Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2021, 2021:44
https://doi.org/10.28919/cmbn/5391
ISSN: 2052-2541

APPLICATION OF BETA MIXTURE DISTRIBUTION IN DATA ON GPA PROPORTION AND COURSE SCORES AT THE MBTI TELKOM UNIVERSITY

NURVITA TRIANASARI[1,2,*], I MADE SUMERTAJAYA[3], ERFIANI[3], I WAYAN MANGKU[4]
[1]Telkom of Economics and Business School, Telkom University, Bandung, Indonesia
[2]Department of Statistics, Student of Doctoral Program at Sekolah Pascasarjana-IPB, Bogor, Indonesia
[3]Department of Statistics, IPB University, Bogor, Indonesia
[4]Department of Mathematics, IPB University, Bogor, Indonesia

Abstract: Cluster analysis is a multivariate analysis that aims to cluster objects or data so that objects or data that are in the same cluster have relatively more homogeneous properties than objects or data in different clusters. Probabilistic clustering method is often based on the assumption that data comes from a mixture of distributions, for examples Poisson, normal, lognormal, and Erlang. Thus the probabilistic clustering problem is transformed into a parameter estimation problem because the data is modeled by a cluster of mixture distribution. Data points that have the same distribution can be defined as one cluster. This distribution is applied to identify users on the community question answering site (CQA). In this paper the distribution of beta mixtures for single variable cases will be applied to the data on the proportion of student's GPA in the subject of Business Statistics and Economic Mathematics of the Informatics Telecommunications Business Management, Faculty of Economics and Business, Telkom University. Based on the results of the analysis on the GPA data, Economic Mathematics and Business Statistics shows the smallest integrated classification likelihood estimation Bayesian criterion (ICL BIC) scores in two clusters for GPA and Business Statistics Value. While the ICL value of BIC in Economic Mathematics shows the smallest ICL BIC value in one cluster. Then it can be concluded that GPA and Business Statistics occur in Mixture 2 clusters.

*Corresponding author
E-mail address: nurvitatrianasari@telkomuniversity.ac.id
Received January 4, 2021

Key word: bivariate beta mixture model; integrated classification; likelihood estimation; Bayesian criterion.
2010 AMS Subject Classification: 62H30.

1. INTRODUCTION

Cluster analysis is a multivariate analysis that aims to cluster objects or data so that objects or data that are in the same cluster have relatively more homogeneous properties than objects or data in different clusters (Zickmund et al. 2010 and Hair et al. 2010).

The concept of cluster formation includes hierarchical methods, non-hierarchical methods and clustering methods that are probable (probabilistic clustering). The hierarchical method starts clustering with two or more objects that have the closest similarity. Then the process is forwarded to other objects that have a second closeness so the cluster will form a kind of tree where there is a clear hierarchy or level between objects, from the most similar to the less similar to form a cluster. The endpoint is a set of clusters, where each cluster is distinct from the other cluster, and the objects within each cluster are broadly similar to each other.

The non-hierarchical method begins by first determining the desired number of groups (two or more groups). After determining the number of clusters, then the grouping process is carried out without following a hierarchical process. The disadvantage in this method is that the number of cluster must be determined in advanced.

In addition to the hierarchical and non-hierarchical methods outlined above, there is other method that is often used, namely the clustering method that has the opportunity to determine the optimal number of groups based on the distribution of the data. This clustering method is called a probabilistic clustering technique which assumes that the data follows a certain distribution. Probabilistic methods have the potential to be widely used in a variety of applications such as market segmentation, image segmentation (Blekas et al. 2005) and (Stauffer et al. 1999), handwriting recognition (Revow et al., 1996), and document clustering (Hoffman, 2001). This clustering method has the opportunity to try to optimize the suitability of the observed data with mathematical models using a probabilistic approach (Anggarwal, 2014). This method is often based on the assumption that data comes from a mixture of distributions of opportunities, for example Poisson, normal, lognormal, and Erlang. Thus the clustering problem is transformed into a parameter estimation problem because the data is modeled by a cluster of mixture distribution. Data points that have the same distribution can be defined as groups.

Sahu et al. (2016) discusses the distribution of beta mixtures of multiple variables where the parameter estimation use the EM algorithm and the determination of the optimal number of groups using the integrated classification likelihood (Bayesian information criterion) determinant method. This distribution is applied to identify users on the community question answering site (CQA). In this paper the distribution of beta mixtures for single variable cases will be applied to the data on the proportion of GPA in the subject of Business Statistics and Economic Mathematics, students of the Informatics Telecommunications Business Management (MBTI), Faculty of Economics and Business, Telkom University.

2. BETA DISTRIBUTION

Let Y be a random variable having beta distribution with the parameters α and β, where-$-\infty < \alpha < \infty$ and $-\infty < \beta < \infty$. The density function of this random variable is:

$$g(y|\alpha,\beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} \quad ; \quad 0 < y < 1. \tag{1}$$

The curve of the beta density for various combinations of parameters is presented in Figure 1.

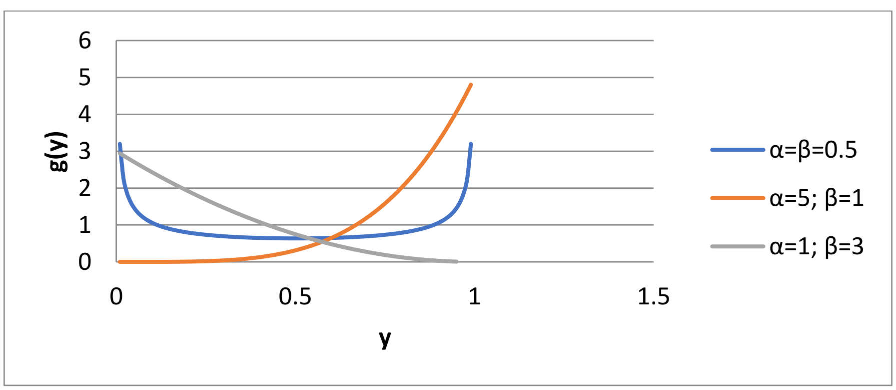


Figure 1. Beta Density Function Curve of Beta Distribution for Various Parameters

The mean and variances of this random variable are:

$E(Y) = \frac{\alpha}{\alpha+\beta}$,

and $Var(Y) = \frac{\alpha\beta}{[(\alpha+\beta)^2(\alpha+\beta+1)]}$.

3. THE MULTIVARIATE BETA MIXTURE MODEL

Sahu et al. (2016) used the i, x_i, $i = 1,2,\cdots,n$, observation data to form a mixture density function

$$f(x_i|\boldsymbol{\alpha}, \boldsymbol{a}, \boldsymbol{b}) = \sum_{c=1}^{C} \alpha_c f_c(x_i|a_c, b_c) \qquad (2)$$

where $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \dots, \alpha_C\}, \sum_{c=1}^{C} \alpha_c = 1$; $\alpha_c > 0$ express the mixture coefficient; C denotes the number of groups in the mix; f_c denotes the density function of probability from the beta distribution of the single c-variable; $\boldsymbol{a} = \{a_1, a_2, \dots, a_C\}$ and $\boldsymbol{b} = \{b_1, b_2, \dots, b_C\}$ where a_C and b_C represent c-cluster parameters.

The density function of the beta distribution of a single variable for the c-class mix of beta is defined as

$$f_c(x_i|a_c, b_c) = \frac{\Gamma(a_c+b_c)}{\Gamma(a_c)\Gamma(b_c)} x_i^{a_c-1}(1-x_i)^{b_c-1} \qquad (3)$$

where $\Gamma(.)$ states the gamma function which is defined as

$$\Gamma(y) = \int_0^\infty t^{y-1}e^{-t}dt; t > 0.$$

4. MAXIMUM LIKELIHOOD ESTIMATION FOR THE MULTIVARIATE BETA MIXTURE MODEL

The parameters of the multivariate BMM can be estimated using maximum likelihood estimation. Suppose that $\Theta = \{\alpha_1, \alpha_2, \dots, \alpha_C;\ a_1, a_2, \dots, a_C;\ b_1, b_2, \dots, b_C\}$ represents the set of unknown mixture parameters of the model and $x = \{x_1, x_2, \dots, x_n\}$ represent the set of the normalized feature vectors. Therefore, the likelihood function corresponding to C components of the mixture can be expressed as (Sahu at al. 2016)

$$L(X|\Theta) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\alpha}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{n}\sum_{c=1}^{C} \alpha_c f_c(x_i|a_c, b_c). \tag{4}$$

The expectation maximization (EM) algorithm is used to estimate the mixture model parameters for maximum likelihood in which each user's feature vector x_i is assigned to C dimensional indication vector $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \dots, z_{iC})^T$ such that

$$z_{ic} = \begin{cases} 1 & ;\ \text{If } x_i \text{ belongs to the component } c \\ 0 & ;\text{otherwise.} \end{cases} \tag{5}$$

Suppose that $Z = \{z_1, z_2, \dots, z_n\}$ denote the set of indication vectors for set of users' $X = \{x_1, x_2, \dots, x_n\}$. The likelihood function of the data set is given by

$$L(X, Z|\Theta) = \prod_{i=1}^{n}\prod_{c=1}^{C}[\alpha_c f_c(x_i|a_c, b_c)]^{z_{ic}} \tag{6}$$

Next, take the logarithm of the likelihood function, which is given by

$$log\big(L(X, Z|\Theta)\big) = \sum_{i=1}^{N}\sum_{c=1}^{C} z_{ic} \log[\alpha_c f_c(x_i|a_c, b_c)] \tag{7}$$

Now, the estimation of Θ is done through EM algorithm with number of iterations I = {0, 1, 2, ...} between the expectation and maximization steps so as to a sequence estimate $\{\widehat{\Theta}\}^{(I)}$ until the change in the value of the log-likelihood function expressed in equation (7) is negligible.

Expectation step: the indication for the *c*-component of feature vectors replaced its expectations as follows

$$z_{ic}^{(I)} = E[z_{ic}|\ x, \Theta] = \frac{\hat{\alpha}_c^{(I)} f_c\big(x_i|\hat{a}_c, \hat{b}_c\big)}{\sum_{k=1}^{C} \hat{\alpha}_k^{(I)} f_k\big(x_i|\hat{a}_c, \hat{b}_c\big)} \tag{8}$$

Maximization steps : the set of unknown parameters $\Theta = \{\alpha_1, \alpha_2, \dots, \alpha_C;\ a_1, a_2, \dots, a_C;\ b_1, b_2, \dots, b_C\}$ of the mixture model are calculated using the estimated z_{ic} values in the expectation step. The mixing coefficients of the model are calculated as

$$\hat{\alpha}_c^{(I+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ic}^{(I)}}{n}\ ; c = 1,2, \dots, C. \tag{9}$$

The gradient derivative of the expectation of the log-likelihood of the dataset a_c and b_c and equated to zero, which is used to find the value $\hat{a}_c$, $\hat{b}_c$ that maximizes the likelihood as follows

$$\begin{bmatrix} \frac{\partial E[\log\,(L(X,Z|\Theta)))]}{\partial a_c} \\ \frac{\partial E[\log\,(L(X,Z|\Theta)))]}{\partial b_c} \end{bmatrix} = 0 \tag{10}$$

where,

$$\frac{\partial E[\log\,(L(X,Z|\Theta)))]}{\partial a_c} = \sum_{i=1}^{n} \hat{z}_{ic}\left[\frac{\Gamma'(a_c+b_c)}{\Gamma(a_c+b_c)} - \frac{\Gamma'(a_c)}{\Gamma(a_c)} + \log(x_i)\right] \tag{11}$$

and

$$\frac{\partial E[\log\,(L(X,Z|\Theta)))]}{\partial b_c} = \sum_{i=1}^{n} \hat{z}_{ic}\left[\frac{\Gamma'(a_c+b_c)}{\Gamma(a_c+b_c)} - \frac{\Gamma'(b_c)}{\Gamma(b_c)} + \log(1-x_i)\right] \tag{12}$$

From equations (11) and (12), equation (10) can be represented as follows

$$\begin{bmatrix} \sum_{i=1}^{n} \hat{z}_{ic}[\psi(a_c+b_c) - \psi(a_c) + \log\,(x_i)] \\ \sum_{i=1}^{n} \hat{z}_{ic}[\psi(a_c+b_c) - \psi(b_c) + \log\,(1-x_i)] \end{bmatrix} = 0 \tag{13}$$

where $\psi(.)$ represents the digamma function defined as $\psi(\lambda) = \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}$. An exact solution to equation (13) as the digamma function is defined through integration. Therefore, the Newton-Raphson (a tangent method for root finding) is used to estimate parameter a_c and b_c iteratively as

$$\begin{bmatrix} a_c^{(I+1)} \\ b_c^{(I+1)} \end{bmatrix} = \begin{bmatrix} a_c^{(I)} \\ b_c^{(I)} \end{bmatrix} - \begin{bmatrix} \frac{\partial E\left[\log\left(L(X,Z|\Theta)\right)\right]}{\partial a_c} \\ \frac{\partial E\left[\log\left(L(X,Z|\Theta)\right)\right]}{\partial b_c} \end{bmatrix}$$
$$\text{x} \begin{bmatrix} \frac{\partial^2 E\left[\log\left(L(X,Z|\Theta)\right)\right]}{(\partial a_c)^2} & \frac{\partial^2 E\left[\log\left(L(X,Z|\Theta)\right)\right]}{\partial a_c \partial b_c} \\ \frac{\partial^2 E\left[\log\left(L(X,Z|\Theta)\right)\right]}{\partial b_c \partial a_c} & \frac{\partial^2 E\left[\log\left(L(X,Z|\Theta)\right)\right]}{(\partial b_c)^2} \end{bmatrix}^{-1} \tag{14}$$

where

$$\frac{\partial^2 E\big[\log\big(L(X,Z|\Theta)\big)\big]}{(\partial a_c)^2} = \sum_{i=1}^{n} \hat{z}_{ic}[\psi'(a_c + b_c) - \psi'(b_c)] \tag{15}$$

where $\psi'(.)$ is a tri-gamma function. The initial value of $a_c^{(0)}$ and $b_c^{(0)}$ needed to start the iteration process expressed in equation (14) is done through estimating the moment of beta distribution. The moment estimates $a_c^{(0)}$ and $b_c^{(0)}$ are defined as

$$\hat{a}_c^{(0)} = \bar{\mu}_c \left[\frac{\bar{\mu}_c(1-\bar{\mu}_c)}{\sigma_c^2} - 1\right] \tag{16}$$

$$\hat{b}_c^{(0)} = (1-\bar{\mu}_c) \left[\frac{\bar{\mu}_c(1-\bar{\mu}_c)}{\sigma_c^2} - 1\right] \tag{17}$$

where $\bar{\mu}_c$ is the sample mean and σ_c^2 is the sample variance of the feature value corresponding to D-dimension of the feature vectors and belongs to the C component of beta distribution. The Newton-Raphson algorithm converges when the change in values of estimates $\hat{a}_c$ and $\hat{b}_c$ is less than a small positive value ξ, with each successive iteration of equations (19) and (20).

The maximum possible estimate of the beta distribution parameters can be done using the EM algorithm. The EM algorithm depends on initialization, Fuzzy C-Means (FCM) is used to initialize. First, the data set $(x_1, x_2, \cdots, x_n)$ is partitioned into a C cluster. Next, the parameters of each component of the dataset is estimated using the method of moment of the beta distribution and setting them as initial parameters which is required in EM algorithm.

5. ESTIMATING THE NUMBER OF COMPONENTS IN THE MIXTURE

Various approaches have been proposed to estimate the number of components in mixture model. Sahu at al. (2016) use deterministic approach based on EM algorithm to obtain a range of values for $C = 1,2,\dots,C_{max}$ which is assumed to have optimal value of C. The number of components is selected according to the following criteria

$$\hat{C} = \arg min_C \left\{MSC\left(\hat{\Theta}(C), C\right), C = 1,2,\dots,C_{max}\right\} \tag{18}$$

where $\hat{\Theta}(C)$ is an estimate of the mixture parameters assuming that it has C components, and MSC $\left(\hat{\Theta}(C), C\right)$ is the model selection criterion. In Sahu at al. (2016), ICL-BIC is used as the model selection criterion defined as follows

$$ICL - BIC(C) = -2\log L_C + p\log(n) - 2\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}\log z_{ic} \qquad (19)$$

where L_C is the logarithm for getting the maximum likelihood solution of the beta mixture model and p is the number of estimated parameters. The detailed procedure for estimating the optimal number of beta components in the mixture of the dataset is illustrated in the following algorithm.

Algorithm: Estimating the number of components in the beta mixture

Input: $(x_1, x_2, \cdots, x_n)$ and C_{max}

Output: The optimal number of components C representing beta mixture

Begin

for C = 1 to C_{max} do

if C == 1 then

Estimates that each parameter pair $\{\hat{a}_c, \hat{b}_c\}$ using equation (14).

Compute the value of ICL-BIC (C) using equation (19).

else

Initialize EM algorithm using FCM clustering algorithm; alternate the following two steps to estimate the mixture parameters as:

E-Step: Compute $z_{ic}^{(I)}$ using equation (8).

M-Step:

(1) Estimate the mixing coefficients using equation (9).

(2) Estimate $\{\hat{a}_c, \hat{b}_c\}$ using equation (14).

Repeat E-Step and M-Step until the change in equation (7) is negligible.

Compute the value of ICL-BIC (C) using equation (19).

end if-else

end for

Select $\hat{C}$ such that $\hat{C} = \arg min_C\{ICL - BIC(C), C = 1,2, \dots, C_{max}\}$

end

6. APPLICATION

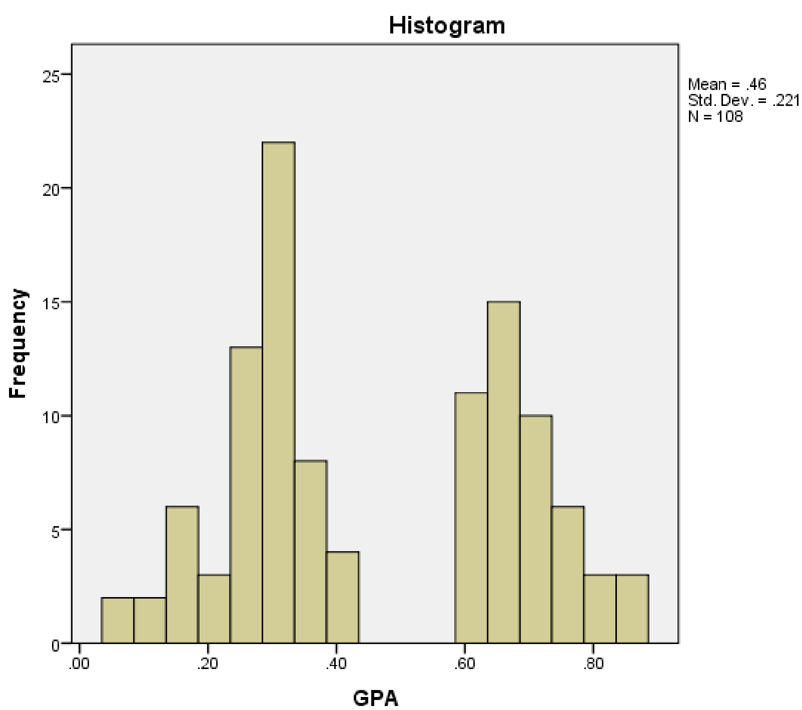
The initial stage of this research was to conduct a study of cluster analysis in the field of business management. The data used is secondary data obtained based on data from Telkom University. Based on data from Telkom University regarding the scores of students and GPA, a univariate beta model will be applied to the data on the proportion of Telkom University students from 2009 to 2017 to classify the scores and GPA of Telecommunication and Informatics Business Management students. The data includes 108 classes for Business Statistics and Economic Mathematics courses and the overall GPA of students. A summary of statistics on proportion data for the three variables is presented in Table 1.

Table 1. Summary of proportion data statistics

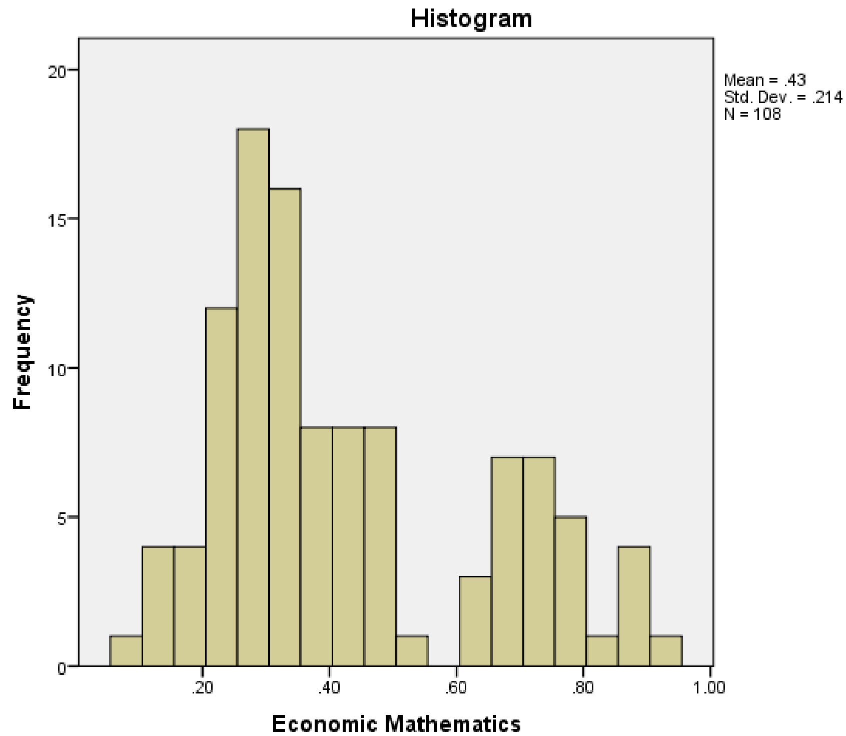
Proportion data statistics	GPA	Economic Mathematics	Business Statistics
Mean	0,4621	0,4273	0,5828
Median	0,3800	0,3350	0,6300
Standard Deviation	0,2205	0,2142	0,1910
Minimum	0,06	0,08	0,22
Maximum	0,85	0,95	0,95

The proportion data for the 3 (three) variables can be illustrated in Figure 2.

GPA



Economic Mathematics



Business Statistics

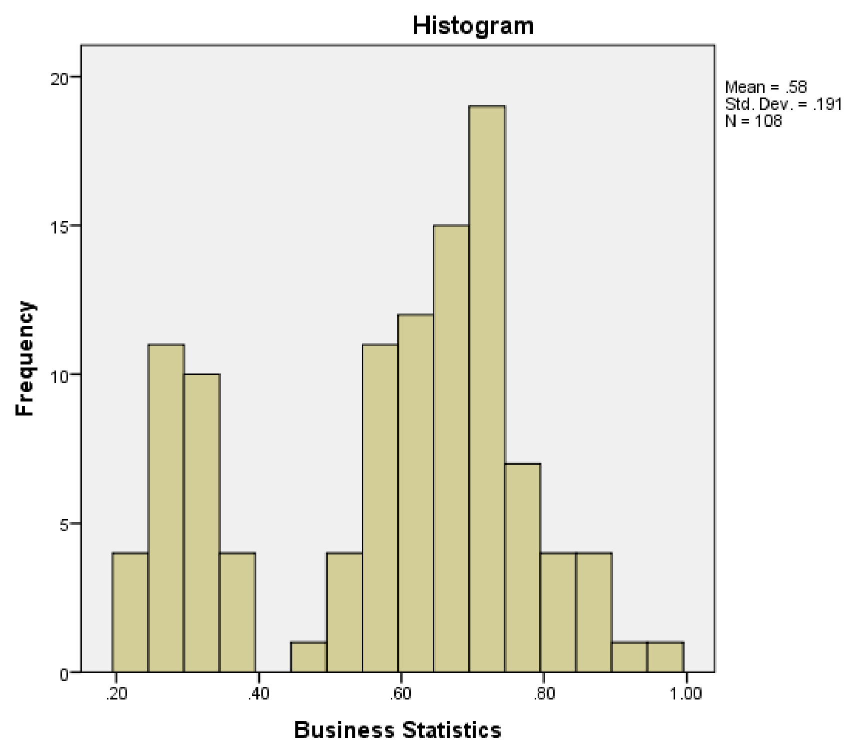


Figure 2: Proportion Histogram Data

7. ANALYSIS RESULTS

GPA

C	**ICL-BIC**
1	-30.0228
2	-66,7877
3	-40.8837
4	30.8365

Economic Mathematics

C	**ICL-BIC**
1	-32.0280
2	-21.5667
3	7.4926
4	101.7882

Business Statistics

C	**ICL-BIC**
1	-55.1038
2	-75.7477
3	-51.1398
4	NaN

Based on the results of the analysis on the GPA data, Economic Mathematics and Business Statistics shows the smallest ICL BIC value in the two clusters for GPA and Business Statistics scores. While the ICL value of BIC in Economic Mathematics shows the smallest ICL BIC value in one cluster. Then it can be concluded that GPA and Business Statistics occur in Mixture 2 clusters.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCE

[1] C.C. Aggarwal, C.K. Reddy (eds), Data clustering. Algorithms and applications. CRC Press, London, (2013).

[2] K. Blekas, A. Likas, N.P. Galatsanos, I.E. Lagaris, A Spatially Constrained Mixture Model for Image Segmentation, IEEE Trans. Neural Netw. 16 (2005), 494–498.

[3] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, Multivariate data analysis (7th ed.). Prentice Hall, Englewood Cliffs, (2010).

[4] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (2001), 177–196.

[5] T.P. Sahu, N.K. Nagwani, S. Verma, Multivariate beta mixture model for automatic identification of topical authoritative users in community question answering sites, IEEE Access. 4 (2016), 5343–5355.

[6] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), IEEE Comput. Soc, Fort Collins, CO, USA, 1999: pp. 246–252.

[7] M. Revow, C.K.I. Williams, G.E. Hinton, Using generative models for handwritten digit recognition, IEEE Trans. Pattern Anal. Machine Intell. 18 (1996), 592–606.

[8] W.G. Zikmund, B.J. Babin, J.C. Carr, M. Griffin, Business research methods, Cengage Learning, Boston, (2009).