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,

Sri Lanka

INTRODUCTION

Binomial outcome data are frequently mentioned in fields of toxicology, biology, clinical medicine, epidemiology and much more. Consider a binomial experiment that a fair coin is being tossed ***n*** number of times. Let the event of falling head be defined as success of probability ***p.*** Then, the number of heads out of ***n*** tosses is considered to be a single Binomial variable, ***Y***. Consider similar binomial experiments occurring in ***N*** number of different places , collection of ***Y1 ,Y2,…,YN*** would form the Binomial Outcome Data. It is crucial to choose an appropriate parametric distribution model which better fits the set of empirical data. The traditional Binomial distribution can be suitably used in modeling the random variable Y only if the Bernoulli trials forming the Binomial random variable are independent and identical. 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 be explained by two reasons. First reason is saying there exists a variation in the probability of success parameter p of those binary trials. Second is saying there exists correlations among the binary trials which make up the binomial. Collett (2003). “Binomial Mixture Distribution” is developed for to solve this issue, simply consider that p as a continuous random variable bounded between 0 and 1 because it randomly varies from place to place of N places. There are six distributions of Binomial Mixture Distribution. Overcome this issue of over dispersion and under dispersion Altham (1978) constructed theoretical distributions such as Correlated Binomial distribution, Additive Binomial distribution and Multiplicative Binomial distribution.

To reach effective and efficient results with less time spent the need of computational software is essential by looking through the equations related to the distributions and tasks related to modeling the empirical BOD . A collection of computational code accumulated will encourage to use the distributions more frequently, and develop new Binomial Mixture distributions. In perspective to achieve this an open platform software is used, R statistical software. A package named “fitODBOD”accumulating all the necessary equations into functions of R software are constructed and guided how to develop a package as well.

LITERATURE REVIEW

The least concerned distribution of all is Uniform Binomial distribution. Horsnell (1957) developed the distribution, it lacks flexibility in modeling BOD. Arnold, B.C et al. (2008) has used the Course data from Athens University of economics to fit the Triangular Binomial distribution, while comparing the results to Beta Binomial distribution. Beta Binomial distribution is used in both classical statistics and Bayesian statistics. In classical statistics this distribution is mainly used to model over dispersed binomial outcome data. Bayesian framework, beta binomial is the posterior predictive distribution of the binomial distribution, when beta distribution is used as the conjugate prior of the probability of success. Person's (1925) on empirical Bayes methods can be regarded as the earliest work in this sense. Wiiliams (1975) and Haseman and Kupper (1979) distribution in toxicological experiments to model "litter effect". Chatfield and Goodhardt (1970) modeled consumer purchasing behaviors in market researches. Lindsey and Altham (1998) used distribution to analyze human sex ratio. Ennis and Bi (1998) importance of beta-binomial distribution in sensory and consumer researches.

Since Kumaraswamy Binomial distribution is relatively recent member to the statistical literature, both Classical statistics and Bayesian Statistics applications of this distribution are yet to be discussed. Rodriguez-avi et.al (2007) used Gaussian Hypergeometric Generalized Beta-Binomial distribution to model Alanko and Lemmens alcohol consumption data and an academic pass rates data. Manoj, C., Wijekoon, P. and Yapa R.D (2013) have initially mentioned the McDonald Generalized Beta Binomial distribution with the process of developing the necessary equations and computational code.

There is ample amount of literature to guide how to create a package but considering the topic of over dispersion there are no mention of any task based packages. Still there are few packages who intend to show the characteristics of the distributions. Such as probability function values, cumulative function values, and parameter estimation.

METHODOLOGY

Figure 1.1 indicates a simple step by step process of explaining how to develop a R package with the accommodation of R statistical software knowledge and basic computing knowledge. There are few supportive packages such as "roxygen", "devtools", "knitr" and rmarkdown to reduce the amount of stressful manual work of the outdated methods. Also figure 1.1 indicates the most crucial parts of the R package to work as a whole and be user friendlier as well. Especially vignettes and Manual two pdf documents to help the user understand how to use the package with examples related to the functions. The conversion of all the necessary equations of the Binomial Mixture distributions and Alternate Binomial distributions is also quite useful because it demonstrates a certain standard of coding that should be maintained.

Situations such as using supplementary packages in order to make computation easy is a simple example where there are certain methods ensure that they do not interfere the developers functions.

RESULTS AND DISCUSSION

Considering the table 1.1 it clearly indicates that Triangular Binomial Distribution does not have the ability to model the Alcohol Consumption data week 1 from Alanko and Lemmens (1996), considering the p - value is less than 0.05 and expected frequencies do not match the observed frequencies. Using similar techniques in the Beta Binomial distribution the most suitable method is Maximum Likelihood Estimation (MLE) even though Moment generation method exists because p – value is 0.0901 for MLE. Considering the Kumarswamy Binomial distribution the distribution converges when the iteration value is 20000 and this leads up to the output of p- value 0.0733, where previously for iteration value 1000 p – value is not greater than 0.05. In the last two distributions the MLE method is used and p - value for Gaussian Hypergeometric Generalized Beta Binomial distribution is 0.8642 where McDonald Generalized Beta Binomial distribution has 0.7131. Therefore considering all the p –values and further looking at the expected frequencies which came close to the observed frequencies the scientific choice is Gaussian Hypergeometric distribution.

Table 1.2 indicates from the considered Alternate Binomial distributions none tend to exceed the p –value of 0.05, directly Correlated Binomial and Additive Binomial distributions produce zero for p –value and multiplicative Binomial distribution has p - value of 0.0037.

The R package development is a wide region of information in order to clearly understand and learn the most simplest to produce an output the developer should be working on task oriented or user oriented. This means that the developer should have the basic idea of what is the purpose of the package and for whom is being developed.

CONCLUSION

Finally to conclude, the amount of knowledge and experience that a developer gains from developing a R package will further improve the package and make it user friendly so that even a non-statistician would be able to use it without any help. Future aspect includes creating a Graphical user interfaces to run the functions needed without any manual coding and to model the BOD data as well.

References

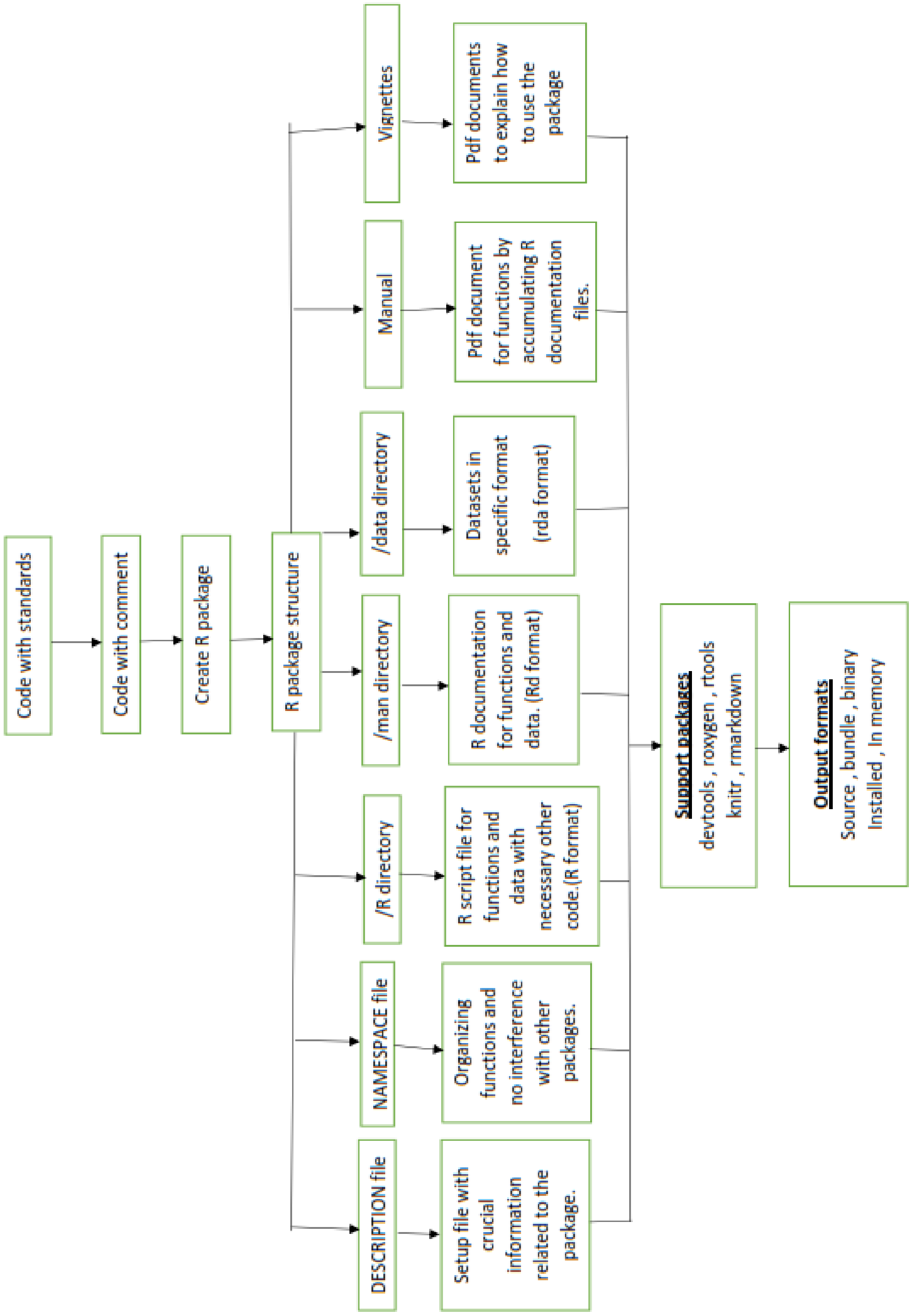
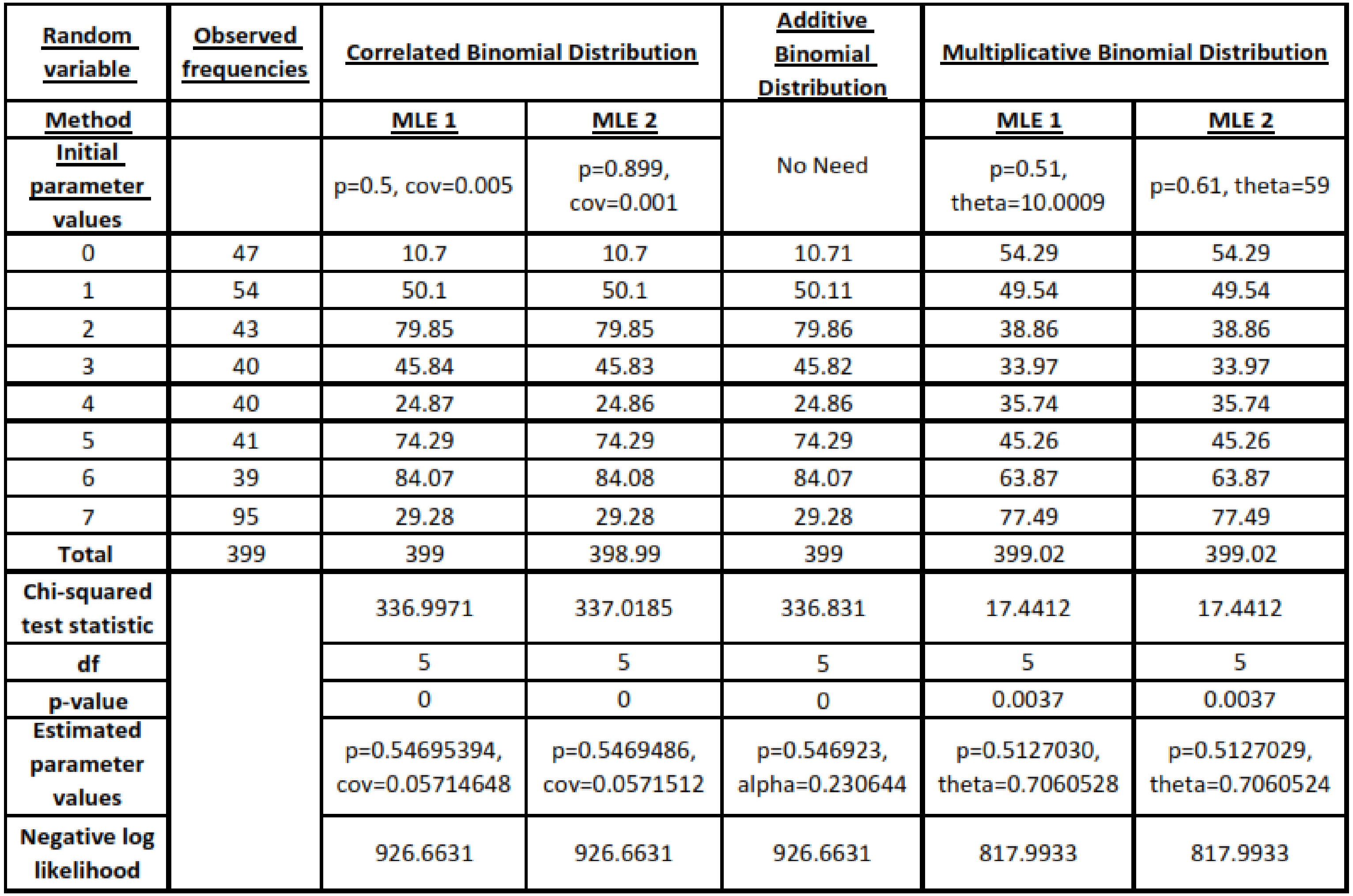

Code with standards
Code with comment
Create R package
R package structure
DESCRIPTION file
Setup file with crucial information related to the package.
NAMESPACE file
Organizing functions and no interference with other packages.
/R directory
R script file for functions and data with necessary other code.(R format)
/man directory
R documentation for functions and data. (Rd format)
/data directory
Datasets in specific format (rda format)
Manual
Pdf document for functions by accumulating R documentation files.
Vignettes
Pdf documents to explain how to use the package
Support packages
devtools , roxygen , rtools
knitr , rmarkdown
Output formats
Source , bundle , binary
Installed , In memory

Figure 1.1

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=15 , b=5 , 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	56.6	54.62	54.62	51.62	53.59	47.88	47.88	51.13	51.24
1	54	23.47	43.01	42	42	42.79	42.05	50.14	50.14	45.65	45.7
2	43	35.21	39.57	38.9	38.9	40.77	39.4	46.52	46.52	43.2	43.23
3	40	46.94	38.97	38.54	38.54	40.88	39.3	42.08	42.08	41.66	41.67
4	40	58.66	40.27	40.07	40.07	42.54	40.97	38.58	38.58	40.59	40.59
5	41	70.2	43.89	44	44	46.2	44.91	37.32	37.32	40.05	40.03
6	39	79.57	52.39	53.09	53.09	54.03	53.66	41.78	41.78	41.71	41.62
7	95	73.21	84.29	87.78	87.78	80.17	85.09	94.71	94.71	95.01	94.92
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	816.2508	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

Table 1.1- Results to modeling BOD using Binomial Mixture distributions



Random variable	**Observed frequencies**	**Correlated Binomial Distribution**		**Additive Binomial Distribution**	**Multiplicative Binomial Distribution**	
Method		**MLE 1**	**MLE 2**	No Need	**MLE 1**	**MLE 2**
Initial parameter values		p=0.5, cov=0.005	p=0.899, cov=0.001		p=0.51, theta=10.0009	p=0.61, theta=59
0	47	10.7	10.7	10.71	54.29	54.29
1	54	50.1	50.1	50.11	49.54	49.54
2	43	79.85	79.85	79.86	38.86	38.86
3	40	45.84	45.83	45.82	33.97	33.97
4	40	24.87	24.86	24.86	35.74	35.74
5	41	74.29	74.29	74.29	45.26	45.26
6	39	84.07	84.08	84.07	63.87	63.87
7	95	29.28	29.28	29.28	77.49	77.49
Total	399	399	398.99	399	399.02	399.02
Chi-squared test statistic		336.9971	337.0185	336.831	17.4412	17.4412
df		5	5	5	5	5
p-value		0	0	0	0.0037	0.0037
Estimated parameter values		p=0.54695394, cov=0.05714648	p=0.5469486, cov=0.0571512	p=0.546923, alpha=0.230644	p=0.5127030, theta=0.7060528	p=0.5127029, theta=0.7060524
Negative log likelihood		926.6631	926.6631	926.6631	817.9933	817.9933

Table 1.2 – Results to modeling BOD using Alternate Binomial distributions