*J.Natn.Sci.Foundation Sri Lanka 2015 43 (4):357-367*
DOI: http://dx.doi.org/10.4038/jnsfsr.v43i4.7970

**RESEARCH ARTICLE**

# Simulation study of a novel method for comparing more than two independent receiver operating characteristic (ROC) curves based on the area under the curves (AUCs)

**A.N. Meyen and M.R. Sooriyarachchi***
*Department of Statistics, Faculty of Science, University of Colombo, Colombo 03.*

Revised: 29 May 2015; Accepted: 25 June 2015

**Abstract:** Receiver operating characteristic (ROC) graphs are useful for organising binary classifiers and visualising their performance. In order to compare classifiers it may be needed to reduce the ROC performance to a single scalar value representing expected performance. Such a commonly used summary statistic is the area under the curve (AUC) of the ROC curve. The AUCs can be estimated either parametrically or non-parametrically. The parametric approach assumes that the signal present (positive) and signal absent (negative) groups can be represented as two overlapping Gaussian distributions. If the observations of two or more ROC curves are obtained from the same region of interest, their AUCs are considered to be correlated.

A novel asymptotic test for comparing multiple AUCs of several ROC curves was proposed by Meyen and Sooriyarachchi in 2014, and it was of interest to study the behaviour of the test statistic for various sample sizes and varying degrees of overlap between the Gaussian distributions *via* a simulation study. Hence this study was carried out to test the properties of the test statistic when the AUCs were estimated parametrically by Dorfman and Alf's method. This simulation was carried out for the case where the AUCs are independent.

Inferences were made regarding the distribution of the test statistic for various sample sizes. The test statistic performed better when the spread between the two Gaussian distributions increased, while the test statistic was valid with respect to sample sizes above 100 when 2 ROC curves were being compared simultaneously.

**Keywords:** Area under the curve (AUC), beta distribution, likelihood ratio test (LRT), receiving operating characteristic (ROC) curve, score test.

## INTRODUCTION

A ROC curve is simply a graphical plot, which illustrates the performance of a binary classifier system as its discrimination threshold is varied. A variety of summary indices have been proposed as a single measure or simplification for considering the entire curve. One of the most common measures used for summarising the performance of the diagnostic modalities is the value of the area under the curve (AUC), which ranges from 0 to 1, where the higher the value of the AUC a better discrimination power is implied (Fawcett, 2006). There are parametric, nonparametric and semi parametric methods of estimating the area under a ROC curve. The method of estimating the AUC depends on whether the data used is either continuous or in rating form.

The application areas of receiver operating characteristic (ROC) curves range from medical imaging and radiology to machine learning and data mining. In practice it is often required to compare several alternative diagnostic tests and this entails the comparison of several AUCs under the ROC curves. At present this is achieved by comparing all pairwise combinations but this procedure has some serious drawbacks.

To make things clear consider an example from medical testing. Suppose it is required to determine whether suspected coronary artery disease (CAD) patients are positive or negative with respect to CAD. The gold standard test for determining this is an angiogram.

* Corresponding author (roshini@mail.cmb.ac.lk)

However an angiogram is a very expensive diagnostic tool and specially in developing countries like Sri Lanka only a few persons can afford this test. Two substitute tests for diagnosing CAD are the coronary stress test (CST) and non-invasive carotid artery ultrasound of the neck (NICAUN). In order to compare the performance of the two substitute tests with respect to that of an angiogram, two independent groups of patients of size m and n, respectively can be selected. One group will be given the angiogram and CST and the other group will be given the angiogram and NICAUN. The comparison of each substitute test with respect to the angiogram will be done using the AUCs of the respective ROC curves, which are constructed by plotting the sensitivity versus (1-specificity) for varying discrimination threshold values (Fawcett, 2006). As the two groups of patients are independent the AUCs will also be independent. This corresponds to the comparison of two independent AUCs.

An asymptotic test for comparing several AUCs under the curves has been proposed by Meyen and Sooriyarachchi (2014) and it is of interest to study the effectiveness of this method and validate its properties. Therefore, the motivation of this study arose to address the need for such proper simulation based analysis.

No proper study concerning the properties of the asymptotic test proposed by Meyen and Sooriyarachchi (2014) has been carried out yet. An important assumption of this test is that the AUCs are multivariate normally distributed. Therefore it was of interest to study the properties of the test under different circumstances, as in many real life applications the AUCs may not be multivariate normally distributed. The study determined the appropriate sample sizes and checked the null distribution of the proposed statistic, whilst identifying its limitations. Here the sample size corresponds to the size of the sample used for generating the ROC curve and its AUC. In the simulation study each sample generates a ROC curve. The type of classifier used here is binary.

To carry out this study it was necessary to implement a programme of the Dorfman and Alf method of maximum likelihood estimation (Dorfman & Alf, 1969) in order to estimate the parameters needed to calculate the AUC. The language used for this implementation was C. Furthermore, the analysis of AUCs of ROC curves was carried out for categorical rating-scale data where 3 categories were considered for uncorrelated ROC curves with two ROC curves being compared at once. As per previous research carried out by Cleeves (2002), it was decided to simulate data for sample sizes of 20, 50, 100, 120,140, 250 and 500 observations in total (i.e. sample sizes of 10, 25, 50, 60, 70, 125 and 250 with respect to the positive and negative groups, respectively). The degree of overlap of the two populations was controlled by generating observations from Gaussian distributions whose means differed by 0.5, 0.75 and 1 standard deviations. Additionally, data were simulated assuming equal variances in the two subpopulations, and assuming distributions with standard deviation ratios of 1:1.5. The mean and the standard deviation of the negative population were taken to be 0.5 and 0.1, respectively. Each of the 42 combinations of sample size, degree of overlap, and standard deviation ratio was replicated 1000 times. After simulation for the various sample sizes, likelihood ratio (LR) and score tests based on the beta distribution, which the proposed test statistic should asymptotically follow, were applied to the sample of test statistics thus formed for the different sample sizes. When the sample size was very small (20) confidence intervals based on the above tests could not be constructed as large sample approximations were invalid.

## METHODS AND MATERIALS

### Binormal ROC curves

In order to understand the construction of ROC curves it is important to understand the signal-detection paradigm from which it is derived. According to Grey and Morgan (1972), the signal-detection paradigm consists simply of a subject successively choosing between signal present (with background noise), SN, or no signal present (just noise), N. The model now usually assumed is that the stimulus sets up a response within the subject that can be represented as a continuous, uni-dimensional random variable $X$ with probability distribution function, $F_N(x)$ as the stimulus noise and $F_{SN}(x)$ as the stimulus signal and noise.

Typically $F_N(x) = F(x)$, $F_{SN}(x) = F(bx - a)$ for some $F$, i.e. the distributions are two-parameter and of the same form. The ROC curves simulated in this study use $F$=**N**(.,.), where **N**(.,.) denotes the normal distribution. If the subject is required to respond just Yes (signal present) or No (no signal present) then the model assumes that the tests for the underlying distribution by setting up a criterion ($z$, say) and responding "Yes" if the response $X$ is, say greater than $z$ and "No" if $X$ is less than $z$. Then it follows that,

$$P(Hit) = P\left(Yes/_{SN}\right) = 1 - F(bz - a) \quad ...(1)$$

$$P(False\ alarm) = P\left(Yes/_{N} = 1 - F(z)\right) \quad ...(2)$$

Varying $z$ gives the ROC curve [$P(Hit)$ against $P(False\ alarm)$], defined by $a$ and $b$ that is taken to represent the subject's performance. In order to estimate $b$, the experiment must be elaborated, the two standard means of doing so being either a repetition of the above with different $z$ values (which could be obtained by manipulating a payoff matrix) or by requiring that the subject grade the response (for example, Yes, Sure or No, Fairly sure). It is the second elaboration that is considered here: by analogy with the above the model then assumes $n$ criteria, $z_k$ , $k = 1,2, \ldots , n$ (In addition it is convenient to define $z_0 = -\infty$ and $z_{n+1} = +\infty$ ) so that the subject makes the overt response $R_i$, if the latent response $X$ falls in the interval:

$$z_{i-1} \leq X \leq z_i \ , \qquad i = 1,2, \ldots , n+1$$

The values of $a$ and $b$ along with other parameters of the ROC curve were estimated using the method of scoring proposed by Dorfman and Alf (1969).

**Problems of iteration**

The start for the initial iteration was used as the parameter estimates of the simple linear regression as outlined in Grey and Morgan (1972). Iteration continues until either, two successive iterates differ by less than $10^{-3}$ in all of their components and the final iterate is a possible solution (i.e. $\hat{z}_1 \leq \hat{z}_2 \leq \ldots \leq \hat{z}_n$ and $b > 0$). The degenerate solution for the parameter estimates of the ROC curve can occur from empty cells in the data matrix. Therefore in order to overcome the problem of degeneracy similar to Dorfman and Berbaum (1995) the programme developed adds a small positive constant in order to avoid degeneracy in the case of empty cells.

**Determination of the threshold of the categories of the rating data**

The threshold of the categories was decided by dividing the false positive fraction (FPF) into equal fractions, so that for example if 3 category rating data were considered the thresholds or cut points of the categories would be 0.33 and 0.66. If the mean and standard deviation for the signal present (with background noise) cases are denoted by $\mu_{SN}, \sigma_{SN}$ , and the mean and the standard deviation for the noise only cases are denoted as $\mu_N, \sigma_N$ , assuming without loss of generality that $\mu_N \leq \mu_{SN}$ and $c$ represent the cut point on the decision variable axis such that a case is classified as 'negative' if $x \leq C$ and positive if $x > C$, then the values for the FPF is given as follows where $\Phi(.)$ denotes the cumulative standard normal distribution (Metz & Pan, 1999):

$$FPF(c) = 1 - F\left(\frac{c-\mu_N}{\sigma_N}\right) = 1 - \Phi\left(\frac{c-\mu_N}{\sigma_N}\right) = \Phi\left(\frac{\mu_N - c}{\sigma_N}\right) \quad ...(3)$$

**Calculation of the AUC and variance of the AUC**

According to Metz *et al*. (1998), when the mean and standard deviation of the signal present (with background noise) cases, and the mean and the standard deviation of the noise only cases are considered in their usual notation, the values of $a$ and $b$ are given as follows in equations (4) and (5).

$$a = \frac{\mu_{SN} - \mu_N}{\sigma_{SN}} \quad ...(4)$$

$$b = \frac{\sigma_N}{\sigma_{SN}} \quad ...(5)$$

It is possible to obtain the AUC of a ROC curve using the following formula where $\Phi(.)$ denotes the cumulative standard normal distribution as shown in equation (6).

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \quad ...(6)$$

In order to calculate the variance of the AUC, the delta method, which is described in detail in Casella and Berger (2002) is made use of, giving the formula as follows for the variance.

$$Var(\widehat{AUC}) = \left(\frac{\partial AUC}{\partial a}\right)^2 var(a) + \left(\frac{\partial AUC}{\partial b}\right)^2 var(b) + 2\left(\frac{\partial AUC}{\partial a}\right)\left(\frac{\partial AUC}{\partial b}\right) cov(a,b) \quad ...(7)$$

**Proposed test statistic**

The test developed by Meyen and Sooriyarachchi (2014) has been developed using various results from multivariate statistics along with the properties of ROC curves. Since the test is derived for the asymptotic distribution of the statistic it is of interest to study the small sample behaviour of the test statistic as well. In the following sections is the derivation of the test developed by Meyen and Sooriyarachchi (2014).

Let, $\underline{AUC} = \begin{pmatrix} AUC_1 \\ AUC_2 \\ \vdots \\ AUC_p \end{pmatrix}$ which is a p×1 vector, where $AUC_i$ denotes the $AUC$ of the $i^{th}$ ROC curve.

Let $\underline{\widehat{AUC}}$ be an estimate of $\underline{AUC}$, let $\mu$ be the expected value of $\underline{\widehat{AUC}}$ and let $\underline{\Sigma}$ be the associated variance-covariance matrix of $\underline{\widehat{AUC}}$. Then as $\underline{\widehat{AUC}}$ is the Dorfman and Alf (1969) maximum likelihood estimate (MLE) of $\underline{AUC}$ and as MLE's are asymptotically normal, for large samples $\underline{\widehat{AUC}} \sim N_p(\mu, \Sigma)$. Furthermore, if the estimate $\underline{\widehat{AUC}}$ of $\underline{AUC}$ of a ROC curve is made up of the sum of $n$ independent quantities where $n$ is a function of $n_1$ (the number of positive responses) and $n_2$ (the number of negative responses) according to Vergara *et al*. (2008). Then $\underline{\widehat{AUC}}$ is made up of $n_1 n_2$ quantities of which $n = \min(n_1, n_2)$ are independent. Thus $n$ is the number associated with $\underline{\widehat{AUC}}$. $\underline{\hat{\Sigma}}$ is the (Dorfman & Alf, 1969) MLE of the covariance matrix $\underline{\Sigma}$ of $\underline{\widehat{AUC}}$. According to Mardia *et al*. (1979), the sampling distribution of the MLE of the $(\underline{\widehat{AUC}} \mid \mu)(\underline{\widehat{AUC}} - \mu)'$ matrix is asymptotically $W_p(\underline{\Sigma}, \mathbf{n})$ as $\underline{\widehat{AUC}}$ has an asymptotic multivariate normal distribution. Therefore, $\underline{\hat{\Sigma}} \sim W_p(\underline{\Sigma}, \mathbf{n})$.

It is needed to test the null hypothesis $H_0$ that all $\underline{AUC}$ s are same on average versus the alternative hypothesis $H_1$ that all the $\underline{AUC}$ s are not the same on average.

i.e. $H_0$: $\underline{\mu} = \underline{K} =$ constant vector versus $H_1$: $\underline{\mu} \neq \underline{K}$ It is possible to estimate $\underline{K}$ as the simple average of $\underline{\widehat{AUC}}$ (i.e. the simple average of the individual $\widehat{AUC}_i'$s.) Therefore $\underline{K}$ can be estimated by $\overline{K}$ (under $H_0$) where,

$$\overline{K} = \frac{\sum_{i=1}^{p} \widehat{AUC}_i}{p} \qquad ...(8)$$

As $\underline{K}$ is not known it has to be estimated. From (Hotelling, 1947) the general form of the Hotelling's $T^2$ statistic is as follows,

$$T_G^2 = (\underline{\widehat{AUC}} - \overline{K})' \underline{\Sigma}^{-1} (\underline{\widehat{AUC}} - \overline{K}) \qquad ...(9)$$

The dimensionality $p$ needs to be reduced by 1 for estimating $\overline{K}$. Therefore taking $q = p - 1$ instead of for large samples gives the following,

$$T_G^2 \frac{n}{(n-1)^2} \sim Beta\left(\frac{q}{2}, \frac{n-q-1}{2}\right) \qquad ...(10)$$

Here $p$ is the number of AUCs and $n$ is the number of independent quantities used to calculate the AUCs. For the case of large samples (large $n_1$ and $n_1$) $n$ will be large. The test statistic $T_G^2$ can be used to test $H_0$.

Confidence intervals for the beta distribution are fit under MATLAB as described in Hahn and Shapiro (1994).

**Application of the likelihood ratio and score tests for the test statistic**

Since the test statistic is supposed to asymptotically follow a beta distribution it is necessary to check the hypothesis that it comes from a beta distribution with given parameters, which can be calculated for the ROC curves generated. Apart from this confidence intervals also need to be constructed for the maximum likelihood estimates (MLEs) fitted to the data under the hypothesis that they are beta distributed.

Consider the standard form of the $beta(p, q)$ distribution with shape parameters $p > 0, q > 0$ and support on $[0,1]$. The beta distribution includes some well-known distributions as special cases, such as the uniform distribution $(p = q = 1)$ and the power distribution $(p = 1 \; or \; q = 1)$. The total log-likelihood function for $p$,$q$ based on a random sample $y_1$, ..., $y_n$ of size $n$ can be written as given in equation (10) (Maia *et al*., 2003).

$$l(p, q) = n(p-1)\log(g_1) + n(q-1)$$

$$\log(g_2) + n\log\left\{\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\right\} \qquad ...(11)$$

where $g_1$ and $g_2$ are the geometric means of the $y_i$ s and $[(1-y]_i)$ s respectively, and $\Gamma(.)$ is the gamma function. The log-likelihood function given above is regular (Maia *et al*., 2003) with respect to all $p$ and $q$ derivatives up to the fourth order. The score function is

$$U = n\begin{pmatrix}\psi(p+q) - \psi(p) + \log(g_1) \\ \psi(p+q) - \psi(p) + \log(g_2)\end{pmatrix} \qquad ...(12)$$

where $\psi(.)$ is the digamma function, and the observed Fisher's information matrix for $p$ and $q$ is given by

$$K = -n\begin{pmatrix}\psi'(p+q) - \psi'(p) & \psi'(p+q) \\ \psi'(p+q) & \psi'(p+q) - \psi'(q)\end{pmatrix} \qquad ...(13)$$

Where $\psi'(x) = d\psi(x)/dx$ is the trigamma function. The maximum likelihood estimates (MLEs) $\hat{p}, \hat{q}$ of $p$ and $q$ have no closed-form expressions and can only be obtained from an iterative process.

Consider testing the null hypothesis $H_0$:$p = p^{(0)}$, $q = q^{(0)}$ where $p^{(0)}$ and $q^{(0)}$ are specified values tested against the alternative hypothesis $H_1$: $H_0$ *is false* . Then according to Maia *et al*. (2003) the likelihood ratio (LR) statistic for testing $H_0$ versus $H_1$ is given by,

$$T_1 = 2n(\hat{p} - p^{(0)})\log(g1) + 2n(\hat{q} - q^{(0)})$$

$$+ 2n\ log\left\{\frac{\Gamma(\hat{p}+\hat{q})\Gamma(p^{(0)})\Gamma(q^{(0)})}{\Gamma(p^{(0)}+q^{(0)})\Gamma(\hat{p})\Gamma(\hat{q})}\right\}$$

In large sample $T_1$ has under the null hypothesis approximately a chi squared distribution with two degrees of freedom.

For testing $H_0$ the score statistic is particularly attractive since it does not involve estimation under the alternative hypothesis $H_1$, but only requires the evaluation of the score function and the observed information matrix under the null hypothesis. The score statistic for the null hypothesis is given by,

$$T_2 = \tilde{U}'\tilde{K}^{-1}\tilde{U}$$

where all quantities with tilde represent estimates evaluated at the null hypothesis. The asymptotic chi-squared distribution of the LR and score statistics is used to test statistical hypotheses since their exact distributions are usually unknown.

## RESULTS

### Analysis of 3 category rating-method data for 20 individuals

When the AUCs of the generated ROC curves were applied to the test statistic in this case it was of interest to note that some of the values obtained were lying outside the uppermost boundary of the beta distribution [ i.e. the value 1 of the interval (0,1)]. This simply means that the large sample theory is not applicable with this sample size. Thus no confidence intervals are calculated for this case.

Table 1 gives the results of the analysis of 3 category rating-method data for 20 individuals.

### Analysis of 3 category rating-method data for 50 individuals

When the AUCs of the generated ROC curves were applied to the test statistic, it was noted in the case when the AUCs were distributed with means 0.75 standard deviations apart and standard deviations in the ratio 1:1, that it was not possible to apply the likelihood ratio and score tests, as the geometric mean of the AUCs could not be numerically represented. Apart from the particular instance it was possible to apply the likelihood ratio test and score tests to the other cases. However in all the other cases the values of the statistics obtained were high, which led to the rejection of the hypothesis that the AUCs came from beta distributions with the given parameters.

**Table 1:** Analysis of 3 category rating-method data for 20 individuals

| | Means 0.5 standard deviations apart | | Means 0.75 standard deviations apart | | Means 1 standard deviation apart | |
|---|---|---|---|---|---|---|
| | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 |
| Observed MLE of $q/2$ | 0.2019 | 0.172 | 0.2422 | 0.2254 | 0.2496 | 0.2775 |
| Observed MLE of $(n-q-1)/2$ | 1.1374 | 0.9705 | 1.4244 | 1.2436 | 1.621 | 1.6658 |
| Expected value of $q/2$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Expected value of $(n-q-1)/2$ | 4 | 4 | 4 | 4 | 4 | 4 |

Table 2 gives the results of the analysis of 3 category rating-method data for 50 individuals.

**Analysis of 3 category rating-method data for 100 individuals**

In this instance it was possible to compute the score statistics along with the likelihood ratio statistics.

Table 3 gives the results of the analysis of 3 category rating-method data for 100 individuals.

**Analysis of 3 category rating-method data for 120 individuals**

Table 4 gives the results of the analysis of 3 category rating-method data for 120 individuals. It should be noted in the case when the AUCs were distributed with means 0.75 standard deviations apart and standard deviations in the ratio 1:1.5 that it was not possible to apply the likelihood ratio and score tests as the geometric mean of the AUCs could not be numerically represented.

**Analysis of 3 category rating-method data for 140 individuals**

Table 5 gives the results of the analysis of 3 category rating-method data for 140 individuals. It is of interest to note that the value of the score statistic is negative when the AUCs were distributed with means 0.5 standard deviations apart and standard deviations in the ratio 1:1.5. According to Morgan *et al*. (2007) this is possible because, when the observed information matrix is used in place of the expected information matrix and if the observed information matrix is negative definite this could result in negative values for the score statistic.

**Analysis of 3 category rating-method data for 250 individuals**

Table 6 gives the results of the analysis of 3 category rating-method data for 250 individuals.

**Table 2:** Analysis of 3 category rating-method data for 50 individuals

| | Means 0.5 standard deviations apart | | Means 0.75 standard deviations apart | | Means 1 standard deviation apart | |
|---|---|---|---|---|---|---|
| | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 |
| Observed MLE of $q/2$ | 0.4782 | 0.4544 | 0.3806 | 0.4655 | 0.3932 | 0.4481 |
| Observed MLE of $(n-q-1)/2$ | 9.9200 | 9.3589 | 8.1448 | 9.8104 | 6.2532 | 9.0766 |
| Expected value of $q/2$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Expected value of $(n-q-1)/2$ | 11.5 | 11.5 | 11.5 | 11.5 | 11.5 | 11.5 |
| 95 % confidence interval for $q/2$ | [0.4461, 0.5125] | [0.4230, 0.4881] | [0.3751, 0.3861] | [0.4344, 0.4988] | [0.3661, 0.4223] | [0.4188, 0.4795] |
| 95 % confidence interval for $(n-q-1)/2$ | [9.1743, 10.7264] | [8.6268, 10.1532] | [7.5215, 8.8198] | [9.1195, 10.5536] | [6.0597, 6.4528] | [8.5619, 9.6223] |
| Value of LR statistic | 21.3131 | 29.0134 | – | 20.6766 | 183.4875 | 35.4873 |
| Value of Score statistic | 26.3091 | 11.4393 | – | 12.8730 | 38.2452 | 11.1723 |
| Value of $\chi^2_{2,5\%}$ | 5.99 | 5.99 | – | 5.99 | 5.99 | 5.99 |

**Table 3:** Analysis of 3 category rating-method data for 100 individuals

| | Means 0.5 standard deviations apart | | Means 0.75 standard deviations apart | | Means 1 standard deviation apart | |
|---|---|---|---|---|---|---|
| | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 |
| Observed MLE of $q/2$ | 0.4878 | 0.4822 | 0.4637 | 0.4882 | 0.4550 | 0.4882 |
| Observed MLE of $(n-q-1)/2$ | 20.8061 | 21.6107 | 19.3811 | 20.9926 | 17.7614 | 21.2555 |
| Expected value of $q/2$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Expected value of $(n-q-1)/2$ | 24 | 24 | 24 | 24 | 24 | 24 |
| 95 % confidence interval for $q/2$ | [0.4528, 0.5255] | [0.4488, 0.5181] | [0.4307, 0.4993] | [0.4533, 0.5259] | [0.4241, 0.4882] | [0.4527, 0.5265] |
| 95 % confidence interval for $(n-q-1)/2$ | [18.7433, 23.0958] | [19.3171, 24.1766] | [17.4650, 21.5074] | [19.0169, 23.1735] | [16.1388, 19.5470] | [19.2218, 23.5043] |
| Value of LR statistic | 7.9503 | 3.4994 | 14.8469 | 6.9136 | 30.8485 | 5.5090 |
| Value of Score statistic | 10.5818 | 1.1315 | 3.8170 | 9.0087 | 14.9949 | 6.6172 |
| Value of $\chi^2_{2,5\%}$ | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 |

**Table 4:** Analysis of 3 category rating-method data for 120 individuals

| | Means 0.5 standard deviations apart | | Means 0.75 standard deviations apart | | Means 1 standard deviation apart | |
|---|---|---|---|---|---|---|
| | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 |
| Observed MLE of $q/2$ | 0.5148 | 0.5015 | 0.5085 | 0.4029 | 0.4667 | 0.5035 |
| Observed MLE of $(n-q-1)/2$ | 31.2776 | 31.2265 | 30.49242 | 23.9617 | 27.2751 | 30.9002 |
| Expected value of $q/2$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Expected value of $(n-q-1)/2$ | 29 | 29 | 29 | 29 | 29 | 29 |
| 95 % confidence interval for $q/2$ | [0.4528, 0.5536] | [0.4488, 0.5387] | [0.4307, 0.5467] | [0.4533, 0.4085] | [0.4241, 0.4999] | [0.4527, 0.5402] |
| 95 % confidence interval for $(n-q-1)/2$ | [27.8719, 35.0995] | [27.9454, 34.8928] | [27.7289, 34.4877] | [21.4851, 26.7237] | [24.5406, 30.3144] | [27.8768, 34.2515] |
| Value of LR statistic | 1.7149 | 2.4962 | 1.3385 | – | 3.6031 | 1.6316 |
| Value of Score statistic | 0.2106 | 4.5961 | 1.0732 | – | 0.7117 | 2.5911 |
| Value of $\chi^2_{2,5\%}$ | 5.99 | 5.99 | 5.99 | – | 5.99 | 5.99 |

**Table 5:** Analysis of 3 category rating-method data for 140 individuals

| | Means 0.5 standard deviations apart | | Means 0.75 standard deviations apart | | Means 1 standard deviation apart | |
|---|---|---|---|---|---|---|
| | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 |
| Observed MLE of $q/2$ | 0.5096 | 0.4866 | 0.4966 | 0.5215 | 0.5081 | 0.5133 |
| Observed MLE of $(n-q-1)/2$ | 34.1440 | 31.9175 | 34.458 | 33.8716 | 35.9667 | 33.4923 |
| Expected value of $q/2$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Expected value of $(n-q-1)/2$ | 34 | 34 | 34 | 34 | 34 | 34 |
| 95 % confidence interval for $q/2$ | [0.4757, 0.5459] | [0.4528, 0.5228] | [0.4757, 0.5459] | [0.4852, 0.5605] | [0.4731, 0.5457] | [0.4759, 0.5536] |
| 95 % confidence interval for $(n-q-1)/2$ | [30.4907, 38.2351] | [29.3911, 35.8820] | [30.4907, 38.2351] | [30.4378, 37.6928] | [32.3627, 39.9722] | [29.8931, 37.5250] |
| Value of LR statistic | 0.3745 | 1.2295 | 0.3745 | 2.3346 | 1.002 | 1.3639 |
| Value of Score statistic | 0.6323 | -0.1499 | 0.6323 | 4.9657 | 0.6751 | 3.3484 |
| Value of $\chi^2_{2,5\%}$ | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 |

**Table 6:** Analysis of 3 category rating-method data for 250 individuals

| | Means 0.5 standard deviations apart | | Means 0.75 standard deviations apart | | Means 1 standard deviation apart | |
|---|---|---|---|---|---|---|
| | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 |
| Observed MLE of $q/2$ | 0.4917 | 0.4959 | 0.5087 | 0.4940 | 0.5070 | 0.5008 |
| Observed MLE of $(n-q-1)/2$ | 61.1898 | 59.7002 | 61.4909 | 59.4911 | 58.0148 | 60.3211 |
| Expected value of $q/2$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Expected value of $(n-q-1)/2$ | 61.5 | 61.5 | 61.5 | 61.5 | 61.5 | 61.5 |
| 95 % confidence interval for $q/2$ | [0.4528, 0.5291] | [0.4488, 0.5352] | [0.4307, 0.5481] | [0.4533, 0.5298] | [0.4241, 0.5444] | [0.4527, 0.5384] |
| 95 % confidence interval for $(n-q-1)/2$ | [54.6205, 68.5491] | [53.8132, 66.2310] | [55.2284, 68.4635] | [53.8145, 65.7664] | [51.9969, 64.7291] | [54.4276, 66.8527] |
| Value of LR statistic | 0.2712 | 0.2866 | 0.3702 | 0.3368 | 2.8390 | 0.2274 |
| Value of Score statistic | 0.4157 | 0.2154 | 0.7478 | 0.0838 | 6.9990 | 0.5065 |
| Value of $\chi^2_{2,5\%}$ | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 | 5.99 |

**Analysis of 3 category rating-method data for 500 individuals**

Table 7 gives the results of the analysis of 3 category rating-method data for 500 individuals. It is of interest to note that the value of the score statistic is negative when the AUCs were distributed with means 0.5 standard deviations apart and standard deviations in the ratio 1:1, and when the AUCs were distributed with means 0.75 standard deviations apart and standard deviations in the ratio 1:1.5.

**Table 7:** Analysis of 3 category rating-method data for 500 individuals

| | Means 0.5 standard deviations apart | | Means 0.75 standard deviations apart | | Means 1 standard deviation apart | |
|---|---|---|---|---|---|---|
| | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 | Standard deviations in the ratio 1:1 | Standard deviations in the ratio 1:1.5 |
| Observed MLE of $q/2$ | 0.5238 | 0.4902 | 0.4213 | 0.4947 | 0.4688 | 0.5149 |
| Observed MLE of $(n-q-1)/2$ | 130.7749 | 122.9364 | 103.4261 | 121.7594 | 119.94 | 127.5198 |
| Expected value of $q/2$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Expected value of $(n-q-1)/2$ | 124 | 124 | 124 | 124 | 124 | 124 |
| 95 % confidence interval for $q/2$ | [0.4860, 0.5646] | [0.4552, 0.5278] | [0.4157, 0.4269] | [0.4597, 0.5324] | [0.4378, 0.5019] | [0.4801, 0.5522] |
| 95 % confidence interval for $(n-q-1)/2$ | [116.496, 146.804] | [109.4625, 138.0687] | [92.974, 115.0531] | [109.4903, 135.4034] | [107.1281, 134.2842] | [114.0848, 142.537] |
| Value of LR statistic | 1.5836 | 0.3499 | – | 0.1121 | 3.6068 | 0.6263 |
| Value of Score statistic | -0.4029 | 0.4459 | – | -0.0612 | 3.9375 | 0.0571 |
| Value of $\chi^2_{2,5\%}$ | 5.99 | 5.99 | – | 5.99 | 5.99 | 5.99 |

## DISCUSSION AND CONCLUSION

In the case of the smallest sample size (i.e. 20) it could be seen that it was not appropriate to use the LR and score statistics as the values of the geometric mean of the AUC vectors, which were used in computing these statistics, could not be numerically represented in Matlab as they were very large. This illustrates that large sample theory does not hold for sample size 20. Therefore no confidence intervals were constructed for this case. The observed and expected parameters for this case were very different from each other indicating that the said theory regarding the beta distribution of the test statistic does not hold in the case of sample size 20.

For cases simulated under the sample size of 50 it could be seen that both the likelihood ratio and score test resulted in rejecting the assertion that the test statistic came from a beta distribution with specified parameters, apart from a single case in which both tests were not applicable. However in that case the confidence interval for the parameters did not include the expected values.

When considering the cases simulated under the sample size of 100 it could be seen that when the means differed by 0.5 and the standard deviation between the signal absent and signal present group were taken to be 1:1.5, the likelihood ratio and score tests indicated that the test statistic indeed came from a beta distribution

with specified parameters, while the expected values for the parameter lay within the confidence interval thereby confirming this assertion. However, in the other cases it could be seen from the results that the assertion that the test statistic came from a beta distribution with specified parameters was rejected. It could be seen however that when the spread between the two Gaussian populations (i.e. signal present and signal absent) was higher (i.e. the standard deviations were in the ratio 1:1.5) the value of the likelihood ratio and score statistics were closer to that of the 5 % significance level of the $\chi^2_2$ distribution, which is indicative of the fact that the test statistic performs better when there is a greater spread between the two populations.

For the cases simulated under the sample size of 120 it could be seen that apart from a single case, the score and likelihood ratio tests could be calculated for all the other cases and they lead to the acceptance of the assertion that the test statistic follows a beta distribution with specified parameters. In this case as well it could be seen by the values of the likelihood ratio and score statistics obtained, that the test statistic appears to perform better when there is a greater spread between the two Gaussian populations.

For all cases simulated under the sample size 140 it could be seen that both the likelihood ratio and score test resulted in acceptance of the assertion that the test statistic comes from a beta distribution with specified parameters. Similar to the case under the sample size of 100 it could be seen that when the spread between the two Gaussian populations were higher, the value of the likelihood ratio and score statistics were closer to that of the 5 % significance level of the $\chi^2_2$ distribution, which serves to strengthen the assertion made towards the end in the previous paragraph.

For all cases simulated under the sample size of 250 it could be seen the results obtained were similar to the case when the sample size of 140 was considered.

When considering the cases simulated under the sample size of 500 it could be seen that apart from a single case, the score and likelihood ratio tests could be calculated for all the other cases and they lead to the acceptance of the assertion that the test statistic follows a beta distribution with specified parameters. In this case too it could be seen by the values of the likelihood ratio and score statistics obtained that the test statistic appears to perform better when there is a greater spread between the two Gaussian populations.

## REFERENCES


1. Casella G. & Berger R.L. (2002). *Statistical Inference*, 2nd edition. Duxbury Press, California, USA.
2. Cleeves A.M. (2002). Comparative assessment of three common algorithms for estimating the variance of the area under the nonparametric reciever operating characteristic curve. *The Stata Journal* **2**(3): 280 − 289.
3. Dorfman D.D. & Alf E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology* **6**: 487 − 496.
4. Dorfman D.D. & Berbaum K.S. (1995). Degeneracy and discrete reciever operating characteristic rating data. *Academic Radiology* **2**(10): 907 − 915.
5. Fawcett T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* **27**: 861 − 874. DOI: http://dx.doi.org/10.1016/j.patrec.2005.10.010
6. Grey D.M. & Morgan B.T. (1972). Some aspects of ROC curve fitting: normal and logistic models. *Journal of Mathematical Psychology* **9**: 128 − 139.
7. Hahn G. J. & Shapiro S.S. (1994). *Statistical Models in Engineering*. John Wiley & Sons, Inc., New Jersey, USA.
8. Hotelling H. (1947). Multivariate quality control. *Techniques of Statistical Analysis* (eds. C. Eisenhart, M.W. Hastay & W.A. Wallis), pp.111 – 184. McGraw-Hill, New York, USA.
9. Maia A.S., Braga junior A.R. & Cordeiro G.M. (2003). Corrected likelihood ratio and score tests for the beta distribution. *Journal of Statistical Computation and Simulation* **73**(8): 585 − 596. DOI: http://dx.doi.org/10.1080/0094965021000033549
10. Mardia K.V., Kent J.T. & Bibby J.M. (1979). *Multivariate Analysis* Academic Press, London, UK.
11. Metz C.E., Herman B.A. & Shen J.H. (1998). Maximum likelihood estimation of reciever operating characteristic (ROC) curves from continuously distributed data. *Statistics in Medicine* **17**: 1033 − 1053. DOI:http://dx.doi.org/10.1002/(SICI)1097-0258(19980515)17:9<1033::AID-SIM784>3.0.CO;2-Z
12. Metz C.E. & Pan X. (1999). Proper binormal ROC curves: theory and maximum-likelihood estimation. *Journal of Mathematical Psychology* **43**: 1 − 33. DOI: http://dx.doi.org/10.1006/jmps.1998.1218
13. Meyen N. & Sooriyarachchi M.R. (2014). Determining the properties of a newly developed test for comparing receiver operating characteristic (ROC) curves. In: Pure Science Track. *Proceedings of the Jaffna University International Research Conference (JUICE-2014)*, 18 − 19 December. pp. 353 – 356.

14. Morgan B.J.T., Palmer K.J. & Ridout M.S. (2007). Negative score test statistic. *The American Statistician* **61**(4): 285 − 288. DOI: http://dx.doi.org/10.1198/000313007X242972

15. Vergara I.A., Norambuena T., Ferrada E., Slater A.W. & Melo F. (2008). StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatic* **9**(265). DOI: http://dx.doi.org/10.1186/1471-2105-9-265

## APPENDIX

**Matlab implementation of the score test for the beta distribution**

Given below is the Matlab implementation of the score test for the beta distribution.

```
% score function for beta distribution function val = score(n,p,q,logg1,logg2)

u = zeros(2,1);

u(1,1) = n*(psi(p+q)-psi(p)+logg1);
u(2,1) = n*(psi(p+q)-psi(q)+logg2);

K = zeros(2,2);

K(1,1) = -n*(psi(1,p+q)-psi(1,p));
K(1,2) = -n*psi(1,p+q);
K(2,1) = -n*psi(1,p+q);
K(1,1) = -n*(psi(1,p+q)-psi(1,q));

inverseK = inv(K);
val = transpose(u)*inverseK*u;

end
```

**Matlab implementation of the likelihood ratio test for the beta distribution**

Given below is the Matlab implementation of the likelihood ratio test for the beta distribution.

```
% Likelihood ratio statistic for beta distributionfunction val = LHR(n,p,q,phat,qhat,logg1,logg2)
val1 = (2*n*(phat -p)*logg1);
val2 =(2*n*(qhat-q)*logg2);
val3 = 2*n*(log(gamma(phat+qhat))+log(gamma(p))+log(gamma(q))-log(gamma(p+q))-log(gamma(phat))
-log(gamma(qhat)));
val = val1 + val2 + val3;

end
```