ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

Statistical Inference on AUC from A Bi-Lognormal ROC Model for Continuous Data

R Amala, Sudesh Pundir
Department of Statistics, Pondicherry University, Puducherry, INDIA

Abstract*—Receiver Operating Characteristic analysis has become a renowned statistical tool for classification purposes and to describe the accuracy of diagnostic tests in medical sciences and in all other related fields. In this paper, a new approach is discussed for the analysis of ROC curve using lognormal distribution. Variance, standard error and confidence interval for Area under the ROC Curve (AUC) is also found for the proposed model. Properties of the lognormal model are also discussed. The theoretical results are validated using three clinical examples.

***Index Terms* — Area under the ROC Curve, Bi-Lognormal Distribution, Confidence Interval, ROC Curve, Standard Error.**

I. INTRODUCTION

In medical diagnosis, continuous test results are available more and more when compared to the categorical data. Hence the smooth ROC curve is needed more than empirical curve for depicting the accuracy of diagnostic tests. When an individual present for a screening test for the diagnosis of disease, a predetermined threshold value or cut off value is considered on the basis of which the individual is distinguished as healthy or diseased.

Let us assume that the random vector X represents the test scores of healthy individuals and Y represents the scores of diseased individuals. The scores are represented by S and it may belong to either X or Y. In distinguishing between subjects, we come across four measures of probability namely, True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR) and False Negative rate (FNR). Sensitivity also called as True Positive Rate (TPR) is the probability of diagnosing the presence of disease when it is actually present i.e. TPR=P(S>t|D). Specificity also called as True Negative Rate (FNR) is the probability of identifying the absence of disease when it is not present i.e. FPR=P(S>t|H). A ROC curve is a tradeoff between 1-Specificity and Sensitivity.

ROC curve is used to compare two or more imaging modalities or biomarkers visually. It can also be used to determine the optimal threshold value with higher sensitivity and lower FPR. It is a unique technique which display all possible threshold value so that one can get the optimum classification. It is also used for the evaluation of Machine learning techniques and various statistical techniques.

Normal distribution is assumed to describe the biomarkers of diagnostic test in the context of ROC analysis in most of the situations [1]-[7]. However, many biological measurements show skewed or asymmetrical distribution. Such skewed distribution closely fits to an important life distribution called lognormal distribution. Moreover, Bi-normal model produces degeneracy when the sample size is small. In such a case, the use of Bi-lognormal model is extremely useful.

Lognormal distribution is a most widely used distribution for positively skewed data sets. It is characterized by log-transformed variable, using the parameter μ (the expected value) and σ^2 (the variance). It is symmetrical at the log level. The probability density function of log-normal distribution is given by [8].

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2}, -\infty < \mu < \infty, \sigma^2 > 0, X > 0, \qquad (1)$$

The estimation of reliability parameter from a bi-variate log-normal data for the case of equal mean assumed and unequal means has been discussed in [9]. Statistical inference for the AUC of bi-normal model in the presence of measurement error has been developed in [10]. The ROC curve has been also studied by assuming other well known distribution viz. Proper bi-gamma model [11], testing the difference of the ROC curves in Bi-exponential model [12], Bi-Rayleigh ROC model [13], a comparison of Bi-Rayleigh ROC model with Bi-Gamma ROC model [14].

In this paper, we have proposed the bi-lognormal ROC model and it's AUC. We have also discussed the statistical inference on AUC of proposed model. If the data fit the lognormal distribution for both the populations, we recommend the use of bi-lognormal ROC model to fit a smooth ROC curve for getting the accuracy of bio-marker. The software algorithm and codes were written in R version 14.2. Statistical test of fitting the lognormal distribution is done using Easy Fit software independently for each of the healthy and diseased sample. The proposed model is validated using three real life examples.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

The paper is prepared in the following way. In Section 2, Bi-lognormal ROC model is discussed. The properties of the ROC model are discussed in Section 3. In Section 4, estimation of AUC is proposed. Confidence interval for AUC is discussed in section 5. Section 6 gives the illustration of the proposed procedure for three real life examples viz. Tuberculosis data, Multiple Sclerosis data and pancreatic cancer data. R coding used for the computation is provided in the appendix.

II. BI-LOGNORMAL ROC MODEL

In this section, the parametric ROC curve is studied by assuming lognormal distribution. It is assumed that the biomarker of the diagnostic test is continuous. Let us assume that the test scores X and Y are independent and log-normally distributed with different parametric values and the mean of Y is greater than mean of X. Notationally, X ~ LN(μ_0,σ_0^2) and Y~ LN(μ_1,σ_1^2).The cumulative distribution function is defined by

$$F(x) = \Phi\left(\frac{\log x-\mu}{\sigma}\right) \quad (2)$$

Where Φ(.) is the standard normal distribution function.
The False positive rate for Bi-lognormal distribution is found to be

$$X(t) = P(S> t\,|H) = 1 - \Phi\left(\frac{\log t-\mu_0}{\sigma_0}\right) \quad (3)$$

One can also estimate the threshold value from FPR as follows

$$t = \exp(\mu_0) + \sigma_0\Phi^{-1}(1 - x(t)) \quad (4)$$

The true positive rate is defined by
y(t)=P(S> t | D)

$$y(t) = 1 - \Phi\left(\frac{\log t-\mu_1}{\sigma_1}\right) \quad (5)$$

and the ROC model

$$y(x) = \Phi\left(\frac{\mu_1-\mu_0}{\sigma_1} - \frac{\sigma_0}{\sigma_1}\Phi^{-1}(1 - x(t))\right) \quad (6)$$
$$= \Phi(a - b\Phi^{-1}(1 - x(t)))$$

Where $a = \frac{\mu_1 - \mu_0}{\sigma_1}$ and $b = \frac{\sigma_0}{\sigma_1}$ (7)

For plotting of ROC curve and for estimation of AUC one can use the Maximum Likelihood (ML) estimates of the parameters μ and σ^2. The ML estimates of the parameters are given by

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n_1} y_i}{n_1},\ \hat{\mu}_0 = \frac{\sum_{i=1}^{n_0} x_i}{n_0},\ \hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n_1}(y_i - \hat{\mu}_1)^2}{n_1}\ \&\ \hat{\sigma}_0^2 = \frac{\sum_{i=1}^{n_0}(x_i - \hat{\mu}_0)^2}{n_0} \quad (8)$$

Where n_0 and n_1 are the number of individuals in healthy and diseased samples respectively.

III. PROPERTIES

As the well-known Bi-normal ROC model satisfies the three properties of ROC curve [15], the proposed lognormal model also satisfies these three basic properties of the ROC curve. The properties are

a. The Lognormal ROC curve model is monotonically increasing function in the positive quadrant lying between 0 and 1.
b. The Lognormal ROC model is a convex function.

Which are explained in the form of theorems?

A. Theorem1: (Monotone property)

Let $y(x) = \Phi\left(\frac{\mu_1-\mu_0}{\sigma_1} - \frac{\sigma_0}{\sigma_1}\Phi^{-1}(1 - x(t))\right)$ be a continuous function on [0, 1] and y(x) is differentiable on [0, 1]. If the first derivative of y(x) exists and it is positive then it is monotonically increasing function.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

Proof: Let $(x) = \Phi\left(\frac{\mu_1-\mu_0}{\sigma_1} - \frac{\sigma_0}{\sigma_1}\Phi^{-1}(1-x(t))\right)$. Differentiating y(x) with respect to FPR, we get,

$$\frac{dy(x)}{dx(t)} = \frac{b\phi\left(a-b\Phi^{-1}(1-x(t))\right)}{\phi\left(\Phi^{-1}(1-x(t))\right)} > 0 \quad (9)$$

Where $\phi(.)$ is the probability density function of standard normal distribution and $\Phi^{-1}(.)$ is the inverse cumulative distribution functions of a normal variate. From equation (11), it is obvious that the proposed model is monotonically increasing function in [0, 1] which is one of the desirable property of the ROC curve.

B. Theorem 2: (Concavity property)

Let $y(x) = \Phi\left(\frac{\mu_1-\mu_0}{\sigma_1} - \frac{\sigma_0}{\sigma_1}\Phi^{-1}(1-x(t))\right)$ be a continuous function on [0, 1] and y(x) is differentiable on [0, 1]. If the second derivative of y(x) exists and it is negative, then it is said to be concave function.

Proof:

For a function y(x) to be a concave function, the second order derivative of y(x) should be negative.

Differentiating (11) with respect to x(t) we get,

$$\frac{dy(x)}{dx(t)} = \frac{b\phi\left(a-b\Phi^{-1}(1-x(t))\right)}{\phi\left(\Phi^{-1}(1-x(t))\right)}$$

$$\frac{d^2y(x)}{dx^2(t)} = \frac{\phi\left[\Phi^{-1}(1-x(t))\right]\frac{d}{dx(t)}\left[b\phi\left(a-b\Phi^{-1}(1-x(t))\right)\right] - \left[b\phi\left(a-b\Phi^{-1}(1-x(t))\right)\right]\frac{d}{dx(t)}\left[\phi\left\{\Phi^{-1}(1-x(t))\right\}\right]}{\left[\phi\left(\Phi^{-1}(1-x(t))\right)\right]^2}$$

$$\frac{d^2y(x)}{dx^2(t)} = \frac{\phi\left(\Phi^{-1}(1-x(t))\right)D_1 - [b\phi\left(a-b\Phi^{-1}(1-x(t))\right)]D_2}{\left[\phi\left(\Phi^{-1}(1-x(t))\right)\right]^2} \quad (10)$$

where $D_1 = \frac{d}{dx(t)}\left[b\phi\left(a-b\Phi^{-1}(1-x(t))\right)\right]$ and $D_2 = \frac{d}{dx(t)}\left[\Phi^{-1}(1-x(t))\right]$

Now evaluating D_1, we get

$$D_1 = \frac{d}{dx(t)}\left[\frac{b}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[a-b\Phi^{-1}(1-x(t))\right]^2\right\}\right]$$

$$= -b^2(a-b\Phi^{-1}(1-x))\frac{\phi(a-b\Phi^{-1}(1-x))}{\phi\left(\Phi^{-1}(1-x)\right)} \quad (11)$$

and $D_2 = \Phi^{-1}(1-x)$ (12)

Substituting (13) and (14) in (12) we get

$$\frac{d^2y(x)}{dx^2(t)} = \frac{-b\phi\left(a-b\Phi^{-1}(1-x(t))\right)\{b[\left(a-b\Phi^{-1}(1-x(t))\right)+\Phi^{-1}(1-x(t))]\}}{\left[\phi\left(\Phi^{-1}(1-x(t))\right)\right]^2} < 0 \quad (13)$$

Hence Bi-lognormal ROC curve is concave in nature.

VI. ESTIMATION OF AREA UNDER THE BI-LOGNORMAL ROC CURVE

Let X and Y be two random and independent vector from the healthy and diseased population and X follows lognormal distribution with parameters $\mu_0 \& \sigma_0^2$ and Y follows lognormal distribution with parameters $\mu_1 \& \sigma_1^2$. We know that AUC is the probability that the classifier will allocate a higher score to a randomly chosen individual from D and it will allocate lower score to a randomly and independently chosen individual from H i.e. AUC = P(Y>X). For the estimation of AUC, we have to use log transformation to the random vector X and Y. We know that if X or Y follows log-normal distribution then logarithm of X and Y follows normal

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

distribution. Let lnY = y and lnX= x for diseased and healthy score respectively, since it is more easy to work with normal as compared to lognormal distribution.

$$AUC = P(Y > X) = P(\ln Y > \ln X) = P(y > x)$$

It immediately follows that $y - x \sim N(\mu_1 - \mu_0, \sigma_1^2 + \sigma_0^2)$implies that

$$= P(y - x > 0) = P\left(Z > \frac{\mu_0 - \mu_1}{\sqrt{\sigma_1^{\,2} + \sigma_0^2}} \right)$$

$$AUC = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^{\,2} + \sigma_0^2}} \right). \qquad (14)$$

where $\mu_1, and\ \mu_0$ are the means and σ_1^2, σ_0^2 are the variances of the scores y and x respectively. The AUC can numerically be obtained by substituting the estimated values of the parameters from the sample data. Alternatively one can obtain the AUC in the following way:

We know that $Z_1 = \frac{y - \mu_1}{\sigma_1} \sim N(0,1)$ and $Z_0 = \frac{x - \mu_0}{\sigma_0} \sim N(0,1)$

$$AUC = P(Y > X) = P(\ln Y > \ln X) = P(y > x) \quad = P(Z_1\sigma_1 + \mu_1 > Z_0\sigma_0 + \mu_0)$$

$$= P(Z_1\sigma_1 - Z_0\sigma_0 > \mu_0 - \mu_1) = P\left(\frac{(Z_1\sigma_1 - Z_0\sigma_0) - 0}{\sqrt{\sigma_0^2 + \sigma_1^2}} > \frac{(\mu_0 - \mu_1) - 0}{\sqrt{\sigma_0^2 + \sigma_1^2}} \right)$$

$$= P\left(Z > \frac{\mu_0 - \mu_1}{\sqrt{\sigma_0^2 + \sigma_1^2}} \right) = 1 - P\left(Z \le \frac{\mu_0 - \mu_1}{\sqrt{\sigma_0^2 + \sigma_1^2}} \right) = 1 - \Phi\left(\frac{\mu_0 - \mu_1}{\sqrt{\sigma_0^2 + \sigma_1^2}} \right)$$

$$= 1 - \Phi\left(-\left[\frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}} \right] \right) = \Phi\left[\frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}} \right] \qquad (15)$$

Where Φ(.) is the cumulative distribution of standard normal distribution.

V. ESTIMATION OF CONFIDENCE INTERVAL FOR BI-LOGNORMAL AUC

In this section, we obtain the asymptotic 100 (1-α) % confidence interval for the area under the Bi-Lognormal ROC curve. In section 2, we have obtained the AUC of bi-lognormal model as

$$AUC = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^{\,2} + \sigma_0^2}} \right) = \Phi(\delta) \qquad (16)$$

The maximum likelihood estimate of δ is given by

$$\hat{\delta} = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\sqrt{\hat{\sigma}_1^{\,2} + \hat{\sigma}_0^2}} \qquad (17)$$

Since $\Phi(\hat{\delta})$ is a monotonic increasing function of $\hat{\delta}$, it is enough to find the variance and standard error of $\hat{\delta}$ for determining the confidence interval for AUC. Since δ is a function of parameter $\theta=(\mu_1,\mu_0, \sigma_1^2, \sigma_0^2)'$, we will adopt the Delta method for finding the approximate variance and standard error for $\hat{\delta}$.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

By definition,

$$V(\hat{\delta}) = \left(\frac{\partial\delta}{\partial\mu_1}\right)^2 V(\hat{\mu}_1) + \left(\frac{\partial\delta}{\partial\mu_0}\right)^2 V(\hat{\mu}_0) + \left(\frac{\partial\delta}{\partial\sigma^2{}_1}\right)^2 V(\hat{\sigma}_1{}^2) + \left(\frac{\partial\delta}{\partial\sigma_0{}^2}\right)^2 V(\hat{\sigma}_0{}^2) + 2\left(\frac{\partial\delta}{\partial\mu_1}\right)\left(\frac{\partial\delta}{\partial\mu_0}\right) Cov(\hat{\mu}_1, \hat{\mu}_0)$$

$$+ 2\left(\frac{\partial\delta}{\partial\sigma^2{}_1}\right)\left(\frac{\partial\delta}{\partial\sigma_0{}^2}\right) \mathrm{cov}(\hat{\sigma}_1{}^2, \hat{\sigma}_0{}^2)$$

$$= \left(\frac{1}{\sqrt{(\sigma_1^2+\sigma_0^2)}}\right)^2 \frac{\sigma_0^2}{n_0} + \left(\frac{-1}{\sqrt{(\sigma_1^2+\sigma_0^2)}}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{-(\mu_1-\mu_0)}{(\sigma_1^2+\sigma_0^2)^{\frac{3}{2}}}\right)^2 \frac{2\sigma_0^4}{n_0-1} + \left(\frac{-(\mu_1-\mu_0)}{2(\sigma_1^2+\sigma_0^2)^{\frac{3}{2}}}\right)^2 \frac{2\sigma_1^4}{n_1-1}$$

Hence the variance expression for $\hat{\delta}$ can be obtained using the following expression

$$V(\hat{\delta}) = \frac{1}{(\sigma_1^2+\sigma_0^2)}\left\{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0}\right) + \frac{(\mu_1-\mu_0)^2}{2(\hat{\sigma}_1^2+\hat{\sigma}_0^2)^3}\left[\frac{\sigma_1^4}{(n_1-1)}+\frac{\sigma_0^4}{(n_0-1)}\right]\right\} \quad (18)$$

Variance of $\hat{\hat{\delta}}$ can be obtained by substituting the estimated values of μ_1, μ_0, $\sigma_1{}^2$ and $\sigma_0{}^2$ in the above expression. Standard error of $\hat{\hat{\delta}}$ is just the square root of equation (18). Then the calculation of Confidence interval for lognormal AUC is straight forward. Hence, one can determine the asymptotic 100(1-α) % confidence interval for AUC is given by

$$\left(\Phi(\hat{\delta} - z_{\alpha/2}\sqrt{V(\hat{\delta})}), \Phi(\hat{\delta} + z_{\alpha/2}\sqrt{V(\hat{\delta})})\right) \quad (19)$$

Where α is the level of significance and $z_{\alpha/2}$ is the critical value of Z for a two tailed test at level of significance α/2.

VI. APPLICATIONS

Here three real life data sets are taken to validate the model.

1. The Tuberculosis data was collected from Sri Venkateswara University of Medical Sciences (SVIMS), a tertiary hospital in Tirupathi. Data consists of 100 samples with 4 variables. Among those variables Adenosine De Aminase (ADA) is the most significant factor to diagnose Tuberculosis. Totally 100 samples were collected, out of which 33 were healthy individuals and 67 were diseased individuals. The data set is affected by outliers and it is removed by 5% trimmed mean.
2. Multiple Sclerosis (MS) is an inflammatory disease in which the fatty myelin sheaths around the axons of the brain and spinal cord are damaged, leading to demyelination and scarring as well as a broad spectrum of signs and symptoms. For detecting Multiple sclerosis, one of the possible laboratory tests is the cerebrospinal fluid (CSF) immunoglobulin G(IgG) index. CSF immunoglobulin index is defined as the ratio of (IgG) in CSF/ IgG in Serum) to CSF-albumin/serum albumin. High values of the CSF IgG index are suspicious for multiple sclerosis. The data has been collected on 40 patients. Among them 20 were affected by other neurological disorders [16].
3. A study on the accuracy of biomarkers, CA19-9 & CA125 for pancreatic cancer has reported in [17]. They collected serum concentrations of CA125 (cancer antigen) & CA19-9 (a carbohydrate antigen) from 51 control patients with pancreatitis and 90 patients with pancreatic cancer [16].

To illustrate the lognormal ROC model, we have used the tuberculosis, multiple sclerosis and pancreatic data set which are described above. The result of goodness of fitness test for all the three data sets has been given in Table II. In all the data sets, the estimates of parameters are obtained through maximum likelihood procedure and are given in Table I.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

Table I ML Estimates of the Parameters

Data Set	μ_1	μ_0	σ_1	σ_0
ADA	3.9002	2.9733	0.2543	0.3421
CSF IgG index	1.6655	0.643	0.8969	0.2891
CA 19-9	5.4152	2.46296	2.34157	0.86805

Table II Details of Fitness of Data Set Used For the Validation of Bi-Lognormal ROC Model

Example	Variable	Test	Statistic		Rank		P-value	
			H	D	H	D	H	D
Tuberculosis data	Adenosine Deaminase Assays	Kolmogorov-Smirnov	0.14875	0.10194	15	32	0.6356	0.55957
		Chi-Square	2.2783	3.5728	34	31	0.32009	0.6124
		Anderson Darling's	0.91621	0.60214	26	29	-	-
MS Data	CSF IgG index	Kolmogorov-Smirnov	0.15833	0.20026	36	36	0.64127	0.35122
		Chi-Square	0.67215	1.0776	21	28	0.71457	0.58345
		Anderson Darling's	0.65108	0.72213	37	22	-	-
Pancreatic	CA 19-9	Kolmogorov-Smirnov	0.13615	0.06451	28	2	0.28513	0.82448
		Chi-Square	8.4041	4.2744	30		0.07785	0.6396
		Anderson Darling's	0.96498	0.57792	21	5	-	-

Fig. 1 and 2 depicts the ROC curve plotted for Tuberculosis data plotted using non-parametric method and parametric method. In this case, there is no major difference in the ROC curve plotted using these two methods due to the nature of the data set. Fig. 3 and 5 shows the ROC curve plotted for MS data and CA 19-9 using non-parametrically. Fig. 4 and 6 shows the ROC curve plotted for MS data and CA 19-9 parametrically and are smooth curves with better accuracy.

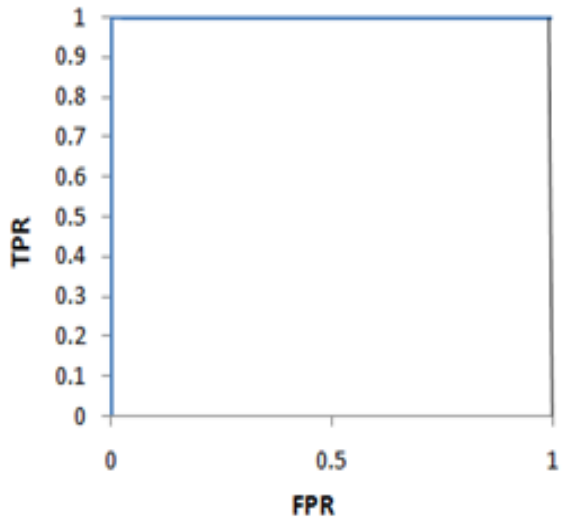

Fig. 1. Non-Parametric ROC Curve Plotted For TB Data

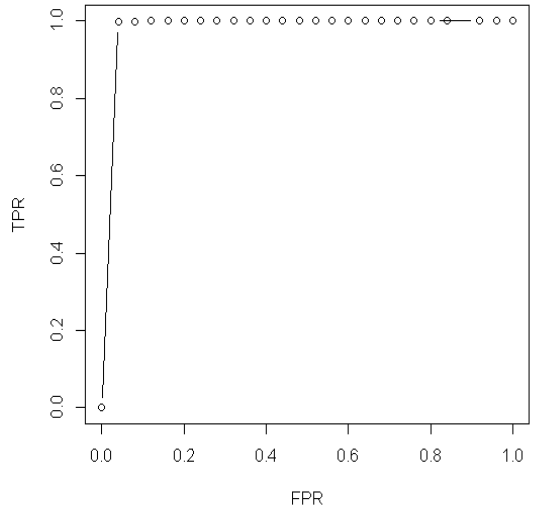

Fig. 2. Parametric Smooth ROC Curve Plotted For TB Data

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

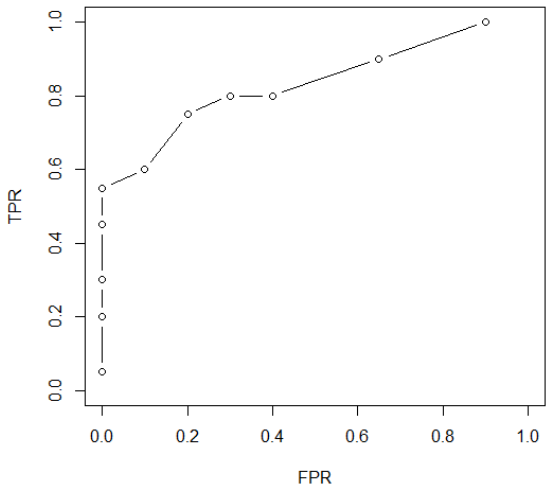

Fig. 3. Non-Parametric ROC Curve Plotted For CSF Igg Index

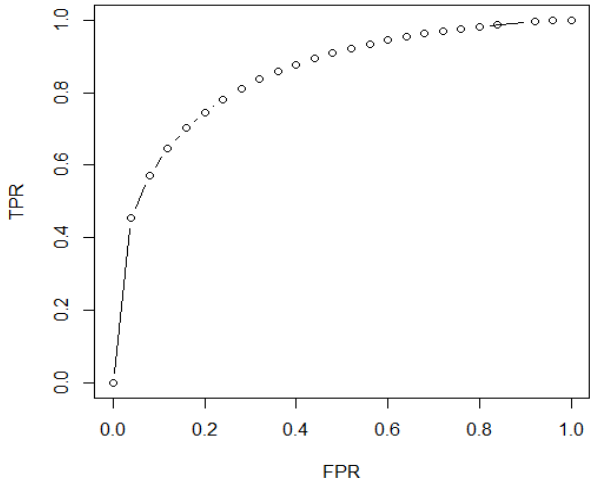

Fig. 4. Parametric Smooth ROC Curve Plotted For CSF Igg Index

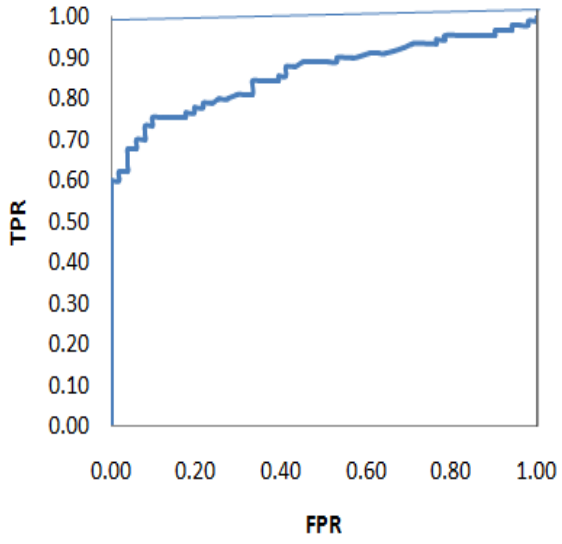

Fig. 5. Non-Parametric ROC Curve Plotted For Pancreatic Cancer Data

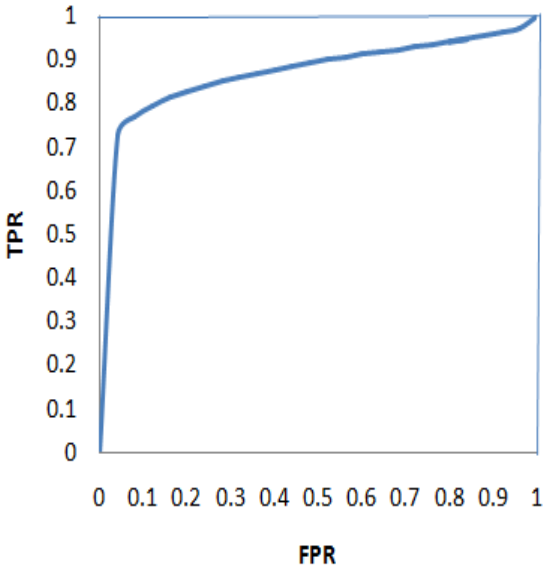

Fig. 6. Parametric Smooth ROC Curve Plotted For Pancreatic Cancer Data

From Table III for TB data, it has been found that accuracy obtained by Non Parametric method is exactly1.0 since it has been calculated approximately and the parametric method has given an accuracy of 0.9987. The ROC curve plotted for CSF IgG index using non-parametric method gives the accuracy as 0.843 whereas a parametric method gives an AUC of 0.853. Similarly, pancreatic cancer data gives an accuracy of 0.862 and 0.8814 for non-parametric and parametric respectively.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

Table III Accuracy of Lognormal AUC for Parametric Vs Non-Parametric Methods

Data Set	Non- parametric Method	Parametric Method
TB	1.0	0.9987
MS	0.843	0.853
CA 19-9	0.862	0.8814

Table IV explains the change of accuracy level and the closeness of the confidence band over various cut-off values of ADA. It is clear that for cut-off values from 31 to 41 the AUC values have become stable with closer confidence band of difference 0.010041. So the choice of cut-off value from 31 to 41 can be considered as a confidence band which maximizes the correct classification. Similar kinds of analysis are done for CSF IgG, CA19-9 and are shown in Table IV & 5 respectively.

Table IV The Variation of Accuracy Of Lognormal AUC With Confidence Band For ADA

Cut-off	AUC	Se ($\hat{\delta}$)	Lower Bound	Upper Bound	Width of the Interval
46	0.995622	0.343119	0.9743470	0.999506	0.025159
45	0.995622	0.343119	0.9743470	0.999506	0.025159
44	0.998047	0.358963	0.9854508	0.999834	0.014383
43	0.998047	0.358963	0.9854508	0.999834	0.014383
42.5	0.998047	0.358963	0.9854508	0.999834	0.014383
42	0.998047	0.358963	0.9854508	0.999834	0.014383
41	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
40	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
39	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
38	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
37	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
36	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
35	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
34	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
33	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
32	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
31.3	**0.998796**	0.363902	0.9898702	0.999911	**0.010041**
27.64	0.998251	0.320515	0.9890483	0.999806	0.010758
24.94	0.998006	0.303372	0.9888283	0.999743	0.010915
22.63	0.997975	0.293895	0.9892246	0.99972	0.010495

Table V reveals the change of accuracy level and the closeness of the confidence band over various cut-off values of CSF IgG index. From the table it is clear that for cut-off values from1.5 to 1.6, the AUC values neutralized with closer confidence band of difference 0.044904. So the choice of cut-off value from 1.5 to 1.6 can be considered as a confidence band which maximizes the correct classification.

Table V The variation of Accuracy of lognormal AUC for CSF IgG index

Cut-off	AUC	Se ($\hat{\delta}$)	Lower Bound	Upper Bound	Width of the Interval
2.1	0.979995	0.310491	0.925783	0.996119	0.070335
1.9	0.990148	0.345347	0.951042	0.998689	0.047646
1.8	0.990148	0.345347	0.951042	0.998689	0.047646

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

1.7	0.990148	0.345347	0.951042	0.998689	0.047646
1.6	**0.991124**	0.350024	0.953977	0.998881	**0.044904**
1.5	**0.991124**	0.350024	0.953977	0.998881	**0.044904**
1.33185	0.989723	0.34454	0.94958	0.998612	0.049031
1.09552	0.98297	0.328797	0.929882	0.997143	0.067261
1.0135	0.976425	0.318856	0.913083	0.995472	0.082389
0.886312	0.971307	0.312225	0.901195	0.994003	0.092808
0.794604	0.967257	0.308829	0.891885	0.992802	0.100917
0.716242	0.967257	0.308829	0.891885	0.992802	0.100917
0.643	0.961395	0.306982	0.878077	0.991077	0.113
0.569758	0.963651	0.317356	0.879544	0.99217	0.112626
0.491396	0.961354	0.320734	0.872436	0.991696	0.11926
0.399688	0.99697	0.366373	0.978642	0.999733	0.021091
0.272504	0.99697	0.366373	0.978642	0.999733	0.021091

Table VI tells the change of accuracy level and the closeness of the confidence band over various cut-off values of serum concentration CA19-9. From the table it is clear that the optimum cut-off point is around 227 with an accuracy of 0.9952.

Table VI The Variation Of Accuracy Of Lognormal AUC For Pancreatic Cancer Data

Cut-off	**AUC**	**Se ($\widehat{\delta}$)**	**Lower Bound**	**Upper Bound**	**Width of the Interval**
28	0.977816	0.18559	0.950205	0.99121	0.041005
31.2	0.979817	0.188905	0.953494	0.992245	0.038751
32.5	0.980724	0.190474	0.955021	0.992703	0.037682
32.6	0.981627	0.192066	0.956564	0.99315	0.036586
39.3	0.983635	0.195711	0.960095	0.994115	0.03402
43.5	0.984381	0.197125	0.961441	0.994464	0.033023
44.2	0.985008	0.198325	0.962591	0.994751	0.032161
59.2	0.989655	0.207664	0.971711	0.996742	0.025031
87.5	0.992831	0.214052	0.978778	0.997936	0.019158
100	0.9938	0.215951	0.981112	0.998271	0.017159
227	**0.995222**	0.214743	0.985018	0.998704	**0.013686**
251	0.995099	0.210832	0.984979	0.998632	0.013654
369	0.995015	0.206635	0.985067	0.998568	0.013501
525	0.994513	0.201506	0.984167	0.998351	0.014184
900	0.994729	0.19811	0.984968	0.99839	0.013422
1600	0.993794	0.193042	0.983075	0.998003	0.014928
2100	0.993293	0.190131	0.982132	0.997781	0.015649
2400	0.993166	0.189391	0.981899	0.997724	0.015824

VII. DISCUSSION

One natural question that may arise in the minds of readers is that one can make use of the bi-normal ROC model directly instead of using lognormal ROC model since it involves the transformation to make it normal in

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

many of the derivation of AUC. For developing the AUC and standard error expression, we are making use of the transformation but the data sets are not transformed originally. If the data is asymmetric or positively skewed then the usage of bi-normal ROC model will not yield an accurate AUC. Hence the need for alternative model is required viz. Bi-lognormal model.

ACKNOWLEDGEMENT

We thank Dr. Alladi Mohan and his team for providing the Tuberculosis data set.

REFERENCES

[1] D.M. Green, and J.A. Swets, Signal detection theory and psychophysics, New York, John Wiley and Sons Inc., 1966.

[2] D.D. Dorfman, and E. Alf, "Maximum likelihood estimation of parameters of signal detection theory & determination of confidence interval Rating method data," Journal of Mathematical Psychology, 6, 487-496, 1969.

[3] C.E. Metz, B.A. Herman, and J. Shen, "Maximum likelihood estimation of Receiver Operating Characteristic (ROC) curves from continuously-distributed data," Statistics in Medicine, 17, 1033–1053, 1998.

[4] K.H. Zou, and W.J. Hall, "Two transformation models for estimating an ROC curve derived from continuous data," Journal of Applied Statistics, 27, 621–631, 2000.

[5] M.S. Pepe, "The Statistical Evaluation of Medical Tests for Classification and Prediction," Oxford Statistical Science Series, Oxford University Press, Oxford, 2003.

[6] R. Cai, and C.S. Moskowitz, "Semi-parametric estimation of the bi-normal ROC curve for a continuous diagnostic test," Biostatistics, 5(4), 573-586, 2004.

[7] J.A. Hanley, "The robustness of the "binormal" assumptions used in fitting ROC curves," Medical decision making, 8, 197-23, and 1988.

[8] J. Aitchison, and J. A. C. Brown, "The Log-normal distribution," Cambridge University Press, 1957.

[9] R.C. Gupta, M.E. Ghitany, and D.K. Al-Mutairi, "Estimation of reliability from a bivariate log-normal data," Journal Statistical Simulation and Computation, DOI:10.1080/00949655.2011.649284, 2012.

[10] E.F. Schisterman, D. Faraggi, , B. Reiser, and M. Trevisan, "Statistical Inference for the Area under the Receiver Operating Characteristics Curve in the presence of Random measurement error", American journal of Epidemiology, 154(2), 174-179, 2001.

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

[12] M. Betinec, "Testing the difference of the ROC Curves in Biexponential model," Tatra Mountains Mathematical Publications, 39, 215-223, 2006.

[13] S. Pundir, and R. Amala, A study on the Bi-Rayleigh ROC model, Bonfring International Journal of Data Mining, 2(2), 42-47, 2012.

[14] S. Pundir, and R. Amala, 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, Coimbatore, India, 196-209, 2012.

[15] W.J. Krzanowski, and D.J. Hand, "ROC curves for continuous data, Monographs on Statistics and Applied Probability," CRC Press, Taylor and Francis Group, NY, 2009.

[16] X.H. Zhou, Obuchowski and D.K. McClish, "Statistical methods in diagnostic medicine," John Wiley and Sons, NY, 2002.

[17] S. Wieand, M. H. Gail, B. R. James, "A family of nonparametric for comparing diagnostic markers with paired or unpaired data," Biometrika, 76, 585-592, 1988.

AUTHOR BIOGRAPHY

Dr. Sudesh Pundir is currently working as an Assistant Professor in the Department of Statistics, Pondicherry University, Puducherry. Her areas of research are Biostatistics, Applied Statistics, and Reliability. She has a no. of published research papers in reputed journals. She has also participated in many conferences in India as well as abroad. She has organized one International Conference and acted as an organizing committee member in many conferences. She has

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

presented many research papers and gave invited talks, special invited talks and acted as a resource person in various National and International Conferences/Workshops. She is a life member of ISPS.

Dr. Sudesh Pundir
Assistant Professor,
Department of Statistics,
Ramanujan School of Mathematical Sciences,
Pondicherry University,
R.V. Nagar,
Kalapet,
Puducherry-605 014
sudeshpundir19@gmail.com

Amala. R did her M.Sc. Statistics with First class distinction at Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University, R.V. Nagar, Kalapet, and Puducherry-605 014 during 2009-2011. She is presently pursuing her Ph.D. in Statistics under the guidance of Dr. Sudesh Pundir, Department of Statistics, Pondicherry University, and Puducherry. Thrust area of her research is Applied Statistics. She has presented two research papers, one in an International Conference on Computational Statistics and Bio-Sciences at Pondicherry University and the other in a National Conference of National Symposium on Statistics and its Application for Young Researcher at Madras University Chennai. She has participated in 3 National conferences, one international conference and two Seminars. She has published three research papers in an international Journal.

Amala. R
Ph.D. Scholar
Department of Statistics
Ramanujan School of Mathematical Sciences
Pondicherry University
R.V. Nagar
Kalapet
Puducherry-605 014
amalar.statistics@gmail.com

APPENDIX

A.1 Creating the estimates for healthy population

```
>h<- c(sample values of healthy individuals separated by commas)

>loglik<-function(param){
+n0<-length(h)
+a0<-param[1]
+b0<-param[2]
+ll<-((-n0)*log(sqrt(b0)*sqrt(2*pi)))-sum(log(h))-(sum((log(h)-a0)^2))/(2*b0)
+ll
}
>M0<-maxNR(loglik,start=c(2,3))
>print(summary(M0))
```

A.2 Creating the estimates for diseased population

```
>d<- c(sample values of diseased individuals separated by commas)

>loglik<-function(param){
+n1<-length(h)
+a1<-param[1]
+b1<-param[2]
+ll<-((-n1)*log(sqrt(b1)*sqrt(2*pi)))-sum(log(d))-(sum((log(d)-a1)^2))/(2*b1)
+ll}
>M1<-maxNR(loglik,start=c(2,3))
>print(summary(M1))
```

A.3 R coding for the computation of ROC and AUC

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

```
z<-read.table("file path\\file name .text") # file should be in the text format
> h<-log(z$H)
> d<-log(z$D)
> m1<-mean(d); m0<-mean(h); s1<-sd(d); s0<-sd(h)
> a<-(m1-m0)/s1; > b<-s0/s1
# Calculating AUC
> lnauc<-pnorm(a/sqrt(1+b^2))
> print(lnauc)
> FPR<-seq(0,1,0.04)
 > N<-length(FPR)
> TPR<-array(dim=length(N))
> for(i in 1:N){
+ TPR[i]<-pnorm( a-(b*qnorm(1-FPR[i]) ) )
+ }
# Plotting of ROC curve
> plot(TPR~FPR,type="b",xlab="FPR",ylab="TPR")
> dt<-data.frame(FPR,TPR)
> print(dt)
```

A.4 R coding for the computation of Variance and standard error of $\hat{\delta}$ and Confidence interval for AUC

```
>z<-read.table("file path\\file name .text") # file should be in the text format
>t<-c ("cut-off values for which the se δ̂ has to be found separated by commas")
>N=length(t) # assigning the size of array t to N
```

Defining arrays of predetermined size N for the calculation of mean (md: for D, mh: for H), standard deviation (sdd: for D, sdh: for H), ROC curve parameter (a, b), AUC, $\hat{\delta}$ (del), standard error of $\hat{\delta}$ (se), lower bound (lb), upper bound (ub), width of the interval (dif), size of H sample (nh), size of D sample (nd).

```
>md<-array(dim=length(N)); mh<-array(dim=length(N)); sdd<-array(dim=length(N))
>sdh<-array(dim=length(N)); >a<-array(dim=length(N)); b<-array(dim=length(N))
>AUC<-array(dim =length(N)) ; cut<-array(dim =length(N)); cons<-array(dim =length(N))
>nd<-array(dim =length(N)); h<-array(dim =length(N)); del<-array(dim =length(N))
>se<-array(dim =length(N)); lb<-array(dim =length(N)); ub<-array(dim =length(N))
>dif<-array(dim =length(N))
```

Initializing for loop for each i the cut-off changed and simultaneously finding the above said features for each dataset defined by i.

```
>for(i in 1:N)
{
        h=z[z<t[i]]; d=z[z>=t[i]]; lnh<-log(h); lnd<-log(d);     nd[i]=length(d)
        nh[i]=length(h); md[i]=mean(lnd); mh[i]=mean(lnh)
        sdd[i]=sd(lnd); sdh[i]=sd(lnh); a[i]=((md[i]-mh[i])/sdd[i])
        b[i]=sdh[i]/sdd[i]
        AUC[i]=pnorm(a[i]/sqrt(1+b[i]^2))
        del[i]<-(a[i]/sqrt(1+b[i]^2))
        cons[i]<- ( (sdd[i]^2/nd[i]) + (sdh[i]^2/nh[i])  )* (1/ (sdh[i]^2+sdd[i]^2) ) +
                ((md[i]-mh[i])^2 /( 2* (sdd[i]^2 + sdh[i]^2)^3 ) )  *  ( (sdh[i]^4/ (nh[i]-1) )+ (sdd[i]^4/ (nd[i]-
        1)) )
        se[i]<-sqrt(cons[i])
        lb[i]<-pnorm(del[i]-(1.96*se[i]))
```

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 1, Issue 2, November 2012

```
ub[i]<-pnorm(del[i]+(1.96*se[i]))
dif[i]<-ub[i]-lb[i]
dt<-data.frame(nd, nh, md, mh, sdd, sdh, cut, AUC, del, se, lb, ub, dif)
print(dt)}
```