Original Article

Probability Model of Allele Frequency of Alzheimer's Disease Genetic Risk Factor

***Afshin Fayyaz-Movaghar**[1] Mohammad Taghi Kamel-Mirmostafaee[1] Tahura Sadat Borhani[1]

1- Department of Statistics, School of Mathematical Sciences, University of Mazandaran, Babolsar, Iran

***a_fayyaz@umz.ac.ir**

(Received: 16 Jan 2016; Revised: 17 Mar 2016; Accepted: 11 May 2016)

Abstract

Background and Purpose: The identification of genetics risk factors of human diseases is very important. This study is conducted to model the allele frequencies (AFs) of Alzheimer's disease.

Materials and Methods: In this study, several candidate probability distributions are fitted on a data set of Alzheimer's disease genetic risk factor. Unknown parameters of the considered distributions are estimated, and some criterions of goodness-of-fit are calculated for the sake of comparison.

Results: Based on some statistical criterions, the beta distribution gives the best fit on AFs. However, the estimate values of the parameters of beta distribution lead us to the standard uniform distribution.

Conclusion: The AFs of Alzheimer's disease follow the standard uniform distribution.

*[*Fayyaz-Movaghar A, Kamel-Mirmostafaee MT, Borhani TS.* **Probability Model of Allele Frequency of Alzheimer's Disease Genetic Risk Factor**. *Iran J Health Sci 2016; 4(2): 43-48] http://jhs.mazums.ac.ir*

***Keywords*:** Allele Frequency, Alzheimer's Disease**,** Exponentiated Kumaraswamy Distribution, Single Nucleotide Polymorphisms

1. Introduction

A major aim of human genetics is to identify genetics risk factors for common, complex diseases such as Type II diabetes and Alzheimer. Alzheimer's disease is a chronic neurodegenerative disease that usually starts slowly and gets worse over time. It is the cause of 60-70% of cases of dementia. The most common early symptom is difficulty in remembering recent events. As the disease advances, symptoms can include problems with language, disorientation (including easily getting lost), mood swings, loss of motivation, not managing self-care, and behavioral issues. As a person's condition declines, they often withdraw from family and society. Gradually, bodily functions are lost, ultimately leading to death. Although the speed of progression can vary, the average life expectancy following diagnosis is 3-9 years. The cause of Alzheimer's disease is poorly understood. About 70% of the risk is believed to be genetic with many genes usually involved (1). Other risk factors include a history of head injuries, depression, or hypertension.

There are different approaches and technologies for identifying and studying genetics risk factors. One of the new tools is the Genome-wide Association Study (GWAS) which verifies genetic variants from across human genome in an effort to identify genetic risk factors for diseases that are common in the population. GWASs typically focus on associations between single nucleotide polymorphisms (SNPs) and traits like major diseases. The single base-pair changes in the DNA sequence is SNPs. These changes occur with high frequency in the human genome (2). There are many SNPs in a large ratio of populations (3). SNPs can occur in two locations (locus), namely, there are two alleles for each SNP. If one allele is more frequent in a population with an especial disease, one says the allele is associated with the disease. Therefore, the associated SNPs are considered to show human genome regions that influence the disease risk. Furthermore, allele frequencies (AFs) show the amount of variation at a particular locus. Hence, studying the frequency of SNPs provides important information about a population (4).

The AFs cannot be generally explained by a regular binomial distribution. It is possible that the proportion of a certain allele (p) changes during data collection. Therefore, it seems reasonable that the parameter of the binomial distribution, P, varies among observations. As the values of P vary over the interval [0,1], one can say that P is distributed as beta distribution. In this paper, we focus on a real data of Alzheimer's disease genetic risk factor and propose a model for the distribution of allelic probability, P. In the next section, we list some distributions with support [0,1]. Section 3 is concerned with the Alzheimer's disease and its genetic risk factors. In Section 4, we fit the proposed distributions in Section 2 to AF of the disease and then compare them using some well-known goodness-of-fit statistics to see which one provides the best fit. Finally, we conclude in Section 5.

2. Materials and Methods

2.1. Materials

2.1.1 Real data

GWAS data set with 1,237,567 SNPs and 1225 individuals is chosen. This data set was previously examined in a study on Alzheimer's disease genetic risk factor (5, 6).

2.1.2 Some distributions over [0,1]

As the allele probability has values over [0,1], its probabilistic behavior can be explained by the distribution with support [0,1]. In what follows, we list several distributions with support [0,1] and present their relations, briefly.

a. Uniform distribution: A very simple distribution for a continuous random variable is the uniform distribution. The random variable X is uniformly distributed over the interval [a,b] if its probability density function (pdf) is defined as follows:

$$f(x) = \frac{1}{b-a} \quad , a \le x \le b \qquad (1.1.1)$$

The expectation and the variance of X are given by:

$$E(x) = \frac{a+b}{2} \text{ and } V(x) = \frac{(b-a)^2}{12},$$

respectively.

Setting a = 0 and b = 1, we attain the so-called standard uniform distribution.

b. Beta distribution: The random variable X with pdf

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \, , 0 < x < 1 \qquad (1.2.1)$$

Is defined to have a beta distribution with parameters a > 0 and b > 0 where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$ is the complete beta function (7). The expectation and the variance of X are, respectively,

$$E(X) = \frac{a}{a+b} \qquad V(X) = \frac{ab}{(a+b+1)(a+b)^2}$$

Note that if we set a = b = 1, then the beta distribution reduces to the standard uniform distribution.

c. Kumaraswamy and Exponentiated Kumaraswamy distribution: The Kumaraswamy distribution was first introduced by Kumaraswamy (8). The random variable X is said to have the Kumaraswamy distribution with parameters and if its pdf is given by

$$f(x) = \alpha \lambda x^{\lambda-1}\left(1-x^{\lambda}\right)^{\alpha} \quad ,0 < x < 1$$
$$\alpha, \lambda > 0 \qquad (1.3.1)$$

Recently, a generalization of the Kumaraswamy model has been introduced (9), called the exponentiated Kumaraswamy distribution with pdf

$$f(x) = \alpha \beta \lambda \left\{1-x^{\lambda}\right\}^{\alpha-1}\left[1-\left\{1-x^{\lambda}\right\}^{\alpha}\right]^{\beta-1} x^{\lambda-1}, 0 < x < 1 \qquad (1.3.2)$$

Where, $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are the shape parameters. The exponentiated Kumaraswamy distribution contains many sub-models as special cases which are listed as follows:

- If $\beta = 1$, then (1.3.2) simplifies to the exponentiated power distribution
- If $\alpha = 1$, then (1.3.2) simplifies to the power distribution with parameter $\beta\lambda$.
- If $\lambda = 1$, then (1.3.2) reduces to the exponentiated generalized uniform distribution
- If $\alpha = \lambda = 1$, then (1.3.2) reduces to the power distribution with parameter β.

- If $\beta = \lambda = 1$, then (1.3.2) reduces to the exponentiated uniform distribution
- If $\alpha = \beta = 1$, then (1.3.2) reduces to the standard uniform distribution.

d. McDonald distribution: McDonald (10) introduced a distribution with pdf:

$$f(x) = \frac{ax^{ap-1}\left(1-\frac{x^a}{b}\right)^{q-1}}{b^{ap}B(p,q)}, 0 < x < b, \quad a > 0, \quad p > 0, \quad q > 0 \qquad (1.4.1)$$

Setting b = 1, we have a distribution over [0,1]. The beta and Kumaraswamy distribution are obtained from McDonald distribution by setting a = b = 1 and b = p = 1, respectively.

e. Marshall-Olkin Uniform distribution: The Marshall-Olkin (11) is defined by

$$f(x) = \frac{\alpha}{[\alpha+(1-\alpha)x]^2}, \quad 0 < x < 1, \alpha > 0 \qquad (1.5.1)$$

Where, $\alpha > 0$ is the shape parameter of distribution (11).

2.2 Modeling the allelic probability

Here, based on the presented distributions in the previous section, we try to find the best distribution which is appropriate for the allelic probability of the real data in Section 2.1.1 from a practical case.

Figure 1 shows the empirical AFs of the data. It is known that depending on the condition, SNPs with low minor AF (MAF) are close to either 0 or 1. These SNPs with very low MAF are excluded from GWAS (Figure 2).

The candidate distributions are then fitted to the allele frequencies after elimination. The next section provides the results of fitting data.

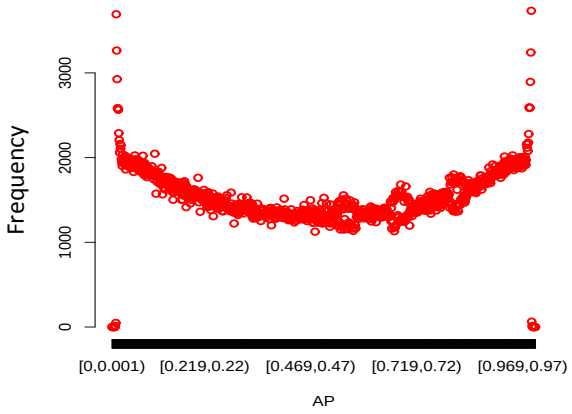

Figure 1. Empirical allele frequency of Genome-wide Association Study dataset with 1,237,567 single nucleotide polymorphisms concerns the Alzheimer's disease. These frequencies are derived over 1000 intervals between [0,1]

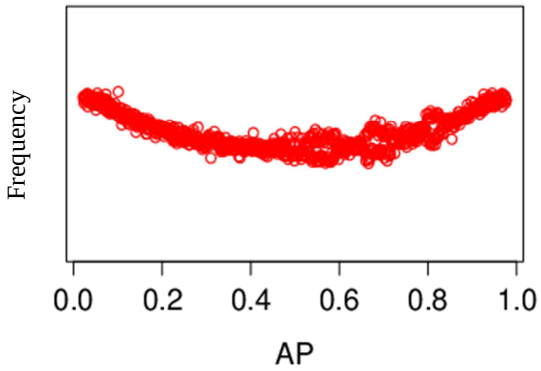

Figure 2. Truncated empirical allelic frequency of figure 1 over [0.021, 0.978]

3. Results

Four distributions are fitted on the data. Table 1 presents the results of fitting. The distribution parameters are estimated based on data using the maximum likelihood method, and consequently, the estimates are called maximum likelihood estimates. Some well-known goodness-of-fit criterions are

Table 1. The results of fitting the candidate distributions on allele frequency (AF)

Distribution	MLE	LL	AIC	BIC
Kumaraswamy	$\hat{\alpha} = 1.0682$ $\hat{\gamma} = 1.0672$	2116.238	−4228.47	−4204.07
McDonald	$\hat{\alpha} = 0.5553$ $\hat{p} = 1.9326$ $\hat{q} = 1.0639$	2234.63	−4567.076	−4530.47
Marshal-Olkin Uniform	$\hat{\alpha} = 1.0004$	−0.0005	2.001	14.199
Exponentiated Kumaraswamy	$\hat{\alpha} = 1.0780$ $\hat{\beta} = 1.0736$ $\hat{\lambda} = 1.4439$	2145.89	−4285.79	−4249.19
Beta	$\hat{\alpha} = 1.0864$ $\hat{\beta} = 1.0851$	3389.58	−6775.17	−6750.78

The parameters of each distribution are estimated via the maximum likelihood method. Furthermore, some goodness-of-fit criterions such as AIC, BIC, and LL are calculated. MLE: Maximum likelihood estimation; AIC: Akaike information criterion; BIC: Bayesian information criterion; LL: Log-likelihood

calculated and reported in table 1. These criterions are the log-likelihood (LL), Akaike information criterion (AIC) (12), and Bayesian information criterion (BIC) (13). The smaller values of AIC and BIC and bigger value of LL provide the better fit.

As it is seen, the beta distribution has the maximum LL, minimum AIC, and BIC which confirms that the best fitting is carried out by the beta distribution. Moreover, the estimates of its two parameters are very close to 1 and it seems that the distances between the estimated parameters of the beta distribution and the number 1 are meaningless (Recall that the standard distribution is a special case of the beta model when its parameters equal 1). Therefore, we may prefer the standard uniform distribution which is the simplest model with support [0,1].

Figure 3 shows the empirical cumulative AF (black dots) fitted by Marshall-Olkin Uniform distribution (a), Kumaraswamy distribution (b), Exponentiated Kumaraswamy distribution (c), McDonald distribution (d), and beta distribution (e).

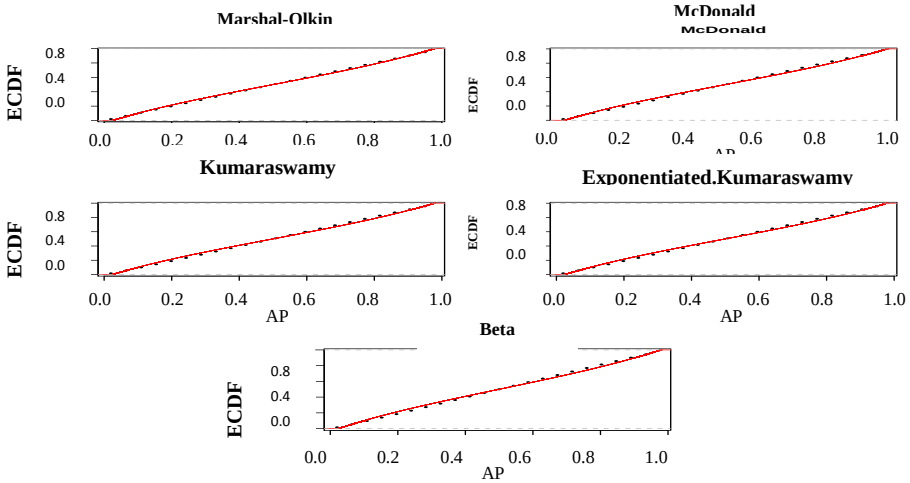

Figure 3. Empirical cumulative allele frequency (ECDF) of Alzheimer disease (black dots) is fitted by the candidate distributions (red lines)

4. Discussion

For any characterization of the allelic distributions in the genome, one needs to model allelic probabilities. As far as we know, there is not any similar work about AFs. In this study, following the numerical results in the previous section, it is observed that the beta distribution explains the behavior of AFs of Alzheimer's disease risk factors quite well. As its estimates of parameters are very close to 1, we may claim that the standard uniform distribution approximately provides the best fit. Therefore, it is claimed that the proportion of a certain allele (p) of Alzheimer's disease changes as a uniform distribution over [0,1]. This means that the proportion of the allele of Alzheimer's disease on a genetic locus behave as a uniform distribution.

Conflict of Interests

The Authors have no conflict of interest.

Acknowledgement

This article is extracted from MSc. thesis at the University of Mazandaran, Iran.

References

1. Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, et al. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron 2007; 54(5): 713-20.
2. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature 2010; 467(7319): 1061-73.
3. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010; 467(7311): 52-8.
4. Lockwood JR, Roeder K, Devlin B. A Bayesian hierarchical model for allele frequencies. Genet Epidemiol 2001; 20(1): 17-33.
5. Antunez C, Boada M, Gonzalez-Perez A, Gayan J, Ramirez-Lorca R, Marin J, et al. The membrane-spanning 4-domains, subfamily A (MS4A) gene cluster contains a common variant associated with Alzheimer's disease. Genome Med 2011; 3(5): 33.
6. Munoz DG, Feldman H. Causes of Alzheimer's disease. CMAJ 2000; 162(1): 65-72.
7. Jones MC. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. Stat Methodol 2009; 6(1): 70-81.
8. Kumaraswamy P. A generalized probability density function for double-bounded random processes. J Hydrol 1980; 46(1): 79-88.
9. Lemonte AJ, Barreto-Souza W, Cordeiro GM. The exponentiated Kumaraswamy distribution and its log-transform. Braz J Probab Stat 2013; 27(1): 31-53.
10. McDonald JB. Some generalized functions for the size distribution of income. Econometrica 1984; 52(3): 647-63.
11. Josea KK. Marshall-Olkin extended uniform distribution. ProbStat Forum 2011; 4: 78-88.
12. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974; 19(6): 716-23.
13. Schwarz G. Estimating the dimension of a model. Ann Stat 1978; 8(2): 461-4.