Estimation of Parameters for a Generalized Hypergeometric Distribution Generated by Birth-Death Process

Davood Farbod[*]

Department of Mathematics, Quchan University of Technology, Quchan, Iran

Introduction

The mechanism of the dynamic of bioinformatics systems are often described using Birth-Death Process with various types of coefficients. The stationary distributions of the process can be used as frequency distributions of different phenomena arising in bioinformatics. Based on Birth-Death Processes several frequency distributions have been proposed for the needs of phenomena in bioinformatics. One of the important frequency distributions which introduced by Danielian and Astola (2006), is a three-parametric regularly varying generalized Hypergeometric distribution. The probability mass function of this distribution is as follows:

$$
\begin{cases}
p_x(\theta) = p_0(\theta) \prod_{k=0}^{x-1} \frac{(p_1+k)(p_2+k)}{(1+k)(q+k)} & x = 1,2,3,\dots \\
p_0(\theta) = \left(1 + \sum_{y=1}^{\infty} \prod_{k=0}^{y-1} \frac{(p_1+k)(p_2+k)}{(1+k)(q+k)}\right)^{-1}
\end{cases}
\tag{1}
$$

where $\theta = (p_1, p_2, q)$. The unknown parameters are $0 < p_1 < \infty$, $0 < p_2 < \infty$, $0 < q < \infty$ and also $q - p_1 - p_2 > 0$.

The lack of closed formulas for the probability mass function and cumulative distribution function, is a major drawback to the use of this distribution. In this paper, we are going to consider some statistical properties for the model (1).

Material and methods

In this paper, maximum likelihood estimation (MLE) of the unknown parameters of the model (1) are obtained. An approximate computation is discussed for estimating the parameters. Using the method of Markov Chain Monte Carlo (MCMC), simulation studies are done.

In order to propose real applications of the model (1), some real data sets in bioinformatics are fitted with this model and then compare with four other frequency distributions. All of computations are done with the help of R statistical software.

Results and discussion

Under satisfying some conditions, the MLE are given for the model (1) which coincides with the solution of the system of likelihood equations. The presented MLE are the same as some moment estimators.

We use an approximate method for MLE of the unknown parameters. Based on MCMC method we see that our method works well. Mean square error (MSE) are presented as well. It has been shown that by increasing the sample size the MSE and the number of iterations are decreased.

To show the applications of the model (1), three real data sets are fitted with this model. According to p-value, logarithm of likelihood function ($\ln L$), Akaike Information Criterion (AIC), AIC with Correction (AICc), we compare the model (1) with four other frequency distributions. It seems that the model (1) gives a better fit with respect to other models.

Conclusion

The following conclusions were obtained from this research.

- Under satisfying some conditions, we obtain the MLE.
- An approximate computation for the MLE of the unknown parameters are given.
- Simulation study is illustrated.
- Some real data sets are fitted with the model (1) which shows a satisfactory fit to these real data sets.
- Based on some well-known statistical criterions, we compare the model (1) with four other frequency distributions arising in bioinformatics. It is readily seen that the model (1) provides a better fit than four other models.

Keywords: Generalized Hypergeometric Distribution, Birth-Death Process, Bioinformatics, Maximum likelihood estimation (MLE), Markov Chain Monte Carlo (MCMC).

*Corresponding author: d.farbod@qiet.ac.ir