# Statistical Analysis of Blood Group Using Maximum Likelihood Estimation

**Nagarajan.D, Sunitha.P, Helen Sheeba.M, Nagarajan.V**

***Abstract*— The researchers have drawn much attention about the genetic systems controlling blood groups in humans. These findings are considerably important in understanding human genetics. The transmission of genes from one generation to the next one takes place biologically through the gametes. The aim of this paper is to estimate the changes in blood group using long run state. This analysis is useful to blood types inheritance diseases.**

***Index Terms*— ABO system, allele frequency, genotype frequency, maximum likelihood estimation.**

## I. INTRODUCTION

Blood group testing plays a key role in medical treatment prior to blood transfusion and child birth. The blood group of a person does not change within one's own life time and so it is considered as a genetic marker for research. The blood group is determined by the genetic make-up of the alleles of a system. The present study was under taken to estimate the existing frequencies of the alleles and the expected frequencies of genotypes.

The most well known and medically important blood types are in the ABO group. All humans and many other primates can be typed for the ABO blood group. There are four principal types: A, B, AB, O. There are two antigens and two antibodies that are mostly responsible for the ABO types. The table below shows the possible permutation of antigens and antibodies with the corresponding ABO type.

**Table1:** Permutation of antigens and antibodies

| ABO blood type | Antigen A | Antigen B | Antibody anti-A | Antibody anti-B |
|---|---|---|---|---|
| A | Yes | No | No | Yes |
| B | No | Yes | Yes | No |
| O | No | No | Yes | Yes |
| AB | Yes | Yes | No | No |

Research carried out in Heidelburg, Germany by Ludwik Hirzfeld and Emil Von Dungem in 1910 and 1911 showed that the ABO blood types are inherited. An individual's ABO type results from the inheritance of 1 of 3 alleles (A, B or O) from each parent. The possible combination of alleles produces blood types as shown in the table below[3].

**Table2:** Combination of alleles

| Parent alleles | A | B | O |
|---|---|---|---|
| A | AA (A) | AB (AB) | AO (A) |
| B | AB (AB) | BB (B) | BO (B) |
| O | AO (A) | BO (B) | OO (O) |

In the above table the offspring receives one of the three alleles from each parent, giving rise to six genotypes (AA, AO, AB, BB, BO, OO) and four possible phenotype (A,B,O,AB). Both A and B are dominant over O. As a result, individuals who have an AO genotype will have an A phenotype. People who have OO genotype will have an O phenotype. In other words, they inherited a recessive O allele from both parents. The A and B alleles are codominant. Therefore, if an A is inherited from one parent and a B from the other, the phenotype will be AB [1] and [7].

Population genetics study frequencies of genotypes and alleles within populations rather than the ratios of phenotypes. Estimates of gene frequencies provide more valuable information on the genetic similarity of different population and to some extent on their ancestral genetic linkage. In a constant environment, genes will continue to sort similarly for generations upon generations. The observation of this constancy led two researches, G.Hardy and W.Weinberg, to express an important relationship in evolution named as Hardy-Weinberg equilibrium which serves as the null model for population genetics. It applies basic rules of probability to a population to make predictions about the next generation.

The allele frequency (and genotype frequency) of a population remains constant over generations, unless a specific factor or combination of factors disrupts this equilibrium. Such factors might include non-random mating, mutation, natural selection, genetic bottle necks leading to increased genetic drift, the immigration or emigration of individuals [2], [4] and [5].

## II. MODEL DESCRIPTION

Collect a sample of individuals having A, B, AB, O phenotypes. Let $n_A$ denote the number of individuals having A phenotype, $n_B$ denotes the number of B, $n_{AB}$ denotes the

Manuscript received June 11, 2014.

**D.Nagarajan** , Math Section , Department of Information Technology, Salalah college of Technology, Salalah, PO Box 608, PC 211 Sultanate of Oman.

**Sunitha.P**,Department of Mathematics, BIT Campus, Anna University, Trichy-24

**M.Helen Sheeba**, Research Scholar, Department of Mathematics, S.T.Hindu College ,Nagercoil -629002

**V.Nagarajan**, S.T.Hindu College ,Nagercoil 629002.

number of AB and $n_O$ denotes the number of individual's with O phenotype so that $n_A + n_B + n_{AB} + n_O = n$

If $P_A$ ,$P_B$ and $P_O$ denote the allele frequencies such that $P_A + P_B + P_O = 1$, then the expected probabilities of the phenotype are as follows.

**Table3**: Expected probabilities

| Phenotype | Genotype | Probability |
|---|---|---|
| A | AA, AO | $P_A^2 + 2P_AP_O$ |
| B | BB, BO | $P_B^2 + 2P_BP_O$ |
| O | OO | $P_O^2$ |
| AB | AB | $2P_AP_B$ |

The likelihood is given as

$$L(P_A, P_B, P_O) = \binom{n}{n_A, n_B, n_{AB}, n_O} (P_A^2 + 2P_AP_O)^{n_A}(P_B^2 + 2P_BP_O)^{n_B}(2P_AP_B)^{n_{AB}}(P_O^2)^{n_O}$$

Since $P_O$ is redundant we substitute $P_O$ by $1 - P_A - P_B$ in the above expression and taking logarithm we obtain the likelihood as

$$l(P_A, P_B) = log\binom{n}{n_A, n_B, n_{AB}, n_O} + n_A log(P_A^2 + 2P_A(1 - P_A - P_B)) + n_B log(P_B^2 + 2P_B(1 - P_A - P_B)) + n_{AB} log(2P_AP_B) + 2n_O log(1 - P_A - P_B)$$

To find the maximizer of the loglikelihood function we differentiate with respect to$P_A$,$P_B$ and then set the derivative to be zero. Hence we can estimate the maximum likelihood estimator by means of EM algorithm [6].

By choosing $P_A$,$P_B$ and $P_O$ we could determine how many individuals with A phenotype have the AA genotype and AO genotype using

$$\hat{n}_{AA} = n_A\left[\frac{P_A^2}{P_A^2 + 2P_AP_O}\right] \text{ and } \hat{n}_{AO} = n_A\left[\frac{2P_AP_O}{P_A^2 + 2P_AP_O}\right]$$

Similarly for B phenotype

$\hat{n}_{BB} = n_B\left[\frac{P_B^2}{P_B^2+2P_BP_O}\right]$ and $\hat{n}_{BO} = n_B\left[\frac{2P_BP_O}{P_B^2+2P_BP_O}\right]$ and

ofcourse $\hat{n}_{AB} = n_{AB}$ , $\hat{n}_{OO} = n_O$ which is the expectation part of the EM algorithm.

After estimating the genotype frequencies we could estimate the allele frequency by gene counting using $\hat{P}_A = \left[\frac{2n_{AA}+n_{AO}+n_{AB}}{2n}\right]$ , $\hat{P}_B = \left[\frac{2n_{BB}+n_{BO}+n_{AB}}{2n}\right]$ and $\hat{P}_O = \left[\frac{2n_{OO}+n_{BO}+n_{AO}}{2n}\right]$

which is the maximization part of the EM algorithm. It is called maximization because we calculate the maximum likelihood estimates of the allele frequency given the observed genotype count. Repeat the whole sequence several times to find the maximum likelihood estimates of the allele frequencies provided the method assumes the genotypes are found in Hardy-Weinberg proportion. Once the maximum likelihood estimator is derived, the general theory of the maximum likelihood estimation provides standard errors, statistical tests and other results useful for statistical inference.

## III. DATA BASE

The present study comprised of 1000 individuals. Data were collected for ABO blood group from the hospitals. The frequencies observed were$n_A = 158, n_B = 628, n_{AB} = 121, n_O = 93$

We know that $n_A + n_B + n_{AB} + n_O = n = 379$ and assume $\hat{P}_A = 0.3333, \hat{P}_B = 0.3333$ and $\hat{P}_O = 0.3333$.

Now

$\hat{n}_{AA} = n_A\left[\frac{P_A^2}{P_A^2+2P_AP_O}\right] = 158\left[\frac{0.3333^2}{0.3333^2+2(0.3333)(0.3333)}\right] =$ 52.6667

$\hat{n}_{AO} = n_A\left[\frac{2P_AP_O}{P_A^2+2P_AP_O}\right] = 158\left[\frac{2(0.3333)(0.3333)}{0.3333^2+2(0.3333)(0.3333)}\right] =$ 105.3333

$\hat{n}_{BB} = n_B\left[\frac{P_B^2}{P_B^2+2P_BP_O}\right] = 628\left[\frac{0.3333^2}{0.3333^2+2(0.3333)(0.3333)}\right] =$ 209.3333

$\hat{n}_{BO} = n_B\left[\frac{2P_BP_O}{P_B^2+2P_BP_O}\right] = 628\left[\frac{2(0.3333)(0.3333)}{0.3333^2+2(0.3333)(0.3333)}\right] =$418 .6667

$\hat{n}_{AB} = n_{AB} = 121$ *and* $\hat{n}_{OO} = n_O = 93$ is the approximate estimation of gene frequency.

Using these estimates approximate estimation of allele frequency are obtained as shown below

$\hat{P}_A = \left[\frac{2n_{AA}+n_{AO}+n_{AB}}{2n}\right] = \left[\frac{2(52.6667)+105.3333+121}{2(1000)}\right] = 0.1658$

$\hat{P}_B = \left[\frac{2n_{BB}+n_{BO}+n_{AB}}{2n}\right] = \left[\frac{2(70)+140+35}{2(379)}\right] = 0.4792$

$\hat{P}_O = \left[\frac{2n_{OO}+n_{BO}+n_{AO}}{2n}\right] = \left[\frac{2(49)+140+56.6667}{2(379)}\right] = 0.3350$

The approximate solution obtained by Hardy-Weinberg method is $\hat{P}_A = 0.1658, \hat{P}_B = 0.4792$ *and* $\hat{P}_O = 0.3350$ . On repeating this process seven more times using the estimated values, better estimation due to Hardy-Weinberg method are $\hat{P}_A = 0.1543, \hat{P}_B = 0.5100$ *and* $\hat{P}_O = 0.3357$ which the maximum loglikelihood estimates. The gene frequency attained by Hardy-Weinberg method are $\hat{n}_{AA} = 29.53$ , $\hat{n}_{AO} = 128.47$ , $\hat{n}_{BB} = 271.10, \hat{n}_{BO} = 356.90, \hat{n}_{AB} = 121$ , $\hat{n}_{OO} = 93$ which means29.53 of the A phenotypes in the sample have AA genotype and the remaining 128.47 have AO genotype. Similarly we have estimated that 271.10 of B phenotypes in the population sample have BB genotype and the remaining 356.90 have BO genotype. By the time we have repeated this process seven more times, the allele frequencies attains stationary value. The above result also reveals that the population allele frequency is closer for B and O type and BO type gene frequency is also found to occur higher.

To investigate whether the genotype frequencies are compatible with the basic principle, we test the assumptions for the genotype frequency using Chi-square test. We first formulate a conservative hypothesis, called the null hypothesis ($H_o$), which states that there is no difference between the observed and expected values; therefore the population is in Hardy Weinberg equilibrium.

The observed frequency and the expected frequency attained by using the maximum loglikelihood estimates are indicated in the table shown below.

**Table4**: Expected value of the phenotype

| Phenotype | Observed Frequency | Probability | Probability value | Expected Frequency |
|---|---|---|---|---|
| A | 158 | $P_A^2 + 2P_A P_O$ | 0.1274 | 127.4 |
| B | 628 | $P_B^2 + 2P_B P_O$ | 0.6027 | 602.7 |
| O | 93 | $P_O^2$ | 0.1127 | 112.7 |
| AB | 121 | $2P_A P_B$ | 0.1574 | 157.4 |

$\psi^2$ for goodness of fit is given by $\psi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i}$=20.27

The object is to establish the significant difference for one degrees of freedom at 5% level. The $\psi^2$ value reveals that there is significant difference in the gene frequencies which is due solely to chance.

## IV. CONCLUSION

From the above result the gene and allele frequency has been attained. It reveals that the population allele frequency is closer for B and O type and BO type gene frequency is also found to occur higher. The probability value also suggests that the B type phenotype occurs highly among the individuals having BB and BO type genotype. It also revealed that the gene frequency is a function of allele frequency and the equilibrium was attained in the seventh generation, independent of the initial genotype frequency. Hence after seventh generation the transmission of genes is completely stopped and the changes in blood group occurs i.e., the offspring receive new alleles from their parents allele and not their ancestor allele. Also the inherited disease will not be generated through the blood during the next generation. This will be helpful in decision making during blood transfusion.

## V. REFERENCES

[1] Singh.B.D, "Fundamentals of Genetics", Kalyani Publishers, New Delhi 1990
[2] Hosking L, et al: "Detection of genotyping errors by Hardy-Weinberg equilibrium testing", Eur J Hum Genet
;12:pp395–399, 2004
[3] Senthamaraikannan .K, Nagarajan.D, Arumugam , "Statistical Analysis of Gene Frequencies for Mothers and
Babies", Anthropologist ,10(2)139-141 ,2008
[4] Emigh T, "A comparison of tests for Hardy–Weinberg equilibrium", Biometrics, 36, pp627– 642, 1980
[5] Rogatko A, Slifker MJ, Babb JS, " Hardy–Weinberg equilibrium diagnostics", Theor Popul Biol, 62, pp251– 257, 2002
[6] John Fox, "Notes on Maximum Likelihood Estimation- Basic ideas"
[7] Supriyo Chakraborty, "Genetic Analysis on Frequency of Alleles for Rh and ABO Blood Group Systems in the Barak Valley Populations of Assam", Not Sci Biol, 2, pp31-34, 2010