International Journal of Research and Review
www.ijrrjournal.com E-ISSN: 2349-9788; P-ISSN: 2454-2237

Research Paper

Comparison of Error of Classification Associated With Gamma Distribution Using Some Selected Discriminant Functions

Nwoye, O. N.[1], Osuji, G. A.[2], Aronu, C. O.[3]

[1,3]Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State
[2]Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria

Corresponding Author: Nwoye, O. N.

ABSTRACT

This study examined the error of classification associated with the gamma distribution for the linear discriminant analysis, the logistic discriminant analysis and the quadratic discriminant analysis. Data were simulated data from Gamma distribution for sample size 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. Fifty simulation for each of the sample size was performed for the methods considered in this study. The findings of the study revealed that for the gamma distribution, the logistic discriminant analysis method performed better followed by the linear discriminant analysis and then the quadratic discriminant analysis. This is because the logistic discriminant analysis method has the least average misclassification error rate across the sample sizes followed by the linear discriminant analysis method while the quadratic discriminant method performed least for gamma distribution.

Keywords: *Classification; Gamma Distribution; Misclassification, Simulation; Error rate*

1.0 INTRODUCTION

Classification of observation in vector form into one of two populations or more often depends on the value of the discriminant function. Fisher's linear discriminant function in literature often serves as a criterion when samples are used to estimate the parameters of the distributions in the two populations. Though other criterion such as the quadratic discriminant function, Likelihood ratio discriminant function, Maximum likelihood discriminant function e.t.c. exist. The discriminant function is also a function of the observations and the parameters of distributions.

Error rates are easily calculated when the population characteristics are known such characteristics of interest to researcher include the mean and variance of the population. These parameters are often estimated from the samples, and most times there seem to be loss of information which might affect the estimation of the probabilities of misclassification, thereby leading to underestimation or overestimation of the parameters. Onyeagu *et al.* (2013) in their study examined the performance of Jackknife and resubstitution method in estimating error rates in classification analysis while Egbo (2016) evaluated some error rate estimators in two group discriminant analysis with multivariate binary variables.

Most literature in the area of discriminant analysis has often assumed multivariate normality whereas very few author(s) have tried to determine the misclassification error rate outside the conventional assumption of normality (Mahmoud and Moustafa, 1995). Hence, a situation may arise when an researcher does not recognize an observation to be normally distributed. The aim of this study is to compare misclassification error rates from Liner discriminant function, Quadratic discriminant function and Likelihood ratio discriminant function for gamma distribution.

2. LITERATURE REVIEW

Ahmed and Lachenbruch (1975) assessed the performance of discriminant analysis in a situation where one or both of the initial samples is contaminated for large sample result. The result of their findings showed that contamination has the potentials to affect the performance of the linear discriminant function. However, the study revealed that some of the simpler estimates improved the overall behavior in the scale contamination problem.

Krzanowski (1975) in his contribution, noted that the utility of an allocation rule in discriminant analysis can be assessed using the probabilities of misclassification approach, or the error rates. The error rates are called optimum error rates in the case of known parameter for the location model this is because they indicate the best results possible with the model. For the unknown parameter, various types of error rates may be distinguished. In particular, once an allocation rule has been derived in practice, it is essential to have an adequate method for estimating the error rates that it incurs to have some measure of its utility and to be able to assess its performance relative to other allocation rules. Glick (1978) examined the D, R, and U methods, as well as the posterior probability estimator and smooth slide estimator, in the univariate normal case. He compared the biases and variances of the estimators considered in the study. The result of the study found that the smoothed estimator did indeed have a smaller bias and variance than the resubstitution estimator.

Snapinn (1983) argued that there is no definitive technique for the estimation of error rates in discriminant analysis. However, he noted that the estimators used often in practice include the parametric plug-in estimator (D method), and the nonparametric resubstitution estimator (R method) and leave-one-out estimator (U method). It was found that the U method has received much attention recently because it is less biased than any other method though it was found to have very large variance.

De Leona *et al* (2011) in their study examined the problem associated with classifying an individual into one of several populations based on mixed nominal, continuous, and ordinal data. They were able to obtain a classification procedure which stands as an extension to the location for linear discriminant function, by specifying a general mixed-data model for the joint distribution of the mixed discrete and continuous variables. They proposed the general mixed-data models (GMDMs) which were found to be effective in correctly classifying individuals, showing relatively superior performance compared with robust minimum distance probability (MDP) method, especially in mixed-data involving ‘high-information’ multi-level ordinal variables. Further result revealed that the error rates can be estimated either by plug-in Monte Carlo or hold-out method. Both approaches yield nearly unbiased estimates for large samples.

Egbo (2016) on decision rule in evaluating the performance of classification rule noted that given the existence of two groups of individuals, interest should be on finding a classification rule for allocating new individuals or observations into one of the existing two groups. Corresponding to each classification rule, there is a probability of misclassifications if that classification rule is used to classify new individuals (observations) into one of the two groups. The best classification rule is the one that leads to the smallest probability of misclassifications, which also called error rates. In addition, Egbo *et al.* (2016) explained that classification systems play an important role in the social and behavioral sciences. The study accessed the relative performance of some well-known classification methods. The result of their findings ranked the eight classification methods considered in the following order: optimal rule, linear discriminant rule, maximum likelihood rule, Predictive rule, Dellion-Goldstein rule, full multinomial, likelihood rule and nearest neighbor rule. The findings of the study showed that for three and four variables, the maximum likelihood rule was the most preferred while for five variables the optimal rule performed better in terms of minimizing the expected error rate.

3. MATERIAL AND METHOD

3.1 Method of Data Collection

The source of data used for this study was simulated data from Gamma distribution for sample size 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100.

3.2 Methodology

Discriminant functions are linear combinations of variables that best separate groups. We shall discuss three discriminant function in this section and they include: Linear Discriminant Function, Quadratic Discriminant Function and the Likelihood Ratio Discriminant Function.

3.2.1 Linear Discriminant Function

Suppose that the individual the individual functions f_k's are multivariate normal densities with different means but with the same covariance, such that

$$X_i / Y_i = k \sim N_{1\times p}(\mu_k, \Sigma) \quad (1)$$

Given that the probability density function as

$$f_k(x \mid \mu_k, \Sigma) = c \times \frac{1}{|\Sigma|^{1/2}} \ell^{-\frac{1}{2}(x-\mu_k)\Sigma^{-1}(x-\mu_k)'}$$

(2)

Recall that $(\mu_k \Sigma^{-1} x')' = x\Sigma^{-1}\mu'_k$

Hence,

$$= c \times \frac{1}{|\Sigma|^{1/2}} \ell^{-\frac{1}{2}(x\Sigma^{-1}x' - 2x\Sigma^{-1}\mu'_k + \mu_k\Sigma^{-1}\mu'_k)}$$

(3)

Since interest is in the highest pdf multiplied by π_k, we shall ignore the factors that are the same in each group which do not depend on k (Marden , 2013).

$$\pi_k \times \ell\left(x\Sigma^{-1}\mu'_k - \frac{1}{2}\mu_k\Sigma^{-1}\mu'_k\right) \quad (4)$$

By taking logs, the k that maximizes (4) is given as

$$d_k^*(x) \equiv x\Sigma^{-1}\mu'_k - \frac{1}{2}\mu_k\Sigma^{-1}\mu'_k + \log(\pi_k) \quad (5)$$

The d_k^*'s in (5) are linear discriminant functions since they are linear in x.

where μ_k is the mean of the sample k and Σ is the pooled covariance.

$$\hat{\mu}_k = \frac{1}{n} \sum_{\{i|y_i=k\}} X_i \quad (6)$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{\{i|y_i=k\}} (X_i - \hat{\mu}_k)' (X_i - \hat{\mu}_k) \quad (7)$$

The estimated discriminant function is written as

$$\hat{d}_k(x) = c_k + x a'_k$$

where,

$$a_k = (\hat{\mu}_k - \hat{\mu}_K)\hat{\Sigma}^{-1}$$

$$c_k = \frac{1}{2}\left(\hat{\mu}_k\hat{\Sigma}^{-1}\hat{\mu}'_k - \hat{\mu}_K\hat{\Sigma}^{-1}\hat{\mu}'_K\right) + \log\left(\frac{\hat{\pi}_k}{\hat{\pi}_K}\right) \quad (8)$$

Hence, the general classification rule based on Fisher's linear discriminant function (FLD) is

$$\hat{c}_{FLD}(X) = k \text{ if } \hat{d}_k(x) > \hat{d}_l(x) \quad \text{for } l \neq k \quad (9)$$

Considering k=2, the corresponding discriminant rule is stated as

$x \rightarrow \pi_1$ if $W_k(x) > 0$

$x \rightarrow \pi_2$ if $W_k(x) \leq 0$

where,

$$W_k(x) = \hat{d}_k(x) - \hat{d}_l(x)$$

$$W_k(x) = \hat{d}_k(x) - \hat{d}_l(x)$$

3.2.2 Quadratic Discriminant Function

When the equality of the covariance matrices is not tenable, we can use a slightly more complicated procedure (Marden, 2013). Here the conditional probabilities are proportional to

$$\pi_k f_k\left(x|\mu_k, \Sigma_k\right) = c \times \pi_k \frac{1}{|\Sigma_k|^{1/2}} \ell^{-\frac{1}{2}(x-\mu_k)\Sigma^{-1}(x-\mu_k)'}$$

(10)

$$= c \times \ell^{-\frac{1}{2}(x-\mu_k)\Sigma^{-1}(x-\mu_k)' - \frac{1}{2}\log|\Sigma_k| + \log(\pi_k)}$$

(11)

Marden (2013) stated that the discriminant functions can be taken to be the terms in the exponents (times 2, for convenience) in (11), or their estimates:

$$\hat{d}_k^Q(x) = -(x-\hat{\mu}_k)\hat{\Sigma}_k^{-1}(x-\hat{\mu}_k)' + c_k \quad (12)$$

where,

$$c_k = -\log\left(\left|\hat{\Sigma}_k\right|\right) + 2\log\left(N_k/n\right) \quad (13)$$

and $\hat{\Sigma}_k$ is the sample covariance matrix from the k^{th} group. The boundaries between regions are quadratic rather than linear, hence the Fisher's quadratic discriminant function defined as

$$\hat{C}_{FQD}(x) = k \text{ if } \hat{d}_k^Q(x) > \hat{d}_l^Q(x) \quad \text{for } l \neq k$$

(14)

Considering k=2, the corresponding quadratic discriminant rule is stated as

$x \rightarrow \pi_1$ if $W_k^Q(x) > 0$

$x \rightarrow \pi_2$ if $W_k^Q(x) \leq 0$

where,

$$W_k^Q(x) = \hat{d}_k^Q(x) - \hat{d}_l^Q(x)$$

3.2.3 Likelihood Ratio Discriminant Function

The likelihood ratio discriminant function was described in Anderson (1984) and was applied to a multivariate normal model. The basic idea was to derive a rule from a generalized likelihood ratio test for the hypothesis

$H_0: = x_{11}, x_{1n_1} \sim f_1(x)$ and $x_{21}, \cdots, x_{2n_2} \sim f_2(x)$

against

$H_1: \neq x_{11}, x_{1n_1} \sim f_1(x)$ and $x_{21}, \cdots, x_{2n_2} \sim f_2(x)$

Considering the multinomial model, the test statistic which is a function of x is written as

$$L(x) = \frac{\left(1 + \frac{1}{n_1(x)}\right)^{n_1(x)} (n_1(x)+1)}{\left(1 + \frac{1}{n_2(x)}\right)^{n_2(x)} (n_2(x)+1)} \times \frac{\left(1 + \frac{1}{n_2}\right)^{n_2} (n_2+1)}{\left(1 + \frac{1}{n_1}\right)^{n_1} (n_1+1)} \quad (15)$$

where, if $n_1(x) = 0$ $(n_2(x) = 0)$ the numerator (denominator) of the first fraction must be replaced by 1. The classification rule derived is to

$x \rightarrow \pi_1$ if $L(x) > 1$

$x \rightarrow \pi_2$ if $L(x) < 1$ and random else

Note that if $n_1 = n_2$ the likelihood rule and the Maximum Likelihood rule becomes equivalent because the numerator of the first fraction in (15) is a strictly increasing succession of $n_{1(x)}$.

A new observation x with $n_1(x) = 0$ will can be classified in π_1 if and only if

$$\left(1 + \frac{1}{n_2(x)}\right)^{n_2(x)} (n_2(x)+1) < c \quad (16)$$

where c is the value of the second fraction in (15).

4. DATA ANALYSIS AND RESULTS

4.1 Result of Misclassification Error Rate for Gamma Distribution

Table 1: Average Misclassification Error Rate from Linear Discriminant Analysis (LDA) for Gamma Distribution

Sample Size	E1	E2	E
10	0.1180	0.2183	0.1681
20	0.1106	0.2151	0.1629
30	0.1118	0.2158	0.1638
40	0.1121	0.2196	0.1658
50	0.1081	0.2146	0.1613
60	0.1133	0.2239	0.1686
70	0.1129	0.2208	0.1669
80	0.1123	0.2191	0.1657
90	0.1105	0.2194	0.1649
100	0.1122	0.2203	0.1662

The result obtained in table 1, found that the least misclassification error rate for the LDF to be 0.1613 at sample 50 while the highest misclassification error rate was obtained to be 0.1686 at sample 60.

Table 2: Misclassification Error Rate from Logistics Discriminant Analysis (LR) for Gamma Distribution

Sample Size	E1	E2	E
10	0.131	0.182	0.1565
20	0.129	0.209	0.169
30	0.1307	0.192	0.1613
40	0.1243	0.1945	0.1594
50	0.1254	0.1992	0.1623
60	0.123	0.199	0.161
70	0.1281	0.2023	0.1652
80	0.1254	0.1975	0.1614
90	0.1284	0.1999	0.1642
100	0.1261	0.2016	0.1639

The result obtained in table 2, found that the least misclassification error rate for the LR to be 0.1565 at sample 10 while the highest misclassification error rate was obtained to be 0.1690 at sample 20.

Table 3: Average Misclassification Error Rate from Quadratic Discriminant Analysis (QDA) for Gamma Distribution

Sample Size	E1	E2	E
10	0.0533	0.2800	0.1725
20	0.0750	0.2780	0.1765
30	0.0717	0.2940	0.1828
40	0.0768	0.2985	0.1876
50	0.0760	0.2912	0.1836
60	0.0708	0.2960	0.1834
70	0.0699	0.3009	0.1854
80	0.0568	0.3178	0.1873
90	0.0591	0.3096	0.1843
100	0.0543	0.3158	0.1851

Also, it was found in table 3 that the least misclassification error rate for the QDA to be 0.1725 at sample 10 while the highest misclassification error rate was obtained to be 0.1876 at sample 40.

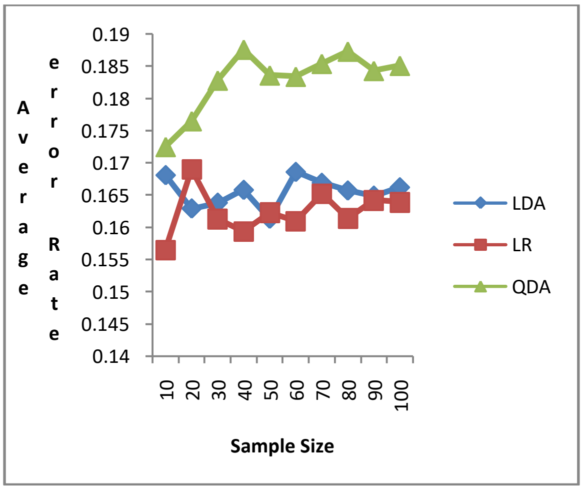

Figure 1: Distribution of Average Misclassification Error Rate from LDA, LR and QDA for Gamma Distribution

The result obtained in figure 1, revealed that the logistic discriminant analysis method performed better than the others, since it has the least average misclassification error rate across the sample sizes followed by the linear discriminant analysis method while the quadratic discriminant method performed least for gamma distribution.

5. CONCLUSION

This study assessed the error of classification associated with the gamma distribution for the linear discriminant analysis, the logistic discriminant analysis and the quadratic discriminant analysis. Fifty simulation for each of the sample size was performed for the methods considered in this study. The findings of the study revealed that for the gamma distribution, the logistic discriminant analysis method performed better followed by the linear discriminant analysis and then the quadratic discriminant analysis. This is because the logistic discriminant analysis method has the least average misclassification error rate across the sample sizes followed by the linear discriminant analysis method while the quadratic discriminant method performed least for gamma distribution.

REFERENCES

- Ahmed, S. and Lachenbruch, P. (1975). Discriminant Analysis when one or both of the Initial Samples is Contaminated: Large Sample Results. *EDV in Medizin and Biologie*, 6: 35–42.
- De Leona, A. R., Soob, A. and Williamson, T. (2011). Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistics*, 38(5): 1021–1032.
- Egbo I. (2016). Evaluation of Error Rate Estimators in Discriminant Analysis with Multivariate Binary Variables. *American Journal of Theoretical and Applied Statistics*, 5(4): 173-179.
- Egbo I., Egbo, M. and Onyeagu, S. I. (2016). A Comparative Assessment of Classification Methods for Binary Variables. *2016 Annual Conference Proceedings of Nigerian Statistical Association, International Conference Centre Abuja, Nigeria, pp. 517-540.*

- Glick, N. (1978). Additive Estimators for Probabilities of Correct Classification. *Pattern Recognition*, 10, 211-222.
- Krzanowski, W. J. (1975). Discrimination and Classification Using Both Binary and Continuous Variables. *Journal of the American Statistical Association*, 70(352): 782-790.
- Mahmoud, M. A. W. and Moustafa, H. M. (1995). Errors of Misclassification Associated with Gamma Distribution. *Mathematical Computing Modelling*, 22(3): 105-119.
- Onyeagu, S.I., Osuji, G.A., Ekezie, D.D., and Ogbonna, C. J. (2013). A Review of Classification Models Using Discrete Variables. *Research Journal of Mathematical and Statistical Sciences*, 1(8): 28-38.
- Snapinn, S. M. (1983). An Evaluation of Smoothed Error Rate Estimators in Discriminant Analysis. PhD Dissertation submitted to the Department of Biostatistics, University of North Carolina, Chapel Hill, USA, Pg 1-132.

How to cite this article: Nwoye, O. N., Osuji, G. A. Aronu, C. O. Comparison of error of classification associated with gamma distribution using some selected discriminant functions. International Journal of Research and Review. 2019; 6(4):29-34.

******