)}\right] dx \qquad (1.7)$$

$f_2(x)$ is the comparison distribution and $f_1(x)$ as the reference distribution and D is the common support range of f_1 and f_2. Here, it should be noted that if $KL\left(f_1, f_2\right)$ and

$KL(f_2, f_1)$ are positive and $KL(f_1, f_2) = KL(f_2, f_1) = 0$, if and only if $f_1(x) = f_2(x)$. These two measure gives the asymmetry of ROC Curve with respect to diagonal. If $KL(f_1, f_2) < KL(f_2, f_1)$, then the ROC Curve is said to be True Positive Rate asymmetric and if $KL(f_1, f_2) > KL(f_2, f_1)$, then the ROC Curve is said to be True Negative Rate asymmetric.

In this paper, there are five sections. In first section, we discussed the review of literature of Mixture distributions, Exponential mixture distribution and ROC curve. In section second, we discussed Exponential mixture ROC model and the properties of the Exponential mixture ROC curve to know the behavior of the ROC curve. In third section, we discussed the Area Under the Curve of Exponential mixture ROC curve to measure its accuracy. The estimates of parameters of exponential mixture ROC model are derived from maximum likelihood method and method of moments. In section four, we derived variance of Area Under the Curve for testing the hypothesis. In the last section, we discussed the simulation studies by taking a numerical example for the validation of the model.

2. EXPONENTIAL MIXTURE ROC MODEL

Suppose that X follows exponential mixture distribution which is coming from healthy controls and Y also follows Exponential mixture distribution which is coming from diseased cases. The probability distribution function of healthy control and diseased cases are

$$f(x) = p\frac{1}{\lambda_1}e^{-\left(\frac{x}{\lambda_1}\right)} + (1-p)e^{-\left(\frac{x}{\lambda_2}\right)}, 0<\lambda_1,\lambda_2<\infty, 0<x\le\infty, p>0, \quad (2.1)$$

$$g(y) = p\frac{1}{\lambda_1}e^{\left(-\frac{y}{\lambda_1}\right)} + (1-p)\frac{1}{\lambda_2}e^{\left(-\frac{y}{\lambda_2}\right)}, 0<\lambda_1,\lambda_2<\infty, 0<y\le\infty, p>0. \quad (2.2)$$

The cumulative distribution function of healthy control and diseased cases are given as

$$F(x) = p\left[1 - e^{\left(-\frac{x}{\lambda_1}\right)}\right] + (1-p)\left[1 - e^{\left(-\frac{x}{\lambda_2}\right)}\right], \quad (2.3)$$

$$G(y) = p\left[1 - e^{\left(-\frac{y}{\lambda_1}\right)}\right] + (1-p)\left[1 - e^{\left(-\frac{y}{\lambda_2}\right)}\right]. \quad (2.4)$$

The ROC model of two component exponential mixture distribution is as follows

$$y(t) = p\left[x(t)\right]^{\frac{\lambda_{10}}{\lambda_{11}}} + (1-p)\left[x(t)\right]^{\frac{\lambda_{20}}{\lambda_{21}}}, 0 \le x(t) \le 1. \quad (2.5)$$

$$x(t) = p\,e^{\left(-\frac{t}{\lambda_{10}}\right)} + (1-p)\,e^{\left(-\frac{t}{\lambda_{20}}\right)},$$

and

$p>0$, $0 < t < \infty$, $0 < \lambda_{ij} < \infty$ and i=1,2 , j=1,0.

where p is the weight of the mixture distribution and t is the cut-off or threshold value. The first subscript in λ_{ij} shows that density function is from first or second distribution and second subscript shows that it is coming from either healthy controls or diseased cases. Here, second subscript 0 means it is coming from healthy controls and 1 means it is coming from diseased cases.

The Exponential mixture ROC curve should follow the following assumptions.

1. The mean of the diseased distribution should be greater than the mean of the healthy distribution i.e. $\mu_1 > \mu_0$

2. Attach more weight to the population with higher mean.

The Exponential Mixture ROC curve satisfies the following properties:

1. The ROC curve is monotonically increasing function.

Proof: A function is said to be a monotone increasing function, if the first derivative of the function is positive. Differentiating (2.1) with respect to x(t), we get

$$\frac{dy(t)}{dx(t)} = p\left(\frac{\lambda_{10}}{\lambda_{11}}\right)\left[x(t)\right]^{\frac{\lambda_{10}}{\lambda_{11}}-1} + (1-p)\left(\frac{\lambda_{20}}{\lambda_{21}}\right)\left[x(t)\right]^{\frac{\lambda_{20}}{\lambda_{21}}-1} > 0. \quad 2.6$$

As, the first derivative is greater than zero, hence the ROC curve is monotonically increasing function.

2. ROC curve is concave function.

Proof: A function is said to be concave if its second derivative is negative. Differentiating (2.3) with respect to x(t), we get

$$\frac{d^2y(t)}{d\{x(t)\}^2} = p\left(\frac{\lambda_{10}}{\lambda_{11}}\right)\left(\frac{\lambda_{10}}{\lambda_{11}}-1\right)\{x(t)\}^{\frac{\lambda_{10}}{\lambda_{11}}-2} + (1-p)\left(\frac{\lambda_{20}}{\lambda_{21}}\right)\left(\frac{\lambda_{20}}{\lambda_{21}}-1\right)\{x(t)\}^{\frac{\lambda_{20}}{\lambda_{21}}-2} < 0 \quad 2.7$$

Hence, ROC curve is concave in nature.

3. The slope of the ROC curve at the threshold t is given by

$$slope = \frac{f(y/t)}{f(x/t)}$$

$$= \frac{\lambda_{10}\lambda_{20}\left[p\lambda_{21}e^{\left(-\frac{t}{\lambda_{11}}\right)} + (1-p)\lambda_{11}e^{\left(-\frac{t}{\lambda_{21}}\right)}\right]}{\lambda_{21}\lambda_{11}\left[p\lambda_{20}e^{\left(-\frac{t}{\lambda_{10}}\right)} + (1-p)\lambda_{10}e^{\left(-\frac{t}{\lambda_{20}}\right)}\right]}. \qquad (2.8)$$

The slope of the ROC curve also gives the Likelihood-Ratio which is useful in hypothesis testing.

4. ROC curve is invariant with respect to monotone increasing transformation of the test scores.

5. The ROC curve is TPR asymmetric.

Proof: The K-L divergence between the distribution of diseased and healthy group with p(x) as the comparison distribution and q(x) as the reference distribution has been given as

$$KL(p,q) = \left[\ln\lambda_{10} + \ln\lambda_{20} + \ln\left\{p\lambda_{21}e^{\left(\frac{t}{\lambda_{11}}\right)} + (1-p)\lambda_{11}e^{\left(\frac{t}{\lambda_{21}}\right)}\right\} - \ln\lambda_{11} - \ln\lambda_{21} - \ln\left\{\begin{matrix} p\lambda_{20}e^{\left(\frac{t}{\lambda_{10}}\right)} + \\ (1-p)\lambda_{10}e^{\left(\frac{t}{\lambda_{20}}\right)}\end{matrix}\right\}\right] * \left[\frac{p}{\lambda_{11}}e^{\left(\frac{t}{\lambda_{11}}\right)} + \frac{(1-p)}{\lambda_{21}}e^{\left(\frac{t}{\lambda_{21}}\right)}\right] \qquad (2.9)$$

Similarly, the K-L divergence between the distribution of healthy and diseased group with q(x) as the comparison distribution and p(x) as the reference distribution has been given as

$$KL(q,p) = \left[\ln\lambda_{21} + \ln\lambda_{11} + \ln\left\{p\lambda_{20}e^{\left(\frac{t}{\lambda_{10}}\right)} + (1-p)\lambda_{10}e^{\left(\frac{t}{\lambda_{20}}\right)}\right\} - \ln\lambda_{10} - \ln\lambda_{20} - \ln\left\{\begin{matrix} p\lambda_{21}e^{\left(\frac{t}{\lambda_{11}}\right)} \\ +(1-p)\lambda_{11}e^{\left(\frac{t}{\lambda_{21}}\right)}\end{matrix}\right\}\right] * \left[\frac{p}{\lambda_{10}}e^{\left(\frac{t}{\lambda_{10}}\right)} + \frac{(1-p)}{\lambda_{20}}e^{\left(\frac{t}{\lambda_{20}}\right)}\right] \qquad (2.10)$$

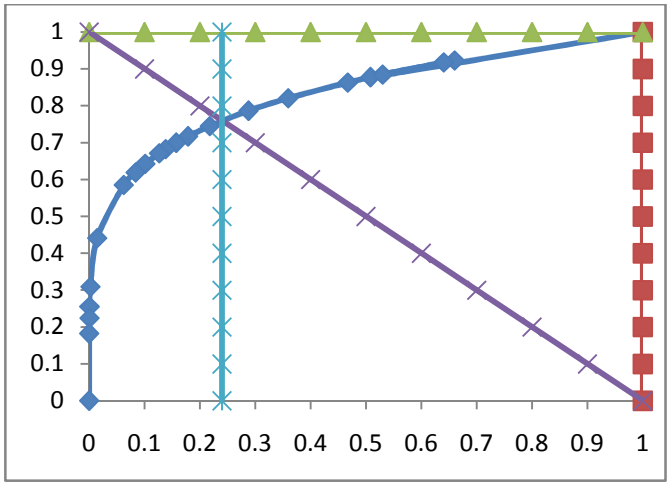

Fig. 2.1: TPR asymmetric Exponential Mixture ROC curve

It was found that KL(q,p) > KL(p,q). These two divergence measures would be zero if the healthy and diseased groups are identical. Hence, we have proved that, the ROC curve is TPR asymmetric. Figure 2.1 shows the asymmetric property of Exponential mixture ROC curve.

3. AREA UNDER THE EXPONENTIAL MIXTURE ROC CURVE

The measure of accuracy of the discrimination between the healthy and diseased distribution is Area Under the ROC curve.The Area Under the ROC curve gives the best measure of accuracy of the classification. Mathematically the Area Under the ROC curve is given as

$$AUC = P(y > x) = \int_0^1 y(t)dt \qquad (3.1)$$

The Area Under the ROC curve lies between 0 and 1. For a perfect diagnostic test, the Area under the ROC curve should be 1 and the curve should lie on the upper left border of the ROC space. For a worthless test, the Area Under the ROC curve will be equal or less than 0.50. Area under the exponential mixture ROC curve is given as

$$AUC = p\left(\frac{\lambda_{11}}{\lambda_{11}+\lambda_{10}}\right) + (1-p)\left(\frac{\lambda_{21}}{\lambda_{21}+\lambda_{20}}\right) \qquad (3.2)$$

The estimates of the parameters of ROC model can be found by using two methods which are discussed below.

(a) Maximum likelihood method

The probability distribution function of exponential mixture distribution which is coming from healthy control is as follows

$$f(x) = pf(x;\lambda_{10}) + (1-p)f(x;\lambda_{20})$$

where

$$f(x;\lambda_{10}) = \frac{1}{\lambda_{10}}e^{-\left(\frac{x}{\lambda_{10}}\right)} \quad and \quad f(x;\lambda_{20}) = \frac{1}{\lambda_{20}}e^{-\left(\frac{x}{\lambda_{20}}\right)}.$$

For estimating the parameter λ_{10}, the likelihood function is given as

$$L(x_{10}, x_{20}, \ldots, x_{m0} / \lambda_{10}) = \left(\frac{1}{\lambda_{10}}\right)^{m_{10}} e^{-\sum_{i=1}^{m_{10}}\frac{x_i}{\lambda_{10}}}.$$

The log-likelihood of the above equation is

$$\log L(\lambda_{10} / x_{10}, x_{20}, \ldots, x_{m0}) = m_{10}\log 1 - m_{10}\log\lambda_{10} - \frac{1}{\lambda_{10}}\sum_{i=1}^{m_{10}} x_i.$$

Differentiating above equation with respect to λ_{10} and equating it to zero, we get

$$\hat{\lambda}_{10} = \frac{1}{m_{10}} \sum_{i=1}^{m_{10}} x_i . \tag{3.3}$$

Similarly, we get

$$\hat{\lambda}_{20} = \frac{1}{m_{20}} \sum_{i=1}^{m_{20}} x_i . \tag{3.4}$$

The probability distribution function of Exponential mixture distribution which is coming from diseased cases is written as

$$g(y) = p\, g(y; \lambda_{11}) + (1 - p)\, g(y; \lambda_{21}).$$

On the similar lines as done above, the maximum likelihood estimate of $\hat{\lambda}_{11}$ and $\hat{\lambda}_{21}$ are given as

$$\hat{\lambda}_{11} = \frac{1}{n_{11}} \sum_{i=1}^{n_{11}} y_i , \tag{3.5}$$

$$\hat{\lambda}_{21} = \frac{1}{n_{21}} \sum_{i=1}^{n_{21}} y_i . \tag{3.6}$$

Substituting all the estimates from (3.3)-(3.6) in the Area Under the Curve of Exponential mixture ROC curve (3.2), the Estimated AUC is given as

$$\hat{AUC} = p\left(\frac{\frac{1}{n_{11}}\sum_{i=1}^{n_{11}} y_i}{\frac{1}{n_{11}}\sum_{i=1}^{n_{11}} y_i + \frac{1}{m_{10}}\sum_{i=1}^{m_{10}} x_i}\right) + (1-p)\left(\frac{\frac{1}{n_{21}}\sum_{i=1}^{n_{21}} y_i}{\frac{1}{n_{21}}\sum_{i=1}^{n_{21}} y_i + \frac{1}{m_{20}}\sum_{i=1}^{m_{20}} x_i}\right). \tag{3.7}$$

(b) Method of moments

Method of moment is another method for estimating the population parameters. It is a simple method for estimating the parameters. The method of moment is introduced by Pearson (1894) for estimating the parameters of two component normal-mixture distribution. Rider (1961) discussed the method of moment for estimating the parameter of exponential mixture distribution. In the method of moments, first we derive the equation which is based on population moments and these population moments are estimated by drawing the sample observation. These equations are solved by using the sample moments in place of population moment. The rth moment about origin of the exponential mixture distribution is given as

$$\mu_r' = p\int_0^\infty x^r f_1(x)dx + (1-p)\int_0^\infty x^r f_2(x)dx \tag{3.8}$$

where $f_1(x)$ and $f_2(x)$ are the density functions of the Exponential distribution which have been discussed earlier. Now (3.8) can be written as

$$\mu_r' = p\int_0^\infty x^r \frac{1}{\lambda_1} e^{-\left(\frac{x}{\lambda_1}\right)} dx + (1-p)\int_0^\infty x^r \frac{1}{\lambda_2} e^{-\left(\frac{x}{\lambda_2}\right)} dx$$

$$\mu_r' = p\Gamma(r+1)(\lambda_1)^r + (1-p)\Gamma(r+1)(\lambda_2)^r \quad for\ \ r = 1, 2, 3. \tag{3.9}$$

The corresponding sample moment is defined as $m_r' = p\Gamma(r+1)(\lambda_1)^r + (1-p)\Gamma(r+1)(\lambda_2)^r$, $for\ \ r = 1, 2, 3.$ (3.10)

The first three sample moments are given as follows

$$m_1' = p\lambda_1 + (1-p)\lambda_2 , \tag{3.11}$$

$$m_2' = 2p\lambda_1^2 + 2(1-p)\lambda_2^2 , \tag{3.12}$$

$$m_3' = 6p\lambda_1^3 + 6(1-p)\lambda_2^3 . \tag{3.13}$$

Using these equations, we can estimate the parameters λ_1 and λ_2 in terms of sample moments as

$$\hat{\lambda}_1 = \frac{2m_1' - 2(1-p)m_1' \pm \sqrt{-4p(1-p)m_1'^2 + 2p(1-p)m_2'}}{2p} \tag{3.14}$$

$$\hat{\lambda}_2 = \frac{2(1-p)m_1' \pm \sqrt{-4p(1-p)m_1'^2 + 2p(1-p)m_2'}}{2(1-p)} \tag{3.15}$$

4. VARIANCE OF AUC AND CONFIDENCE INTERVAL OF AUC

For finding the variance of AUC of Exponential mixture ROC curve, we will use the delta method. It gives the approximate expression of variance of AUC and it is based on Taylor series expansion.

$$\text{var}\,iance(AUC) = p\,\text{var}\,ince\left(\frac{\lambda_{11}}{\lambda_{11}+\lambda_{10}}\right) + (1-p)\,\text{var}\,iance\left(\frac{\lambda_{21}}{\lambda_{21}+\lambda_{20}}\right) \tag{4.1}$$

$$\text{var}iance(AUC) = p\,\text{var}iance(\delta_1) + (1-p)\,\text{var}iance(\delta_2) \quad (4.2)$$

where, $\delta_1 = \frac{\lambda_{11}}{\lambda_{11} + \lambda_{10}}$ and $\delta_2 = \frac{\lambda_{21}}{\lambda_{21} + \lambda_{20}}$.

First, we find the variance of δ_1 and variance of δ_2 as

$$\text{var}iance(\delta_1) = \left(\frac{\partial \delta_1}{\partial \lambda_{11}}\right)^2 \text{var}iance(\lambda_{11}) + \left(\frac{\partial \delta_{10}}{\partial \lambda_{10}}\right)^2 \text{var}iance(\lambda_{10}) + 2\left(\frac{\partial \delta_1}{\partial \lambda_{11}}\right)\left(\frac{\partial \delta}{\partial \lambda_{10}}\right)\text{cov}ariance(\lambda_{11}, \lambda_{10}) \quad (4.3)$$

$$\text{var}iance(\delta_2) = \left(\frac{\partial \delta_2}{\partial \lambda_{21}}\right)^2 \text{var}iance(\lambda_{21}) + \left(\frac{\partial \delta_2}{\partial \lambda_{20}}\right)^2 \text{var}iance(\lambda_{20}) + 2\left(\frac{\partial \delta_2}{\partial \lambda_{21}}\right)\left(\frac{\partial \delta_2}{\partial \lambda_{20}}\right)\text{cov}ariance(\lambda_{21}, \lambda_{20}) \quad (4.4)$$

For finding the variance of λ_{11}, variance of λ_{10} and also covariance between λ_{11} and λ_{10}, we will use fisher information matrix. The fisher information matrix is given as

$$I(\theta) = -\begin{bmatrix} E\left(\frac{\partial^2 LogL}{\partial \lambda_{11}^2}\right) & E\left(\frac{\partial^2 LogL}{\partial \lambda_{11}\partial \lambda_{10}}\right) \\ E\left(\frac{\partial^2 LogL}{\partial \lambda_{10}\partial \lambda_{11}}\right) & E\left(\frac{\partial^2 LogL}{\partial \lambda_{10}^2}\right) \end{bmatrix}$$

where

$$L = \left(\frac{1}{\lambda_{11}}\right)^{n_{11}} e^{-\sum_{i=1}^{n_{11}}\left(\frac{y_i}{\lambda_{11}}\right)}.$$

Taking log on both sides, we get

$$\log L = n_{11}\log 1 - n_{11}\log \lambda_{11} - \frac{1}{\lambda_{11}}\sum_{i=1}^{n_{11}} y_i. \quad (4.5)$$

Differentiating (4.5), twice with respect to λ_{11}, we get

$$\frac{\partial^2 LogL}{\partial \lambda_{11}^2} = -\frac{n_{11}}{\lambda_{11}^2}.$$

Similarly

$$\frac{\partial^2 LogL}{\partial \lambda_{10}^2} = -\frac{m_{10}}{\lambda_{10}^2}.$$

The covariance terms are zero due to independent identically distributed random variables. Substituting these values in Fisher information matrix, we get

$$I(\theta) = -\begin{bmatrix} E\left(-\frac{n_{11}}{\lambda_{11}^2}\right) & 0 \\ 0 & E\left(-\frac{m_{10}}{\lambda_{10}^2}\right) \end{bmatrix}.$$

The diagonal terms of the inverse of the Fisher information matrix gives

$$variance(\lambda_{11}) = \frac{\lambda_{11}^2}{n_{11}} \text{ and } \text{var}iance(\lambda_{10}) = \frac{\lambda_{10}^2}{m_{10}}.$$

We know that $\delta_1 = \frac{\lambda_{11}}{\lambda_{11} + \lambda_{10}}$. Now differentiating δ_1 with respect to λ_{11} and λ_{10}, we get

$$\frac{\partial \delta_1}{\partial \lambda_{11}} = \frac{\lambda_{10}}{(\lambda_{10} + \lambda_{11})^2} \text{ and } \frac{\partial \delta_1}{\partial \lambda_{10}} = -\frac{\lambda_{11}}{(\lambda_{10} + \lambda_{11})^2}.$$

Substituting all these values in (4.3), we get

$$\text{var}iance(\delta_1) = \left[\frac{\lambda_{10}}{(\lambda_{10}+\lambda_{11})^2}\right]^2 \frac{\lambda_{11}^2}{n_{11}} + \left[-\frac{\lambda_{11}}{(\lambda_{10}+\lambda_{11})^2}\right]^2 \frac{\lambda_{10}^2}{m_{10}}$$

and

$$\text{var}iance(\delta_1) = \frac{\lambda_{11}^2\lambda_{10}^2}{(\lambda_{10} + \lambda_{11})^2}\left[\frac{m_{10} + n_{11}}{m_{10}n_{11}}\right]. \quad (4.6)$$

Similarly, we get

$$\text{var}iance(\delta_2) = \frac{\lambda_{21}^2\lambda_{20}^2}{(\lambda_{21} + \lambda_{20})^2}\left[\frac{m_{20} + n_{21}}{m_{20}n_{21}}\right]. \quad (4.7)$$

Putting the values of $\text{var}iance(\delta_1)$ and $\text{var}iance(\delta_2)$ in (4.2), we get

$$\text{var}iance(AUC) = p\left[\frac{\lambda_{11}^2\lambda_{10}^2}{(\lambda_{11}+\lambda_{10})^2}\left(\frac{m_{10}+n_{11}}{m_{10}n_{11}}\right)\right] + (1-p)\left[\frac{\lambda_{21}^2\lambda_{20}^2}{(\lambda_{21}+\lambda_{20})^2}\left(\frac{m_{20}+n_{21}}{m_{20}n_{21}}\right)\right]. \quad (4.8)$$

For testing the hypothesis,

$$H_0 : AUC = AUC_0 \text{ vs } H_1 : AUC \neq AUC_0,$$

the test statistic of AUC is given as

$$Z = \frac{AUC - \hat{AUC}}{\sqrt{V\left(\hat{AUC}\right)}} \sim N(0,1). \quad (4.9)$$

The 100(1-α)% confidence interval of AUC is given as

$$AUC \pm SE\left(\hat{AUC}\right)Z_{\alpha/2} \quad (4.10)$$

where α is the level of significance and $Z_{\alpha/2}$ is the critical value of the confidence interval.~

5. SIMULATION STUDIES

In this paper, we use Monte Carlo Simulation technique for validation of our results by taking an example. Monte Carlo Simulation method is very useful when the population is heterogeneous. It is based on the repeated random sampling. The following steps define the Monte Carlo Simulation algorithm

- First define a possible range of unit.
- Second generate random numbers for a given possible distribution over the range.
- Third to execute a deterministic computation from the unit.
- Fourth get the result.

We estimate the population parameters of exponential mixture distribution of six different sample sizes 10, 20, 30, 100, 200, 300. We take the population parameter of exponential mixture distribution viz. $\lambda_{10}=1$, $\lambda_{20}=2$, $\lambda_{11}=5$ and $\lambda_{21}=4$. The sample size of each density function is same $m_{10}=m_{20}=n_{11}=n_{21}=n$. The Area Under the Exponential mixture ROC curve is 78.33% using the population parameters.

The estimated value of the parameters of exponential mixture ROC model using maximum likelihood method and method of moments are given in Tables 5.1 and 5.2. The estimated AUC, its variance for the exponential mixture ROC curve using maximum likelihood method and method of moments are given in the Tables 5.2 and 5.4. The bias of parameters is given in parenthesis.

Table 5.1. The estimated value of parameters of Exponential mixture ROC curve using Maximum Likelihood method

n	$\hat{\lambda}_{10}$	$\hat{\lambda}_{20}$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{21}$
10	0.905 (-0.094)	0.905 (-1.094)	4.058 (-0.941)	4.058 (0.058)
20	0.813 (-0.186)	1.795 (-0.205)	4.258 (-0.741)	0.173 (4.173)
30	1.472 (0.472)	1.472 (-0.528)	4.125 (-0.874)	4.125 (0.125)
100	0.552 (-0.448)	3.60 (1.60)	4.464 (-0.536)	1.683 (-2.317)
200	1.400 (0.400)	1.400 (-0.6)	5.112 (0.112)	5.111 (1.111)
300	0.777 (0.223)	1.477 (-0.522)	6.045 (1.045)	3.656 (-0.344)

Table 5.2. Estimated variance, standard error and confidence interval of AUC of Exponential mixture ROC curve using Maximum Likelihood method

n	$\hat{AUC}$	$v(\hat{AUC})$	$SE(\hat{AUC})$	$\hat{AUC}-Z_{\alpha/2}SE(\hat{AUC})$	$\hat{AUC}+Z_{\alpha/2}SE(\hat{AUC})$
10	0.8175	0.0044	0.0667	0.686	0.9482
20	0.8337	0.0007	0.0271	0.7805	0.8868
30	0.7370	0.0025	0.0500	0.639	0.835
100	0.7185	0.0003	0.0197	0.6798	0.7571
200	0.7850	0.0002	0.0168	0.7520	0.8179
300	0.8339	0.0002	0.0152	0.8041	0.8636

Table 5.3. The estimated value of parameters of Exponential mixture ROC curve using Method of moments

n	$\hat{\lambda}_{10}$	$\hat{\lambda}_{20}$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{21}$
10	0.905 (-0.094)	0.905 (-1.094)	2.847 (-2.153)	3.684 0.316
20	1.194 (0.194)	1.153 (-0.847)	4.782 (-0.218)	14.641 (10.641)
30	1.469 (0.469)	1.472 (-0.528)	3.496 (-1.504)	3.355 (-0.645)
100	1.231 (0.231)	5.984 (3.984)	3.984 (-1.016)	6.125 (2.125)
200	1.399 (0.399)	1.400 (-0.600)	5.112 (0.112)	5.115 (1.115)
300	1.283 (0.283)	4.163 (2.163)	2.905 (-2.095)	5.136 (1.136)

Table 5.4. Estimated variance, standard error and confidence interval of AUC of Exponential mixture ROC curve using Maximum Likelihood method

n	$\hat{AUC}$	$v(\hat{AUC})$	$SE(\hat{AUC})$	$\hat{AUC}-Z_{\alpha/2}SE(\hat{AUC})$	$\hat{AUC}+Z_{\alpha/2}SE(\hat{AUC})$
10	0.771	0.0036	0.6002	0.654	0.889
20	0.838	0.0001	0.0131	0.812	0.864
30	0.701	0.0031	0.0562	0.591	0.811
100	0.682	0.0003	0.0193	0.647	0.725
200	0.785	0.0002	0.0168	0.752	0.818
300	0.651	0.0001	0.0114	0.628	0.673

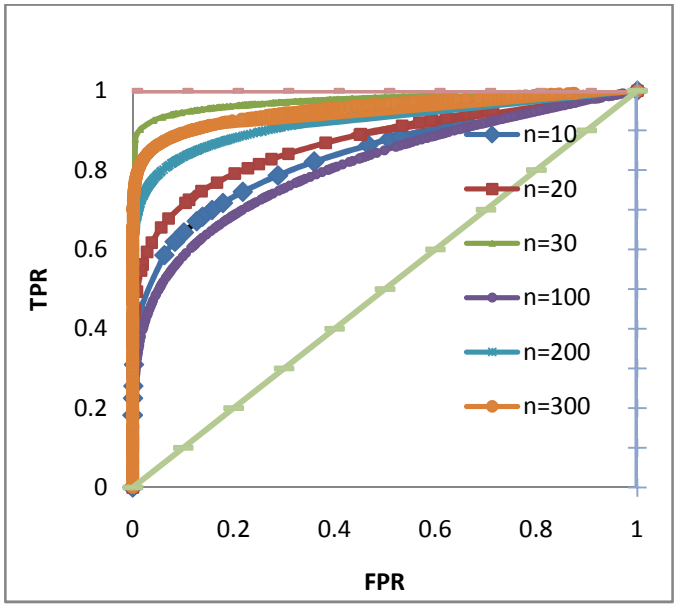

Figure 5.1. Exponential mixture ROC Curves

From Tables 5.1 to 5.4, it is observed that estimate become closer to the parameters with increase in sample size and bias get reduced. It is clear that estimates given by maximum likelihood method are more closer to the true parameter as compared to the estimates by method of moments. It is also observed that the Area under the ROC curve (AUC) is more closer to the true value of AUC using maximum likelihood method as compared to the method of moments.

Figure 5.1 shows Exponential mixture ROC curve for different sample sizes and as the sample size increases the exponential mixture ROC curve is much closer to the perfect ROC curve.

6. CONCLUSION

The Exponential mixture distribution is a very useful distribution in life testing specially in the case of presence of heterogeneity in the population. In this paper, the expression for the Exponential mixture ROC curve is found and its properties are discussed. We also derived the expressions for accuracy, variance and confidence interval of AUC. The

parameters of Exponential mixture ROC model are derived by using the maximum likelihood method and method of moments. It was found that the maximum likelihood method gives better estimates in terms of less bias as compared to the estimates by method of moments. We also observed that the accuracy increased and variance of AUC decreased by increasing the sample size. Hence, when the population have heterogeneity and also follows exponential distribution then one should use Exponential mixture ROC model instead of Exponential ROC model.

REFERENCES

[1] Dass, S.C. and Seong, W.K., "*A Semi-parameteric Approach to Estimation of ROC Curves for Multivariate BinormalMixtures*", 2011, www.stt.msu.edu/Links/Research_Memoranda/RM/RM_683.pdf.

[2] Everitt, B. and Hand, D.J., *Finite Mixture Distributions*, Chapman and Hall, New York, 1981.

[3] Gonen, M., "Mixtures of Receiver Operating Characteristic Curves", *Memorial Sloan-Katter in Cancer,* 2013, http//works.bepress.com/mithat_gonen/23.

[4] Hughes, G. and Bhattacharya B., "*Symmetry properties of Bi-normal and Bi-gamma Receiver operating characteristic curves are described by Kullback-Leibler divergences*", *Entropy*, 15(4), 2013, pp.1342-1356.

[5] McLachlan, G. and Peel D., *Finite Mixture Models*, John Wiley, New York, 2003.

[6] Newcomb, S., "A generalized theory of the combination of observations so as to obtain the best result", *American Journal of Mathematics*, 8, 1886, pp. 343-346.

[7] Pearson, K., "Contribution to the mathematical theory of evolution", *Philosophical Transactions of Royal Society*, *A*, 185, 1894, pp. 71-110.

[8] Rider, Paul R., "Estimating the parameters of mixed Poisson, Binomial and Weibull Distribution by the method of moments", *Bulletin de I'institue International de Statistique,* 1961, part2, 38.

[9] Schlattmann, P., *Medical Application of Finite Mixture Models*, Springer-Verlag, Berlin, Heidelberg, 2009.

[10] Titterington, D.M., Smith, A.F.M. and Makov, U.E., *Statistical Analysis of Finite Mixture Distributions*, John Wiley and Sons Ltd., New York, 1985.

ACKNOWLEDGEMENT

This work was supported by UGC-Major Research Project from University Grants Commission (UGC), New Delhi, INDIA. The authors wish to express their gratitude to UGC for the financial support to carry out this research work.

PROGRAM

In this paper, we did the simulation by using Mathematica 8 software. The commands are as follows:

```
Dist=MixtureDistribution[{p,1-p},{ExponetialDistribution[1/λ1],ExponentialDistribution[1/λ2]}]

Data=RandomVariate[Dist,sample size]

FindDistributionParameters[Data,MixtureDistribution[{p,1-p},{ExponentialDistribution[1/λ1],
ExponentialDistribution[1/λ2]}],ParameterEstimator->"MaximumLikelihood"]
FindDistributionParameters[Data,MixtureDistribution[{p,1-p},{ExponentialDistribution[1/λ1],
ExponentialDistribution[1/λ2]}],ParameterEstimator->"MethodOfMoments"]
```