IJESMR

International Journal OF Engineering Sciences & Management Research

AREA UNDER THE BINORMAL ROC CURVE USING CONFIDENCE INTERVAL OF MEANS BASING ON QUARTILES

Suresh Babu. Nellore
Department of Statistics, S.V. University, Tirupati, India

Keywords:Binormal ROC cureve.

ABSTRACT

Roc curve analysis is a best statistical tool to assess the performance of test accuracy by an area under the curve (AUC). In binormal model, let X and Y be two normal populations with means µx and µy for diseased population (D) and healthy population (H). This paper emphasis, area under the binormal roc curve model and comparisons are made with the help of different AUCs basing on various possible distances (difference between population means) Dj; j=1, 2 ….9. These nine possible distances can be calculated by taking lower and upper limits of confidence interval of means, which can be computed from first and third quartiles (i.e. Q1 and Q3) from their respective normal populations. Estimation of three new method of averages namely i) Simple Average Method ii) Fixed Weights Method (FW-Method) and iii) Proportional Weights Method (PW-Method) are briefly discussed also comparisons made between them and normality is tested by P-P plot.

INTRODUCTION

The immense importance of classification problems have been extensively increasing day by day in several areas viz., engineering, medical, biological sciences, radiology, epidemiology etc. In the recent years there exists a necessity not only in classification rule but also performance of diagnostic test accuracy. Generally any classification is made on the basis of markers. Especially markers play an imperative role in the medical science, so the name gets biomarkers. However markers are used to merely classification rule but not for the performance of test accuracy. In this scenario performance of the test accuracy can be measured by a very popular statistical tool - *Roc Curve Analysis*. Receiver Operating Characteristic (ROC) curve analysis plays a vital role as a classifier performance assist tool in medicine or health related areas. ROC curve is a graphical representation of sensitivity and

(1-specificity) on XY plane. In many diagnostic problems one needs to assess the performance of a classifier or more than one classifier, roc curve analysis accomplish all requirements as one of the best statistical tool.

One of the problems in ROC curve analysis is that of evaluating the Area under the ROC Curve (AUC). The AUC can be used to know the overall accuracy of the diagnostic test. Estimation of AUC arises from different approaches such as i) Parametric approaches ii) Non-Parametric approaches and iii) Semi-Parametric approaches. This approach makes an ROC curve smoothed and also estimates AUC using binormal model when the diagnostic test results are continuous from the D and H normal populations.

In parametric approach, if we have considered two normal populations, let $X \sim N(\mu_x, \sigma^2_x)$ and $Y \sim N(\mu_Y, \sigma^2_y)$ be the two random variables (continuous) and are normally distributed with different means. Then the binormal ROC curve of the form is given by

$$ROC(t) = \Phi(a + b\,\Phi^{-1}(t))$$

where Φ is the cumulative normal distribution function and $a = \frac{\mu_x - \mu_y}{\sigma_x}$, $b = \frac{\sigma_y}{\sigma_x}$

In this parametric approach, the AUC equals the probability that a randomly selected diseased subject has diagnostic higher than a randomly selected non diseased.

Earlier, I (*Suresh Babu .N* et al.,) have proposed a new method of estimating the Binormal AUC basing on the possible distances between the means of the two normal populations using 100(1-α) % confidence intervals of means. These estimates are based on the weighted average of 9 possible estimates that arise from the confidence intervals.

In this paper , I reexamine the estimation of new possible AUC estimates from the binormal ROC model basing on 9 possible distances by replacing lower and upper limits with respective first and third quartiles and arrive more or less new improved AUC estimates attain Target AUC. And also prove that Fixed Weight Method provides very accurate approximation to the true AUC. Normality was tested among the AUC estimates by P-P plot when sample size increases.

IJESMR
International Journal OF Engineering Sciences & Management Research

AREA UNDER THE BINORMAL ROC CURVE

Let X and Y be two continuous random variables representing the test variables in D and H groups respectively, such that $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$.

In parametric approach, area under the binormal ROC curve is the summary index of the performance of the diagnostic test denoted by AUC. If X and Y are the scores allotted to randomly and independently chosen individuals from D and H populations respectively.

Then AUC can be defined as $\quad AUC = P(X > Y)$

$$AUC = P(X - Y > 0)$$

If $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ then $X - Y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$. Hence if Z denotes a standard normal random variate,

$$AUC = P\left(Z > 0 - \left(\frac{\mu x - \mu y}{\sqrt{\sigma x2 + \sigma y2}}\right)\right)$$

$$= 1 - \Phi\left(\frac{-\mu x + \mu y}{\sqrt{\sigma x2 + \sigma y2}}\right)$$

$$= \Phi\left(\frac{\mu x - \mu y}{\sqrt{\sigma x2 + \sigma y2}}\right)$$

Dividing numerator and denominator by σ_x, then

$$AUC = \Phi\left(\frac{\frac{\mu x - \mu y}{\sigma x}}{\frac{\sqrt{\sigma x2 + \sigma y2}}{\sigma x}}\right)$$

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b2}}\right)$$

The above expression is the simplest form of AUC in parametric approach.

The binormal model gives an expression for the TPR (Sensitivity) as a function of the FPR (1- Specificity) expressed in terms of cumulative normal probability.

Faraggi and Reiser (2002) have shown that with binormal model, the AUC is given by

$$AUC = \Phi\left(\frac{\mu_X - \mu_y}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right) \tag{1}$$

Let $\hat{\mu}_x$ and $\hat{\mu}_y$ represent the estimated means and s_x^2, s_y^2 represent the sample variances of X and Y respectively. Then the estimated AUC is obtained by replacing the parameters with their sample estimates. This gives

$$\widehat{AUC} = \Phi\left(\frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{s_x^2 + s_y^2}}\right) \tag{2}$$

Suppose, we wish to develop an ROC curve such that it has a predetermined AUC denoted by AUC*, like 0.9, 0.8 etc. If $Z^* = \Phi^{-1}(AUC^*)$ will be denotes the standard normal deviate corresponding to AUC^* then Faraggi and Reiser (2002) have shown that the estimated mean of the diseased group is

$$\hat{\mu}_X = \hat{\mu}_y + Z^* \sqrt{s_x^2 + s_y^2} \tag{3}$$

Using this $\hat{\mu}_X$ the AUC can be estimated.

Let D denotes the distance between the means of the two normal populations. Then $D = (\hat{\mu}_X - \hat{\mu}_Y)$ and estimated AUC takes the form

$$\Phi\left(\frac{D}{\sqrt{s_x^2 + s_y^2}}\right) \tag{4}$$

Since the true population mean can be anywhere in confidence interval (CI) for each of the two populations the true difference between the means depends on the upper and lower limits of the intervals.

THE POSSIBLE DISTANCES BETWEEN THE MEANS BASING ON QUARTILES

The true mean of the population most probably may lie at the middle of confidence interval. Suppose including the lower and upper confidence limits can be taken into the consideration then there exists different distances between them which provides probable AUC estimates as same as estimation from the difference between true means. Lower and upper confidence limits of a confidence interval for means can be obtained from *first quartile* (Q1) and *third quartile* (Q3) from their respective normal populations.

For a diseased normal population (D), the first and third quartiles are given by

IJESMR
International Journal OF Engineering Sciences & Management Research

$$\text{Mean} \pm (0.6745)\ .\text{Standard Error of mean i.e } \hat{\mu}_x \pm (0.6745)\ \text{S.E} (\hat{\mu}_x)$$

For a healthy normal population (H), the first and third quartiles are given by

$$\text{Mean} \pm (0.6745)\ .\text{Standard Error of mean i.e } \hat{\mu}_y \pm (0.6745)\ \text{S.E} (\hat{\mu}_y)$$

For the H group let the confidence interval be (L_1, U_1) and for the D groups (L_2, U2) where the limits are as defined below.

Limit	Value
L_1	$Q_{1Y} = \hat{\mu}_y - (0.6745)\left(\frac{s_Y}{\sqrt{n_1}}\right)$
U_1	$Q_{3Y} = \hat{\mu}_y + (0.6745\left(\frac{s_Y}{\sqrt{n_1}}\right)$
L_2	$Q_{1X} = \hat{\mu}_x - (0.6745)\left(\frac{s_X}{\sqrt{n_2}}\right)$
U_2	$Q_{3X} = \hat{\mu}_x + (06745)\left(\frac{s_X}{\sqrt{n_2}}\right)$

where n_1, n_2 are the sizes and s_X, s_y are the s.d's of the healthy and diseased groups respectively.

The 9 possible true distances between the means basing on the limits are shown inTable-1.

D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
$(L_2 - L_1)$	$(L_2 - \hat{\mu}_y)$	$(L_2 - U_1)$	$(\hat{\mu}_x - L_1)$	$(\hat{\mu}_x - \hat{\mu}_y)$	$(\hat{\mu}_x - U_1)$	$(U_2 - L_1)$	$(U_2 - \hat{\mu}_y)$	$(U_2 - U_1)$

Table-1: List of possible distances between the means

both the groups, the true value of the mean can be anything within confidence interval.

These 9 distances represent nine mutually exclusive cases and each one occurs with some probability. The usual method of computing AUC is based on D5.

Now with the j^{th} distance the estimated AUC, denoted by A_j becomes

$$A_j = \Phi\left(\frac{D_j}{\sqrt{s_X^2 + s_Y^2}}\right), j = 1, 2 \ldots\ldots 9 \qquad (5)$$

Lasko et. al (2005) have shown that the estimated variance of the AUC is given by

$$\hat{V}(A_j) = \frac{1}{n_1 + n_2}\left[\frac{1}{s_X + s_Y}\left(\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}\right) + \frac{A_j^2}{2(s_X^2 + s_Y^2)^2}\left(\frac{s_X^4}{n_1 - 1} + \frac{s_Y^2}{n_2 - 1}\right)\right] \qquad (6)$$

NEW SUMMARIZED ESTIMATES OF AUC

Sarma, et al. (2010) has proposed three new estimates based on weighted average of the 9 individual A_j values. The three methods differ only in terms of defining the weights as discussed below.

i) Simple Average Method

Among three methods, it is the simplest one. The estimation of new AUC through this method is basing on weights. Then the estimated AUC is the sum of the product of the weights and the corresponding AUCs and is given by

$$AUC_{AVG} = \sum_{j=1}^{9} W_j A_j \quad \text{where } W_j = 1/9 \ \forall\, j = 1, 2 \ldots 9$$

where A_j for j = 1, 2, ….., 9 are the AUCs obtained from the confidence intervals of means given in (3.5). It can be seen that $\sum_{j=1}^{9} W_j$

The variance of estimated AUC is computed as below

IJESMR
International Journal OF Engineering Sciences & Management Research

V (AUC$_{AVG}$) = $\sum_{j=1}^{9} W_j^2 V(A_j)$

ii) Fixed Weights Method (FW-Method)

In this method, among all weights, one is to be fixed as $W_5 = 0.5$ because Δ_5 occurs with high probability and the remaining 8 weights are to be calculated by the following formula

$$W_j = \frac{0.5}{8} \ \forall j \neq 5$$

The new estimator by the fixed weights method (FW-Method) is the sum of the product of the weights including W_5 with their corresponding AUCs and is defined as below

$AUC_{FW} = \sum_{j\neq 1}^{9} W_j A_j + \frac{0.5}{8} A_5$

where A_j ; j = 1, 2, ….., 9 are s are the different AUCs obtained from the confidence intervals of means. It can be seen that $\sum_{j=1}^{9} W_j = 1$.For fixed weights method (FW-Method), the estimation of the variance is similar to that of the computation of the variance in the simple average method. So, the variance of the estimated AUC i.e. V (AUC$_{FW}$) is the given by

V (AUC$_{Fw}$) = $\sum_{j=1}^{9} W_j^2 V(A_j)$

iii) Proportional Weights Method (PW-Method)

This method differs from both the simple average method and fixed weight method. In this method the weights are taken as proportional to the difference in the means used in equation(5) such that Wj = $\frac{1}{\Delta_j}$ $\forall$ j = 1, 2,……, 9. It means the weights are increase by proportional to Δ. The new estimator by propositional weight method (PW-Method) is defined by the formula $AUC_{PW} = \frac{\sum_{j=1}^{9} W_j A_j}{\sum_{j=1}^{9} W_j}$ where Aj for j = 1, 2, ….., 9 are the different AUCs obtained from the confidence intervals of means. The variance of the estimated new estimator of the AUC by propositional weights method will be

V (AUC$_{AVG}$) = $\sum_{j=1}^{9} r_j^2 V(A_j)$ where $r_j = W_j \{\sum_{j=1}^{9} W_j\}^{-1}$

EMPIRICAL ILLUSTRATIONS

In binormal model, the AUC can be estimated by taking difference between the means of two normal populations i.e. from the means of diseased (D) population μ_x and healthy(H) population μ_Y. While estimating the AUC, rather than the consideration of only one difference between the means, 9 possible distances, from the true means basing on confidence limits, can be computed from first quartile(Q1), second quartile(Q2 or Mean) and third quartile (Q3) are to be considered to estimate new improved AUCs (A_j ; j=1,2,…….9). Later basing on all possible A_j's, an estimation of three new average methods can be carried out by using weights i.e. $\sum_{j=1}^{9} W_j$ and to make the comparisons between them. Regarding normality between the obtained AUCs is explained by P-P plot with R^2 value.

The following table shows Normality fit for the estimated AUC's at Target AUC = 0.9

Target AUC = 0.9, μ_Y = 85			
S. No	**Inputs**	**Estimated AUC values**	**Normality Fit by P-P Plot**
1	n_1 = 5 ,n_2 = 5 μ_x = 103.12 s_x = 10, s_y = 10	{A_1..A_4} = {0.9000,0.8573,0.8037,0.9325}, {A5} = 0.9000 {A_6..A_9} = {0.8573,0.9562,0.9325,0.9000} AUC_{AVG} = 0.8933, AUC_{FW} = 0.8962 , AUC_{PW} = 0.8841	Y= 0.046X+0.893 R^2 = 0.934

2	$n_1 = 10, n_2 = 15$ $\mu_x = 103.12$ $s_x = 10, s_y = 10$	$\{A_1 \ldots A_4\} = \{0.8951, 0.8709, 0.8432, 0.9199\}$ $\{A5\} = 0.9000$ $\{A_6 \ldots A_9\} = \{0.8767, 0.9401, 0.9240, 0.9048\}$ $AUC_{AVG} = 0.8972$, $AUC_{FW} = 0.8984$, $AUC_{PW} = 0.8936$	$Y=0.03X+0.897$ $R^2= 0.978$
3	$n_1 = 15, n_2 = 10$ $\mu_x = 103.12$ $s_x = 10, s_y = 10$	$\{A_1 \ldots A_4\} = \{0.9048, 0.8767, 0.8432, 0.9240\}$ $\{A5\} = 0.9000$ $\{A_6 \ldots A_9\} = \{0.8709, 0.9401, 0.9199, 0.8951\}$ $AUC_{AVG} = 0.8972$, $AUC_{FW} = 0.8984$, $AUC_{PW} = 0.8936$	$Y=0.03X+0.897$ $R^2= 0.978$
4	$n_1 = 20, n_2 = 15$ $\mu_x = 103.12$ $s_x = 10, s_y = 10$	$\{A_1 \ldots A_4\} = \{0.9029, 0.8800, 0.8535, 0.9199\}$ $\{A5\} = 0.9000$ $\{A_6 \ldots A_9\} = \{0.8767, 0.9346, 0.9175, 0.8971\}$ $AUC_{AVG} = 0.8980$, $AUC_{FW} = 0.8989$, $AUC_{PW} = 0.8955$	$Y=0.024X+0.898$ $R^2= 0.986$
5	$n_1 = 20, n_2 = 25$ $\mu_x = 103.12$ $s_x = 10, s_y = 10$	$\{A_1 \ldots A_4\} = \{0.8908, 0.8762, 0.8602, 0.9125\}$ $\{A5\} = 0.9000$ $\{A_6 \ldots A_9\} = \{0.8862, 0.9308, 0.9203, 0.9086\}$ $AUC_{AVG} = 0.8984$, $AUC_{FW} = 0.8991$, $AUC_{PW} = 0.8964$	$Y=0.022X+0.898$ $R^2= 0.990$

Table-2 : Normality fit for the Estimated AUC's at Target AUC = 0.9

From the above table, it can be seen that R^2 values have been increased as sample sizes increases accordingly and AUC at Fixed Weight Method appears as best average among the others.

CONCLUSIONS

At fixed true mean values μ_x, μ_Y and standard deviations S_x, S_y , I have been derived new improved AUC (A_j; j=1,2,…….9) estimates using confidence interval of means basing on confidence limits which can be which can be obtained from first quartile(Q1), second quartile(Q2 or Mean) and third quartiles(Q3). Among all obtained AUCs, at least more than 5 AUCs toward the Target AUC = 0.9. Among the new summarized estimates, AUC at Fixed Weight Method provides better average than other two methods. Between all possible sample AUC estimates, the normality is tested by P-P plot, which shows R^2 value increases as the sample sizes (n_1 and n_2) are increased and also found that by interchanging n_1 and n_2, R^2 would remain same. However the simple average method has the lowest variance among the three.

REFERENCES

1. Metz, C. E., 1978, "Basic Principles of ROC analysis", Seminars in Nuclear Medicine, 8, pp. 283 – 298
2. Hanley, A., James, and Barbara, J. Mc., Neil, 1982, "A Meaning and Use of the area under a Receiver Operating Characteristic (ROC) curves", Radiology, 143, pp. 29 - 36.
3. Lloyd, C. J., 1998, "The Use of smoothed ROC curves to summarize and compare diagnostic systems", Journal of American Statistical Association, 93, pp. 1356 – 1364.
4. Metz et al., 1998, "Maximum Likelihood Estimation of receiver operating characteristic (ROC) curves continuously distributed data", Statistics in Medicine, 17, pp. 1033-1053
5. Pepe, MS., 2000, "Receiver operating characteristic methodology", Journal of American Statistical Association, 95, pp. 308 – 311.
6. Farragi David, and Benjamin Reiser, 2002, "Estimation of the area under the ROC curve, Statistic in Medicine", 21, pp. 3093 - 3106.
7. Lasko, T. A., 2005, "The use of receiver operating characteristics curves in biomedical informatics", Journal of Biomedical Informatics, vol. 38, pp. 404 –415.
8. Krzanowski, J. Wojtek., and David, J. Hand., 2009, "ROC Curves for Continuous Data".

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR

International Journal OF Engineering Sciences & Management Research

9. Sarma K.V.S.,et., al 2010, "Estimation of the Area under the ROC Curve using Confidence Intervals of Mean", ANU Journal of Physical Sciences.
10. Suresh Babu.N. and Sarma, K. V. S., 2011. "On the Comparison of Binormal ROC Curves using Confidence Interval of Means", International Journal of Statistics and Analysis, Vol. 1, pp. 213-221.
11. Suresh Babu. N and Sarma K.V.S. (2013), On the Estimation of Area under Binormal ROC curve using Confidence Intervals for Means and Variances. International Journal of Statistics and Systems**,** Volume 8, Number 3 (2013), pp. 251-260.
12. Suresh Babu. **N** and Sarma K. V. S. (2013), Estimation of AUC of the Binormal ROC curve using confidence intervals of Variances. International Journal of Advanced Computer and Mathematical Sciences ,Volume 4, issue 4, (2013), pp. 285-294.