

STATISTICS AND DATA SCIENCE



EDITORS

Prin. Dr. M. M. Rajmane
Mrs. S. V. Mahajan
Dr. Mrs. S. P. Patil

**Certified as
TRUE COPY**


Principal

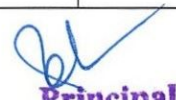
Ramniranjan Jhunjhunwala College,
Ghatkopar (W), Mumbai-400086.

Prarup Publication, Kolhapur

Index

| Sr. No | Title of Paper | Author | P.No |
|--------|--|--|-----------|
| 1. | Application of Logistic Regression Analysis in Microfinance for Women Empowerment | Vanita Suresh Lingayat and Dr. H. S. Lunge | 1 |
| 2. | Comparative Analysis of IT-based Equity Stocks Based on Volatility Using LSTM Technique | Dr. Deepali M. Gala Dr. Bhaskar V. Patil and Dr. Rajesh U. Kanthe | 9 |
| 3. | Trading in the Foreign Exchange Market: A Critical Review Of Literature | Mr. Jayant S. Kadam and Dr. Bhaskar V. Patil | 18 |
| 4. | Analysis of stock market data using Principal component analysis & CART | Ahmed M. Al-Hammadi and Dr. C. D. Sonar | 25 |
| 5. | Statistical Survey of Impact of Online Gaming among Youth | Miss. Aparna Shinde, Miss. Pratiksha Jadhav Miss. Shaheen Sayyad | 53 |
| 6. | A Study on Investment Decisions in Stock Market by Utilizing Predictive Model | Dr. Bhaskar V. Patil, Dr. Deepali M. Gala and Manisha Shinde-Pawar | 59 |
| 7. | A Comparative Study of Data Mining Classification Techniques for Predicting Soil Fertility: Review | Miss. Ashvini G. Patil Dr. Sampada S. Gulwani | 66 |
| 8. | A Conway-Maxwell-Binomial distribution for modelling cluster binary responses | <u>Rahul Ramlalit Tiwari</u> | <u>73</u> |
| 9. | Study of Machine Learning for Rule based Expert System | Dr. J.S. Jadhav | 79 |
| 10. | Study of Machine Learning Algorithms and its Real - World Applications | D.C. Falle, Dr. J.S. Jadhav | 85 |
| 11. | Study of Graph Theory based Machine Learning | S.U.Yadav, Dr.J.S.Jadhav | 92 |
| 12. | To Study the Car Price Prediction Using Machine Learning Algorithm | Miss. Anuja Kumbhar | 100 |
| 13. | A study of Customer Personality Analysis Using Machine learning algorithms | Mr. Lohar Ganesh Ananda | 109 |
| 14. | Application of Chi-square test: Physical Health- Fitness- Diet | Sanjay Karande and Mugdha Tembe | 121 |
| 15. | Prediction of Water Quality by using Data Mining Techniques | Miss. Aishwarya Yalvatkar | 129 |
| 16. | Time Series Analysis And Forecast For Cardiovascular Disease Based On Arima Model | Shashi Rekha B, Manasa B Gowda, Chaithra N | 135 |
| 17. | A statistical study of analysis of India Gross Domestic Product | Miss. Aishwarya G. Desai | 146 |

Certified as
TRUE COPY



Principal
Ramniranjan Jhunjunwala College,
Ghatkopar (W), Mumbai-400086.

ISBN - 978-81-956739-9-5

**A CONWAY-MAXWELL-BINOMIAL DISTRIBUTION FOR
MODELING CLUSTER BINARY RESPONSES**

Rahul Ramlalit Tiwari

Ramniranjan Jhunjhunwala College of Art, Science and Commerce, Mumbai, Maharashtra

Email: rahultiwari@rjcollege.edu.in

Abstract:

In cluster binary response analysis, the Binomial model is not suitable due to the interdependence of the data. In such cases, the Beta-Binomial (BB) model is recommended. An alternative to the BB model is the Conway-Maxwell-Binomial (CMB) distribution. This article focuses on the maximum likelihood estimation of the parameters of the CMB model. There is no analytic solution for likelihood equations, so using a self-written R program based on NR method is used to estimate MLEs iteratively. A simulation study has been conducted to see the behavior of the MLEs.

Keywords: Conway-Maxwell- Binomial distribution, Binomial model, MLEs, NR method.

Introduction:

Cluster binary response refers to a type of data that consists of binary outcomes (i.e., either 0 or 1) for multiple observations within a cluster or group. The observations within a cluster are often dependent or correlated, and as a result, traditional binary response models like the Binomial distribution may not be appropriate for analyzing this type of data. In such cases, alternative models like the Beta-Binomial is often used to capture the dependence structure among the observations within a cluster.


In toxicological studies, the number of occurrences of a certain kind of event in a litter of fetuses which may be death, malformation or mental disorder is recorded. In this example, outcome of an experiment is binary in nature. Mainly, the occurrence or non-occurrence of an event where event can be death of fetus or occurrence of certain malformation in fetus and response of interest is total count of such events. The response variable can be expressed as the sum of Bernoulli random variables for each object under study. To establish this in notation, suppose there are k litters or groups having m_i number of fetuses or objects in i^{th} group. Define $i = 1, 2, \dots, k$; and $j = 1, 2, \dots, m_i$.

$$X_{ij} = \begin{cases} 1 ; & \text{if } j^{th} \text{ subject shows the occurrence of an event in } i^{th} \text{ group.} \\ 0 ; & \text{otherwise} \end{cases}$$

**Certified as
TRUE COPY**

Thus $Y_i = \sum_{j=1}^{m_i} X_{ij}$ represents the total number of occurrences of an event in i^{th} group.

Department of Statistics, S. G.M. College, Karad


Principal
Ramniranjan Jhunjhunwala College,
Ghatkopar (W), Mumbai-400086.

Application of CMB distribution in cluster binary response data can be found in various fields, for e.g, In clinical trials, cluster binary response is observed when track the success or failure of a medical treatment on different patient groups is considered.

In this paper I consider a generalization of the Binomial distribution to a two-parameter distribution which is known as Conway-Maxwell-Binomial (CMB) distribution [2]

The CMB distribution consists of an extra parameter, which we denote by ϑ , and which governs the rate of decay of successive ratios of probabilities such that $P(Y = y - 1)/P(Y = y) = \frac{1}{\theta} \left[\frac{y}{m-y+1} \right]^\vartheta$. The CMB distribution is appealing from a theoretical point of view since it belongs to the exponential family as well as to the two-parameter power series of distributions family. As such, it allows for sufficient statistics and other properties to be elegantly derived [2].

The parameter estimation of the CMB model using the maximum likelihood method brings some challenges, since there are not explicit solutions for the maximum likelihood estimation (MLE) and it is necessary to use iteration methods. Thus, the main objective of this study is to present to estimate the parameters of the CMB model. A simulation study has been conducted to see the behavior of the MLEs.

The CMB distribution and its properties:

A random variable Y is said to follow binomial distribution if assumes only non- negative values and its probability mass function is

$$P(Y = y) = \begin{cases} \binom{m_i}{y} p^y (1 - p)^{m_i - y} & ; y = 0,1,2, \dots, m_i ; 0 < p < 1 ; i= 1, 2, \dots, k \\ 0 & ; \text{ Otherwise} \end{cases} \quad (1)$$

The Conway- Maxwell- Binomial (CMB) distribution is a convenient two parameter family that generalize the binomial distribution and models both positive and negative association among the Bernoulli r.vs.

The probability mass function (pmf) of the CMB distribution [3] is denoted by $Y \sim CMB(m_i, p, \vartheta)$ and given by

$$P(Y = y) = \begin{cases} \frac{\binom{m_i}{y}^\vartheta p^y (1 - p)^{m_i - y}}{Z(p, \vartheta)} & ; y = 0,1,2, \dots, m_i ; 0 < p < 1 ; -\infty \leq \vartheta \\ 0 & ; \text{ Otherwise} \end{cases} \leq \infty \quad (2)$$

Where $Z(p, \vartheta) = \sum_{k=0}^{m_i} \binom{m_i}{k}^\vartheta p^k (1-p)^{m_i-k}$

In particular , when $\vartheta = 1$, the pmf in equation (2) reduces to the binomial distribution, when $\vartheta > 1$, it represent under-dispersion and when $\vartheta < 1$ over-dispersion with respect to the binomial distribution.

The CMB distribution can be expressed as a sum of equi-correlated Bernoulli random variables [3] . The Compoission distribution [1] is approximated to the CMB distribution when m is getting large.

The following figure 1 presents the pmf of the CMB distribution for m = 5 and different values of p and ϑ .

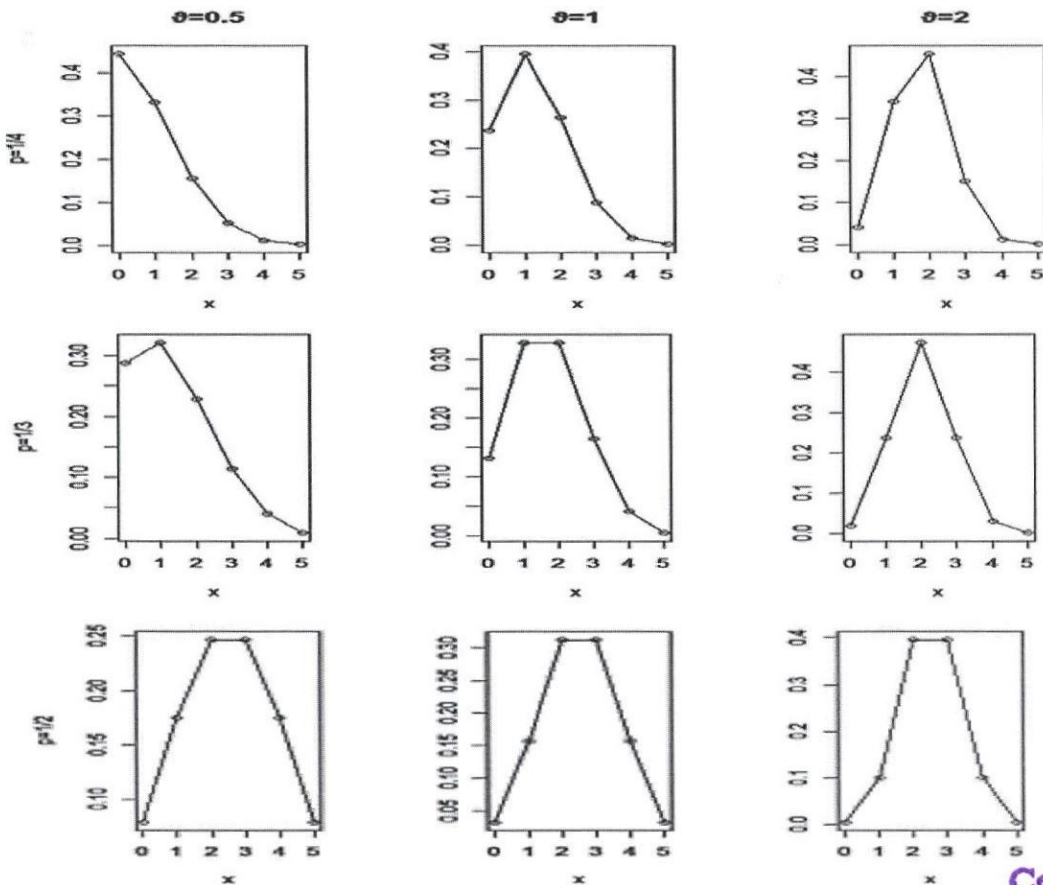


Figure 1 pmf of the CMB distribution

Certified as TRUE COPY

[Signature]
Principal

Ramniranjan Jhunjhunwala College,
Ghatkopar (W), Mumbai-400086.

Considering the reparameterization, $\theta = \frac{p}{1-p}$, the pmf of the CMB distribution is given by

$$P(Y = y) = \begin{cases} \frac{1}{Z(\theta, \vartheta)} \frac{\theta^y}{[y!(m_i - y)!]^\vartheta} & ; y = 0, 1, 2, \dots, m_i; \theta > 0; -\infty \leq \vartheta \leq \infty \quad (3) \\ 0 & ; \text{Otherwise} \end{cases}$$

Where $Q_i = Z(\theta, \vartheta) = \sum_{k=0}^{m_i} \frac{\theta^k}{[k!(m_i - k)!]^\vartheta}$

Maximum Likelihood Estimation of the Parameters:

Let (Y_1, Y_2, \dots, Y_f) , be independent vectors, where each vector has exchangeable binary components.

The Likelihood function $L = L(\theta, \vartheta | y_1, y_2, \dots, y_f)$ will be

$$L = \prod_{i=1}^f P(Y = y_i) \\ = \prod_{i=1}^f \frac{1}{Q_i} \frac{\theta^{y_i}}{[y_i!(m_i - y_i)!]^\vartheta}$$

The log-likelihood function is

$$l = \sum_{i=1}^f \{ y_i \log \theta - \vartheta \log [y_i!(m_i - y_i)!] - \log Q_i \} \text{-----(4)}$$

The partial derivatives are given by

$$\frac{\partial l}{\partial \theta} = \sum_{i=1}^f \left\{ \frac{y_i}{\theta} - \left(\frac{\sum_{k=1}^{m_i} \frac{k\theta^k}{[k!(m_i - k)!]^\vartheta}}{Q_i} \right) \right\} \quad (5)$$

$$\frac{\partial l}{\partial \vartheta} = \sum_{i=1}^f \left\{ -\log [y_i!(m_i - y_i)!] - \left(\frac{\sum_{k=0}^{m_i} \frac{\theta^k \log [k!(m_i - k)!]}{[k!(m_i - k)!]^\vartheta}}{Q_i} \right) \right\} \quad (6)$$

Solving the equations (5) and (6) simultaneously, mle of θ and ϑ can be obtained.

However, these likelihood equations cannot be solved analytically. Hence an iterative method such as Newton-Raphson (NR) iterative method, has to be used to solve likelihood equations.

I have written R program to solve the equations (5) and (6) using NR method.

Simulation Study:

In this section a simulation study has been conducted to see the performance of the estimated parameters. Here, we have generated random observations from CMB with different cluster sizes $K = (20, 50, 80, 150, 200)$ and sample sizes $m_i = 1, 2, 3, 4, 6$ respectively with different combination of true values of parameters θ and ϑ . By using inverse transform technique observation from CMB distribution are generated. Finally, MLEs are computed based on 1000 iterations using self-written R program. Bias and MSE of the parameters given in the Table 1, they are computed using the following formulae.

$$\text{Bias}(\hat{\theta}_1) = E(\hat{\theta}_1) - \theta_1$$

$$\text{MSE}(\hat{\theta}_1) = E(\hat{\theta}_1 - \theta_1)^2$$

$\hat{\theta}_1 = (\hat{\theta}, \hat{\vartheta})$ is estimated parameter and $\theta_1 = (\theta, \vartheta)$ is true parameter .

Here; R = Number of replications = 1000

Table 1. Results of Simulation

| | θ | 0.5 | 0.6 | 0.7 | 0.8 |
|-------------|----------------------------|-----------|-----------|-----------|-----------|
| ϑ | | | | | |
| 0.1 | Bias ($\hat{\theta}$) | -0.001582 | -0.002726 | 0.001176 | -0.002933 |
| | MSE($\hat{\theta}$) | 0.001306 | 0.000886 | 0.000819 | 0.000781 |
| | Bias ($\hat{\vartheta}$) | 0.018667 | 0.003190 | 0.001036 | 0.004220 |
| | MSE ($\hat{\vartheta}$) | 0.007497 | 0.005779 | 0.003995 | 0.003396 |
| 0.2 | Bias ($\hat{\theta}$) | 0.000097 | -0.008027 | -0.003478 | -0.000092 |
| | MSE($\hat{\theta}$) | 0.001169 | 0.000965 | 0.000942 | 0.001044 |
| | Bias ($\hat{\vartheta}$) | -0.010122 | 0.022223 | 0.004813 | 0.002248 |
| | MSE ($\hat{\vartheta}$) | 0.005787 | 0.005399 | 0.005244 | 0.003760 |
| 0.3 | Bias ($\hat{\theta}$) | -0.003050 | -0.005379 | 0.001718 | -0.001319 |

| | | | | | |
|-----|----------------------------|-----------|-----------|-----------|-----------|
| | MSE($\hat{\theta}$) | 0.001092 | 0.000868 | 0.001252 | 0.001293 |
| | Bias ($\hat{\theta}$) | 0.003555 | 0.001344 | -0.007627 | -0.002689 |
| | MSE ($\hat{\vartheta}$) | 0.007814 | 0.005436 | 0.005156 | 0.003559 |
| 0.4 | Bias ($\hat{\theta}$) | 0.004792 | -0.000239 | -0.001968 | 0.003344 |
| | MSE($\hat{\theta}$) | 0.001161 | 0.001079 | 0.000778 | 0.001069 |
| | Bias ($\hat{\vartheta}$) | -0.000503 | 0.013884 | 0.024210 | 0.000919 |
| | MSE ($\hat{\vartheta}$) | 0.005840 | 0.004409 | 0.005784 | 0.004501 |
| 0.5 | Bias ($\hat{\theta}$) | 0.001287 | -0.004993 | 0.002772 | 0.003979 |
| | MSE($\hat{\theta}$) | 0.001187 | 0.001363 | 0.000955 | 0.001239 |
| | Bias ($\hat{\vartheta}$) | 0.007625 | 0.014888 | 0.000682 | 0.004635 |
| | MSE ($\hat{\vartheta}$) | 0.009036 | 0.005867 | 0.005870 | 0.004729 |

From the table we can observed that estimated parameters are quite closer to true parameter and MSE is much smaller.

Conclusion and FutureWork:

In this study, the Conway-Maxwell-Binomial (CMB) distribution is applied in cluster binary data. The parameters are estimated using the method of maximum likelihood estimators. A simulation study has been conducted to see the behavior of the MLEs.

A future prospect of this study, the appropriateness of the fitting distribution is carried out based on the goodness of fit test and some information criteria and applies to the real-life data set and compare with other appropriate distribution.

References:

1. Conway, R. W. and Maxwell, W. L. (1962): A queuing model with state dependent service rates. Journal of Industrial Engineering 12,132-6
2. Kadane, J. B. (2016): Sums of possibly associated Bernoulli variables: The Conway Maxwell-binomial distribution. Bayesian analysis, 11(2), 403-420.
3. Shmueli G., Borle S., and Boatwright P. (2005): A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution, Applied Statistics, 54(1), 127-142.
4. George, E. O. and Bowman, D. (1995): A full likelihood procedure for analysing exchangeable binary data. Biometrics, 51, 512-523.



Rayat Shikshan Sanstha's

SADGURU GADAGE MAHARAJ COLLEGE, KARAD

(AN AUTONOMOUS COLLEGE - Affiliated to Shivaji University, Kolhapur)

Accredited 'A+' with CGPA 3.63 by NAAC @ ISO 9001 : 2015 Certified & NAAC Designed Mentor College

Department of Statistics

INTERNATIONAL CONFERENCE ON

Statistics And Data Science
(SDS-2023)



Sponsored by

RASHTRIYA UCHCHATAR SHIKSHA ABHIYAN (RUSA)

CERTIFICATE

This is to certify that, Prof./Dr./Mr./Ms. Rahul Ramlalit Tiwari
of RT college of Arts & Science
has Participated/Presented a Paper (Oral / Poster) entitled A Conway - Maxwell -
Binomial distribution for Modeling Cluster Binary Responses
in **International Conference on Statistics and Data Science (SDS-2023)** organized by
Dept. of Statistics, Sadguru Gadage Maharaj College, Karad on 24th and 25th February 2023.

S.P. Patil
Dr. Mrs. S. P. Patil
Coordinator

S. V. Mahajan
Mrs. S. V. Mahajan
Convener

M. M. Rajmane
Prin. Dr. M. M. Rajmane
S.G.M. College, Karad

**Certified as
TRUE COPY**

[Signature]
Principal

**Ramniranjan Jhunjhunwala College,
Ghatkopar (W), Mumbai-400086.**