# R PACKAGE DEVELOPMENT TO MODEL OVER DISPERSED BINOMIAL OUTCOME DATA WITH THE USE OF BINOMIAL MIXTURE DISTRIBUTIONS AND ALTERNATE BINOMIAL DISTRIBUTIONS.

**MAHENDRAN AMALAN**

**Department of Statistics and Computer Science, University of Peradeniya,**

**Peradeniya Sri Lanka**

Binomial Outcome Data has vast amount use in the field of biology, medicine and epidemiology. Modeling such data relative to Binomial Mixture Distributions were discussed because Binomial distribution fails to model the data. The reason is actual observed variance of the data is greater than the assumed theoretical variance, which is defined as over dispersion. Alcohol Consumption data is an over dispersed Binomial outcome data. The most flexible distributions to model over dispersed data are Beta Binomial distribution, Kumaraswamy Binomial distribution, Gaussian Hypergeometric Generalized Beta Binomial distribution and McDonald Generalized Beta Binomial distribution. Using the p-values from chi-squared tests and comparing the expected frequencies with the observed frequencies the best model can be decided. Highest p value of 0.8642 among these distributions occurs for Gaussian Hypergeometric Generalized Beta Binomial distribution, with an over dispersion of 0.4325. Other Binomial Mixture Distributions can be modeled but with less effectivity. The Alternate Binomial distributions namely Correlated Binomial distribution. Additive Binomial distribution and Multiplicative Binomial distribution fail to model the alcohol consumption data. The p- values generated are zero for Correlated and Additive Binomial distributions where p-value is 0.0037 for Multiplicative Binomial distribution, which is less than 0.05.

Computational efficiency and effectivity to generate results required were achieved by using R statistical software. Computational code was accumulated into an R package and “fitODBOD” was developed with the use of basic computing and statistical knowledge. Packages “devtools”, “knitr”, “rmarkdown”, “roxygen” and “rtools” were used in the process of development for the Skelton and “hypergeo” and “bbmle” were used to simplify calculations and preserve time. Results comparison is much more useful and easy comparing to the past where manually calculations were conducted. The characteristics and properties of the Binomial Mixture distributions and Alternate Binomial Distributions are also computed into code such as probability mass function, cumulative mass function, and negative log likelihood value and parameter estimations.

Keywords: Binomial Outcome data, over dispersion, Binomial Mixture Distribution, Alternate Binomial Distribution, R package, fitODBOD, bbmle, hypergeo, roxygen, devtools, knitr, rmarkdown.