International Journal of Scientific Engineering and Science
Volume 3, Issue 11, pp. 11-18, 2019. ISSN (Online): 2456-7361

Estimation of AUC of ROC Curve Underlying Two Parameter Rayleigh Distribution

S. Chivukula, C. Uma Shankar
Department of Statistics (OR &SQC), Rayalaseema University, Kurnool

Abstract— *Modelling ROC curves have gained attention over the years and noticed various manifestations in estimation the Area under the Curve, sensitivity and specificity. All these were proposed on Binormal ROC curve and Non-parametric forms, however, in the recent past, many non-normal distributions have been proposed. In this paper, we propose an ROC curve that assumes the data is follow Rayleigh distribution. Certain characteristics are derived and are supported using simulated datasets.*

Keywords— *ROC Curve, Rayleigh distribution, AUC.*

I. INTRODUCTION

The Advanced technological developments over the years, especially in the field of medicine, made the time frequency analysis widely used, due to the advent of tools like magneto encephalogram, functional magnetic resonance imaging, electroencephalogram, and so on. Signals of these equipment's are recorded from 1sec to 1millisec to 1microsec time points. Many software applications are available to detect and analyse Signals. Rayleigh distribution is often observed in studies with frequency analysis, for example Fourier analysis, Wavelet analysis, and so on. This distribution is usually given by two-dimensional uniform distribution.

The Statistical Distribution of the maxima and minima (peaks and troughs) of a continuous random function representing a stationary stochastic process is of great importance in defining many physical phenomena. If one considers the sea surface as a summation of an infinite number of infinitesimally small sinusoids, each having its own frequency, amplitude, and direction that are combined in random phase, such an irregular function can be described by means of an energy spectrum. The distribution of maxima of a random process described by an energy spectrum was first treated by Rice (1944 & 1945). The probability density function suggested by Rice for the maxima of a random stationary Gaussian process was quite general and could be applied regardless of the shape or width of the spectrum. Hence, it was found to be applicable to ocean waves and ship responses. In 1880 Lord Rayleigh derived the aforementioned probability density function in connection with determining the resultant intensity of a large number of independent sounds, and Rice (1944, 1945) developed his theory in relation to electrical noise signals. Cartwright and Longuet-Higgins (1956) developed the distribution of maxima further, with particular emphasis on marine applications. The application of these theories towaves and ship response indicates the variety of phenomena which can be solved, as long as the process described conforms to the conditions of being a linear, stationary, Gaussian, ergodic process, as subsequently discussed.

The Rayleigh distribution has since been found by many investigators (Watters, 1953; Bailey et al., 1963; Cartwright et al. (1956) to correctly represent observed short-term distributions of the heights of sea waves and many other related phenomena, such as ship motions and bending moment responses.

The starting point for signal detection theory is that nearly all reasoning and decisionmaking takes place in the presence of some uncertainty. Signal detection theory provides a precise language and graphic notation for analyzing decision making in thepresence of uncertainty. The general approach of signal detection theory has directapplication in terms of sensory experiments same and also offer corresponding challenging situations for decision outcomes. ROC curve analysis has gained importance during the last seven to eight decades (Krzanowski and Hand, 2009& Zou et al., 2011). There is a vast growth in both theoretical and practical approach of ROC curves in diversified fields. However, its origin for theoretical development and practical orientation was in analysing radar signals which come under the theory of signal detection. With the above note on Rayleigh distribution and its applications in Signal processing and as well with the ROC platform in reading and assessing the signal lead to the thought of imputing the mathematical functionality of Rayleigh distribution in the theory of classification. The model proposed here is by considering two parameter Rayleigh Distribution, i.e., location and scale parameters.

II. TWO PARAMETER BI-RAYLEIGH ROC CURVE

Let X be a random variable follows two parameter Rayleigh Distribution with location and scale. Let there be two populations, namely Normal (H) and Abnormal (D), here the main goal is to assign the objects into one of two populations using a threshold (classification rule). The identification of the value of the threshold should be chosen in such a way that it should minimize the error classification.Let $x_1$and $x_2$be the test scores which are distributed according to two parameter Rayleigh Distribution in H and D populations respectively.

The Probability Density Function for two parameter Rayleigh Distribution with scale parameter σ and location parameter $\mu$is given in (1) and (2)

$$f(x;\sigma,\mu) = \frac{x-\mu}{\sigma^2} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2} \quad (1)$$

The Corresponding Cumulative Distribution Function (CDF) for $x > \mu$ is a follows;

International Journal of Scientific Engineering and Science
Volume 3, Issue 11, pp. 11-18, 2019. ISSN (Online): 2456-7361

$$F(x;\mu,\sigma) = 1 - e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2} \quad (2)$$

Since both x_1 and x_2 are continuous, every data value acts as a possible threshold. As usual let t be threshold so that in the usual notation $x(t) = 1 - F(t)$ and $y(t) = 1 - G(t)$.

The False Positive Rate $(x(t))$ is defined as

$$\text{FPR} = P(s > t/H) = 1 - P(s \le t/H)$$

$$x(t) = 1 - (1 - e^{\frac{-1}{2}(\frac{t-\mu_H}{\sigma_H})^2})$$

$$x(t) = e^{\frac{-1}{2}(\frac{t-\mu_H}{\sigma_H})^2} \quad (3)$$

The True Positive Rate $(y(t))$ is defined as

$$\text{TPR} = P(s > t/D) = 1 - P(s \le t/D)$$

$$y(t) = 1 - (1 - e^{\frac{-1}{2}(\frac{t-\mu_D}{\sigma D})^2})$$

$$\text{y(t)} = e^{\frac{-1}{2}(\frac{t-\mu_D}{\sigma_D})^2} \quad (4)$$

Now we find expression for t from $x(t))$

$$x(t) = e^{\frac{-1}{2}(\frac{t-\mu_H}{\sigma_H})^2}$$

$$\text{Log}[x(t)] = -\frac{(t-\mu_H)^2}{2\sigma_H}$$

$$t - \mu_H = \sqrt{2\sigma_H \text{Log}[\frac{1}{x(t)}]}$$

$$t = \mu_H + \sqrt{2\sigma_H \text{Log}[\frac{1}{x(t)}]} \quad (5)$$

Substituting t in (4) , we have

$$y(x(t)) = e^{\frac{-1}{2}\left(\frac{\mu_H + \sqrt{-2\sigma_H \text{Log}[x(t)]} - \mu_D}{\sigma_D}\right)^2}$$

$$y(x(t)) = e^{\frac{-1}{2\sigma_D{}^2}\left((\mu_H-\mu_D)^2 + 2\sigma_H \text{Log}[\frac{1}{x(t)}] - 2(\mu_H-\mu_D)\sqrt{-2\sigma_H \text{Log}[x(t)]}\right)}$$

$$y(x(t)) = e^{\frac{-1}{2\sigma_D{}^2}\left((\mu_H-\mu_D)^2 + 2\sigma_H \text{Log}[\frac{1}{x(t)}] - 2(\mu_H-\mu_D)\sqrt{2\sigma_H \text{Log}[\frac{1}{x(t)}]}\right)}$$

$$y(x(t)) = e^{\frac{-1}{2\sigma_D{}^2}(\mu_H-\mu_D)^2} e^{\frac{-1}{\sigma_D{}^2}(\sigma_H \text{Log}[\frac{1}{x(t)}])} e^{\frac{-1}{\sigma_D{}^2}(-2(\mu_H-\mu_1)\sqrt{2\sigma_0 \text{Log}[\frac{1}{x(t)}]})}$$

$$y(x(t)) = e^{\frac{-1}{2\sigma_D{}^2}(\mu_H-\mu_D)^2} e^{(\text{Log}[x(t)]^{\frac{\sigma_H}{\sigma_D{}^2}})} e^{2(\mu_D-\mu_H)\sqrt{2\sigma_H \text{Log}[\frac{1}{x(t)}]}}$$

$$y(x(t)) = e^{\frac{-1}{2\sigma_D{}^2}(\mu_H-\mu_D)^2} x(t)^{\frac{\sigma_H}{\sigma_D{}^2}} e^{2(\mu_D-\mu_H)\sqrt{2\sigma_H \text{Log}[\frac{1}{x(t)}]}} \quad (6)$$

The expression in (6) is the ROC curve obtained through the distributional imputation of Rayleigh form and hence it is referred as *Bi-Rayleigh ROC Curve (ROC$_{VH}$ Curve)* with location and scale parameters.

In order to obtain the expression for Area under the ROC$_{VH}$ curve, we need to integrate $y(x(t))$ between 0 to 1, i.e.,

$$\text{AUC} = \int_0^1 y(x(t))\, dx(t)$$

$$\text{AUC} = e^{\frac{-1}{2\sigma_D{}^2}(\mu_H-\mu_D)^2} \int_0^1 x(t)^{\frac{\sigma_H}{\sigma_D{}^2}} e^{2(\mu_D-\mu_H)\sqrt{2\sigma_H \text{Log}[\frac{1}{x(t)}]}}\, dx(t) \quad (7)$$

The expression (7) does not have a closed form solution and hence it is evaluated using numerical integration.

III. Random Number Generation

- Generate random samples, say 'n' from u ~ U(0,1)
- Using the expression (8), we obtain the random samples of Rayleigh distribution

$$u = F(x)$$

$$u = 1 - e^{\frac{-1}{2}(\frac{t-\mu}{\sigma})^2}$$

$$t = \mu + \sqrt{2\sigma^2 \text{Log}[\frac{1}{1-u}]} \quad (8)$$

- Repeat the above two steps for generating random samples of different sizes
- In order to generate random samples for H and D populations, the expression (8) can be rewritten as
- H population $t = \mu_H + \sqrt{2\sigma_H^2 \text{Log}[\frac{1}{1-u_1}]}$ and (9)
- D population $t = \mu_D + \sqrt{2\sigma_D^2 \text{Log}[\frac{1}{1-u_2}]}$ (10)

Here u_1 and u_2 are the uniform random numbers generated between 0 and 1.

IV. Excel Templates

4.1 Generate Random Samples of Different Sample Sizes

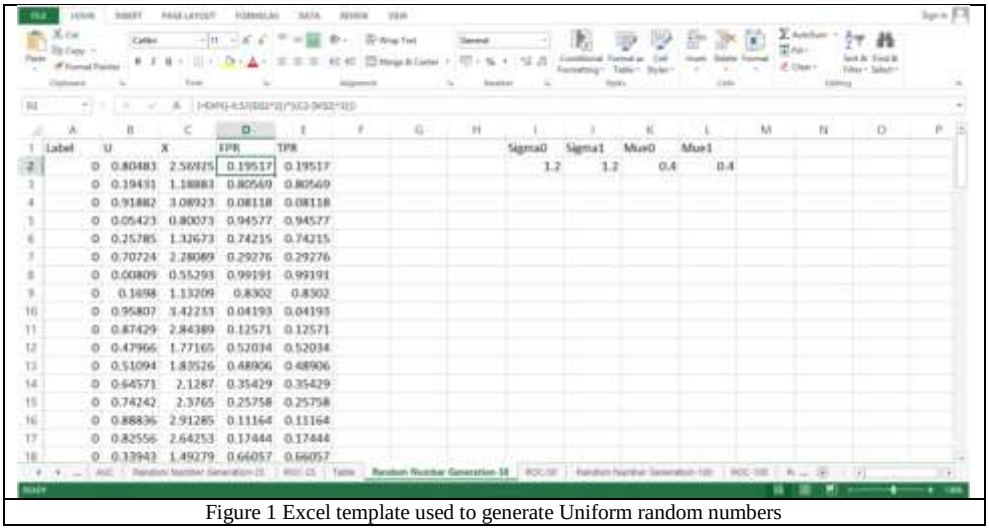
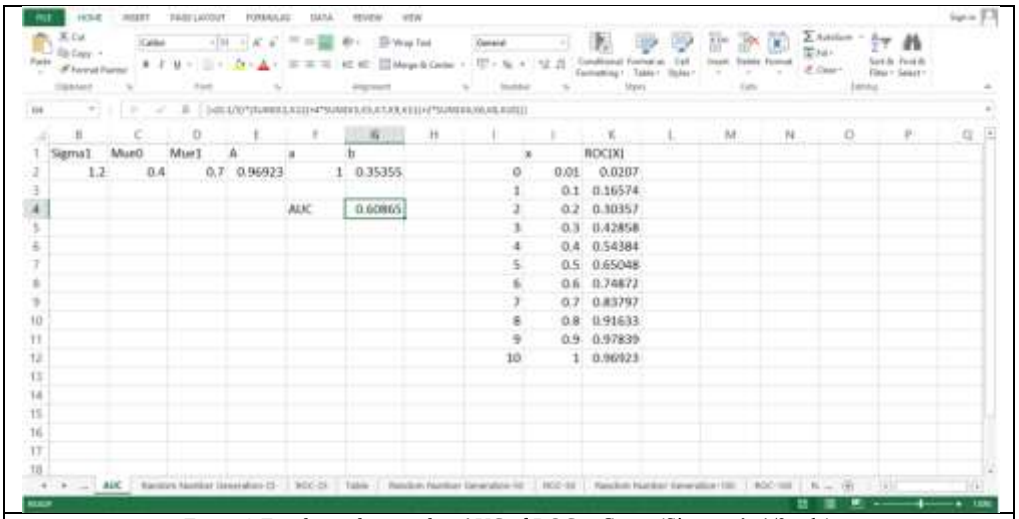


Figure 1 Excel template used to generate Uniform random numbers

International Journal of Scientific Engineering and Science
Volume 3, Issue 11, pp. 11-18, 2019. ISSN (Online): 2456-7361

4.2 Computation of Area under the ROC_{VH} curve



Sigma1	Mue0	Mue1	A	a	b		x	ROC(X)
1.2	0.4	0.7	0.96923	1	0.35355		0	0.01 0.0207
							1	0.1 0.16574
				AUC	0.60865		2	0.2 0.30357
							3	0.3 0.42858
							4	0.4 0.54384
							5	0.5 0.65048
							6	0.6 0.74872
							7	0.7 0.83797
							8	0.8 0.91633
							9	0.9 0.97839
							10	1 0.96923

Figure 2 Excel template used to AUC of ROC_{VH} Curve (Simpson's 1/3 rule)

4.3 Construction of ROC_{VH} curve

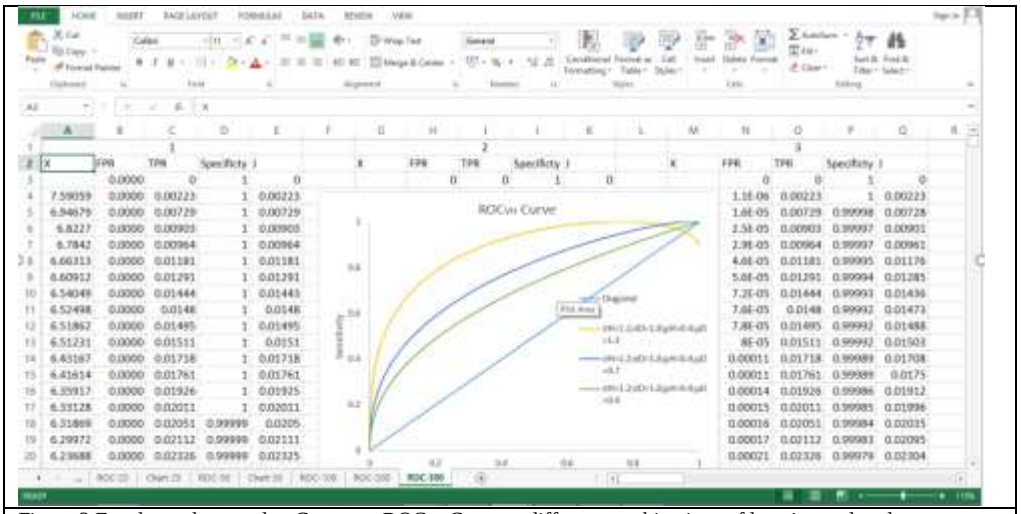


Figure 3 Excel template used to Construct ROC_{VH} Curve at different combinations of location and scale parameters

Algorithm to construct ROC_{VH} Curve

Step 1: Random numbers are generated from U(0,1)

Step 2: Two parameter Rayleigh random variables are computed for H and D population using equations (9) and (10)

Step 3: FPR(1-Specificity) and TPR(Sensitivity) values are computed from the two parameter Rayleigh random variable for H and D population using equations (3) and (4)

Step 4: FPR and TPR values are sorted in ascending order for FPR as the ROC curve is monotonically increasing.

Step 5: The ROC_{VH} Curve is plotted by using the FPR on x-axis and TPR on y-axis.

V. Results and Discussions

In this section, an attempt is made to explain the behaviour of the proposed ROC_{VH} Curve through simulation studies. All the simulations are done at different sample sizes.

Random numbers are generated for healthy and diseased population with following parameter values. μ_D ={0.4,0.7,1.3}; μ_H ={0.4}; σ_D ={1.2,1.5,1.8} and σ_H ={1.2}.Initially random numbers are generated using $U(0,1)$ and then converted using equations (9) and (10).These simulations are carried out at $n_D = n_H$={25, 50, 100, 200 and 500}.

The AUC value has been computed using expression given in (7).The AUC values are computed through Numerical interpretation. All the computations and plots are done in MS-Excel. The simulation Studies are carried out in such a way that the three typical forms of an ROC curve can be shown and demonstrated in detail. Table 1, depicts the AUC values for various combinations ofμ_H,μ_D,σ_H and σ_D. As it is known that a combination with huge difference in location parameter and large clariability in scale parameters should attain better accuracy resulting to greater amount of correct classification. In contrast, a combination with minimum mean difference and similar lower variability should result in low accuracy intern

explaining the random classification. We can consider another case where the combination will provide an accuracy with a moderate amount of correct classification.

The above mentioned three cases can be viewed virtually with overlapping areas of density curves of two Rayleigh populations. The first case resemblance minimum overlapping meaning to less percentage of false negative and false positive cases; second case reveals the scenario of complete overlapping meaning to the almost same percentage of trace positive, true negatives, false positive and false negative and third case explains the situation of moderate overlapping of density curves.

TABLE 1. AUC values at different combination scale and location parameters

σ_H	σ_D	μ_H	μ_D	AUC
1.2	1.8	0.4	1.3	0.8799
1.2	1.8	0.4	0.7	0.7686
1.2	1.8	0.4	0.4	0.6931
1.2	1.5	0.4	1.3	0.8407
1.2	1.5	0.4	0.7	0.7017
1.2	1.5	0.4	0.4	0.6104
1.2	1.2	0.4	1.3	0.7820
1.2	1.2	0.4	0.7	0.6086
1.2	1.2	0.4	0.4	0.5003

TABLE 2.

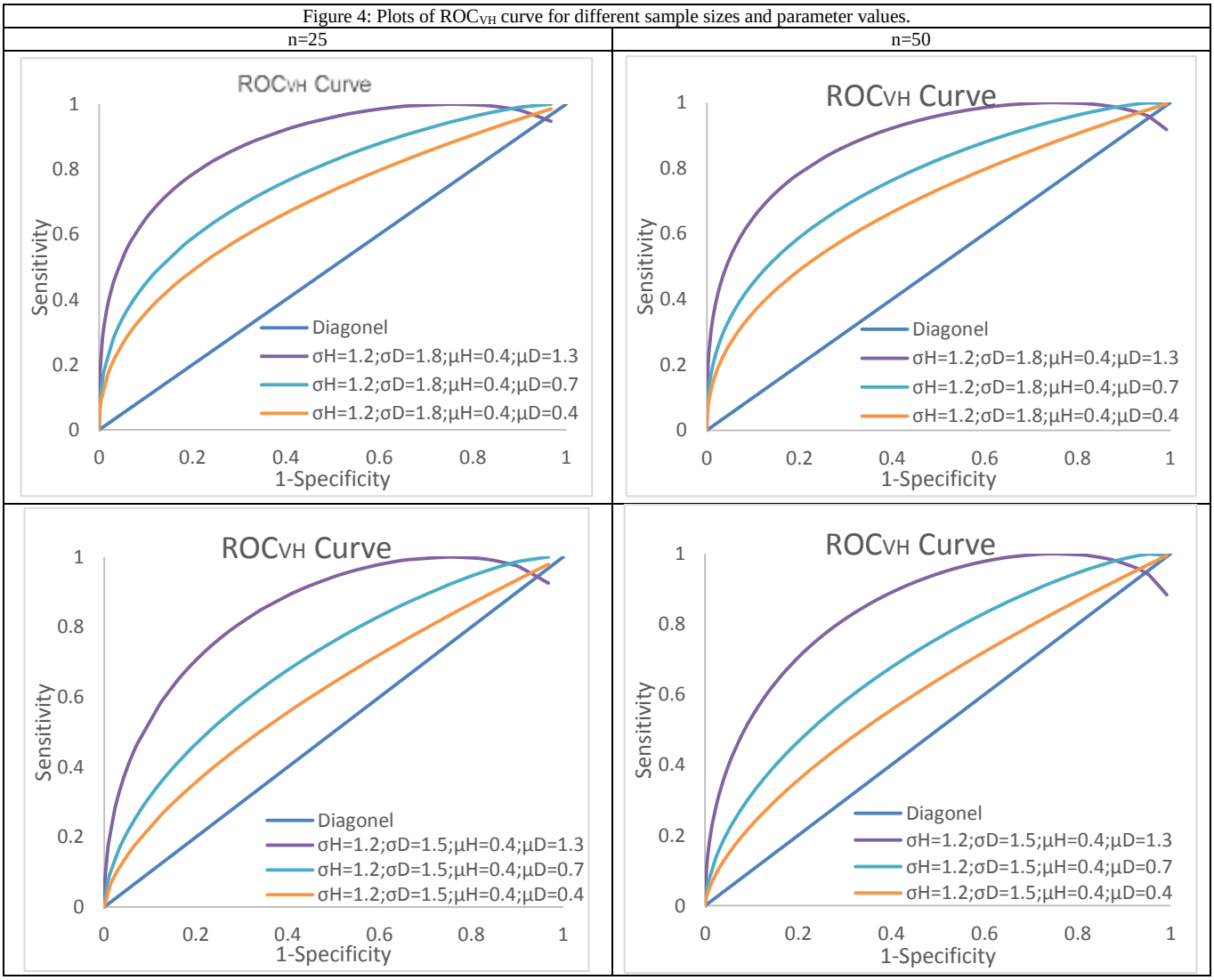
No. of Samples	σ_H	σ_D	μ_H	μ_D	AUC	(FPR,TPR)	Youden's J	Cut-off
25	1.2	1.8	0.4	1.3	0.8799	(0.1945,0.7792)	0.5847	2.5716
	1.2	1.8	0.4	0.7	0.7686	(0.2334,0.6244)	0.391	2.447
	1.2	1.8	0.4	0.4	0.6931	(0.2334,0.5238)	0.2904	2.447
	1.2	1.5	0.4	1.3	0.8407	(0.2636,0.7792)	0.5156	2.3597
	1.2	1.5	0.4	0.7	0.7017	(0.3326,0.6144)	0.2818	2.1806
	1.2	1.5	0.4	0.4	0.6104	(0.2881,0.4509)	0.1628	2.2932
	1.2	1.2	0.4	1.3	0.782	(0.4041,0.8372)	0.4331	2.0155
	1.2	1.2	0.4	0.7	0.6086	(0.5198,0.6706)	0.1508	1.7727
50	1.2	1.8	0.4	1.3	0.8799	(0.1952,0.7799)	0.5847	2.5693
	1.2	1.8	0.4	0.7	0.7686	(0.2296,0.6204)	0.3909	2.4587
	1.2	1.8	0.4	0.4	0.6931	(0.2227,0.5130)	0.2903	2.4799
	1.2	1.5	0.4	1.3	0.8407	(0.2731,0.7888)	0.5156	2.3334
	1.2	1.5	0.4	0.7	0.7017	(0.3095,0.5912)	0.2819	2.2379
	1.2	1.5	0.4	0.4	0.6104	(0.2928,0.4556)	0.1628	2.2809
	1.2	1.2	0.4	1.3	0.782	(0.3794,0.8137)	0.4342	2.0706
	1.2	1.2	0.4	0.7	0.6086	(0.5203,0.6712)	0.1508	1.7716
100	1.2	1.8	0.4	1.3	0.8799	(0.2046,0.7894)	0.5849	2.5378
	1.2	1.8	0.4	0.7	0.7686	(0.2398,0.6308)	0.391	2.428
	1.2	1.8	0.4	0.4	0.6931	(0.2277,0.5180)	0.2904	2.4645
	1.2	1.5	0.4	1.3	0.8407	(0.2696,0.7853)	0.5157	2.343
	1.2	1.5	0.4	0.7	0.7017	(0.3234,0.6053)	0.2819	2.2031
	1.2	1.5	0.4	0.4	0.6104	(0.2932,0.4560)	0.1628	2.2799
	1.2	1.2	0.4	1.3	0.782	(0.3719,0.8061)	0.4342	2.0879
	1.2	1.2	0.4	0.7	0.6086	(0.5378,0.6886)	0.1508	1.7365
200	1.2	1.8	0.4	1.3	0.8799	(0.2042,0.7891)	0.5849	2.5389
	1.2	1.8	0.4	0.7	0.7686	(0.2329,0.6239)	0.391	2.4485
	1.2	1.8	0.4	0.4	0.6931	(0.2311,0.5215)	0.2904	2.454
	1.2	1.5	0.4	1.3	0.8407	(0.2734,0.7891)	0.5156	2.3325
	1.2	1.5	0.4	0.7	0.7017	(0.3203,0.6022)	0.2819	2.2108
	1.2	1.5	0.4	0.4	0.6104	(0.2870,0.4498)	0.1628	2.2961
	1.2	1.2	0.4	1.3	0.782	(0.3751,0.8094)	0.4343	2.0804
	1.2	1.2	0.4	0.7	0.6086	(0.5297,0.6805)	0.1508	1.7528
500	1.2	1.8	0.4	1.3	0.8799	(0.1179,0.6799)	0.5619	2.8813
	1.2	1.8	0.4	0.7	0.7686	(0.2156,0.6058)	0.3902	2.5022
	1.2	1.8	0.4	0.4	0.6931	(0.2331,0.5235)	0.2904	2.448
	1.2	1.5	0.4	1.3	0.8407	(0.2696,0.7853)	0.5157	2.3429
	1.2	1.5	0.4	0.7	0.7017	(0.3213,0.6032)	0.2819	2.2082
	1.2	1.5	0.4	0.4	0.6104	(0.2896,0.4524)	0.1628	2.2893
	1.2	1.2	0.4	1.3	0.782	(0.3756,0.8099)	0.4343	2.0793
	1.2	1.2	0.4	0.7	0.6086	(0.5292,0.6800)	0.1508	1.7539

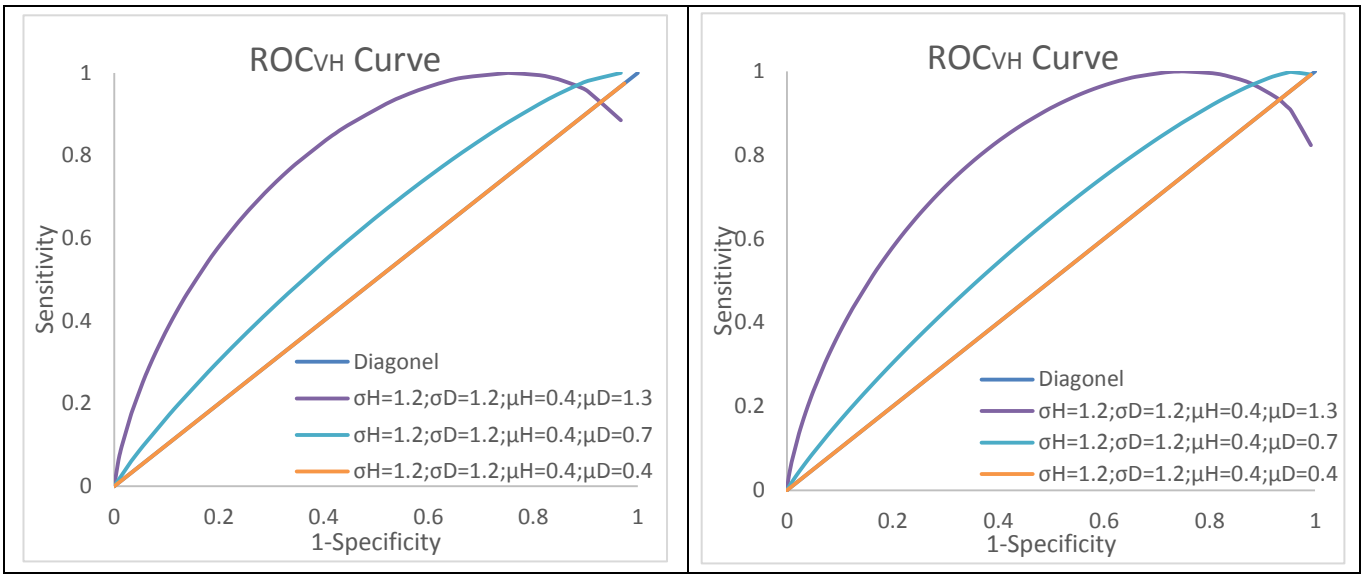
International Journal of Scientific Engineering and Science
Volume 3, Issue 11, pp. 11-18, 2019. ISSN (Online): 2456-7361

The above situation can be understood in a better way with the values reported in Table 1. Now consider combination μ_H=0.4;μ_D=1.3;σ_H=1.2 and σ_D=1.8, this provides an accuracy of 0.8799. This means that 87.99% of cases can be correctly classified. However to interpret the AUC we need to take support of intrinsic measures and cut-off. Similarly take the combination, where $\mu_H = \mu_D$=0.4;$\sigma_H = \sigma_D$=1.2, AUC attains a value equal to 0.5 which is an illustration for random classification.

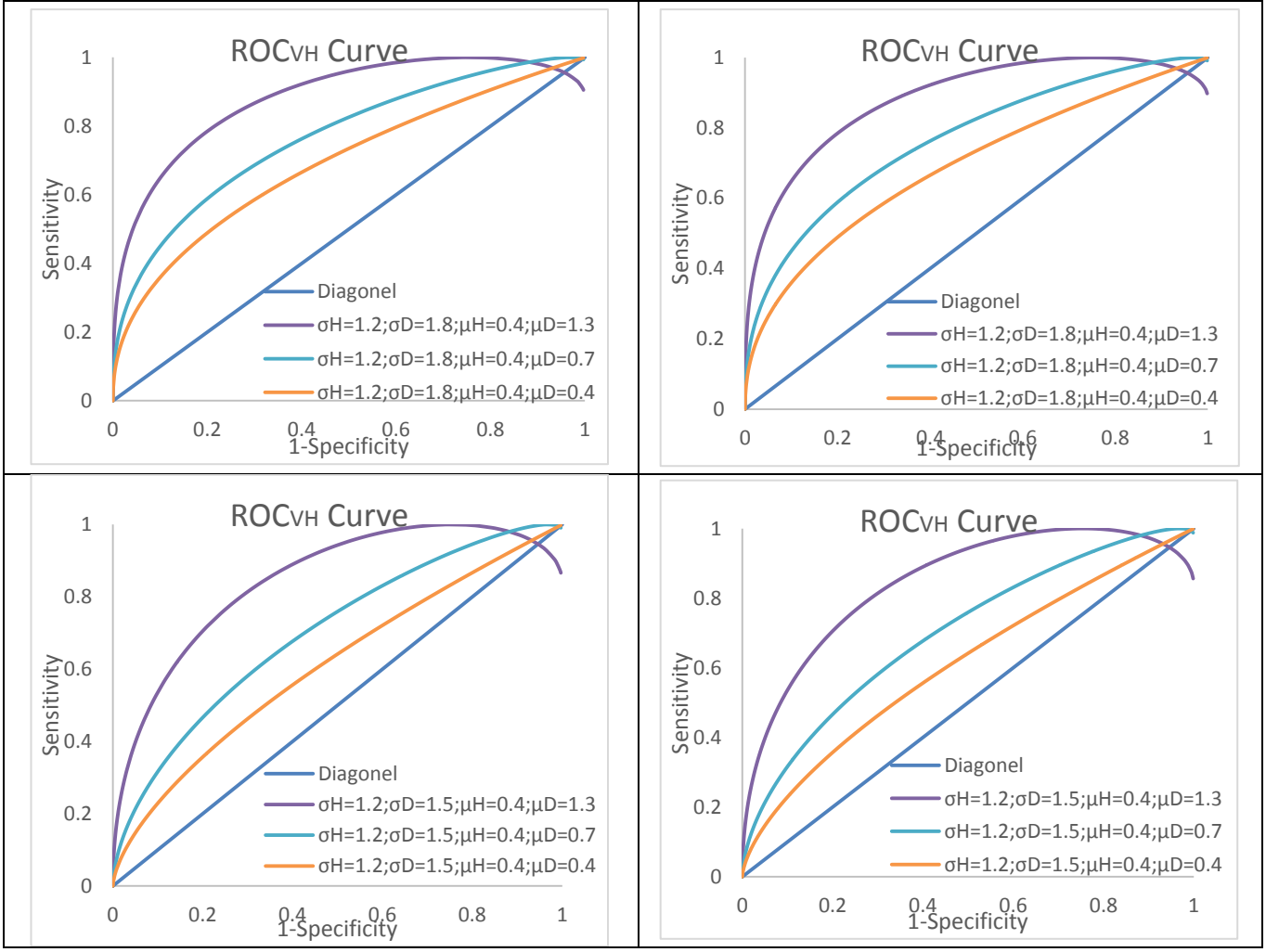
In Table 2, the values of the ROC coordinates, namely FPR and TPR, Youden's J and cut-off 'c' are reported at every combination and sample sizes. The main purpose of using and the Youden's index 'J' is to identify the cut-off which is optimal and also to note the co-ordinates of ROC curve at the optimal cut-off. The pair (FPR,TPR) will provide the necessary information about the experiment or test conducted. Apart from that it reveals the prominence of the cut-off in terms of AUC. In other words, if we considers the first combination in Table 1, the AUC is observed to be 0.8799, this means that the cut-off 2.5 is able to identify around 88% of correct cases with a sensitivity 77.9% and specificity 79.56%.Similarly interpretations can be made for the remaining combinations at different sample sizes. We can notice a minute variation in the intrinsic measures at all sample sizes.

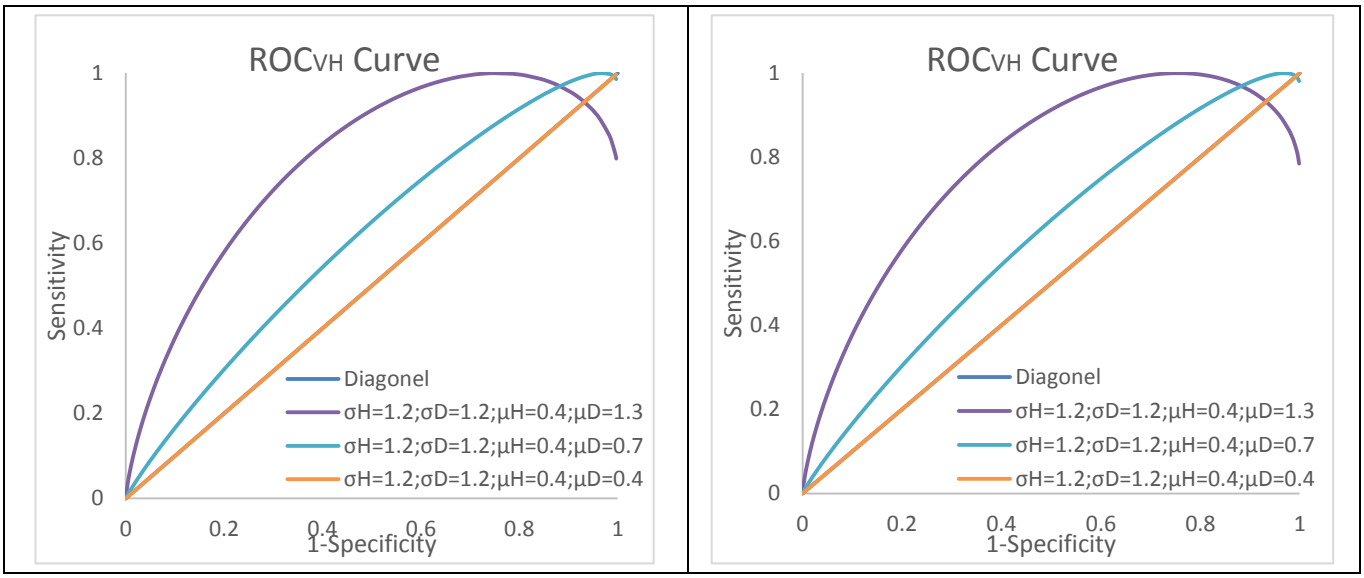
Figure 4: Plots of ROC_{VH} curve for different sample sizes and parameter values.



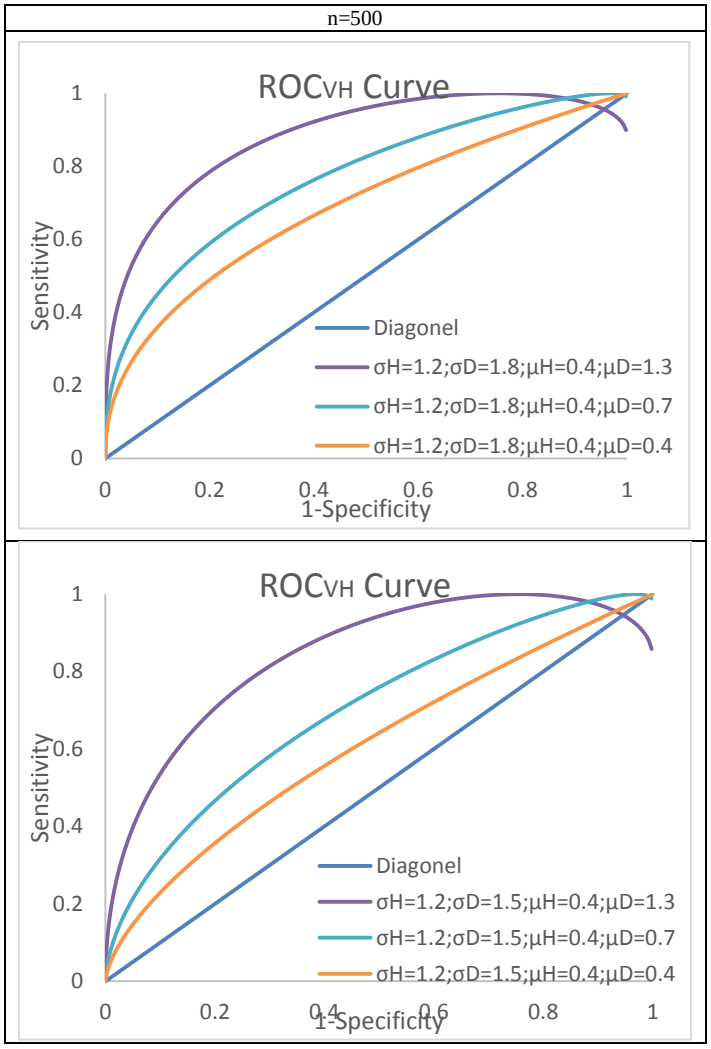


n=100	n=200

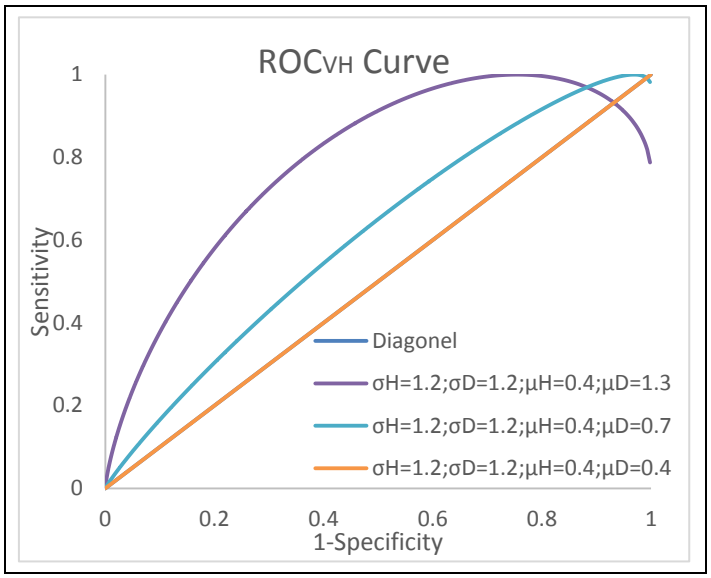




n=500



International Journal of Scientific Engineering and Science
Volume 3, Issue 11, pp. 11-18, 2019. ISSN (Online): 2456-7361



Further the ROC_{VH} Curves are plotted using the coordinates (FPR, TPR) and are depicted for all sample sizes in figure 1. The curve which is closer to the chance line illustrates the case of worst classification, the curve on the top explains the better curve scenario and the curve which is between these two depicts the moderate case. Here we have to make a note on the properties of ROC curve, i.e. the ROC curve should be concave and the coordinates (FPR,TPR) at every value should be monotonically increasing. Any curve which is not satisfying the above properties can be termed as "not proper" ROC curve and the information such as AUC, Cut-off, FPR and TPR cannot be considered for further evaluation and classification purpose.

In the simulation studies of the present work, we come across such situation, where, one of the ROC_{VH} Curve obtained at a huge mean difference possess such characteristic. The curve obtained non concavity and non-monotonic increasing is called as "Not proper" ROC_{VH} curve. However, this generates an interest in focusing towards tractability and importing restrictions on the location and scale parameters of the considered distribution.

VI. Summary and Conclusions

As there are many bi-distributional ROC models available in the literature, an attempt is made to observe the practical importance of two parameter Rayleigh distribution with location and scale parameters. The main reason for considering this distribution is that it has lot of applications in reading/analyzing the problem of signal process of different set independent wave lengths (populations). The other part is that the origin of ROC curve also started in analyzing radar signals tuning to signal detection theory. Hence the above two points gave the platform and thought to impart the mathematics of Raleigh parameters into the theory of classification. With this attempt, the mathematical expression for ROC_{VH} curve, AUC, FPR, TPR, Youden's index and cut-off are designed. Through simulation studies the behavior of the proposed curve is explained. These studies are conducted to illustrate the better, moderate and worst case scenarios in classification.

However, during such experimentation, interesting issues were observed. One violates the properties like concavity and monotonic increasing function of co-ordinates and secondly its characterization. With the above two issues it is clear that when a curve crosses the chance line and fails to attain monotonically then such a curve can be claimed as "Not proper/Improper" ROC curve, this refers to that the measures related to it cannot be taken for granted in interpreting the results and the other is to think about the characterization of the probability density function. "Not proper/Improper" ROC curves can be addressed by defining through inflection point and crossing points and the other way is to impart some restrictions on the scale parameter.

References

1. Bailey, F. C., Fritch, D. J., and Wise, N. S., (1963) "Preliminary Analysis of Bending Moments Data for Ships at Sea," Ship Structure Committee Report SSC-158.
2. Cartwright, D. E. and Longuet-Higgins (1956), M. S., "The Statistical Distribution of the Maxima of a Random Function," Proceedings, Royal Society of London, Series A, Vol. 237.
3. Kelly H. Zou, Aiyi Liu, Andriy I. Bandos, LucilaOhno-Machado, Howard E. Rockette, (2011), Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis (Chapman & Hall/CRC Biostatistics Series)
4. Krzanowski, W.J, Hand, D.J, (2009) ROC curves for Continuous Data, CRC Press
5. Lord Rayleigh, F.R.S. (1879), On the Stability, or Instability, of certain Fluid Motions, *Proceedings of the London Mathematical Society*, Volume s1-11, Issue 1, November 1879, Pages 57–72, https://doi.org/10.1112/plms/s1-11.1.57
6. Rice, S. O. (1944, 1945), "Mathematical Analysis of Random Noise," Bell System Technical Journal.
7. Watters, J. K. A., "Distribution of the Heights on Ocean Waves,"New Zealand Journal of Science and Technology, Section B., Vol. 34, No. 5, March 1953.