Aligarh Journal of Statistics

Vol. 34 (2014), 67-84

STATISTICAL INFERENCE ON AUC IN NORMAL-EXPONENTIAL ROC MODEL

Sudesh Pundir and R. Amala

ABSTRACT

The accuracy of discrimination between the two populations namely healthy and diseased in a medical diagnosis can be assessed through the renowned statistical technique called Receiver Operating Characteristic (ROC) curve. The Area Under the ROC Curve (AUC) is the traditional index to measure the diagnostic accuracy. Several parametric distributions are assumed to plot a parametric ROC curve viz. Normal, Exponential, Gamma, Lognormal, Rayleigh, etc. But in all these cases, single distribution is assumed for both the populations. This paper deals with the problem of estimating ROC, AUC and standard error of AUC based on healthy test scores follow Normal distribution and diseased test scores follow Exponential distribution, we call it as *Normal-Exponential* ROC curve. The proposed model is explored using simulation as well as real life example.

1. INTRODUCTION

In a medical diagnosis, a *biomarker* which is strongly related to the disease is often assumed to be effective for screening and diagnosis of a particular disease. For assessing the accuracy of diagnosis, we have two measures namely the biomarker value often referred to as test score or risk score and the true status i.e. whether the individual belongs to healthy H or diseased D group determined by the "*Gold Standard*", where the gold standard test refers to the best performing test available. For example, gold standard test for diagnosis of aortic dissection is Magnetic Resonance Angiogram (MRA). In a general condition, the test scores of higher values corresponds to diseased and the test scores of lower values corresponds to healthy. Hence, the mean of diseased scores will be higher than the mean of healthy scores.

In a diagnostic process, a subject is regarded as "*healthy/negative*" or "*diseased/positive*" depending on the fact that the biomarker value is "*less than*" or "*greater than or equal*" to a gold standard cut-off point t. Let X and Y represent the test scores from H and D respectively determined from the gold standard. In order to assess the accuracy of selected biomarker in predicting the status, a renowned statistical tool called Receiver Operating Characteristic (ROC) curve has long been used. The cut-off point is varied within the range of test scores in order to get a ROC plot.

For a selected cut-off t, if the test score is greater than or equal to t given that the test scores from D is regarded as True Positive (TP). The *True Positive Proportion* (TPP) is defined as $P(Y \geq t)$. TPP is also called as *sensitivity*. If the test score is less than t given that the test scores is from H is regarded as True Negative (TN). The *True Negative Proportion* (TNP) is defined as $P(X < t)$. TNP is also called as *Specificity*. If the test score is greater than or equal to t given that the test scores is from H is regarded as False Positive (FP). The *False Positive Proportion* (FPP) is defined as $P(X \geq t)$. If the test score is less than t given that the test score is from D is regarded as False Negative (FN). The *False Negative Proportion* (FNP) is defined as $P(Y < t)$. For different values of t, we will get different values of these four probabilities. By plotting each pair of sensitivity and 1-specificity one can get the ROC plot.

Let the random variable Y denotes the test results of diseased subject with Probability Density Function (PDF), $g_Y(y)$ and Cumulative Distribution Function (CDF), $G_Y(y)$ Similarly, let the random variable X denotes the test results of healthy subject with PDF, $f_X(x)$ and CDF, $F_X(x)$. Assume that X and Y are independent and continuous. Mathematically, *Sensitivity* of the diagnostic test is defined as

$$TPP = y(t) = \int_{-\infty}^{t} g(y)dy,\ 0 \leq \text{y(t)} \leq 1 \qquad (1.1)$$

Specificity of the diagnostic test is defined as

$$TNP = 1 - x(t) = \int_{-\infty}^{t} f(x)dx,\ 0 \leq \text{x(t)} \leq 1 \qquad (1.2)$$

Receiver Operating Characteristic (ROC) curve is a graphical plot of FPP against TPP for different values of t. The mathematical model representing the ROC curve can be written in the form of

$$y[x(t)] = 1 - G[F^{-1}\{1 - x(t)\}]; 0 \leq x(t) \leq 1 \qquad (1.3)$$

where $x(t)$ and $y(t)$ are defined in equation (1.1) and (1.2). The area under the co-ordinates $[0,0],[0,1],[1,1]$ correspond to the ROC space. The ROC curve that falls near to $[0,1]$ has maximum accuracy 1. A completely randomized classification lies on the line joining $[0,0]$ and $[1,1]$. The area under the ROC curve given in equation (1.3) is defined as the probability that the scores of a randomly chosen diseased individual have higher values than the scores of a randomly chosen healthy individual i.e.

$$AUC = P(Y > X) = \int_0^1 y[x(t)]dx(t) \tag{1.4}$$

In ROC curve analysis, estimation of AUC and its statistical inference is the primary interest. In general, the ROC curve should satisfy the following properties.

1. The test values of Y are higher than X.
2. ROC curve is invariant with respect to monotone increasing transformation of the test scores (Krzanowski and Hand, 2009).
3. $y\left[x(t)\right]$ is monotonically increasing function i.e. the first order derivative of $y\left[x(t)\right]$ with respect to $x(t)$ should be positive i.e. $y^{'}[x(t)] > 0$.
4. $y\left[x(t)\right]$ is said to be concave, if the second order derivative of $y(x)$ with respect to $x(t)$ is negative i.e. $y^{''}[x(t)] < 0$ and convex, if $y^{''}[x(t)] > 0$.
5. The slope of ROC curve at any operating point corresponding to a cut-off t is equal to the ratio of PDF of diseased to that of PDF of healthy which is given by
$$Slope = \frac{g(t)}{f(t)} \tag{1.5}$$
6. Let $KL(f,g)$ denote the Kullback – Leibler $(K-L)$ divergence between the distributions of healthy and diseased group with $f(x)$ as the comparison distribution and $g(y)$ as the reference distribution. Then
$$KL(f,g) = \int_D f(x)\ln\left[\frac{f(x)}{g(y)}\right]dz \tag{1.6}$$

where $z \in x \cap y$; $-\infty < \mathrm{x} < \infty$; $0 < \mathrm{y} < \infty$ and D is based on z, let us represent x and y by z.

Similarly, $KL(g,f)$ denote the $K-L$ divergence between the distribution of diseased and healthy population with $g(y)$ as the comparison distribution and $f(x)$ as the reference distribution, then

$$KL(g,f) = \int_D g(y)\ln\left[\frac{g(y)}{f(x)}\right]dz \tag{1.7}$$

It is to be noted that $KL(f,g)$ and $KL(g,f)$ are positive and

$$KL(f,g) = KL(g,f) = 0, \text{ if and only if } f(x) = g(y).$$

These two measures tell us about the asymmetry of ROC curve about the negative diagonal of the ROC plot. If $KL(f,g) < KL(g,f)$, then the ROC curve is said to be TPP asymmetric and if $KL(f,g) > KL(g,f)$, then the ROC curve is said to be TNP asymmetric.

The one parameter Bi-Exponential ROC model has been studied by Betinec (2008). Bi-Normal model is the most commonly used ROC model for rating data. But it produces non-proper ROC curve i.e. it crosses the chance line because of degeneracy in the data set. As a solution to this problem, Dorfman *et al.* (1996) proposed a proper ROC analysis using Gamma distribution. Campbell and Ratnaparkhi (1993) have developed Bi-Lomax ROC model by assuming Lomax distribution, Generalized Bi-Exponential ROC model had proposed by Hussain (2011), Bi-Lognormal ROC model and its inference on AUC has been studied by Amala and Pundir (2012), Bi-Rayleigh ROC model that make use of Rayleigh distribution had been worked by Pundir and Amala (2012(a), (b)) and Pundir and Amala (2014) have reviewed some of the parametric ROC models in case of continuous data. Symmetric properties of ROC curves in terms of Kullback-Leibler divergence has been studied by Hughes and Bhattachariya (2013).

In all the above parametric ROC models, same distribution is assumed for both the populations H and D. In real life situations, it may happen that healthy scores follow one distribution and diseased scores may follow another distribution. In such a situation, we need to develop a model from two different distributions. In this paper, we propose Normal-Exponential ROC model by assuming Normal distribution to X and exponential distribution to Y to see the behavior of ROC curve and study its properties.

The paper is organized in the following way: In Section 2, Normal-Exponential ROC model, its properties and Maximum Likelihood Estimation (MLE) of AUC have been discussed. In Section 3, the asymptotic variance and confidence interval are derived for estimated AUC of Normal-Exponential ROC curve. In Section 4, the proposed model is applied to real life example and simulated data set. Section 5 discusses the concluding remarks.

2. NORMAL-EXPONENTIAL ROC MODEL

Let us assume that X is distributed as Normal with parameter μ and σ^2 and Y is distributed as exponential with inverse scale parameter θ.

The PDF of X and Y are given by

$$f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}Exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], -\infty<x,\mu<\infty,\ \sigma>0. \tag{2.1}$$

$$f_Y(y)=\theta Exp[-\theta y],\ \mathrm{y},\ \theta>0. \tag{2.2}$$

The $CDF's$ of X and Y are given by

$$F_X(x)=\Phi\left(\frac{x-\mu}{\sigma}\right) \tag{2.3}$$

$$G_Y(y) = 1 - e^{-\theta y} \tag{2.4}$$

The Specificity of the biomarker at *t* is defined as

$$1 - x(t) = FNP = \int_{-\infty}^{t} f(x)dx = \Phi\left(\frac{t-\mu}{\sigma}\right) \tag{2.5}$$

Similarly, the Sensitivity of the biomarker at *t* is defined as

$$y(t) = TPP = \int_{t}^{\infty} g(y)dy = e^{-\theta t} \tag{2.6}$$

The theoretical ROC model based on sensitivity (2.6) and specificity (2.5) is obtained as

$$y(x(t)) = Exp[-\mu\theta + \theta\sigma\Phi^{-1}(x(t))],\ 0 \le \mathrm{x(t)} \le 1 \tag{2.7}$$

where $\Phi(.)$ is the CDF of normal distribution.

1.1 Properties of Normal-Exponential ROC model

1. Normal-Exponential ROC curve is monotonically increasing function.

Proof: A function is said to be a monotonically increasing function, if the first derivative is positive. Since, first derivative of Normal-Exponential ROC curve with respect to $x(t)$ is positive i.e.

$$y'[x(t)] = \sqrt{2\pi}\theta\sigma Exp\left\{-\theta\mu + \theta\sigma\Phi^{-1}[x(t)] + \left[\frac{\Phi^{-1}[x(t)]}{\sqrt{2}}\right]^2\right\} \tag{2.8}$$

Equation (2.8) is positive since exponential function is always positive. Hence, Normal-Exponential ROC curve is monotonically increasing function.

2. Normal-Exponential ROC curve is concave and partially proper.

Proof: From equation (2.8), the second derivative of $y[x(t)]$ is obtained as

$$y''[x(t)] = \left(\theta\sigma + \Phi^{-1}[x(t)]\right)2\pi\theta\sigma Exp\left\{-\theta\mu + \theta\sigma\Phi^{-1}[x(t)] + \left(\Phi^{-1}[x(t)]\right)^2\right\}$$

$$y''[x(t)] = \begin{cases} < 0 & \text{for} \quad 0 \le \mathrm{x(t)} \le 0.5 \\ > 0 & \text{for} \quad 0.5 < \mathrm{x(t)} \le 1 \end{cases} \tag{2.9}$$

Hence, Normal-Exponential ROC curve is partially concave and partially convex in nature. Now, let us prove that it is partially proper.

ROC curve is said to be proper ROC curve if it never crosses the chance line or the decision variable is a strictly increasing function of the likelihood ratio (Dorfman *et al.*, 1996). Consider any two points 'a' and 'b' (say, $b > a$) where

$0 < a,b < 0.5$ on Normal-Exponential ROC curve. Since we have proved that the Normal-Exponential ROC curve is concave partially, the line segment connecting the point a and b never lies above the curve. So the property of proper ROC curve retains as long as the ROC curve is concave. The ROC curve may or may not cross the chance line near the convex region of the curve. Hence, we have proved that the Normal-Exponential ROC curve is partially proper.

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

$$Slope = \theta\sigma\sqrt{2\pi}Exp\left\{\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2 - t\theta\right\} \tag{2.10}$$

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

5. Normal-Exponential ROC curve is TPP asymmetric.

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

$$KL(f,g) = \left\{(2\theta\mu - 1 - \ln\left(2\pi[\sigma\theta]^2\right)\right\}\frac{1}{2}\Phi\left(\frac{\mu}{\sigma}\right) + \frac{e^{\frac{-\mu^2}{2\sigma^2}}}{\sqrt{2\pi}}\left(\frac{\mu}{2\sigma} + \theta\sigma\right) \tag{2.11}$$

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

$$KL(g,f) = \ln(\theta\sigma\sqrt{2\pi}) - 1 + \frac{2+\theta^2\mu^2 - 2\theta\mu}{2\theta^2\sigma^2} \tag{2.12}$$

It was found that $KL(f,g) < KL(g,f)$. These two divergence measures would be zero, if the healthy and diseased group is identical. Hence, we have proved that, the Normal-Exponential ROC curve is TPP asymmetric.

1.2 Estimation of AUC

The area under the Normal-Exponential ROC curve is obtained as

$$AUC = P(Y > X) = Exp\left[-\mu\theta + \frac{\sigma^2\theta^2}{2}\right] \tag{2.13}$$

where μ, σ^2 and θ are the parameters of healthy and diseased group respectively.

To estimate the AUC, we need the MLE of μ, σ^2 and θ and it is discussed in the following section.

2.3 Maximum Likelihood Estimator of AUC

Let $X_1, X_2, \ldots, X_m$ be a random sample of size m from $N(\mu, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from $Exp(\theta)$, then the log likelihood function of the joint density can be written as

$$\ln L = -m \ln \sigma - m \ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (x_i - \mu)^2 + n \ln \theta - \theta \sum_{j=1}^{n} y_j \tag{2.14}$$

Differentiating equation (2.14) with respect to μ and σ^2 we get

$$\frac{\partial \ln L}{\partial \mu} = \frac{\sum_{i=1}^{m} (x_i - \mu)}{\sigma^2} \tag{2.15}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{-m}{2\sigma^2} + \frac{\sum_{i=1}^{m} (x_i - \mu)}{2\sigma^4} \tag{2.16}$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} - \sum_{j=1}^{n} y_j \tag{2.17}$$

By equating equations (2.15), (2.16) and (2.17) to zero, we will get the ML estimates of parameters given by

$$\left.\begin{aligned} &\hat{\mu} = \frac{\sum_{i=1}^{m} x_i}{m} = \bar{x}, \\ &\hat{\sigma}^2 = \frac{\sum_{i=1}^{m} (x_i - \hat{\mu})^2}{m} \\ &\text{and} \\ &\hat{\theta} = \frac{\mathrm{n}}{\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{y_j}} = \frac{1}{\bar{y}} \end{aligned}\right\} \tag{2.18}$$

By substituting the estimates in equation (2.13), we will get the ML estimator of AUC, i.e.

$$A\hat{U}C = Exp\left\{ -\frac{\bar{x}}{\bar{y}} + \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2}{2m\bar{y}^2} \right\} \tag{2.19}$$

3. ASYMPTOTIC VARIANCE OF $A\hat{U}C$ FROM NORMAL-EXPONENTIAL ROC CURVE AND CONFIDENCE INTERVAL OF $A\hat{U}C$

In this section, we will derive the asymptotic variance of $A\hat{U}C$ and confidence interval of $A\hat{U}C$ and it is given in the form of a theorem.

Theorem 3.1: The area under the Normal-Exponential ROC curve will converge in distribution to a Normal random variable with mean zero and variance (τ)

$$e^{\theta^2\sigma^2-2\mu\theta}\left(\frac{\theta^2\sigma^2}{m}+\frac{\theta^4}{4m\left(\frac{\mu^2+\sigma^2}{\sigma^2}-\frac{1}{4\sqrt{\sigma^3}}-\mu^2\right)}+\frac{(\theta\sigma^2-\mu)^2\theta^2}{n}\right)$$

for large $\mathrm{N}(=\mathrm{m}+\mathrm{n})$

Proof: Let $L(\mu,\sigma,\theta / x,y)$ be the likelihood function of the sample observations from X and Y which is given in equation (2.14). We know that the consistent solution of the likelihood equation is asymptotically normally distributed about the true value θ_0 where $\theta_0=(\mu,\sigma^2,\theta)$, i.e.

$$\hat{\theta} \sim \mathrm{N}(\theta_0,\mathrm{I}^{-1}(\theta_0)) \tag{3.1}$$

$$\Rightarrow \sqrt{\mathrm{N}}(\hat{\theta}_0-\theta_0)\rightarrow \mathrm{N}(0,\mathrm{I}^{-1}(\theta_0)) \tag{3.2}$$

where $I(\theta)$ is the Fisher Information matrix which is given by

$$I(\theta)=-\begin{bmatrix} E\left(\frac{\partial^2 \ln L}{\partial\mu^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial\mu\partial\sigma^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial\mu\partial\theta}\right) \\ E\left(\frac{\partial^2 \ln L}{\partial\sigma^2\partial\mu}\right) & E\left(\frac{\partial^2 \ln L}{\partial(\sigma^2)^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial\sigma^2\partial\theta}\right) \\ E\left(\frac{\partial^2 \ln L}{\partial\theta\partial\mu}\right) & E\left(\frac{\partial^2 \ln L}{\partial\theta\partial\sigma^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial\theta^2}\right) \end{bmatrix}=\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{3.3}$$

where

$$a_{11}=\frac{m}{\sigma^2},\ a_{22}=m\left(\frac{\mu^2+\sigma^2}{\sigma^2}-\frac{1}{4\sqrt{\sigma^3}}-\mu^2\right),$$

$$\mathrm{a}_{33}=\frac{n}{\theta^2},\ \mathrm{a}_{12}=a_{21}=a_{13}=a_{31}=a_{23}=a_{32}=0$$

The $I^{-1}(\theta)$ is calculated as

$$I^{-1}(\theta)=\begin{bmatrix} V(\hat{\mu}) & Cov(\hat{\mu},\hat{\sigma}^2) & Cov(\hat{\mu},\hat{\theta}) \\ Cov(\hat{\sigma}^2,\hat{\mu}) & V(\hat{\sigma}^2) & Cov(\hat{\sigma}^2,\hat{\theta}) \\ Cov(\hat{\theta},\hat{\mu}) & Cov(\hat{\theta},\hat{\sigma}^2) & V(\hat{\theta}) \end{bmatrix}$$

$$=\begin{bmatrix} \frac{\sigma^2}{m} & 0 & 0 \\ 0 & \dfrac{1}{m\left(\dfrac{\mu^2+\sigma^2}{\sigma^2}-\dfrac{1}{4\sqrt{\sigma^3}}-\mu^2\right)} & 0 \\ 0 & 0 & \frac{\theta^2}{n} \end{bmatrix} \tag{3.4}$$

Since area under the ROC curve is a function of parametersμ, σ^2 and θ. We will adopt the Delta method (Powell, 2007) for finding the approximate variance which is given as follows:

$$V(A\hat{U}C)=\left(\frac{\partial AUC}{\partial\theta}\right)^2 V(\hat{\theta})+\left(\frac{\partial AUC}{\partial\mu}\right)^2 V(\hat{\mu})+\left(\frac{\partial AUC}{\partial\sigma^2}\right)^2 V(\hat{\sigma}^2)$$

$$+\ 2Cov(\hat{\mu},\hat{\theta})\left(\frac{\partial AUC}{\partial\mu}\right)\left(\frac{\partial AUC}{\partial\theta}\right)+\ 2Cov(\hat{\theta},\hat{\sigma}^2)\left(\frac{\partial AUC}{\partial\theta}\right)\left(\frac{\partial AUC}{\partial\sigma^2}\right)$$

$$+\ 2Cov(\hat{\mu},\hat{\sigma}^2)\left(\frac{\partial AUC}{\partial\sigma^2}\right)\left(\frac{\partial AUC}{\partial\mu}\right) \tag{3.5}$$

$$e^{\theta^2\sigma^2-2\mu\theta}\left\{\frac{\theta^2\sigma^2}{m}+\frac{\theta^4}{4m\left(\dfrac{\mu^2+\sigma^2}{\sigma^2}-\dfrac{1}{4\sqrt{\sigma^3}}-\mu^2\right)}+\frac{(\theta\sigma^2-\mu)^2\theta^2}{n}\right\}$$

for large $\mathrm{N}(=\mathrm{m}+\mathrm{n})$

The estimate of variance is obtained by substituting the estimates of the parameters μ, σ^2 and θ.

Hence, the estimate of AUC follows that

$$\frac{\sqrt{N}(A\hat{U}C - AUC)}{\sqrt{V(A\hat{U}C)}} \to N(0,\ 1). \tag{3.6}$$

Hence, it is proved that

$$A\hat{U}C \sim N[0,\ \tau].$$

The standard error of $A\hat{U}C$ can be obtained by taking square root of $V(A\hat{U}C)$ in equation (3.5). The $100(1-\alpha)$ % confidence interval is obtained by

$$\left[A\hat{U}C \pm Se(A\hat{U}C)Z_{\alpha/2}\right] \tag{3.7}$$

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

One can also find the ROC model by assuming that X is distributed as Exponential with parameter θ and Y is distributed as Normal with parameters μ and σ^2.

ROC model is obtained as

$$y[x(t)] = \Phi\left\{\frac{\mu}{\sigma} + \frac{\ln[x(t)]}{\theta\sigma}\right\}, 0 \le x(t) \le 1 \tag{3.8}$$

where $\Phi(.)$ is the CDF of normal distribution.

The explicit function of AUC of the model is given in equation (3.8) is not possible. But one can evaluate the AUC by substituting the estimated values of parameters.

4. NUMERICAL EXAMPLE

In this section, we provide the results of asymptotic variance of $A\hat{U}C$ and confidence interval using simulated and real life datasets.

4.1 Simulation Studies

(i) Asymptotic Variance Method

In this section, we did simulation studies to observe how the asymptotic variance of AUC behaves using simulated data sets. Let us generate four samples of size $m = 30$ for healthy from a normal population i.e. $X \sim N(\mu, \sigma^2)$ with μ taking the values (15, 20, 15, 10) with σ = (6, 8.5, 7, 6).

Similarly, let us generate four samples of size $n = 30$ from an exponential population i.e. $Y \sim Exp(\theta)$ with θ taking the values (0.0254, 0.01311, 0.0094, 0.0054).

The estimated parameters, $A\hat{U}C$, $V(A\hat{U}C)$, and Se(AÛC) and 95% Confidence Interval for $A\hat{U}C$ are shown in Table 1 which is given below.

Table 1: AÛC, Se(AÛC) and 95% confidence interval for AÛC based on Normal-Exponential ROC through asymptotic variance method

S.No	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\theta}$	$A\hat{U}C$	$V(A\hat{U}C)$	$Se(A\hat{U}C)$	95% Confidence Interval
1	14.2443	5.535	0.0284	0.676	0.00257	0.05070	[0.5762, 0.7750]
2	20.7288	7.6385	0.0117	.7877	0.00130	0.03612	[0.7169, 0.8585]
3	13.3369	6.728	0.01128	.8628	0.00066	0.02574	[0.8124, 0.9134]
4	9.3820	5.5799	0.0056	.9489	0.00011	0.01045	[0.9288, 0.9698]

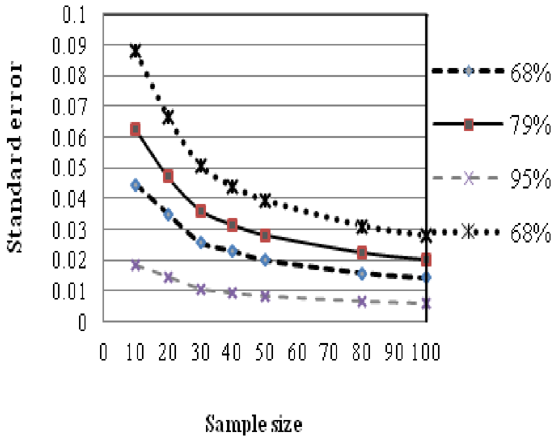
Table 2: AÛC, Se(AÛC), 95% confidence interval for AÛC and coverage area of confidence band (W)

Sample Size (m, n)		10,10	20, 20	30,30	40,40	50,50	80,80	100,100
$X \sim N$ $(\mu = 15, \sigma = 6)$ Y~Exp $(\theta = 0.0254)$	AUC Var Se LCI UCI W	0.6756 0.0077 0.0878 0.5035 0.8477 0.3443	0.6756 0.0044 0.0665 0.5453 0.8059 0.2606	0.6756 0.0026 0.0507 0.5762 0.7750 0.1988	0.6756 0.00193 0.0439 0.58952 0.7616 0.17213	0.6756 0.00154 0.03927 0.5986 0.7616 0.16304	0.6756 0.0096 0.0311 0.6147 0.7364 0.1217	0.6756 0.0008 0.0278 0.6212 0.7300 0.1089
$X \sim N$ $(\mu = 20, \sigma = 8.5)$ Y~Exp $(\theta = 0.01312)$	AUC Var Se LCI UCI W	0.7877 0.0039 0.0626 0.6651 0.9103	0.7877 0.0022 0.0470 0.6956 0.8797	0.7877 0.0013 0.0361 0.7169 0.8585	0.7877 0.00978 0.03128 0.72637 0.84898	0.7877 0.00078 0.02797 0.73284 0.84897	0.7877 0.0005 0.0222 0.7443 0.8310	0.7877 0.00039 0.01978 0.7489 0.8264

		0.2452	0.1841	0.1416	0.12261	0.11613	0.0867	0.0775
$X \sim N$ $(\mu=15,\sigma=7)$ Y~Exp $(\theta=0.0094)$	**AUC** **Var** **Se** **LCI** **UCI** **W**	.8628 .0020 .0446 .7754 .9502 .1747	.8628 .0012 .0348 .7947 .9309 .1362	.8628 .0007 .0257 .8124 .9133 .1009	0.8628 0.0005 0.0229 0.8191 0.9065 0.0874	0.8628 0.0004 0.0199 0.8237 0.9065 0.0828	0.8628 0.0003 0.0158 0.8320 0.8937 0.0618	0.8628 0.0002 0.0141 0.8352 0.8904 0.0553
$X \sim N$ $(\mu=10,\sigma=6)$ Y~Exp $(\theta=0.054)$	**AUC** **Var** **Se** **LCI** **UCI** **W**	0.9489 0.0003 0.0181 0.9138 0.9848 0.071	0.9489 0.0002 0.0144 0.9210 0.9775 0.0565	0.9489 0.0001 0.0105 0.9288 0.9698 0.0410	0.9489 0.00008 0.00905 0.93154 0.9670 0.0355	0.9489 0.00007 0.00809 0.9334 0.9670 0.0336	0.9489 0.00004 0.0064 0.9367 0.9618 0.0251	0.9489 0.00003 0.0057 0.9381 0.9605 0.0224

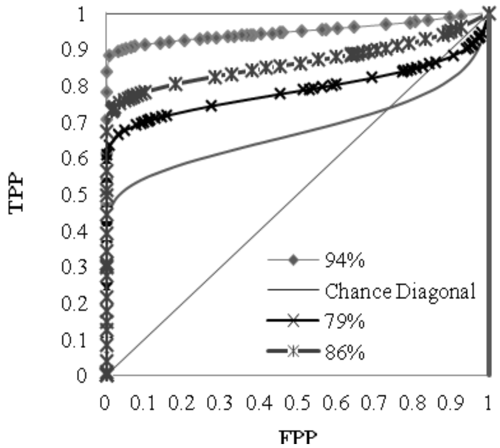
From Table 1, we observe that, variance and Se of AÛC decreases as the accuracy increase. In Table 2, the behavior of asymptotic variance is studied by varying the sample size viz. (10, 20, 30, 40, 50, 60, 80, and 100) for different values of parameters. From Table 2, it is observed that $Se(A\hat{U}C)$ decreases with increase in sample size and accuracy. The behavior is depicted in Figure 1.

Fig. 1 Standard error versus sample size



In figure 2, the property 2 from Section 2.1 is well explained. The non-proper ROC curves have occurred for $\hat{\mu}$ =(14.2443, 20.7288), $\hat{\sigma}$ = (5.535, 7.6385) and $\hat{\theta}$ =(0.0284, 0.0117) within the region of $0.5 \leq x\ (t) \leq 1$. The ROC curve is proper as long as the concavity property holds using property 2.

Fig. 2 *ROC* curve for Normal-Exponential with different parametric values



4.2 Real Life Example

A study on the relative accuracy of biomarkers viz. $CA19-9$ and $CA125$ for pancreatic cancer has been reported in Wieand *et al.* (1989). Serum concentrations of $CA125$ (cancer antigen) and $CA19-9$ (a carbohydrate antigen) have been collected from 51 control patients with pancreatitis and 90 patients with pancreatic cancer.

The data has been given in Zhou, Obuchowski, and McClish (2002). $CA125$ is not fitting both of the distribution. So, we have applied the Normal-Exponential ROC model on the Bio-marker $CA19-9$ to observe the accuracy provided the model and its behavior.

The original data set is not fitting either Normal or Exponential distribution. In order to fit the specific distribution, logarithmic transformation is done for healthy scores and square root transformation is adopted for diseased scores.

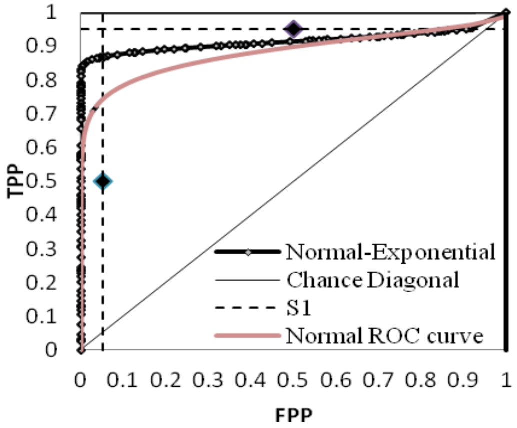
We have fitted normal distribution to healthy and exponential distribution to diseased test scores and presented the statistic, $P-$value and ranks for goodness of fit tests like Kolmogorov-Smirnov, χ^2 and Anderson- Darling tests. The results are as follows.

Table 3: Results of Goodness of Fit test

	Test	Statistic	P-value	Rank	α %
Healthy	Kolmogorov-Smirnov	0.11987	0.42311	44	20, 10, 5, 2, 1
	χ^2	5.6643	0.34026	41	20, 10, 5, 2, 1
	Anderson- Darling	0.82119	-	37	20, 10, 5, 2, 1
Diseased	Kolmogorov-Smirnov	0.10259	0.28017	28	20, 10, 5, 2, 1
	χ^2	6.9887	0.32189	27	20, 10, 5, 2, 1
	Anderson- Darling	1.2918	-	22	20, 10, 5, 2, 1

By using equation (2.18), the estimated parameters are $\hat{\mu} = 2.4723$, $\hat{\sigma} = 0.8648$ and $\hat{\theta} = 0.03621$.The AUC and standard error are estimated as 0.9148 and 0.0094 respectively. The 95% asymptotic confidence interval for AUC becomes [0.8963, 0.9148]. The test's sensitivity and specificity are found to be 88% and 88% respectively. The ROC curve plotted for the given data set is shown in Figure 3.

Fig. 3 Normal-Exponential ROC curve for using CA1 19-9 data



Now, let us discuss the asymmetry property of Normal-Exponential ROC curve plotted in Figure 3. The line segment connecting (0,1) and (1,0) is called the negative diagonal and it is obtained by plotting FPP on $X-$axis and 1-FPP on $Y-$axis.

The dashed vertical line segment S1 (say) corresponds to the co-ordinate [$FPP = a$ (0.09, say), $0 \le TPP \le 1$]. The dashed horizontal line segment $S2$ (say) corresponds to the co-ordinate

[$0 \le FPP \le 1$, $TPP = 1 - a$ (0.95)].

Let $A = [a, 0.5]$, $B = [0.5, 1 - a]$ and $C = \left[a^* > a, 1 - a^*\right]$.

A ROC curve is said to be symmetric if it passes through the co-ordinate A, B and C. Any ROC curve is said be TPP asymmetric if it passess through $S2$ after the co-ordinate B and the one that passes though S2 before the co-ordinate B is called TNP asymmetric. From Figure 3, it is an evident that Normal-Exponential ROC curve is TPP asymmetric.

When the data is applied to Bi-Normal ROC model (Krzanowski and Hand, 2009) and the accuracy and standard error are found to be 0.8793 and 0.08322 respectively. The 95% asymptotic confidence interval for AUC becomes [0.716, 1.000].

The sensitivity and specificity are found to be 82% and 82% respectively. By comparing the accuracy of the proposed Normal-Exponential and Bi-Normal model, the proposed model proves to be the best.

5. Conclusion

In some situations, it may happen that healthy population will follow normal distribution and diseased population will follow exponential distribution. In that case, Normal-Exponential ROC model should be used. Some of the properties of the model have been discussed. It was found that Normal-Exponential ROC curve is monotonically increasing, TPP asymmetric, invariance under monotone transformation and partially proper. AUC of Normal-Exponential ROC curve has been estimated. Asymptotic variance and confidence interval of estimated AUC have been computed. It is observed that Se(AUC) decreases with increase in sample size and accuracy. In the real life example $CA19-9$, it is observed that, the proposed Normal-Exponential ROC model is giving better accuracy than the conventional Bi-Normal ROC model.

Acknowledgement

The authors wish to express their gratitude and thanks to the University Grants Commission through UGC Major Research Project and Pondicherry University through University Research Fellowship for the financial assistance to carry out this research work.

REFERENCES

W. J. Krzanowski, D. J. Hand (2009): *ROC curves for continuous data. Monogr. Statist. Appl. Probab.*, CRC Press, Taylor and Francis Group, NY.

M. Betinec (2008): Testing the difference of the *ROC* Curves in Biexponential model. *Tatra Mt. Math. Publ.*, **39**, 215-223.

D. D. Dorfman, K. S. Berbaum, C.E. Metz, R. V. Lenth, J. A. Hanley, H. A. Dagga (1996): Proper Receiver Operating Characteristics Analysis: The bigamma model. *Academic Radiology*, **4**, 138-149.

G. Campbell, M. V. Ratnaparkhi (1993): An application of lomax distributions in receiver operating characteristic (*roc*) curve analysis. *Comm. Statist. Theory Methods*, **22**(**6**), 1681–1687.

E. Hussain (2011): The *ROC* Curve Model from Generalized-Exponential Distribution. *Pakistan Journal of Statistics and Operations Research*, **7**(**2**), 323-330.

R. Amala, S. Pundir (2012): Statistical Inference on AUC from a Bi-Lognormal *ROC* model for Continuous data. *International Journal of Engineering Science and Innovative Technology*, **1**(**2**), 283-295.

S. Pundir, R. Amala (2012a): A study on the Bi-Rayleigh *ROC* model. *Bonfring International Journal of Data Mining*, **2**(**2**), 42-47.

S. Pundir, R. Amala (2012b): A study on the comparison of Bi-Rayleigh *ROC* model with Bi-Gamma *ROC* model, Edited volume, Application of Reliability Theory and Survival Analysis, *Bonfring Publication*, 196-209.

S. Pundir, R. Amala (2014): Parametric receiver operating characteristic modeling for continuous data: A Glance, *Model Assisted Statistics and Application*, **9**(**2**), 121-135.

G. Hughes, B. Bhattacharya (2013): Symmetry Properties of Bi-Normal and Bi-Gamma Receiver Operating Curves are described by Kullback-Leibler Divergences. *Entropy*, **15**, 1342-1356.

L. A. Powell (2007): Approximating variance of Demographic parameters using the Delta method: A reference for avian Biologists. *The Condor*, **109**, 949-954.

S. Wieand, M. H. Gail, B. R. James (1989): A family of nonparmatric for comparing diagnostic markers with paired or unpaired data. *Biometrika*, **76**, 585-592.

X. H. Zhou, N. A. Obuchowski, D. K. McClish (2002): *Statistical Methods in Diagnostic Medicine*, John Wiley and Sons, New York.

Received : 20.09.2013

Revised : 30.03.2014

Sudesh Pundir and R. Amala

Department of Statistics
Pondicherry University
R.V. Nagar, Kalapet
Puducherry, INDIA

email:amalar.statistics@gmail.com