Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

fitODBOD: An R Package to Model Binomial Outcome Data using Binomial Mixture Distributions

Amalan Mahendran
University of Peradeniya.

Prof. Pushpakanthie Wijekoon
University of Peradeniya.

Abstract

The **fitODBOD** package can be used to identify the best-fitted model for Over-dispersed Binomial Outcome Data (BOD). The Triangular Binomial (TriBin), Beta-Binomial (BetaBin), Kumaraswamy Binomial (KumBin), Gaussian Hypergeometric Generalized Beta-Binomial (GHGBB) and McDonald Generalized Beta-Binomial (McGBB) distributions in the Family of Binomial Mixture Distributions (FBMD) are considered for model fitting in this package. Further, Probability Mass Function (PMF), Cumulative Probability Mass Function (CPMF), Negative Log Likelihood, Over-dispersion and parameter estimation (shape and distribution distinct parameters) can be explored for each fitted model by using **fitODBOD** package.

Keywords: R, **fitODBOD**, BOD, BMD, Over-dispersion.

1. Introduction

Statistical methods are widely used in the areas of research in most of the current disciplines. There is more focus towards fitting distributions to given data since the distributions of data depends on the method of data collection. For example, consider a Binomial experiment that a fair coin is being tossed n number of times. Let the event of falling head be defined as the success of probability p. Then, the number of heads out of n tosses is considered to be a single Binomial variable, Y. Also if similar Binomial experiments occur in N number of different clusters, a collection of $Y_1, Y_2, Y_3, ..., Y_N$ would form the BOD. Such data are frequently mentioned in fields of Toxicology, Biology, Clinical Medicine, Epidemiology and much more. One may attempt to fit the BOD using the traditional Binomial distribution, as it is characterized using the number of identical trials n and the probability of success parameter p. The parameter p ($p \in [0, 1]$) is usually assumed to be a constant from trial to trial and the

trials are independent. In many empirical situations, it has been frequently observed that the actual observed variance of the BOD is greater than the assumed theoretical Binomial variance. This outcome is typically known as "Over-dispersion" ((Cox 1983); (Anderson 1988)). Over-dispersion in BOD can occur either with a probability of success parameter p varying from trial to trial or there is a correlation among binary trials. However, (Collett 1991) argued that the above two cases of Over-dispersion are frequently the same.

According to the literature, the researchers introduced Binomial Mixture distributions as one of the solutions to tackle the problem of Over-dispersion. In this case, p is considered as a continuous random variable bounded between *zero* and *one* by assuming that it randomly varies from cluster to cluster of N clusters. The six members of the Binomial Mixture distributions considered under this study are Uniform Binomial distribution, Triangular Binomial distribution, Beta-Binomial distribution, Kumaraswamy Binomial distribution, Gaussian Hypergeometric Generalized Beta-Binomial distribution and McDonald Generalized Beta-Binomial distribution. The theoretical and simulation studies have shown that the BOD fits these distributions more accurately.

Although Binomial Mixture distributions contain several members, only a few of them were widely used in practice due to computational problems that may appear during estimation procedure in real data situations due to their complex model structures. However, such computations can be done currently through R statistical software.

Conditional on the unknown and random probability of success parameter p, suppose the random variable Y follows a Binomial distribution given $Bin(n, P)$, which is denoted by $Y|p \sim Bin(n, P)$, where P denotes the random variable describing p (Johnson, Kemp, and Kotz (2005)). Then the unconditional distribution of Y is known as the Binomial Mixture distribution. The unconditional Probability Mass Function (PMF) of the Y can be obtained by

$$P_Y(y) = \int P_{Y|p}(y) g_p(p|\psi) dp \tag{1}$$

$$P_{Y|p}(y) = \binom{n}{y} p^y (1-p)^{n-y} \tag{2}$$

for $n = 1, 2, 3, ...,\ y = 0, 1, 2, ..., n.$, where $g_P(p|\psi)$ is the mixing distribution. If Y be a Binomial mixture distributed random variable, then the mean and the variance of the Y are given by

$$E[Y] = n\pi \quad and \quad var[Y] = n\pi(1-\pi) + n(n-1)\sigma^2 \tag{3}$$

where π and σ^2 are mean and variance of the mixing distribution of the random variable P. Let Y be a Binomial mixture distributed random variable such that $Y = \sum_{i=1}^{n} Z_i$ where Z_i is a binary trial, then covariance ($\sum$) and correlation (ρ) between the binary trials are given by

$$\sum = cov(Z_i, Z_j) = \sigma^2 \quad and \quad \rho = corr(Z_i, Z_j) = \frac{\sigma^2}{\pi(1-\pi)} \tag{4}$$

Since σ^2 and π are positive, the correlation between the binary trials is always positive (Z_i's are positively correlated).

Binomial mixture distribution can only model Over-dispersed Binomial outcomes (sum of positively correlated binary trials), but not the Under-dispersed Binomial outcomes (sum of negatively correlated binary trials). The correlation between the binary trials ρ is commonly known as "Over-dispersion parameter" or "Intraclass Correlation parameter" of the Binomial mixture distribution. A way to represent the variance of the Binomial mixture distribution in terms of mean parameter π and Over-dispersion parameter ρ can be obtained by substituting $\sigma^2 = \rho\pi(1-\pi)$ to var[Y] in equation 3 and is given as

$$var[Y] = n\pi(1-\pi)[1+(n-1)\rho]. \tag{5}$$

None of the existing packages **VGAM**, **gamlss.dist**, **ActuDistns**, **BinaryEPPM**, **triangle** and **extraDistr** have an extensive amount of focus towards FBMD or its applications. The objective of introducing the package **fitODBOD** is to resolve some existing computational issues and to provide adequate solutions. In order to build this package successfully, the packages **bbmle**(Bolker and Team 2017) and **hypergeo**(Hankin 2016) were used. The package **bbmle** is used to estimate the parameters using Maximum Log Likelihood method, and the package **hypergeo** was applied for constructing Gaussian Hypergeometric Generalized Beta-Binomial distribution and McDonald Generalized Beta-Binomial distribution.

To fit a Binomial Mixture distribution for a raw BOD set, the following steps have to be used when using this package.

1. Extract the data in a meaningful way (`BODextract` function).
2. Check whether the Binomial distribution can be fitted and if not test the Over-dispersion (`fitBin` function) by using Pearson Chi-square Goodness of Fit test.
3. If Over-dispersion exists, estimate the parameters for each distribution for the given data separately (`EstTriBin`, `EstMLEBetaBin`, `EstMGFBetaBin`, `EstMLEKumBin`, `EstMLEGHGBB`, `EstMLEMcGBB` functions).
4. Based on the above estimated parameters corresponding models can be fitted (`fitTriBin`, `fitBetaBin`, `fitKumBin`, `fitGHGBB`, `fitMcGBB` functions).
5. Finally, compare the results and choose the best-fitted distribution for the data.

To demonstrate the above process the Alcohol Consumption Data, which is the most commonly used dataset by the researchers to explain Over-dispersion will be taken (Lemmens, Knibbe, and Tan 1988). In this dataset, the number of alcohol consumption days in two reference weeks is separately self-reported by a randomly selected sample of 399 respondents from the Netherlands in 1983. Here, the number of days a given individual consumes alcohol out of seven days a week can be treated as a Binomial variable. The collection of all such variables from all respondents would be defined as "Binomial Outcome Data".

The Alcohol consumption data is already in the necessary format to apply steps 2 to 5 and hence, step 1 can be avoided. The steps 2 to 5 can be applied only if the dataset is in the form of a frequency table as follows.

```
R> suppressMessages(library("bbmle"))  #Loading packages
R> suppressMessages(library("fitODBOD"))
R> print(Alcohol_data)  #print the alcohol consumption data set
```

```
  Days week1 week2
1    0    47    42
2    1    54    47
3    2    43    54
4    3    40    40
5    4    40    49
6    5    41    40
7    6    39    43
8    7    95    84
```

```
R> sum(Alcohol_data$week1)  #No of respondents or N
```

```
[1] 399
```

```
R> Alcohol_data$Days  #Binomial random variables or x
```

```
[1] 0 1 2 3 4 5 6 7
```

Suppose your dataset is not a frequency table as shown in the following dataset called `datapoints`. Then the function `BODextract` can be used to prepare the appropriate format as follows.

```
R> datapoints = sample(0:7, 340, replace = TRUE)  #creating a set of raw BOD
R> head(datapoints)  #first few observations of datapoints dataset
```

```
[1] 5 1 4 4 1 7
```

```
R> # extracting and printing BOD in a usable way for the package
R> new_data = BODextract(datapoints)
R> matrix(c(new_data$RV, new_data$Freq), ncol = 2, byrow = FALSE)
```

```
     [,1] [,2]
[1,]    0   37
[2,]    1   49
[3,]    2   31
[4,]    3   33
[5,]    4   40
[6,]    5   44
[7,]    6   61
[8,]    7   45
```

As in the second step we test whether the Alcohol Consumption data follows the Binomial distribution based on the hypothesis given below:

Null Hypothesis : The data follows Binomial Distribution.
Alternate Hypothesis: The data does not follow Binomial Distribution.

Alcohol Consumption data consists of frequency information for two weeks but only the first week is considered for computation. By doing so the researcher can verify if the results acquired from the functions are similar to the results acquired from previous researchers work.

```
R> BinFreq <- fitBin(Alcohol_data$Days, Alcohol_data$week1)
R> print(BinFreq)

Call:
fitBin(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1)

Chi-squared test for Binomial Distribution

      Observed Frequency :  47 54 43 40 40 41 39 95

      expected Frequency :  1.59 13.41 48.3 96.68 116.11 83.66 33.49 5.75

      estimated probability value : 0.5456498

      X-squared : 2911.434   ,df : 6   ,p-value : 0
```

Looking at the p-value it is clear that null hypothesis is rejected at 5% significance level. This indicates that data doesn't fit the Binomial distribution. The reason for a Warning message is that one of the expected frequencies in the results is less than five. Now we compare the actual and the fitted Binomial variances.

```
R> # Actual variance of observed frequencies
R> var(rep(c(0, 1, 2, 3, 4, 5, 6, 7), times = c(47, 54, 43, 40, 40, 41,
R+     39, 95)))

[1] 6.253788

R> # Calculated variance for frequencies of fitted Binomial
R> # distribution
R> var(rep(c(0, 1, 2, 3, 4, 5, 6, 7), times = fitted(BinFreq)))

[1] 1.696035
```

The variance of observed frequencies and the variance of fitted frequencies are 6.253788 and 1.696035 respectively, which indicates Over-dispersion. Since the Over-dispersion exists in the data now it is necessary to fit the Binomial Mixture distributions Triangular Binomial, Beta-Binomial, Kumaraswamy Binomial, GHGBB and McGBB using the package, and select the best-fitted distribution using Negative Log likelihood value, p-value and by comparing

observed and expected frequencies. Definitions and applications of fitting these distributions are given in the next sections.

2. Triangular Binomial Distribution

Suppose $Y|p \sim Bin(n, P)$ and if p has a Triangular distribution with mode parameter c $(0 < c < 1)$ then $P \sim Tri(c)$. Mixing the Triangular distribution with Binomial distribution the simplified resulting PMF and CPMF of Triangular Binomial distribution are given by

$$d_{TriBin}(y) = 2\binom{n}{y}\left(c^{-1}B_c(y+2, n-y+1) + (1-c)^{-1}B(y+1, n-y+2) - (1-c)^{-1}B_c(y+1, n-y+2)\right) \tag{6}$$

and

$$p_{TriBin}(y \leq k) = \sum_{y=0}^{k} 2\binom{n}{y} \left(c^{-1}B_c(y+2, n-y+1) + (1-c)^{-1}B(y+1, n-y+2) - (1-c)^{-1}B_c(y+1, n-y+2)\right) \tag{7}$$

respectively where $n = 1, 2, 3, ...,$ $y = 0, 1, 2, .., n.$ and $k = 0, 1, 2, ..., n.$. Integrals in the form $B_c()$ are in fact incomplete Beta integrals (Abramowitz and Stegun (1964)) and defined as :

$$B_x(\alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt \tag{8}$$

Note that if y follows a Triangular Binomial distribution with parameter c then $x = (1-y)$ follows a Triangular Binomial distribution with parameter $(1-c)$ (Karlis and Xekalaki (2008)).

Let $Y = [Y_1, Y_2, ..., Y_N]^T$ be a random sample of size of N from a Triangular Binomial distribution and n is fixed for all clusters then the unknown parameter c (*mode*) should be estimated as $\hat{c}$. The Likelihood function for $\hat{c}$ is defined as follows:

$$L(\hat{c}|y) = \prod_{i=1}^{N} 2\binom{n}{y_i} \left(c^{-1}B_c(y_i+2, n-y_i+1) + (1-c)^{-1}B(y_i+1, n-y_i+2) - (1-c)^{-1}B_c(y_i+1, n-y_i+2)\right) \tag{9}$$

Then the Negative Log Likelihood function for $\hat{c}$

$$l(\hat{c}|y) = Nlog(2) + \sum_{i=1}^{N} log\binom{n}{y_i}$$
$$+\sum_{i=1}^{N} log\left(c^{-1}B_c(y_i+2, n-y_i+1) + (1-c)^{-1}B(y_i+1, n-y_i+2) - (1-c)^{-1}B_c(y_i+1, n-y_i+2)\right) \tag{10}$$

The estimation of the mode parameter c can be done by using the `EstMLETriBin` function, and then the estimated value has to be applied to `fitTriBin` function to check whether the data fit the Triangular Binomial distribution.

```
R> # estimating the mode
R> modeTB = EstMLETriBin(Alcohol_data$Days, Alcohol_data$week1)
R> coef(modeTB)  #printing the estimated mode

 mode
 0.944444

R> # printing the Negative log likelihood value which is minimized
R> NegLLTriBin(Alcohol_data$Days, Alcohol_data$week1, modeTB$mode)

[1] 880.6167
```

To fit the Triangular Binomial distribution for estimated mode parameter the following hypothesis is used

Null Hypothesis :	The data follows Triangular Binomial Distribution.
Alternate Hypothesis:	The data does not follow Triangular Binomial Distribution.

```
R> # fitting the Triangular Binomial Distribution for the estimated
R> # mode value
R> fitTriBin(Alcohol_data$Days, Alcohol_data$week1, modeTB$mode)

Call:
fitTriBin(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1,
    mode = modeTB$mode)

Chi-squared test for Triangular Binomial Distribution

      Observed Frequency :  47 54 43 40 40 41 39 95

      expected Frequency :  11.74 23.47 35.21 46.94 58.66 70.2 79.57 73.21

      estimated Mode value: 0.944444

      X-squared : 193.6159   ,df : 6   ,p-value : 0

      over dispersion : 0.2308269
```

Since the $p-value$ is 0 which is less than 0.05 it is clear that the null hypothesis is rejected, and the estimated Over-dispersion is 0.2308269. Therefore, it is necessary to fit a better flexible distribution than the Triangular Binomial distribution.

Further, if a researcher needs to construct PMF, CPMF, Negative Log likelihood stated in the equations 6, 7 and 10 he/she can use the `dTriBin`, `pTriBin` and `NegLLTriBin` functions.

3. Beta-Binomial Distribution

Beta-Binomial distribution is acquired by mixing the Beta distribution with the Binomial distribution. The Beta distribution is based on two shape parameters which we denote by α and β. The distribution is denoted by $Beta(\alpha, \beta)$, and depending on the shape parameters there is a high flexibility to obtain a wide variety of shapes.

Suppose $Y|p \sim Bin(n, P)$ and $P \sim Beta(\alpha, \beta)$. The PMF and CPMF of Beta-Binomial distributed random variable Y are given by

$$d_{BetaBin}(y) = \binom{n}{y} \frac{B(\alpha + y, n + \beta - y)}{B(\alpha, \beta)} \tag{11}$$

where $n = 1, 2, 3, ...,\ y = 0, 1, 2, ..., n$ and $\alpha, \beta > 0$, and

$$p_{BetaBin}(y \leq k) = \sum_{y=0}^{k} \binom{n}{y} \frac{B(\alpha + y, n + \beta - y)}{B(\alpha, \beta)} \tag{12}$$

respectively, where $k = 0, 1, 2, ..., n$..

PMF of Beta-Binomial distribution is often defined by three parameters, Binomial number of trials n, the two shape parameters of the Beta distribution, α and β. Further, using the PMF we can construct the Likelihood function for $\Omega_{BB} = (\alpha, \beta)^T$ as given below:

$$L(\Omega_{BB}|y) = \prod_{i=1}^{N} \binom{n}{y_i} \frac{B(\alpha + y_i, n + \beta - y_i)}{B(\alpha, \beta)} \tag{13}$$

The Negative Log Likelihood function is given as

$$l(\Omega_{BB}|y) = \sum_{i=1}^{N} log\binom{n}{y_i} + \sum_{i=1}^{N} log(B(\alpha + y_i, n + \beta - y_i)) - Nlog(B(\alpha, \beta)) \tag{14}$$

To estimate the two shape parameters of the Beta-Binomial distribution Methods of Moments or Maximum Likelihood estimation can be used. In this article the Methods of Moments is discussed only for Beta-Binomial distribution. First Maximum Likelihood estimates of Ω_{BB} are obtained by minimizing the Negative Log Likelihood function with respect to Ω_{BB} .

Minimizing the Negative Log Likelihood value when there are two or more parameters to be estimated in the distribution requires the `"mle2"` function of **bbmle** package. In this article, parameters are estimated twice for such distributions from here onwards. Which means two sets of initial random values in the domain of the parameters are used as inputs and results are obtained. If the user wishes he/she can estimate parameters more than twice for more

accuracy and clarity. For a further understanding of the "mle2" function, refer the **bbmle** manual.

Negative Log Likelihood value can be minimized by using the function EstMLEBetaBin and the function "mle2" of package **bbmle**. In order to estimate the shape parameters α and β (a and b), initial shape parameter values have to be given by the user to this "mle2" function. These initial values have to be in the domain of the shape parameters. Then, by adding random values we can estimate these parameters multiple times and compare them to see which are best fitted. Below given is the first pair of estimates for initial values where $a = 0.1$ and $b = 0.1$.

```
R> # estimating the shape parameters a,b
R> estimate = bbmle::mle2(EstMLEBetaBin, start = list(a = 0.1, b = 0.1),
R+     data = list(x = Alcohol_data$Days, freq = Alcohol_data$week1))
R> estimate@min  #extracting the minimized Negative log likelihood value

[1] 813.4571

R> # extracting the estimated shape parameter a
R> a1 = bbmle::coef(estimate)[1]
R> # extracting the estimated shape parameter b
R> b1 = bbmle::coef(estimate)[2]
R> print(c(a1, b1))  #printing the estimated shape parameters

        a         b
0.7229420 0.5808483
```

To fit the Beta-Binomial distribution for estimated (Maximum Likelihood Estimation method) shape parameters the following hypothesis is used

Null Hypothesis :	The data follows Beta-Binomial Distribution by the Maximum Likelihood Estimates.
Alternate Hypothesis:	The data does not follow Beta-Binomial Distribution by the Maximum Likelihood Estimates.

```
R> # fitting Beta Binomial Distribution for estimated shape
R> # parameters
R> fitBetaBin(Alcohol_data$Days, Alcohol_data$week1, a1, b1)

Call:
fitBetaBin(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1,
    a = a1, b = b1)

Chi-squared test for Beta-Binomial Distribution

          Observed Frequency :  47 54 43 40 40 41 39 95

          expected Frequency :  54.62 42 38.9 38.54 40.07 44 53.09 87.78
```

```
          estimated a parameter : 0.722942   ,estimated b parameter : 0.5808483

          X-squared : 9.5171   ,df : 5   ,p-value : 0.0901

          over dispersion : 0.4340673
```

The $p - value$ of $0.0901 > 0.05$ indicates that the null hypothesis is not rejected. Current estimated shape parameters fit the Beta-Binomial distribution. Note that the estimated Over-dispersion parameter is 0.4340673. Below is the results for the randomly selected second set of initial shape parameter values where $a = 10$ and $b = 5$ while minimizing the Negative Log Likelihood function.

```
R> # estimating the shape parameters
R> estimate = bbmle::mle2(EstMLEBetaBin, start = list(a = 10, b = 5),
R+     data = list(x = Alcohol_data$Days, freq = Alcohol_data$week1))
R> estimate@min  #extracting the minimized negative log likelihood value

[1] 813.4571

R> # extracting the estimated shape parameter a
R> a2 = bbmle::coef(estimate)[1]
R> # extracting the estimated shape parameter b
R> b2 = bbmle::coef(estimate)[2]
R> print(c(a2, b2))  #printing the estimated shape parameters

        a         b
0.7229428 0.5808493

R> # fitting Beta-Binomial Distribution to estimated shape parameters
R> fitBetaBin(Alcohol_data$Days, Alcohol_data$week1, a2, b2)

Call:
fitBetaBin(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1,
    a = a2, b = b2)

Chi-squared test for Beta-Binomial Distribution

          Observed Frequency :  47 54 43 40 40 41 39 95

          expected Frequency :  54.62 42 38.9 38.54 40.07 44 53.09 87.78

          estimated a parameter : 0.7229428   ,estimated b parameter : 0.5808493

          X-squared : 9.5171   ,df : 5   ,p-value : 0.0901

          over dispersion : 0.4340669
```

Here the $p-value$ is 0.0901 which is greater than 0.05. The expected frequencies, chi-squared values, Negative Log Likelihood values are the same while comparing previous results obtained from the first set of initial input parameters for Beta-Binomial distribution, but there is a negligible difference for Over-dispersion and estimated shape parameters. Considering the two different pairs of initial input shape parameters, the decision is still the null hypothesis not rejected at 5% significance level.

Now using the Method of Moments, Ω_{MM} is obtained by equating the first two population moments with the corresponding sample moments. Let $Y = [Y_1, Y_2, ..., Y_N]^T$ be a random sample of size N from a Beta-Binomial distribution. Here, n is fixed for all clusters. The estimated unknown parameter space for $\Omega_{MM} = (\alpha, \beta)^T$ is then given by

$$\hat{\Omega}_{MM} = \begin{bmatrix} \hat{\alpha}_{MM} \\ \\ \hat{\beta}_{MM} \end{bmatrix} = \begin{bmatrix} \left(\frac{(nm1-m2)m1}{n(m2-m1-m1^2)+m1^2}\right) \\ \\ \left(\frac{(nm1-m2)(n-m1)}{n(m2-m1-m1^2)+m1^2}\right) \end{bmatrix} \tag{15}$$

where

$$m1 = \sum_{i=1}^{N} \frac{y_i}{N} \quad and \quad m2 = \sum_{i=1}^{N} \frac{y_i^2}{N} \tag{16}$$

are the first two sample moments. Function `EstMGFBetaBin` is built based on the equations 15 and 16, and used as below.

```
R> # estimating the shape parameter a
R> a = EstMGFBetaBin(Alcohol_data$Days, Alcohol_data$week1)$a
R> # estimating the shape parameter b
R> b = EstMGFBetaBin(Alcohol_data$Days, Alcohol_data$week1)$b
R> print(c(a, b))  #printing the estimated parameters a,b

[1] 0.7161628 0.5963324

R> # finding the minimized negative log likelihood value
R> print(NegLLBetaBin(Alcohol_data$Days, Alcohol_data$week1, a, b))

[1] 813.5872
```

To fit the Beta-Binomial distribution for estimated (Method of Moments) shape parameters the following hypothesis is used

Null Hypothesis :	The data follows Beta-Binomial Distribution by the Method of Moments.
Alternate Hypothesis:	The data does not follow Beta-Binomial Distribution by the Method of Moments.

```
R> # fitting Beta-Binomial Distribution to estimated shape parameters
R> fitBetaBin(Alcohol_data$Days, Alcohol_data$week1, a, b)
```

```
Call:
fitBetaBin(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1,
    a = a, b = b)

Chi-squared test for Beta-Binomial Distribution

          Observed Frequency :  47 54 43 40 40 41 39 95

          expected Frequency :  56.6 43.01 39.57 38.97 40.27 43.89 52.39 84.29

          estimated a parameter : 0.7161628   ,estimated b parameter : 0.5963324

          X-squared : 9.7362   ,df : 5   ,p-value : 0.0831

          over dispersion : 0.4324333
```

Results from Method of Moments to estimate the parameters have led to a $p-value$ of 0.0831 which is greater than 0.05 indicates that the null hypothesis is not rejected. The parameters estimated through Method of Moments fit the Beta-Binomial distribution for an estimated Over-dispersion of 0.4324333. *Note that by using this method there can be only a unique pair of shape parameters for a given dataset.*

Further, if a researcher needs to construct PMF, CPMF, Negative Log Likelihood stated in equations 11, 12 and 14 he/she can use use the `dBetaBin`, `pBetaBin` and `NegLLBetaBin` functions.

4. Uniform Binomial Distribution

Uniform distribution has a very simple probabilistic structure, and hence the distribution does not have the ability to model a wide variety of shapes. Therefore, it is not generally used as a mixing distribution to the probability of success random variable P. (Horsnell 1957) proposed and used both Uniform and Triangular distributions as mixing distributions to p in economical acceptance sampling studies. By following (Horsnell 1957) the Uniform Binomial distribution can be derived from the Beta-Binomial distribution when shape parameters of the Beta distribution are predefined such as $\alpha = 1$ and $\beta = 1$ ($a = 1$ and $b = 1$).

Derivation of the Uniform Binomial distribution through Beta-Binomial distribution is given below:

$$Y \sim BetaBin(n, \alpha = 1, \beta = 1) \sim UniBin(n) \tag{17}$$

After simplifying Beta and Gamma functions of equations 11 and 12, a simpler form for the PMF and CPMF can be obtained as below:

$$d_{UniBin} = \frac{1}{(n+1)} \quad and \quad p_{UniBin} = \sum_{y=0}^{k} \frac{1}{(n+1)} \tag{18}$$

,where $n = 1, 2, 3, ...,\ y = 0, 1, 2, ..., n.$ and $k = 0, 1, 2, ..., n.$
With its inability to provide a variety of shapes, the functions `NegLLUniBin`, `EstMLEUniBin` and `fitUniBin` are not constructed for Uniform Binomial distribution. The usefulness of the other functions `dUniBin` and `pUnibin` related to the Uniform Binomial distribution are discussed below.

```
R> dUniBin(0:10, 10)$pdf  #extracting the pdf values

 [1] 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909
 [7] 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909

R> dUniBin(0:10, 10)$mean  #extracting the mean

[1] 5

R> dUniBin(0:10, 10)$var  #extracting the variance

[1] 10

R> dUniBin(0:10, 10)$over.dis.para  #extracting the over dispersion

[1] 0.3333333

R> pUniBin(0:15, 15)  #acquiring the cumulative probability values

 [1] 0.0625 0.1250 0.1875 0.2500 0.3125 0.3750 0.4375 0.5000 0.5625 0.6250
[11] 0.6875 0.7500 0.8125 0.8750 0.9375 1.0000
```

5. Kumaraswamy Binomial Distribution

Mixing the Kumaraswamy distribution ($P \sim Kum(\alpha, \beta)$) with the Binomial distribution ($Y|p \sim Bin(n, P)$) we can generate the Kumaraswamy Binomial distribution. The PMF of Kumaraswamy Binomial distributed random variable Y is then given by,

$$d_{KumBin}(y) = \alpha\beta\binom{n}{y}\sum_{j=0}^{\infty}(-1)^j\binom{\beta-1}{j}B(y+\alpha+\alpha j, n-y+1) \tag{19}$$

where $n = 1, 2, 3, ...,\ y = 0, 1, 2, ..., n.$ and $\alpha, \beta > 0.$
Then the CPMF is given by

$$p_{KumBin}(y \leq k) = \sum_{y=0}^{k}\alpha\beta\binom{n}{y}\sum_{j=0}^{\infty}(-1)^j\binom{\beta-1}{j}B(y+\alpha+\alpha j, n-y+1) \tag{20}$$

for $k = 0, 1, 2, ..., n.$

Similarly to the Beta-Binomial distribution, PMF of Kumaraswamy Binomial distribution is also often defined by three parameters, the Binomial number of trials n, and the two shape parameters of the Kumaraswamy distribution α and β. Hence, a wide variety of discrete shapes can be modeled by using this distribution (Li, Huang, and Zhao 2011). Method of Maximum Likelihood estimation can be applied to estimate the two unknown parameters α and β (a and b). However, the Method of Moments estimation may not be directly applied to estimate the parameters since it is not possible to obtain explicit expressions for α and β. Let $Y = [Y_1, Y_2, ..., Y_N]^T$ be a random sample of size N from a Kumaraswamy Binomial distribution and the number of binomial trials n be fixed for all clusters. If the unknown parameter space is $\Omega_{KB} = (\alpha, \beta)^T$, then the Likelihood function of Ω_{KB} can be obtained as follows

$$L(\Omega_{KB} \mid y) = \prod_{i=1}^{N} \alpha\beta \binom{n}{y_i} \sum_{j=0}^{\infty} (-1)^j \binom{\beta - 1}{j} B(y_i + \alpha + \alpha j, n - y_i + 1) \tag{21}$$

The Negative Log Likelihood function for Ω_{KB} is given as

$$l(\Omega_{KB} \mid y) = Nlog(\alpha\beta) + \sum_{i=1}^{N} log \binom{n}{y_i} + \sum_{i=1}^{N} log \left(\sum_{j=0}^{\infty} (-1)^j \binom{\beta - 1}{j} B(y_i + \alpha + \alpha j, n - y_i + 1) \right) \tag{22}$$

When calculating the above expressions the upper limit of the above summation (i.e ∞) is replaced by assigning a variable `"it"`. The default value of `"it"` is taken as 25000. Therefore, when estimating parameters the value of `"it"` also has to be estimated in addition to a and b. Note that the initial `"it"` value needs to be very large and it is safe to set `"it"` above 10000. To understand how the `"it"` parameter affects the time for parameter estimation and fitting the distribution several tasks are done in the code. First, the system time when each function begins to run is recorded and when it completes execution again the system time is noted. Then for different `"it"` values, the time differences are calculated. Note, this task is only done for Kumaraswamy Binomial distribution.

Using the **bbmle** package and its `"mle2"` function, the shape parameters α, β and `"it"` are estimated and fitted below. Suppose the first set of randomly selected input parameters are $a = 10.1$, $b = 1.1$ and $it = 10000$.

```
R> # time before beginning the estimation
R> begin <- Sys.time()
R> # estimating the shape parameters and iteration value
R> estimate = bbmle::mle2(EstMLEKumBin, start = list(a = 10.1, b = 1.1,
R+     it = 10000), data = list(x = Alcohol_data$Days, freq = Alcohol_data$week1))
R> # time after finishing the estimation
R> end <- Sys.time()
R> print(end - begin)  #time for estimation
R> estimate@min  #extracting the minimized negative log likelihood value
R>
R> # extracting the shape parameter a
```

```
R> a1 = bbmle::coef(estimate)[1]
R> # extracting the shape parameter b
R> b1 = bbmle::coef(estimate)[2]
R> # extracting the 'it' value which replaces infinity
R> it1 = bbmle::coef(estimate)[3]
R> print(c(a1, b1, it1))  #print shape parameters and iteration value
```

To fit the Kumaraswamy Binomial distribution for estimated shape parameters the following hypothesis is used

Null Hypothesis :	The data follows Kumaraswamy Binomial Distribution.
Alternate Hypothesis:	The data does not follow Kumaraswamy Binomial Distribution.

```
R> # time before beginning the fitting
R> begin <- Sys.time()
R> # fitting Kumaraswamy Binomial distribution to estimated shape
R> # parameters and it1*100
R> KBFreq <- fitKumBin(Alcohol_data$Days, Alcohol_data$week1, a1, b1,
R+     it1 * 100)
R> print(KBFreq)
R>
R> # time after finishing the fitting
R> end <- Sys.time()
R> print(end - begin)  #time for fitting
R>
R> KBFreq$NegLL
```

The null hypothesis is not rejected at 5% significance level ($p - value = 0.0711$)for the estimated parameters $it = 999999.6$, $a = 0.7302994$, $b = 0.616263$ and the estimated Over-dispersion=0.4230119. According to the results it takes 14.43116 minutes to estimate the parameters and 5.728181 minutes to fit the distribution. These values are very high with compared to the previous two Binomial mixture distributions. The above estimated time durations depend on the estimated and used it values.

Now, we randomly select a second set of initial parameters as $a = 10$, $b = 1$ and $it = 20000$. As before, time to estimate the parameters and fitting the distribution are calculated

```
R> # time before beginning the estimation
R> begin <- Sys.time()
R> # estimating the shape parameters and iteration value
R> estimate = bbmle::mle2(EstMLEKumBin, start = list(a = 10, b = 1, it = 20000),
R+     data = list(x = Alcohol_data$Days, freq = Alcohol_data$week1))
R> # time after finishing the estimation
R> end <- Sys.time()
R> print(end - begin)  #time for estimation
R> estimate@min  #extracting the minimized negative log likelihood value
R>
R> # extracting the shape parameter a
```

```
R> a2 = bbmle::coef(estimate)[1]
R> # extracting the shape parameter b
R> b2 = bbmle::coef(estimate)[2]
R> # extracting the 'it' value which replaces infinity
R> it2 = bbmle::coef(estimate)[3]
R> print(c(a2, b2, it2))  #print the shape parameters and iteration value
R>
R> # time before beginning the fitting
R> begin <- Sys.time()
R> # fitting Kumaraswamy Binomial distribution to estimated shape
R> # parameters and iteration value*100
R> KBFreq <- fitKumBin(Alcohol_data$Days, Alcohol_data$week1, a2, b2,
R+     it2 * 100)
R> print(KBFreq)
R> # time after finishing the fitting
R> end <- Sys.time()
R> print(end - begin)  #time for fitting
R>
R> # calculating the negative log likelihood value for it2*100
R> KBFreq$NegLL
```

Note that the null hypothesis is now not rejected at 5% significance level ($p-value = 0.0733$) for the selected initial values of the parameters. The estimated values for the current set of initial input parameters are $it = 1999998$, $a = 0.7231563$ $b = 0.6100055$ and estimated Over-dispersion=0.4252777. According to the results, its takes 1.750839 hours to estimate the parameters and 46.06927 minutes to fit the distribution. The above estimated time durations depend on the estimated it value which is 1999998.

Further, if a researcher needs to construct PMF, CPMF and Negative Log Likelihood stated in equations 19, 20 and 22 he/she can use `dKumBin`, `pKumBin` and `NegLLKumBin` functions.

6. Gaussian Hypergeometric Generalized Beta-Binomial Distribution

The PMF and CPMF of a Gaussian Hypergeometric Generalized Beta-Binomial distributed random variable Y are obtained from (Kemp 1975) and (Johnson, Kotz, and Balakrishnan 1995). They are given below in their respective order

$$d_{GHGBB}(y) = \frac{1}{2F1(-n, \alpha; -\beta - n + 1; \lambda)} \binom{n}{y} \frac{B(y+\alpha, n-y+\beta)}{B(\alpha, \beta+n)} \lambda^y \tag{23}$$

, and

$$p_{GHGBB}(y \leq k) = \sum_{y=0}^{k} \frac{1}{2F1(-n, \alpha; -\beta - n + 1; \lambda)} \binom{n}{y} \frac{B(y+\alpha, n-y+\beta)}{B(\alpha, \beta+n)} \lambda^y \tag{24}$$

where $n = 1, 2, 3, ...$, $y = 0, 1, ..., n.$, $\alpha, \beta, \lambda > 0$, $k = 0, 1, 2, ..., n.$, and

$$2F1(-n, \alpha; -\beta - n + 1; \lambda) = \sum_{i=0}^{\infty} \frac{(-n)_i (\alpha)_i}{(-\beta - n + 1)_i} \frac{\lambda^i}{i!}. \tag{25}$$

In equation 25 set $(\alpha)_i = \alpha(\alpha+1)...(\alpha+i-1) = \frac{\Gamma(\alpha+i)}{\Gamma(\alpha)}$, and the term $2F1(-n, \alpha; -\beta-n+1; \lambda)$ is the Gaussian Hypergeometric function $(-n, \alpha; -\beta - n + 1; \lambda)$ which can be written as $GHD(-n, \alpha; -\beta - n + 1; \lambda)$. Beta-Binomial belongs to the family of distributions that are generalized by the Gaussian Hypergeometric function. Therefore, Beta-Binomial distribution ($BetaBin(n, \alpha, \beta)$) can be derived by constraining $\lambda = 1$ of GHGBB distribution ($GHGBB(n, \alpha, \beta, \lambda = 1)$ where $GHD(-n, \alpha, -\beta - n + 1, \lambda = 1)$) .

The PMF of the Gaussian Hypergeometric Generalized Beta-Binomial distribution is defined by four parameters, Binomial number of trials n, two shape parameters of the Beta-Binomial distribution α, β, and an additional scalar parameter λ. Mean, variance and Over-dispersion parameter of the GHGBB distribution are computed by means of integration procedures. Complexity of the $GHGBB(n, \alpha, \beta, \lambda)$ distribution and how the Maximum Likelihood estimation method is used to estimate the three unknown parameters α, β and λ (a, b and c) are given in (Rodriguez-Avi, Conde-Sanchez, Saez-Castillo, and Olmo-Jiminez 2007).

Let $Y = [Y_1, Y_2, ..., Y_N]^T$ be a random sample of size N from a GHGBB distribution. Here, n is fixed for all clusters. Now we define the unknown parameter space as $\Omega_{GHGBB} = (\alpha, \beta, \lambda)^T$. The Likelihood function of Ω_{GHGBB} can be obtained as follows

$$L(\Omega_{GHGBB}|y) = \prod_{i=1}^{N} \binom{n}{y_i} \frac{1}{2F1(-n, \alpha; -\beta - n + 1; \lambda)} \frac{B(y_i + \alpha, n - y_i + \beta)}{B(\alpha, \beta + n)} \lambda^{y_i} \tag{26}$$

Then, Negative Log Likelihood function for Ω_{GHGBB} is given as

$$l(\Omega_{GHGBB}|y) = \sum_{i=1}^{N} \left(log\binom{n}{y_i} + log(B(y_i + \alpha, n - y_i + \beta)) + y_i log(\lambda) \right) - Nlog\left(B(\alpha, \beta + n)\right) - Nlog\left(2F1(-n, \alpha; -\beta - n + 1; \lambda)\right) \tag{27}$$

Since GHGBB distribution contains $2F1(-n, \alpha; -\beta - n + 1; \lambda)$ it depends on the GHD function. However, not all input parameters produce valid values for the GHD function. The `"hypergeo_powerseries"` function from the package **"hypergeo"** is used to calculate the valid GHD function values. Since this function produces complex numbers extraction methods are applied only to extract valid (real and positive) values.

Now we estimate the shape parameters and fit the GHGBB distribution for the first set of randomly selected initial input shape parameters of $a = 10.1$, $b = 1.1$ and $c = 5$.

```
R> # estimating the shape parameters
R> estimate = bbmle::mle2(EstMLEGHGBB, start = list(a = 10.1, b = 1.1,
R+     c = 5), data = list(x = Alcohol_data$Days, freq = Alcohol_data$week1))
R> estimate@min  #extracting the minimized negative log likelihood value
```

```
[1] 809.2767

R> # extracting the shape parameter a
R> a1 = bbmle::coef(estimate)[1]
R> # extracting the shape parameter b
R> b1 = bbmle::coef(estimate)[2]
R> # extracting the shape parameter c
R> c1 = bbmle::coef(estimate)[3]
R> print(c(a1, b1, c1))  #printing the shape parameters

        a         b         c
1.3506836 0.3245421 0.7005210
```

To fit the GHGBB distribution for estimated shape parameters the following hypothesis is used.

Null Hypothesis :	The data follows Gaussian Hypergeometric Generalized Beta-Binomial Distribution.
Alternate Hypothesis :	The data does not follow Gaussian Hypergeometric Generalized Beta-Binomial Distribution.

```
R> # fitting GHGBB distribution for estimated shape parameters
R> fitGHGBB(Alcohol_data$Days, Alcohol_data$week1, a1, b1, c1)

Call:
fitGHGBB(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1,
    a = a1, b = b1, c = c1)

Chi-squared test for Gaussian Hypergeometric Generalized Beta-Binomial Distribution

      Observed Frequency :  47 54 43 40 40 41 39 95

      expected Frequency :  47.88 50.14 46.52 42.08 38.58 37.32 41.78 94.71

      estimated a parameter : 1.350684   ,estimated b parameter : 0.3245421  ,estimated c

      X-squared : 1.2835   ,df : 4   ,p-value : 0.8642

      over dispersion : 0.4324874
```

The null hypothesis is not rejected at 5% significance level ($p - value = 0.8642$). The estimated shape parameters are $a = 1.3506836$, $b = 0.3245421$ and $c = 0.7005210$, where the estimated Over-dispersion=0.4324874. Below is the second set of randomly selected initial shape parameters, where $a = 1$, $b = 21$ and $c = 1.1$.

```
R> # estimating the shape parameters
R> estimate = bbmle::mle2(EstMLEGHGBB, start = list(a = 1, b = 21, c = 1.1),
R+     data = list(x = Alcohol_data$Days, freq = Alcohol_data$week1))
R> estimate@min  #extracting the minimized negative log likelihood value
```

```
[1] 809.2767

R> # extracting the shape parameter a
R> a2 = bbmle::coef(estimate)[1]
R> # extracting the shape parameter b
R> b2 = bbmle::coef(estimate)[2]
R> # extracting the shape parameter c
R> c2 = bbmle::coef(estimate)[3]
R> print(c(a2, b2, c2))  #printing the shape parameters

        a         b         c
1.3506541 0.3245486 0.7005286

R> # fitting GHGBB distribution for estimated shape parameters
R> fitGHGBB(Alcohol_data$Days, Alcohol_data$week1, a2, b2, c2)

Call:
fitGHGBB(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1,
    a = a2, b = b2, c = c2)

Chi-squared test for Gaussian Hypergeometric Generalized Beta-Binomial Distribution

      Observed Frequency :  47 54 43 40 40 41 39 95

      expected Frequency :  47.88 50.14 46.52 42.08 38.58 37.32 41.78 94.71

      estimated a parameter : 1.350654   ,estimated b parameter : 0.3245486  ,estimated c

      X-squared : 1.2835   ,df : 4   ,p-value : 0.8642

      over dispersion : 0.4324875
```

Here, the expected frequencies, chi-squared values, p-values, Over-dispersion parameters and Negative Log Likelihood values are the same as the results received when using the first set of initial parameters for GHGBB. The null hypothesis is not rejected at 5% significance level ($p - value = 0.8642 > 0.05$).

Further, if a researcher needs to construct PMF, CPMF and Negative Log Likelihood stated in equations 23, 24 and 27 he/she can use `dGHGBB`, `pGHGBB` and `NegLLGHGBB` functions.

7. McDonald Generalized Beta-Binomial Distribution

The McDonald Generalized Beta-Binomial distribution ($Y \sim McGBB(n, \alpha, \beta, \gamma)$) is constructed by mixing the Generalized Beta distribution of First kind ($P \sim GB1(\alpha, \beta, \gamma)$) to the probability of success parameter P of the conditional Binomial distribution ($Y|p \sim Bin(n, P)$) (Manoj, Wijekoon, and Yapa 2013). Therefore, the random variable Y has the McGBB distribution with parameters n, α, β and γ. Here, n is the Binomial number of trials and α, β, λ are the three shape parameters of the Generalized Beta distribution of the First kind.

Let Y be a discrete random variable that follows a McGBB as defined above. Then, the PMF of McGBB distributed random variable Y is given by:

$$d_{McGBB}(y) = \binom{n}{y} \frac{\gamma}{B(\alpha,\beta)} \sum_{i=0}^{\infty} (-1)^i \binom{\beta-1}{i} B(y+\alpha\gamma+\gamma i, n-y+1) \tag{28}$$

, where $n = 1, 2, 3, ...$, $y = 0, 1, 2, ..., n.$ and $\alpha, \beta, \gamma > 0$.

The above PMF is valid as it is obtained by means of a well-known compound formula, but an infinite series occurs inside the PMF. A rearranged PMF of McGBB distributed random variable Y can be constructed by representing the Infinite series as a Finite series. Now, a modified PMF of McGBB is given below for Y as

$$d_{McGBB}(y) = \binom{n}{y} \frac{1}{B(\alpha,\beta)} \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} B\left(\frac{y}{\gamma}+\alpha+\frac{j}{\gamma},\beta\right) \tag{29}$$

and CPMF is given by

$$p_{McGBB}(y \le k) = \sum_{y=0}^{k} \binom{n}{y} \frac{1}{B(\alpha,\beta)} \sum_{j=0}^{n-y} (-1)^j \binom{n-y}{j} B\left(\frac{y}{\gamma}+\alpha+\frac{j}{\gamma},\beta\right) \tag{30}$$

,where $n = 1, 2, 3, ...$, $y = 0, 1, 2, ..., n.$, $\alpha, \beta, \gamma > 0$ and $k = 0, 1, 2, ..., n$.

Due to McGBB distribution's three shape parameters it can be used to model a variety of Over-dispersed BOD. This new Binomial mixture distribution includes both Beta-Binomial distribution and Kumaraswamy Binomial distribution as its nested distributions. If $Y \sim McGBB(n, \alpha, \beta, \gamma)$ then by setting $\gamma = 1$ we obtain the Beta-Binomial distribution with parameters n, α and β and by setting $\alpha = 1$ we obtain the Kumaraswamy Binomial distribution with parameters n, β and γ. It can be shown that the first three moments of the Beta-Binomial and Kumaraswamy Binomial distribution can be obtained using the first three moments of the McGBB distribution.

The Methods of Moments estimation cannot be applied to estimate the parameters of McGBB distribution, due to the very complex form of first three population moments of the McGBB distribution. Let $Y = [y_1, y_2, ...y_N]^T$ be a random sample of size N from a McGBB distribution with the unknown parameter vector space $\Omega_{McGBB} = (\alpha, \beta, \gamma)^T$. Here, n is fixed for all clusters. It is suitable to use the re-arranged PMF in equation 29 to find the Negative Log Likelihood.

Now the Likelihood function of Ω_{McGBB} is

$$L(\Omega_{McGBB}|y) = \prod_{i=1}^{N} \binom{n}{y_i} \frac{1}{B(\alpha,\beta)} \sum_{j=0}^{n-y_i} (-1)^j \binom{n-y_i}{j} B\left(\frac{y_i}{\gamma}+\alpha+\frac{j}{\gamma},\beta\right) \tag{31}$$

Then, the Negative Log Likelihood function for Ω_{McGBB} is

$$l(\Omega_{McGBB}|y) = Nlog\left(\frac{1}{B(\alpha,\beta)}\right) + \sum_{i=1}^{N} log\binom{n}{y_i} + \sum_{i=1}^{N} log\left(\sum_{j=0}^{n-y_i} (-1)^j \binom{n-y_i}{j} B\left(\frac{y_i}{\gamma}+\alpha+\frac{j}{\gamma},\beta\right)\right) \tag{32}$$

Equation 32 is used to construct the functions `NegLLMcGBB`, `EstMLEMcGBB` in their respective order to calculate the Negative Log Likelihood value and estimate the shape parameters α, β and λ (a, b and c). Here also, the function `"mle2"` of package **bbmle** is used to estimate the shape parameters. Given below is the results generated for the first set of randomly selected initial input parameters where $a = 1.1$, $b = 1.1$ and $c = 1.5$.

```
R> # estimating the shape parameters
R> estimate = bbmle::mle2(EstMLEMcGBB, start = list(a = 1.1, b = 1.1,
R+     c = 1.5), data = list(x = Alcohol_data$Days, freq = Alcohol_data$week1))
R> estimate@min  #extracting the negative log likelihood value which is minimized

[1] 809.6707

R> # extracting the shape parameter a
R> a1 = bbmle::coef(estimate)[1]
R> # extracting the shape parameter b
R> b1 = bbmle::coef(estimate)[2]
R> # extracting the shape parameter c
R> c1 = bbmle::coef(estimate)[3]
R> print(c(a1, b1, c1))  #printing the shape parameters

          a           b           c
 0.03509023  0.18713142 25.43961892
```

To fit the McGBB distribution for estimated shape parameters the following hypothesis is used

Null Hypothesis : The data follows McDonald Generalized Beta-Binomial Distribution.
Alternate Hypothesis: The data does not follow McDonald Generalized Beta-Binomial Distribution.

```
R> # fitting the MCGBB distribution for estimated shape parameters
R> fitMcGBB(Alcohol_data$Days, Alcohol_data$week1, a1, b1, c1)

Call:
fitMcGBB(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1,
    a = a1, b = b1, c = c1)

Chi-squared test for Mc-Donald Generalized Beta-Binomial Distribution

      Observed Frequency :  47 54 43 40 40 41 39 95

      expected Frequency :  51.13 45.65 43.2 41.66 40.59 40.05 41.71 95.01

      estimated a parameter : 0.03509023   ,estimated b parameter : 0.1871314  ,estimated

      X-squared : 2.1352   ,df : 4   ,p-value : 0.7109

      over dispersion : 0.4350398
```

The null hypothesis is not rejected at 5% significance level ($p - value = 0.7109 > 0.05$). The estimated shape parameters are $a = 0.03509023$ $b = 0.18713142$ and $c = 25.43961892$, and the estimated Over-dispersion=0.4350398. Now, for the randomly selected second set of initial input parameters $a = 0.1$, $b = 10$ and $c = 20$ results are obtained below.

```
R> # estimating the shape parameters
R> estimate = bbmle::mle2(EstMLEMcGBB, start = list(a = 0.1, b = 10,
R+     c = 20), data = list(x = Alcohol_data$Days, freq = Alcohol_data$week1))
R> estimate@min  #extracting the negative log likelihood value which is minimized

[1] 809.6659

R> # extracting the shape parameter a
R> a2 = bbmle::coef(estimate)[1]
R> # extracting the shape parameter b
R> b2 = bbmle::coef(estimate)[2]
R> # extracting the shape parameter c
R> c2 = bbmle::coef(estimate)[3]
R> print(c(a2, b2, c2))  #printing the shape parameters

          a           b           c
 0.03435423  0.18419747 25.96049262

R> # fitting the McGBB distribution for estimated shape parameters
R> fitMcGBB(Alcohol_data$Days, Alcohol_data$week1, a2, b2, c2)

Call:
fitMcGBB(x = Alcohol_data$Days, obs.freq = Alcohol_data$week1,
    a = a2, b = b2, c = c2)

Chi-squared test for Mc-Donald Generalized Beta-Binomial Distribution

      Observed Frequency :  47 54 43 40 40 41 39 95

      expected Frequency :  51.24 45.7 43.23 41.67 40.59 40.03 41.62 94.92

      estimated a parameter : 0.03435423   ,estimated b parameter : 0.1841975  ,estimated

      X-squared : 2.1235   ,df : 4   ,p-value : 0.7131

      over dispersion : 0.4351125
```

For the second set of estimated parameters, there is a slight difference in the values of expected frequencies, chi-squared values, p-values, Negative Log Likelihood values, and Over-dispersion parameters. The null hypothesis is not rejected at 5% significance level ($p - value = 0.7131 > 0.05$). If we closely analyze the observed and expected frequencies according to the above results the difference between them is very narrow.

Further, if a researcher needs to construct PMF, CPMF and Negative Log Likelihood stated in equations 29, 30 and 32 he/she can use `dMcGBB`, `pMcGBB` and `NegLLMcGBB` functions.

8. Summary of the Results

Table 1 presents the expected frequencies, chi-squared values, p-values, estimated parameters, Negative Log Likelihood values and Over-dispersion of the Binomial Mixture distributions obtained above for the Alcohol Consumption data. Further, for the above fitted distributions difference values (the difference between observed frequency and expected frequency) are calculated and included in the table in brackets with next to the most fitted expected frequencies of that distribution.

Table 1: The results of expected frequencies, difference values, Chi-squared test values, the degree of freedoms, p-value, estimated parameters, Negative Log Likelihood values and Over-dispersion parameters of the Triangular Binomial distribution, Beta-Binomial distribution, Kumaraswamy Binomial distribution, Gaussian Hypergeometric Generalized Beta-Binomial distribution and McDonald Generalized Beta-Binomial distribution

Random Variable	**Observed Frequencies**	**Triangular Binomial Distribution**		**Beta-Binomial Distribution**		**Kumaraswamy Binomial Distribution**		**GHGBB Distribution**		**McGBB Distribution**	
Method		**MLE**	**MGF**	**MLE 1**	**MLE 2**	**MLE 1**	**MLE 2**	**MLE 1**	**MLE 2**	**MLE 1**	**MLE 2**
Initial parameter Values		**No Need**	**No Need**	**a=0.1, b=0.1**	**a=10, b=5**	**a=10.1, b=1.1, it=1000**	**a=10, b=1, it=20000**	**a=10.1, b=1.1, c=5**	**a=1, b=21, c=1.1**	**a=1.1, b=1.1, c=1.5**	**a=0.1, b=10, c=20**
0	47	11.74, (35.26)	56.60	54.62	54.62, (-7.62)	51.62	53.59, (-6.59)	47.88	47.88,(-0.88)	51.13	51.24,(-4.24)
1	54	23.47, (30.53)	43.01	42.00	42.00, (12.00)	42.79	42.05, (11.95)	50.14	50.14, (3.86)	45.65	45.70, (8.30)
2	43	35.21, (7.79)	39.57	38.90	38.90, (4.10)	40.77	39.40, (3.60)	46.52	46.52,(-3.52)	43.20	43.23,(-0.23)
3	40	46.94, (-6.94)	38.97	38.54	38.54, (1.46)	40.88	39.30, (0.70)	42.08	42.08,(-2.08)	41.66	41.67,(-1.67)
4	40	58.66,(-18.66)	40.27	40.07	40.07, (-0.07)	42.54	40.97, (-0.97)	38.58	38.58, (1.42)	40.59	40.59,(-0.59)
5	41	70.20,(-29.20)	43.89	44.00	44.00, (-3.00)	46.20	44.91, (-3.91)	37.32	37.32, (3.68)	40.05	40.03, (0.97)
6	39	79.57,(-40.57)	52.39	53.09	53.09,(-14.09)	54.03	53.66,(-14.66)	41.78	41.78,(-2.78)	41.71	41.62,(-2.62)
7	95	73.21, (21.79)	84.29	87.78	87.78, (7.22)	80.17	85.09, (9.91)	94.71	94.71, (0.29)	95.01	94.92, (0.08)
Total	**399**	**399**	**398.99**	**399**	**399**	**399**	**398.97**	**399.01**	**399.01**	**399**	**399**
Chi-Squared test Statistic		193.6159	9.7362	9.5171	9.5171	11.1524	10.0705	1.2835	1.2835	2.1352	2.1235
df		6	5	5	5	5	5	4	4	4	4
p-value		0	0.0831	0.0901	0.0901	0.0484	0.0733	0.8642	0.8642	0.7109	0.7131
Estimated parameter values		mode= 0.944444	a=0.7161628, b=0.5963324	a=0.7229420, b=0.5808483	a=0.7229428, b=0.5808493	a=0.7755977, b=0.6558136, it=100000435.16	a=0.7231563, b=0.6100055, it=1999998	a=1.3506836, b=0.3245421, c=0.7005210	a=1.3506541, b=0.3245486, c=0.7005286	a=0.03509023, b=0.18713142, c=25.43961892	a=0.03435423, b=0.18419747, c=25.96049262
Negative Log Likelihood		880.6167	813.5872	813.4571	813.4571	814.2771	813.7668	809.2767	809.2767	809.6707	809.6659
Over-Dispersion		0.2308269	0.4324333	0.4340673	0.4340669	0.4091329	0.4252777	0.4324874	0.4324875	0.4350398	0.4351125

The best-fitted distribution is chosen by comparing three main measurements extracted from the results shown in Table 1, which are p-value, Negative Log Likelihood value and the count of difference between expected and observed frequencies in the range of ±5. Then the following three criteria will be considered for the selection procedure

1. The maximum p-value > 0.05 from the hypothesis test.
2. The minimum Negative Log Likelihood value.
3. The number of difference values within the range of ±5.

The following Table 2 shows the respective values of the above criteria and the corresponding Over-dispersion value for each distribution constructed for Alcohol Consumption data.

Table 2: Summary of results corresponding to the three criteria and Over-dispersion value.

Distribution	P-value	Negative Log Likelihood Value	Count of difference values	Over-dispersion
GHGBB	**0.8642**	**809.2767**	**8**	**0.4324875**
McGBB	0.7131	809.6659	7	0.4351125
KumBin	0.0733	813.7668	4	0.4252777
BetaBin	0.0901	813.4571	4	0.4340669
TriBin	0	880.6167	0	0.2308269

The fitted distributions are GHGBB, McGBB, Beta-Binomial distribution and Kumaraswamy Binomial distribution in the order of decreasing p-values. Triangular Binomial distribution cannot be fitted since its p-value < 0.05. The Negative Log Likelihood values of GHGBB and McGBB distributions are the lowest and are quite similar. Based on the count of difference values for the Beta-Binomial distribution it is four out of eight and for Kumaraswamy Binomial distribution also it is four out of eight, but for the McGBB distribution it is seven out of eight counts. Further, Over-dispersion parameters of all four fitted distributions are same for the second decimal point (Over-dispersion=0.43).The best-fitted distribution GHGBB has the highest p-value of 0.8642, the lowest Negative Log Likelihood value of 809.2767, the count of difference values is eight out of eight and indicates an estimated Over-dispersion of 0.4324875.

9. Conclusion

The **fitODBOD** package is constructed for the main purpose of fitting the given BOD and being able to choose the best-fitted Binomial Mixture Distribution. The package has functions to calculate PMF, CPMF and Negative Log Likelihood of Triangular Binomial, Beta-Binomial, Kumaraswamy Binomial, GHGBB and McGBB distributions. Further, there are functions for probability density, cumulative density and moment about zero values for Triangular, Beta, Kumaraswamy, Gaussian Hypergeometric Generalized Beta and Generalized Beta of First kind distributions. Using the steps mentioned on page 3 of the article, the most fitted Binomial Mixture distribution is decided.

It is clearly noted that Alcohol Consumption data can be best fitted by using GHGBB distribution more effectively and efficiently. However McGBB, Kumaraswamy Binomial, Beta-Binomial distributions can also be used to fit this data set, but the efficiency is less with compared to the GHGBB. Therefore, in a practical situation a researcher can make a decision after fitting all five distributions for a BOD set.

References

Abramowitz M, Stegun IA (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Ninth Dover Printing, Tenth GPO Printing edition. Dover, New York.

Anderson DA (1988). "Some models for Overdispersed Binomial Data." *Australian & New Zealand Journal of Statistics*, **30**(2), 125–148. URL https://doi.org/10.1111/j.1467-842X.1988.tb00844.x.

Bolker B, Team RDC (2017). *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.20, URL https://CRAN.R-project.org/package=bbmle.

Collett D (1991). *Modelling Binary Data*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall. ISBN 9780412387906.

Cox DR (1983). "Some Remarks on Overdispersion." *Biometrika*, **70**(1), 269–274. URL https://doi.org/10.1093/biomet/70.1.269.

Hankin RKS (2016). *hypergeo: The Gauss Hypergeometric Function*. R package version 1.2-13, URL https://CRAN.R-project.org/package=hypergeo.

Horsnell G (1957). "Economical Acceptance Sampling Schemes." *Journal of the Royal Statistical Society. Series A (General)*, **120**(2), 148–201. ISSN 00359238. URL https://doi.org/10.2307/2342822.

Johnson NL, Kemp AW, Kotz S (2005). *Univariate Discrete Distributions*, volume 444. John Wiley & Sons.

Johnson NL, Kotz S, Balakrishnan N (1995). *Continuous Univariate Distributions*. Wiley, New York,.

Karlis D, Xekalaki E (2008). "The Polygonal Distribution." pp. 21–33. URL https://doi.org/10.1007/978-0-8176-4626-4_2.

Kemp Adrienne Wand Kemp CD (1975). "Models for Gaussian Hypergeometric Distributions." pp. 31–40. URL https://doi.org/10.1007/978-94-010-1842-5_4.

Lemmens P, Knibbe RA, Tan F (1988). "Weekly Recall and Diary Estimates of Alcohol Consumption in a General Population Survey." *Journal of Studies on Alcohol*, **49**(2), 131–135. URL https://doi.org/10.15288/jsa.1988.49.131.

Li X, Huang Y, Zhao X (2011). "The Kumaraswamy Binomial Distribution." *Chinese Journal of Applied Probability and Statistics*, **27**(5), 511–521.

Manoj C, Wijekoon P, Yapa RD (2013). "The McDonald Generalized Beta-Binomial Distribution: A New Binomial Mixture Distribution and Simulation Based Comparison with Its Nested Distributions in Handling Overdispersion." *International Journal of Statistics and Probability*, **2**(2), 24–41. ISSN 1927-7040. URL https://doi.org/10.5539/ijsp.v2n2p24.

Rodriguez-Avi J, Conde-Sanchez A, Saez-Castillo AJ, Olmo-Jiminez MJ (2007). "A Generalization of the Beta-binomial Distribution." *Journal of the Royal Statistical Society. Series C: Applied Statistics*, **56**(1), 51–61. ISSN 00359254. URL https://doi.org/10.1111/j.1467-9876.2007.00564.x.

Affiliation:

Amalan Mahendran
University of Peradeniya.
Kandy, Sri Lanka
E-mail: amalan0595@gmail.com
URL: https://amalan-con-stat.netlify.com/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd