ABSTRACT

SAMANTA, SUVAJIT. A Statistical Characterization of the Genetic Structure of Populations. (Under the direction of Bruce Spencer Weir).

In a random mating population the second order descent measure which is also known as coancestry coefficient (θ) characterizes population structure. This structure provides information regarding the history of a population. Characterizing inbred populations, however, requires third (γ) and fourth order (δ and $\Delta_{2,2}$) descent measures. These descent measures are also used to find different expressions in DNA profile matching. In literature there are several estimators for second order descent measure but no estimator exists for the higher order descent measures. In this research we find estimators for second and third order descent measures using a Method of Moments approach and a Probabilistic approach. Simulation studies show that our new estimators for second order descent measure are more accurate than the existing moment estimators while the estimators for third order descent measure are stable in terms of bias and standard error. We also derive the sampling properties of our estimators. Later, we relax the constraint that the descent measures have the same value in different populations and find estimators for population-specific θ and γ. We implement our methods on HapMap SNP data and find the estimates of the descent measures for the human populations.

The expressions for genetic parameters generally depend on the coancestry coefficient (θ) and become very simple when the coancestry coefficient is zero. In this thesis we propose different testing techniques for testing $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$ for random population. For small sample sizes we propose a parametric bootstrap test that has higher power than the non-parametric bootstrap test proposed by Dodds (1986). When the sample sizes are large we find an asymptotically chi square test that works for any number of allelic forms in a particular locus. For more than two alleles per locus our test is better than the asymptotic test proposed by Li (1996). We implement our testing procedures on HapMap SNP data and find that the coancestry coefficient for humans is strictly positive.

The probability of identity by descent simultaneously at two or more loci is a gener-

alization of Wright's inbreeding coefficient. The two-locus identity is a useful parameter in predicting the joint ancestry of pair of loci which is frequently used in mapping studies and in finding variances and covariances of quantitative traits. Weir and Cockerham (1969) extended the inbreeding coefficient concept for two loci to evaluate a measure of identity of descent for genes at each of two linked loci. In this research we show that the two-locus descent measures are not estimable but we can estimate the product of linkage disequilibrium and two-locus descent measures. We find the estimators of different components of the two-locus descent measures multiplied with linkage disequilibrium using a Method of Moments approach. We use haplotype data.

Estimates of heterozygosity and gene diversity have been used in many fields, including conservation and evolutionary biology and forensic studies. In published analyses researchers frequently overlook the sampling properties of these estimators although this affects the resulting inferences. This dissertation characterizes the estimators of heterozygosity and gene diversity by evaluating the sampling properties. Properties of several methods for inferring the variance of sample heterozygosity are evaluated, including the use of a new generalized linear mixed model for the total variance of sample heterozygosity. We have observed a difference in result with the previous linear model. We implement the methods on one published data set and compare the estimates of the variance of sample heterozygosity. Using different variance component methods we can get different estimates of total variance of sample heterozygosity for unbalanced data while for balanced data all the methods are identical.

A STATISTICAL CHARACTERIZATION OF THE GENETIC STRUCTURE OF POPULATIONS

by
Suvajit Samanta

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh
2006

APPROVED BY:

Dr. Subhashis Ghoshal	Dr. Dahlia Michelle Nielsen
Dr. Eric Alan Stone	Dr. Bruce Spencer Weir (Chair of Advisory Committee)

DEDICATION

To my family

BIOGRAPHY

Suvajit Samanta was born on January 30, 1981 in Kashigram, West Bengal, India. He was raised in Kashigram and at the age of 12 years moved to Uttarpara, West Bengal. After successfully finishing higher secondary from Uttarpara Amarendra Vidyapith in 1998, Suvajit attended Indian Statistical Institute, Kolkata and majored in Statistics. He next obtained a Masters degree in Statistics with a special concentration on Biostatistics and Data Analysis from the same Institute. Because of his interest in applied Statistics, Suvajit joined the doctorial program of the Department of Statistics at North Carolina State University. In his PhD, Suvajit worked on the theoretical aspects of genetics under the direction of Dr. Bruce Weir. Suvajit will join Merck & Co., Inc. at their Rahway, New Jersey location as a Biometrician after he receives his doctoral degree.

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my advisor Dr. Bruce Weir. I am especially grateful for his incredible patience and understanding. Through my interactions with him I have learned a great deal not only about research but also about life. I am grateful for the opportunity to work with him.

I wish to express my appreciation to my committee members, Dr. Subhashis Ghoshal, Dr. Dahlia Nielsen, and Dr. Eric Stone for their valuable input and service. These wonderful people helped me in both academic and non-academic matters.

I would like to acknowledge Dr. Sujit Ghosh for his guidance and help. I also would like to thank Dr. Keneth Olsen and Dr. Barbara Schaal for kindly responding to my request for their previously published data set.

I thank everyone at Statistics Department and Bioinformatics Research Center for their friendship and help through my studies. I would like to thank Prasenjit Kapat for always being ready to help with any statistical problems specially with statistical computing and R, Sunil Suchindran for helping me with any genetics problems, and Arin Chaudhuri for his invaluable help and guidance during the first two years of my studies. My doctorate would have been impossible without my room mates and Indian friends who helped in in all possible ways.

Most importantly, I would like to thank my family for their love, encouragement and moral support through my life. Their faith and confidence in me has helped me to achieve all that I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Review

1.1 Introduction

This research covers two different subjects. One is related to decent measures and the other one is related to heterozygosity. The first part of the research evaluates the adequacy of Dirichlet and Normal distribution for allele frequencies. Then it covers different estimation procedures for decent measures and finds the sampling properties of the estimators. It also covers different procedures for testing hypotheses about the coancestry coefficient including the analysis of variance and parametric bootstrap method. The second part of the research is related to the sampling properties of the estimator of heterozygosity. The motivation and the goal of the different parts of the research will be described separately in the following chapters.

The population structure can be characterized by two different set of parameters (i) F-statistics and (ii) descent measures. F-statistics were proposed by Wright (Wright, 1951) and advocated by many others (Cockerham, 1973; Balding, 2003). On the other hand, decent measures were proposed by Malécot (Malécot, 1948) and used by other scientist (Weir and Cockerham, 1984; Weir, 1994; Weir and Hill, 2002). Both F-statistics and descent measures are parameters that characterize the population structure. We can make inferences about the history of a population (age of the population, effective size of the population, the mutation rate present in the population, etc.) if we know

the population structure. The F-statistics and descent measures are conceptually the same under a random mating system. This research is developed under a random mating population. It uses descent measures to characterize a population's structure. Geneticists are interested in finding the genetic distance between different populations. This distance involves the knowledge of decent measures. In current forensic studies, decent measures become a tool to characterize the DNA profile matching.

Heterozygosity and gene diversity are basic tools for summarizing the pattern of genetic variation in a group of populations. Characteristics of population genetic variation are of key interest in studies of evolution. The amount of variation present in a population or species determines the capacity of the heritable change of the group. So the estimate of heterozygosity and gene diversity can be very helpful descriptive measures for populations. To know the population better we also need to find the sampling properties such as bias and variance of these estimators.

In this chapter we introduce all the necessary parameters in detail and review the development of these parameters. Then we discuss different population models that will be used in this research. We also draw conclusions about the relationship between different parameters for a random mating population under different mutation models.

1.2 F-statistics

The population structure is frequently modeled as associations between alleles. These associations can occur on different levels, and to different extents. This association can occur within individuals, between individuals within a population, and between individuals in different populations. The inbreeding coefficient was first proposed by Wright (1921). Later Wright extended his work for hierarchial population model and introduced three F-statistics, F_{IS}, F_{ST} and F_{IT} which represent three different levels of association (Wright, 1951). These F-statistics can be defined as the correlations between alleles sampled from different levels in the population. The subscripts of the F-statistics refer to the level they are concerned with, where I stands for individuals,

S stands for sub-populations and T for the total population. F_{ST} is commonly referred to as the coancestry coefficient, and it measures the degree of relationship, between the individuals within populations relative to the amount of relationship found in the total population. F_{IT} is called the fixation index because it measures the progress of a neutral locus towards fixation to a single allele under the influence of random genetic drift. F_{IS} measures the amount of departure from the Hardy-Weinberg equilibrium of a population.

Cockerham (1969) renamed the F-statistics as $f = F_{IS}$, $\theta = F_{ST}$ and $F = F_{IT}$ in his work. This is done to reduce confusion between the F-statistics and the F distribution. He also wanted to emphasize that these measures are population parameters rather than statistics that are functions of observed data. Since these three parameters, f, θ and F are correlation between alleles, the range of these parameters is from -1 to 1.

Inbreeding within population occurs when some particular individuals are more related to each other than the relatedness of a random set of individuals from the total population. Two factors contribute to the total amount of inbreeding in a set of populations. Generally one factor contributes to θ; the other factor contributes to f. The random genetic drift results in differences among sub-populations that descended from a founder population which contributes to the value of θ. For a single population, drift can be described as the inbreeding coefficient. The assortive mating within populations increase the value of f. Both the factors, θ and f can increase the amount of inbreeding of the whole group of populations. The total inbreeding coefficient, F, is related to θ and f and the relation is (Weir, 1994)

$$F = f + \theta(1 - f).$$

The above relation demonstrates that the total variation, F, is a sum of the variation due to genes that are alike in individuals, summarized in f, and the variation due to unrelated genes in the total population, summarized in θ.

1.3 Descent Measures

Descent measures are parameters that also describe the association among alleles. The development of descent measures is based on the concept of identical by descent (ibd). A set of alleles is called identical by descent if all the alleles are descended from a common allele in some ancestral population and no allele has gone thorough mutation. The definition of ibd explicitly implies that the ibd alleles have the same allelic form. The probability that two or more alleles are identical by descent is called descent measures. For two allele case, the alleles can come from one individual or two different individuals in the same population. Malécot (1948) defined the descent measures for two alleles. He used the same notations as F, θ and f for defining three different descent measures. Since these parameters are probabilities of different events, the value of these parameters is always non-negative. Malécot (1948) defined the three parameters

$$F = \mathrm{E}_{sub-pop}\big[Pr(\text{Two alleles from an individual are ibd})\big], \tag{1.1}$$

$$\theta = \mathrm{E}_{sub-pop}\big[Pr(\text{Two alleles from two individuals within a population are ibd})\big], \tag{1.2}$$

$$f = Pr(\text{Two alleles from an individual in a population are ibd}), \tag{1.3}$$

where "$\mathrm{E}_{sub-pop}$" is to mean that the parameters are defined for random population set up. In this set up we have more than one populations and every population has evolved from the same founder population. In our set up, F and θ are parameters for the random population set up and provide information about the history of the populations. On the other hand, the parameter f is defined for a single population.

At this point, we have two different definitions of the parameters, θ, F and f. One set of definitions of these parameters is provided by Wright and Cockerham (Wright, 1921; Cockerham, 1969) while the other set of definitions is given by Malécot (Malécot, 1948). Wright defined the parameters as the correlation of alleles and Malécot defined as the descent measures. According to Wright's definition the parameters can take negative value while Malécot's definition guarantees non-negative value of the parameters.

Although the two definitions are different theoretically, the definitions are conceptually equivalent for a random mating system. When random mating occurs within the population, the proportion of heterozygosity reduces over time and the correlation between two alleles is always positive. This implies that the parameters defined by Wright take positive values under a random mating system. According to Wright's definition, the coancestry coefficient, θ, measures the differentiation between populations. If the value of θ (defined by Wright) is negative then, two alleles are more related if they are from different populations than if they are from same population. If the populations have been isolated since the base population and mate randomly within sub-populations, differentiation between sub-populations will increase over time. Therefore, the value of θ (defined by Wright) will be positive all the time under a random mating system. So in random mating the F-statistics always take positive values which is the case with descent measures. So the F-statistics and descent measures are conceptually the same. This research will adopt the concept of descent measures for inferring population history or the relatedness between individuals in a population. From now onwards this research will work with the parameters f, θ, and F that are defined in equations (1.1), (1.2) and (1.3) respectively. The parameters f, θ, and F will be considered as descent measures for remaining part the research.

The parameter f measures the amount of local inbreeding present in a population. Generally f gives information about a particular population, while θ and F give long-term effects of demographic and evolutionary forces of a population (Cockerham, 1973). Since we are interested in the long-term history of populations, we focus our interest on estimating the parameters θ and F rather than f. For a random mating population, there is no need to distinguish the cases that which individual contains the alleles. The probability of two alleles within an individual being ibd is the same as two alleles from two different individuals within a population. Therefore, the total inbreeding coefficient F, and the coancestry coefficient, θ are the same in a random mating population.

Descent measures have been extended to three and four alleles. Third and fourth order descent measures can be used to find covariances of inbred relatives (Gillois,

1966) and to characterize a population that has selfing or biparental inbreeding such as a plant population (Ritland, 1987). Third order descent measure is the probability that three alleles are identical by descent. These three alleles can come from either two or three individuals. There are two types of fourth order descent measures. First one is the probability of four alleles are identical by descent and the second one is the probability of two pairs of alleles are identical by descent. Four alleles can come from three different ways from a population while two pairs of alleles can come from five different ways. The list of possible ways in which the alleles can come from a population is given in Table 1.1.

Table 1.1: Possible arrangements for different alleles

Descent Measure	Possible Arrangements [1][2][3]
F_X	$a_X \equiv a'_X$
θ_{XY}	$a_X \equiv a_Y$
$\gamma_{\ddot{X}Y}$	$a_X \equiv a'_X \equiv a_Y$
γ_{XYZ}	$a_X \equiv a_Y \equiv a_Z$
$\delta_{\ddot{X}\ddot{Y}}$	$a_X \equiv a'_X \equiv a_Y \equiv a'_Y$
$\delta_{\ddot{X}YZ}$	$a_X \equiv a'_X \equiv a_Y \equiv a_Z$
δ_{XYZW}	$a_X \equiv a_Y \equiv a_Z \equiv a_W$
$\Delta_{XY.ZW}$	$a_X \equiv a_Y, a_Z \equiv a_W$
$\Delta_{\ddot{X}.YZ}$	$a_X \equiv a'_X, a_Y \equiv a_Z$
$\Delta_{\ddot{X}.\ddot{Y}}$	$a_X \equiv a'_X, a_Y \equiv a'_Y$
$\Delta_{\ddot{X}+YZ}$	$a_X \equiv a_Y, a'_X \equiv a_Z$
$\Delta_{\ddot{X}+\ddot{Y}}$	$a_X \equiv a_Y, a'_X \equiv a'_Y$

[1] One allele from an individual is denoted by a
[2] Two alleles from an individual are denoted by a and a'
[3] The subscript of a indicates the individual that contributes the allele

In a random mating population the descent measures can not be distinguished based the source of alleles. For example, the probability that three alleles from two individuals are identical by descent is the same as the probability that three alleles from three individuals are identical by descent. So for a random mating system, we can group the different parameters defined in Table 1.1 and get the identities

$$
\begin{aligned}
&F_X = \theta_{XY},\\
&\gamma_{\ddot{X}Y} = \gamma_{XYZ},\\
&\delta_{\ddot{X}\ddot{Y}} = \delta_{\ddot{X}YZ} = \delta_{XYZW}, \quad \text{and}\\
&\Delta_{XY.ZW} = \Delta_{\ddot{X}.YZ} = \Delta_{\ddot{X}.\ddot{Y}} = \Delta_{\ddot{X}+YZ} = \Delta_{\ddot{X}+\ddot{Y}}.
\end{aligned}
$$

The above equations suggest that for a random mating population only one parameter characterize the third-order descent measure and two parameters describe the fourth-order descent measures. These three parameters are

$$
\begin{aligned}
\gamma &= \mathrm{E}_{sub-pop}\big[Pr(\text{Three random alleles are identical by descent})\big], && (1.4)\\
\delta &= \mathrm{E}_{sub-pop}\big[Pr(\text{Four random alleles are identical by descent})\big], \quad \text{and} && (1.5)\\
\Delta_{2,2} &= \mathrm{E}_{sub-pop}\big[Pr(\text{Any two random pairs are identical by descent})\big]. && (1.6)
\end{aligned}
$$

There are four different inbreeding coefficients for three alleles. For any four alleles there are 15 arrangements of identity between any of the six pairs of alleles (Cockerham, 1971; Gillois, 1966). In inbred populations to calculate the relatedness between specific individuals we need all the components. But in random mating populations, only two descent measures, θ and, γ describe the third-order inbreeding coefficients while, only four measures, θ, γ, δ, and $\Delta_{2,2}$ describe the fourth-order inbreeding coefficients (Lynch, 1988). Table 1.2 describes the different inbreeding coefficients for two, three and four alleles. It also shows the relation between these inbreeding coefficients with the descent measures θ, γ, δ and $\Delta_{2,2}$ under a random mating population.

Table 1.2: Inbreeding coefficients for two, three and four alleles and their relation with descent measures under a random mating population

Number of alleles	IBD alleles	Inbreeding coefficients under RMP†
Two alleles	$a_1 \not\equiv a_2$	$\delta_0 = 1 - \theta$
(a_1, a_2)	$a_1 \equiv a_2$	$\delta_{a_1,a_2} = \theta$
Three alleles	$a_1 \not\equiv a_2 \not\equiv a_3$	$\delta_0 = 1 - 3\theta + 2\gamma$
(a_1, a_2, a_3)	$a_1 \equiv a_2 \not\equiv a_3$	$\delta_{a_1,a_2} = \theta - \gamma$
	$a_1 \equiv a_3 \not\equiv a_2$	$\delta_{a_1,a_3} = \theta - \gamma$
	$a_2 \equiv a_3 \not\equiv a_1$	$\delta_{a_2,a_3} = \theta - \gamma$
	$a_1 \equiv a_2 \equiv a_3$	$\delta_{a_1,a_2,a_3} = \gamma$
Four alleles	$a_1 \not\equiv a_2 \not\equiv a_3 \not\equiv a_4$	$\delta_0 = 1 - 6\theta + 8\gamma + 3\Delta_{2,2} - 6\delta$
(a_1, a_2, a_3, a_4)	$a_1 \equiv a_2 \not\equiv a_3 \not\equiv a_4$	$\delta_{a_1,a_2} = \theta - 2\gamma - \Delta_{2,2} + 2\delta$
	$a_1 \equiv a_3 \not\equiv a_2 \not\equiv a_4$	$\delta_{a_1,a_3} = \theta - 2\gamma - \Delta_{2,2} + 2\delta$
	$a_1 \equiv a_4 \not\equiv a_2 \not\equiv a_3$	$\delta_{a_1,a_4} = \theta - 2\gamma - \Delta_{2,2} + 2\delta$
	$a_2 \equiv a_3 \not\equiv a_1 \not\equiv a_4$	$\delta_{a_2,a_3} = \theta - 2\gamma - \Delta_{2,2} + 2\delta$
	$a_2 \equiv a_4 \not\equiv a_1 \not\equiv a_3$	$\delta_{a_2,a_4} = \theta - 2\gamma - \Delta_{2,2} + 2\delta$
	$a_3 \equiv a_4 \not\equiv a_1 \not\equiv a_2$	$\delta_{a_3,a_4} = \theta - 2\gamma - \Delta_{2,2} + 2\delta$
	$a_1 \equiv a_2 \equiv a_3 \not\equiv a_4$	$\delta_{a_1,a_2,a_3} = \gamma - \delta$
	$a_1 \equiv a_2 \equiv a_4 \not\equiv a_3$	$\delta_{a_1,a_2,a_4} = \gamma - \delta$
	$a_1 \equiv a_3 \equiv a_4 \not\equiv a_2$	$\delta_{a_1,a_3,a_4} = \gamma - \delta$
	$a_2 \equiv a_3 \equiv a_4 \not\equiv a_1$	$\delta_{a_2,a_3,a_4} = \gamma - \delta$
	$(a_1 \equiv a_2) \not\equiv (a_3 \equiv a_4)$	$\delta_{a_1,a_2:a_3,a_4} = \Delta_{2,2} - \delta$
	$(a_1 \equiv a_3) \not\equiv (a_2 \equiv a_4)$	$\delta_{a_1,a_3:a_2,a_4} = \Delta_{2,2} - \delta$
	$(a_1 \equiv a_4) \not\equiv (a_2 \equiv a_3)$	$\delta_{a_1,a_4:a_2,a_3} = \Delta_{2,2} - \delta$
	$a_1 \equiv a_2 \equiv a_3 \equiv a_4$	$\delta_{a_1,a_2,a_3,a_4} = \delta$

† RMP is Random Mating Population

1.4 Heterozygosity

Heterozygosity is simply the proportion of individuals with a heterozygous genotype in a population at a single locus. Heterozygosity is often termed observed heterozygosity or H_o. If we have more than one locus then we take the average of observed heterozygosity over loci. For the case of species with some degree of selfing, heterozygosity can be inadequate to describe the amount of genetic variation in a population. In this case, it is very common that many different types of homozygous genotypes are present in the population and would not be captured by the frequency of heterozygote. To solve this problem, Nei (1973) proposed another method of gene diversity that captures the diversity at the allelic level. If there are s alleles in a particular locus with frequencies $p_1, p_2, \cdots p_s$, then the gene diversity for this locus is defined as $1 - \sum_{i=1}^{s} p_i^2$. If there is more than one locus, then we take the average of gene diversity over locus. As a measure of genetic variation, Nei's gene diversity should be particularly used for selfing species. The expected value of observed heterozygosity and the value of gene diversity are the same in a random mating population not undergoing selfing. For this reason, gene diversity has been frequently and incorrectly termed average heterozygosity, or H_e in the literature. The relationship between gene diversity and heterozygosity and the coancestry coefficient θ can be expressed exactly for certain specific population and mutation models but may more complicated in real life.

To have a better idea about the genetic variation of a population it is important to find the sampling properties of the observed heterozygosity and gene diversity. Weir (1989) and Weir et al. (1990) developed extensive theory for the variances of sample gene diversity and observed heterozygosity respectively. Later, other scientists proposed different methods for finding the variances of sample gene diversity and observed heterozygosity. We also propose a new method for estimating the variance of sample heterozygosity. We discuss the development of these methods in Chapter 5. Then in Chapter 6, we compare our method with other existing methods for estimating the variance of observed heterozygosity.

1.5 Wright-Fisher Model

The Wright-Fisher model assumes that the alleles in the current generation are derived by sampling with replacement from the previous generation. In this research we always assume that all the loci that we are interested in are neutral. This means all alleles in a particular locus are equally likely to survive and be transmitted to the next generation. There may be any number of allelic types at a particular locus. Basically given the allele frequencies of the previous generation at reproduction the allele counts in the present generation follow a Multinomial distribution. The index of the distribution is $2N$ (N is the size of the present population) and the probability vector is the allele frequencies of the previous generation at reproduction. If we have only two allelic types, then the allele counts follow a Binomial distribution with appropriate parameters.

The Wright-Fisher model can be implemented with different assumptions about mutation. We generally assume (i) No mutation and (ii) Both-way mutation. In the first case where no mutation occurs within alleles, eventually one allele becomes fixed in the population. The probability that one particular allele will fix in the population is the initial frequency of that particular allele. The fixation probability and mean fixation time can be found using a diffusion process (Ewens, 1979). In the second case we assume that any allelic type can mutate to any other allelic type that already exists in the population with some positive rate. We assume that the mutation rate per generation remains same over generations. In this case the allele frequencies eventually follow a joint stationary distribution. The stationary distribution is a Dirichlet distribution with appropriate parameter values. When we have two alleles with both-way mutation then the stationary distribution reduces to a Beta distribution. Sometimes we consider an infinite allele mutation model that assumes any mutation generates a new allelic type. Under this mutation model the allele frequencies have a joint stationary distribution and that is Dirichlet. The stationary distribution and mean time to reach the stationary distribution can be found using a diffusion process (Ewens, 1979).

1.6 Theoretical Values of Descent Measures

In this section we discuss how the descent measures change over generations in a random mating population. We consider the Wright-Fisher model for transmitting alleles over generations. The value of the descent measures in a generation depends on the assumption of mutation of alleles, the value of the descent measures in the previous generation and the size of the previous generation. The descent measures at a particular locus do not depend on the number of alleles and allele frequencies in the locus. We denote the value assumed by θ, γ, δ, $\Delta_{2,2}$ at generation t by θ_t, γ_t, δ_t and $\Delta_{2,2,t}$. We assume that there are N individuals i.e. $2N$ alleles in the population at time t. We derive expressions for θ_{t+1}, γ_{t+1}, δ_{t+1} and $\Delta_{2,2,t+1}$. These values can be expressed in terms of N, θ_t, γ_t, δ_t and $\Delta_{2,2,t}$. But we get different expressions for different assumptions about mutations. In the next two sections we find expressions for the descent measures under different assumptions about mutations.

1.6.1 No Mutation

In this section we assume no mutation among alleles and discuss the behavior of the descent measures. Take two alleles from the population at time $t+1$. These two alleles can be descended from one allele or two different alleles at generation t with probabilities $\frac{1}{2N}$ and $1-\frac{1}{2N}$ respectively. If these two alleles are descended from a single allele then they are always ibd. If the alleles are descended from two different alleles then the probability that they are ibd is θ_t. So we have

$$\theta_{t+1} = \frac{1}{2N} + (1-\frac{1}{2N})\theta_t. \tag{1.7}$$

Now we consider three different alleles from $(t+1)^{th}$ generation. We will find the probability that these alleles are identical by descent which is denoted by γ_{t+1}. These three alleles can be descended from one, two, and three different alleles in the previous generation with probabilities $\frac{1}{4N^2}$, $\frac{3(2N-1)}{4N^2}$ and $\frac{(2N-1)(2N-2)}{4N^2}$ respectively. These alleles

are always ibd if they come from the same allele in the previous generation. If they are descended from two different alleles in the previous generation then the probability that these three alleles are ibd is the same as the probability that the two alleles in the previous generation are ibd which is θ_t. Similarly if the alleles are descended from three different alleles then the alleles at $(t+1)^{th}$ generation are ibd with probability γ_t. So we get

$$\gamma_{t+1} = \frac{1}{4N^2} + \frac{3(2N-1)}{4N^2}\theta_t + \frac{(2N-1)(2N-2)}{4N^2}\gamma_t. \tag{1.8}$$

For four alleles, similar arguments as above lead us to the equations

$$\begin{aligned} \delta_{t+1} &= \frac{1}{8N^3} + \frac{7(2N-1)}{8N^3}\theta_t + \frac{6(2N-1)(2N-2)}{8N^3}\gamma_t \\ &\quad + \frac{(2N-1)(2N-2)(2N-3)}{8N^3}\delta_t \text{ and} \qquad (1.9) \\ \Delta_{2,2,t+1} &= \frac{2N}{8N^3} + \frac{2(4N^2-1)}{8N^3}\theta_t + \frac{4(2N-1)(2N-2)}{8N^3}\gamma_t \\ &\quad + \frac{(2N-1)(2N-2)(2N-3)}{8N^3}\Delta_{2,2,t}. \qquad (1.10) \end{aligned}$$

The transition equations (1.7)-(1.10) had been derived by Weir (1994). If the initial population consists of non-inbred and unrelated individuals, then the four descent measures have explicit solutions (Weir, 1994)

$$\begin{aligned} \theta_t &= 1 - \lambda_1^t, \\ \gamma_t &= 1 - \frac{3}{2}\lambda_1^t + \frac{1}{2}\lambda_2^t, \\ \delta_t &= 1 - \frac{1}{5}(9\lambda_1^t - 5\lambda_2^t + \lambda_3^t) - \frac{3}{20(5N-3)}\lambda_1^t \\ &\quad + \frac{1}{12(N-1)}\lambda_2^t + \frac{8N-3}{30(5N-3)(N-1)}\lambda_3^t, \text{ and} \qquad (1.11) \\ \Delta_{2,2,t} &= 1 - \frac{1}{15}(24\lambda_1^t - 10\lambda_2^t + \lambda_3^t) - \frac{1}{5(5N-3)}(\lambda_1^t - \lambda_2^t), \end{aligned}$$

where,

$$\begin{aligned}\lambda_1 &= 1-\frac{1}{2N},\\ \lambda_2 &= (1-\frac{1}{2N})(1-\frac{2}{2N}), \text{ and}\\ \lambda_3 &= (1-\frac{1}{2N})(1-\frac{2}{2N})(1-\frac{3}{2N}).\end{aligned} \tag{1.12}$$

The above solutions assume that the population size remains the same over generations and equal to N. If the population size changes over generations and the effective population size is N_e, then the above equations approximately hold good if N is replaced by N_e.

Now we approximate the expressions in the equation (1.11) by assuming $N \to \infty$. We also re-scale the time by assuming a unit time is equal to $2N$. In other words we are assuming t is of the order $O(N)$. For notational benefit we denote $c = \lim_{N\to\infty} \frac{t}{2N}$. So for large N and large t we get

$$\lambda_1^t = \exp(-c), \quad \lambda_2^t = \exp(-3c), \quad \text{and} \quad \lambda_3^t = \exp(-6c). \tag{1.13}$$

When N is large, using the above approximation we get the relations (Robertson, 1952)

$$\gamma = \frac{3}{2}\theta^2 - \frac{1}{2}\theta^3, \tag{1.14}$$

$$\delta = 3\theta^3 - 3\theta^4 + \frac{6}{5}\theta^5 - \frac{1}{5}\theta^6, \text{ and} \tag{1.15}$$

$$\Delta_{2,2} = \theta^2 + \frac{2}{3}\theta^3 - \theta^4 + \frac{2}{5}\theta^5 - \frac{1}{15}\theta^6 = \frac{2}{3}\gamma + \frac{1}{3}\delta. \tag{1.16}$$

Due to a finite drift the population becomes more inbreed under a random mating system. Eventually the value of all the descent measures, θ, γ, δ and $\Delta_{2,2}$ after some generations converge to 1. When N is very large some authors assume (Weir and Hill, 2002) normal distribution for allele frequencies. In this case the higher order descent

measures become functions of θ and they can be expressed as (Weir, 1994)

$$\gamma_n \;=\; 0, \;\; \delta_n = 0, \;\; \text{and} \;\; \Delta_{2,2,n} = \theta^2. \tag{1.17}$$

1.6.2 Both-Way Mutation

In this section we allow both way mutation. We assume any allele mutates to another allele with a positive rate u. Using the same arguments that we have used in the previous section get the following transition equations

$$\begin{aligned}
\theta_{t+1} &= (1-u)^2[\frac{1}{2N} + (1-\frac{1}{2N})\theta_t], \\
\gamma_{t+1} &= (1-u)^3[\frac{1}{4N^2} + \frac{3(2N-1)}{4N^2}\theta_t + \frac{(2N-1)(2N-2)}{4N^2}\gamma_t], \\
\delta_{t+1} &= (1-u)^4[\frac{1}{8N^3} + \frac{7(2N-1)}{8N^3}\theta_t + \frac{6(2N-1)(2N-2)}{8N^3}\gamma_t \\
&\quad + \frac{(2N-1)(2N-2)(2N-3)}{8N^3}\delta_t], \;\; \text{and} \\
\Delta_{2,2,t+1} &= (1-u)^4[\frac{2N}{8N^3} + \frac{2(4N^2-1)}{8N^3}\theta_t + \frac{4(2N-1)(2N-2)}{8N^3}\gamma_t \\
&\quad + \frac{(2N-1)(2N-2)(2N-3)}{8N^3}\Delta_{2,2,t}].
\end{aligned} \tag{1.18}$$

The mutation rate u is generally very small. It is safe to assume that the higher orders of u (u^2, u^3 etc) are negligible. Now we assume population sizes are also very large which says u/N, u/N^2, u/N^3, and $1/N^2$ are very close to 0. So we can omit them while doing the algebra. This approximation leads to

$$\begin{aligned}
\theta_{t+1} &= \frac{1}{2N} + (1 - 2u - \frac{1}{2N})\theta_t, \\
\gamma_{t+1} &= \frac{3}{2N}\theta_t + (1 - 3u - \frac{3}{2N})\gamma_t, \\
\delta_{t+1} &= \frac{6}{2N}\gamma_t + (1 - 4u - \frac{6}{2N})\delta_t, \;\; \text{and} \\
\Delta_{2,2,t+1} &= \frac{2}{2N}\theta_t + \frac{4}{2N}\gamma_t + (1 - 4u - \frac{6}{2N})\Delta_{2,2,t}.
\end{aligned} \tag{1.19}$$

Let us define a new parameter $\phi = 4Nu$. Then at equilibrium the descent measures will be

$$\begin{aligned} \theta_e &= \frac{1}{1+\phi}, \\ \gamma_e &= \frac{2}{(1+\phi)(2+\phi)}, \\ \delta_e &= \frac{6}{(1+\phi)(2+\phi)(3+\phi)}, \quad \text{and} \\ \Delta_{2,2,e} &= \frac{6+\phi}{(1+\phi)(2+\phi)(3+\phi)}. \end{aligned} \tag{1.20}$$

From the above set of equations we get another set of relations between descent measures for large population size and generation (Balding and Nichols, 1995)

$$\begin{aligned} \gamma_e &= \frac{2\theta_e^2}{1+\theta_e}, \\ \delta_e &= \frac{6\theta_e^3}{(1+\theta_e)(1+2\theta_e)}, \quad \text{and} \\ \Delta_{2,2,e} &= \frac{\theta_e^2(1+5\theta_e)}{(1+\theta_e)(1+2\theta_e)}. \end{aligned} \tag{1.21}$$

All the results for the Wright-Fisher model with a both-way mutation also hold for the Wright-Fisher model with an infinite-alleles mutation model. In the infinite allele mutation model u is the mutation rate and each mutation generates a new allelic type.

Now we check the adequacy of the Normal and Dirichlet approximation for the descent measures. Figure 1.1 shows that the normal approximations of γ, δ and $\Delta_{2,2}$ do not work at all. The transition equations provide positive values for γ and δ where the normal distribution approximates these parameters to 0. For $\Delta_{2,2}$ the normal distribution gives positive value but it is smaller than the exact value of $\Delta_{2,2}$. These are true for both a pure drift and a both-way mutation model. The Dirichlet approximation works well for $\Delta_{2,2}$ under both the mutation models but does not work for γ and δ. Figure 1.1 shows that the Dirichlet approximation of γ and δ is always higher than the

true values of γ and δ that can be calculated using transition equations. The difference is smaller in the both-way mutation model but still there is a significant difference. So we conclude that in general we can not use the normal distribution to approximate γ and δ. The Dirichlet distribution is not appropriate for pure drift model but it can be used for both-way mutation model (although some small errors are involved in the approximation) for calculating the value of γ and δ.

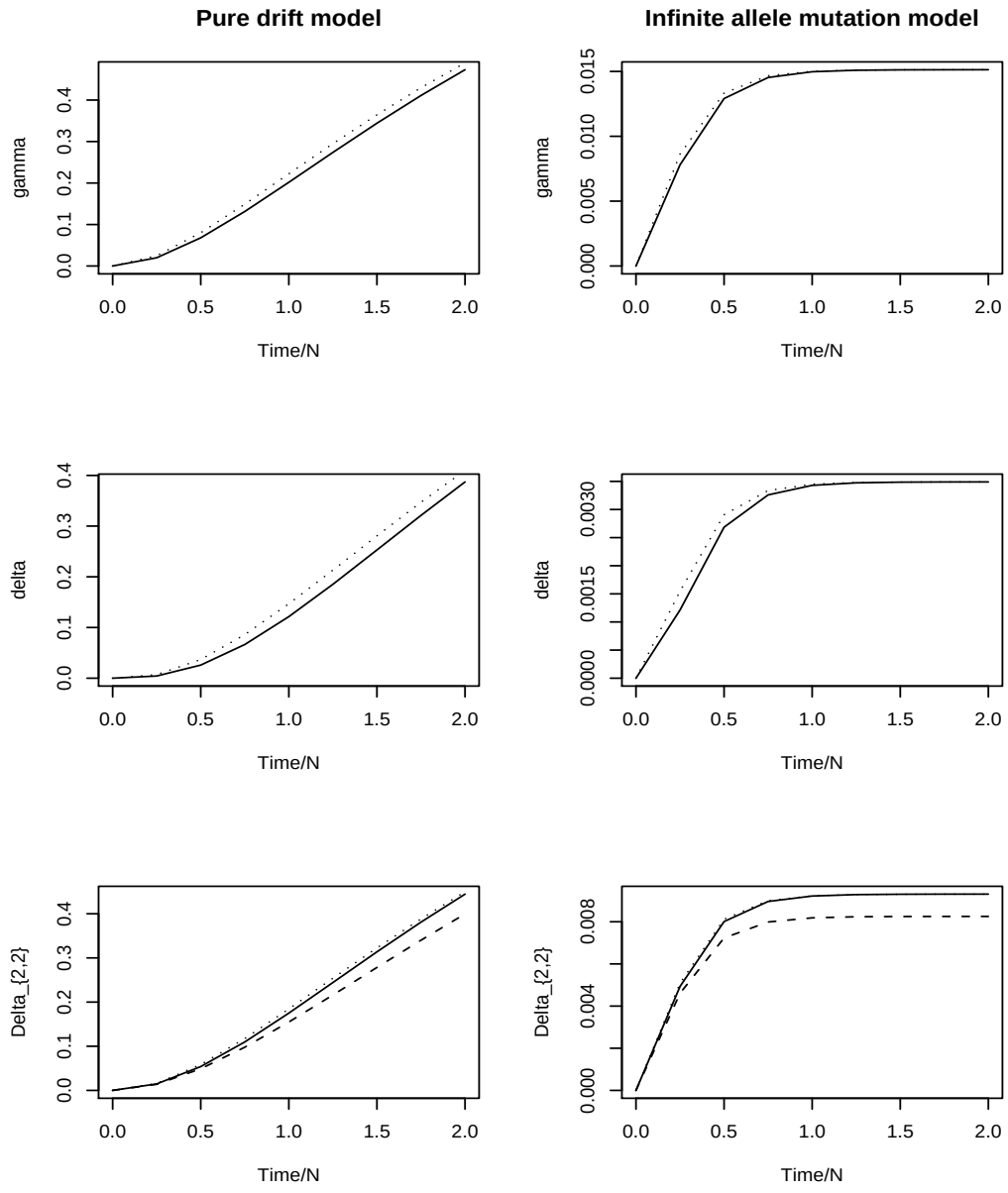

Figure 1.1: γ_e, δ_e, Δ_e from equations (1.21) (dotted line) and γ_n, δ_n, $\Delta_{2,2,n}$ from equations (1.17) (dashed line) compared to exact value of γ, δ, $\Delta_{2,2}$ under a pure drift model and a both-way mutation model (same as an infinite allele model). N = 5,000 and the mutation rate of the both-way mutation model is 0.0005.

Chapter 2

Estimation of Decent Measures

2.1 Introduction

Population structure is a great interest in Genetics. We can make inferences about the history of a population if we know the population structure. The population structure is generally modeled as association between alleles and can be characterized by descent measures. We can also make inferences about the relatedness between two arbitrary individuals in a population using descent measures. There are two types of second order descent measures, total inbreeding (F) and coancestry coefficient (θ). For a random mating population these two parameters are the same. The coancestry coefficient measures the degree of relationship between the individuals within the sub-populations relative to the amount of relationship found in the total population. Higher-order descent measures are useful in special situations. In general, k^{th}-order ($k \geq 2$) descent measures are characterized by the probability of different arrangements of identity between k alleles sampled from different levels in a population. The third (γ) and fourth order (δ and $\Delta_{2,2}$) descent measures can be used to find covariances of inbred relatives (Gillois, 1966). The third and fourth order descent measures are also needed to characterize a population that has selfing or biparental inbreeding such as a plant population (Ritland, 1987). The expression of joint allele frequencies also involves the descent measures. So the second, third, and fourth order descent measures are

frequently used to calculate different expressions in DNA profile matching (Weir, 1994). These descent measures are also useful in affected-relative tests.

The estimation of the descent measure θ and analog of θ have been discussed widely in the literature (Nei, 1973; Weir and Cockerham, 1984; Robertson and Hill, 1984; Slatkin, 1995). To estimate θ, the frequentist approaches, method of moments and maximum likelihood, and Bayesian methods have been used. The frequentist methods are computationally less intensive while the Bayesian approaches have the benefit of systematic incorporation of prior information about the data which increases the ability to capture important information about parameters in complex cases.

Weir and Cockerham (1984) first obtained the moment estimator of θ and Robertson and Hill (1984) followed their method. These two estimators are known as bivariate estimators. Later a multivariate estimator of θ was proposed by Long (1986). The bivariate estimators are constructed through combining individual alleles linearly over all alleles and loci, while the multivariate estimator is combined only over loci. Long's estimator is equivalent to the Robertson-Hill and Weir-Cockerham estimators for bi-allelic data from a single locus. Yang (1998) generalized Weir and Cockerham's estimator to an arbitrary number of levels in a population hierarchy. The above methods do not account for the linkage disequilibrium between loci in combining the information over loci. The best possible way to combine the bivariate estimators over alleles has remained an issue. Weir and Cockerham (1984) combined the estimates by taking the ratio of the sum of the numerators of each estimator to the sum of the denominators of each estimator. Alternatively, Robertson and Hill (1984) combined the estimates by taking weighted average of the ratio estimators over all alleles. Different weights have been proposed for multiple alleles and loci by minimizing the variance of the estimator for different ranges of the true value of θ. When the true value of θ is high then the variance of the estimator minimized for the Weir and Cockerham estimator while the Robertson and Hill approach minimized the variance for low to medium value of θ (Raufaste and Bonhomme, 2000). Raufaste and Bonhomme thus recommended the use of different estimators be governed by the true value of the parameter.

Applying Bayesian methods to the problem of inferring population structure has increased in last few years due to affordable computing power. By using the knowledge about populations gained in the past, the robustness of estimates from extreme data sets sampled from the present can be increased (Lange, 1995). The Bayesian methods also make simultaneous inferences about other parameters of interest such as model fit, number of distinct populations in a group of populations etc. The sensitivity and performance of Bayesian estimates depend on the choice of prior. Bayesian approaches to estimation of θ involve the assumption of hierarchial models including the forms of prior parameter and likelihood distributions. The Dirichlet (Balding and Nichols, 1995; Lange, 1995; Holsinger, 1999) and the multivariate normal (Weir and Hill, 2002) are two commonly used forms for the distribution of population allele frequencies with multiple alleles at a locus. For bi-allelic data such as SNP loci, the bivariate forms of these distributions reduce to a Beta and a normal distribution (Smouse and Willams, 1982; Holsinger, 1999; Balding, 2003; Nicholson and Donnelly, 2002). In Bayesian approaches the estimates of θ are the posterior mean of the conditional distribution of the parameters generated by using MCMC based rejection sampling.

The distributions of allele frequencies vary with population models and the time since divergence of populations. The stationary distribution of allele frequencies for most of the stochastic process models, such as, island model (Wright, 1931) and finite stepping-stone model (Maruyama, 1977) is Beta. The normal distribution has been justified by the appeal to large sample theory (Weir and Hill, 2002; Nicholson and Donnelly, 2002) rather than stationary distribution. The normal distribution has been used for non-equilibrium population which are likely to have shorter time since divergence (Nicholson and Donnelly, 2002) while a Beta or a Dirichlet distribution is a poor fit. For populations with weak drift and migration, the Dirichlet distribution may be a poor fit for stationary distribution because this increases the time to reach equilibrium. A Dirichlet distribution does also not fit in the population with high stepwise mutation rates (Graham et al., 2000).

Long and Kittles (2003) discussed the problems with classical analysis of population

structure introduced by simplifying assumption of a common value of θ across all populations. Their results showed that overall estimates of θ from global human data sets were meaningless and the estimates failed to describe the important local patterns and amount of genetic variance. Balding (Balding, 2003) also pointed out that the usual demographic variation and different population sizes in a collection of populations make the value of θ population specific.

Several estimators have now relaxed the constraint that the value of θ is the same across populations. Weir and Hill (2002) proposed a new parametrization of population model that defined a parameter specific to each population. This allows different amount of coancestry for different population. The first estimator obtained through a method of moments approach which is a direct extension of the previous Weir and Cockerham (1984) moment estimator of θ. This estimator is a ratio of unbiased estimates and therefore expected to be unbiased but it has large sample variance. This method does not assume any form for the distributions of allele frequencies. The second estimator of population-specific θ described by Weir and Hill (2002) was a maximum likelihood estimator. This estimator was developed under the assumption that the sample allele frequencies are multivariate normally distributed. This estimator has several desirable properties, such as invariance to transformation. However, this likelihood estimator is highly unstable when the likelihood function is flat.

In contrast to the frequentist approaches Nicholson and Donnelly (2002) approached this expanded parametrization from a Bayesian point of view, in the context of an application to SNP data. They assumed that the allele frequencies are normally distributed. The authors justify this model as having a reasonable fit to recently diverged, non-equilibrium populations. Balding (2003) also worked with a Bayesian approach but he assumed a Beta distribution for the allele frequency. Holsinger and Wallace (2004) extended Balding's model for the hierarchical model by describing a summary statistic that compared the posterior and prior distribution of the coancestry parameters.

The estimate of the third order descent measure (γ) is not well known. Weir (1994) proposed a moment estimator of γ for a very restrictive case. The performance of this

estimator is yet to be evaluated. If we assume a Dirichlet or a normal distribution for the allele frequencies then γ becomes a function of θ. In this case we can infer γ based on the estimate of θ. Unfortunately, these distributions are not realistic for natural populations, such as human populations. So the parameter γ has to be estimated independently. The demographic and size variation in a set of populations make the descent measures population specific. So the value of γ varies over different population. There is no estimator for a population-specific γ. There are two fourth order descent measures. We will show that these two parameters can not be estimated separately.

In this chapter we propose different estimators of descent measures. First we assume the same value of descent measures across the populations and propose a set of moment estimators of θ and γ based on the third order Analysis of Variance statistics. Then we propose another set of estimators of θ and γ using a direct probabilistic interpretation. We also compare method of moments estimator with probabilistic estimator analytically. Later we assume a population-specific value of descent measures and extend our estimators. To have a better idea about the estimates we also calculate the sampling properties of the estimators. We give the expressions for biases and variances of different estimators. Our estimators have large variances but these variances can be reduced by gathering more information from independent loci.

2.2 Replication of Evolution

In this section we discuss the history of the group of populations that we are interested in. We typically assume that the ancestral population has infinitely many individuals. The populations are evolved independently from the same ancestral population. The different populations have been generated through the replications of the same evolutionary process. Since the replications of the evolutionary process are independent, they will result independent populations. Even if all quantities such as population size, mating structure, and mutation rate were kept the same, a different population would result if evolution were repeated. So the parameter values for different populations will

be different. The averages can be taken over all possible outcomes of the evolutionary process to get the final value of the parameters. For example, the value of F (or θ) is averaged over sub-populations, and therefore requires an evolutionary model that predicts the levels of variation among sub-populations. Similarly, the value of higher order descent measures γ, δ and $\Delta_{2,2}$, are also averaged over sub-populations. So estimating these parameters require observations from more than one sub-population in order to quantify the variation between sub-populations. These parameters give information about long-term effects of demographic and evolutionary forces of the population. Our expectations will always be over different replications of the populations and any parameter value will be the average value over sub-populations.

2.3 Data

We have data from the r independent present-day populations. These populations have evolved from a common ancestral population. We will work with locus A. We assume that there are s different allelic forms in locus A namely, $A_1, A_2, \cdots, A_s$. The expected allele frequencies in each population are the same and they are $p_1, p_2, \cdots, p_s$ respectively. We have n_i sampled alleles from the i^{th} population. So there are total $\sum_{i=1}^{r} n_i = S$ sampled alleles. Define a set of indicator functions that describe our frequency data at locus A as follows:

$$x_{ij,k} = \begin{cases} 1 & \text{if the } j^{th} \text{ allele in } i^{th} \text{ population at locus } A \text{ is } A_k \\ 0 & \text{otherwise} \end{cases}$$

The observed frequency of the allele A_k in the i^{th} population is

$$\tilde{p}_{i,k} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij,k}, \tag{2.1}$$

and the overall weighted and un-weighted observed frequency of the allele A_k are

$$\tilde{p}_{w,k} = \frac{1}{S}\sum_{i=1}^{r} n_i \tilde{p}_{i,k} \quad \text{and} \quad \tilde{p}_{uw,k} = \frac{1}{r}\sum_{i=1}^{r} \tilde{p}_{i,k}. \tag{2.2}$$

When the sample sizes are equal i.e. $n_1 = n_2 = \cdots = n_r$, then $\tilde{p}_{w,k} = \tilde{p}_{uw,k}$.

2.4 Review of Existing Estimators of θ

2.4.1 Method of Moments Estimator

A moment estimator of θ was first found by Weir and Cockerham (1984). They assumed the same value of θ across different populations. In this section we describe the method of moments (MOM) estimator of θ proposed by Weir and Cockerham (1984). They defined the two mean square statistics based on the frequency of the allele A_k. The mean square statistics are

$$MSP_k = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\tilde{p}_{i,k} - \tilde{p}_{w,k})^2 \text{ and} \tag{2.3}$$

$$MSG_k = \frac{1}{\sum_{i=1}^{r}(n_i - 1)}\sum_{i=1}^{r} n_i \tilde{p}_{i,k}(1 - \tilde{p}_{i,k}). \tag{2.4}$$

The expectation of the statistics can be found using the theory

$$\mathrm{E}(x_{ij,k}x_{i'j',k}) = \begin{cases} p_k^2 + p_k(1-p_k)\theta & \text{if } i = i',\ j \neq j' \\ p_k^2 & \text{if } i \neq i'. \end{cases} \tag{2.5}$$

The expectations of the mean square statistics are

$$\mathrm{E}(MSP_k) = p_k(1-p_k)[1 + (n_{c_1} - 1)\theta] \text{ and} \tag{2.6}$$

$$\mathrm{E}(MSG_k) = p_k(1-p_k)(1-\theta), \tag{2.7}$$

where

$$n_{c_1} = \frac{1}{r-1}(\sum_{i=1}^{r} n_i - \frac{\sum_{i=1}^{r} n_i^2}{\sum_{i=1}^{r} n_i}) = \frac{1}{r-1}(S - \frac{\sum_{i=1}^{r} n_i^2}{S}).$$

Using (2.6) and (2.7), Weir and Cockerham (1984) led to their moment estimator of θ,

$$\hat{\theta}_{WC,k} = \frac{MSP_k - MSG_k}{MSP_k + (n_{c_1} - 1)MSG_k}. \tag{2.8}$$

The above estimator is based on the frequency data for the allele A_k. There are different estimators corresponding to different alleles at different loci. After combining information over alleles and loci, Weir and Cockerham (1984) proposed the overall estimator of θ,

$$\hat{\theta}_{WC} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(MSP_{lk} - MSG_{lk})}{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(MSP_{l,k} + (n_{c_1} - 1)MSG_{lk})}. \tag{2.9}$$

where MSP_{lk} and MSG_{lk} are the two mean square statistics for the k^{th} allele at the l^{th} locus. They assumed that there are L independent loci and the l^{th} locus has s_l alleles.

2.4.2 ML Estimator Based on Normal Distribution

Weir and Hill (2002) proposed an estimator of θ assuming a multivariate normal distribution for allele frequencies. They justified the assumption of normal distribution using large sample sizes and central limit theorem. The normal distribution is an approximate distribution and it works well for small values of θ. Here the authors assumed that $n_i \to \infty$ which is equivalent to assume $n_i = n$ and $n \to \infty$. There are s alleles at locus A which give $s-1$ independent allele frequencies. Define a new vector of observed allele frequencies as $\tilde{\mathbf{p}}_i = (\tilde{p}_{i,1}, \tilde{p}_{i,2}, \cdots, \tilde{p}_{i,s-1})' \ \forall i = 1, 2, \cdots, r$. $\tilde{\mathbf{p}}_i$'s are independent and identically distributed. Weir and Hill (2002) assumed

$$\tilde{\mathbf{p}}_i \sim \mathrm{MVN}_{s-1}(\mathbf{p}, C), \tag{2.10}$$

where

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{s-1} \end{bmatrix}, \; C = \phi \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_{s-1} \\ -p_1p_2 & p_2(1-p_2) & \dots & -p_2p_{s-1} \\ \vdots & \vdots & \dots & \vdots \\ p_1p_{s-1} & -p_2p_{s-1} & \dots & p_{s-1}(1-p_{s-1}) \end{bmatrix},$$

where $\phi = \frac{1}{n}[1 + (n-1)\theta]$. From standard theory (see Chapter 3), the quadratic form

$$Q \;=\; \frac{1}{\phi} \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(\tilde{p}_{i,k} - \tilde{p}_{w,k})^2}{\tilde{p}_{w,k}} \sim \chi^2_{(r-1)(s-1)}. \tag{2.11}$$

The equation (2.11) gives the maximum likelihood estimate of θ as

$$\hat{\theta}_N \;=\; \frac{1}{n-1}\Big[\frac{n}{(r-1)(s-1)} \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(\tilde{p}_{i,k} - \tilde{p}_{w,k})^2}{\tilde{p}_{w,k}} - 1\Big]. \tag{2.12}$$

If there is more than one loci then the final estimator is the average of the locus specific estimators of θ.

2.4.3 Bayesian Estimator

In the Bayesian set up scientists assume that the allele frequencies follow a joint distribution (Lange, 1995; Balding and Nichols, 1995). The parameters of this distribution depend on θ. The conditional distribution of the allele counts given the allele frequencies is Multinomial. The joint prior distribution of allele frequencies is generally a Multivariate normal or Dirichlet. Now using these two facts we find the posterior distribution of θ. The posterior mean is the Bayesian estimate of θ. Most of the times we can not recognize the full posterior distribution. In these cases we use MCMC method to find the posterior mean of θ.

2.5 New Estimators of θ and γ

2.5.1 Method of Moments Estimator

In this section we propose new moment estimators for θ and γ. We find a set of statistics whose expectations depend on the parameters θ and γ. Then we equate the theoretical moments with sample moments and get the estimators of the parameters. In the expectation of second moment allele frequency, the parameter γ does not appear. This can be seen from the equation (2.5) which does not have γ in its expression. But the parameter γ appears in the expression of third or higher order moments of allele frequencies. The parameter θ appears in second or higher order moments of allele frequencies. Here consider two statistics that are based on third order sample moments of allele frequencies. To find the third order moments of the frequency of the allele A_k, we need to use the relation

$$\mathrm{E}(x_{ij,k}x_{i'j',k}x_{i''j'',k}) = \begin{cases} \gamma p_k + 3(\theta-\gamma)p_k^2 + (1-3\theta+2\gamma)p_k^3 & \text{if } i = i' = i'' \\ \theta p_k^2 + (1-\theta)p_k^3 & \text{if } i = i' \neq i'' \\ p_k^3 & \text{if } i \neq i' \neq i''. \end{cases} \tag{2.13}$$

The above equation (2.13) can also be written as

$$\mathrm{E}(x_{ij,k}x_{i'j',k}x_{i''j'',k}) = \begin{cases} p_k^3 + 3p_k^2(1-p_k)\theta + p_k(1-p_k)(1-2p_k)\gamma & \text{if } i = i' = i'' \\ p_k^3 + p_k^2(1-p_k)\theta & \text{if } i = i' \neq i'' \\ p_k^3 & \text{if } i \neq i' \neq i''. \end{cases} \tag{2.14}$$

The above equation (2.14) shows that when an allele frequency is 0.5, then the third order moment does not depend on γ but it depends on θ. So an allele with frequency 0.5 does not provide any information about γ.

Now we propose three statistics based on the frequency of the allele A_k to find a

moment estimate of θ and γ. The three statistics are

$$\begin{aligned} S_{1,k} &= \frac{S}{(r-1)(r-2)} \sum_{i=1}^{r} n_i (\tilde{p}_{i,k} - \tilde{p}_{w,k})^3, \\ S_{2,k} &= \frac{r}{(r-1)\sum_{i=1}^{r}(n_i-1)} \sum_{i=1}^{r} n_i^2 \tilde{p}_{i,k}(1-\tilde{p}_{i,k})(\tilde{p}_{i,k} - \tilde{p}_{w,k}), \quad \text{and} \\ S_{3,k} &= \frac{1}{\sum_{i=1}^{r}(n_i-1)(n_i-2)} \sum_{i=1}^{r} n_i^2 \tilde{p}_{i,k}(1-\tilde{p}_{i,k})(1-2\tilde{p}_{i,k}). \end{aligned} \tag{2.15}$$

Using the equations (2.14) we find the estimate of these three statistics and they are

$$\begin{aligned} &\mathrm{E}(S_{1,k}) = u_k(n_{c_2} + 3n_{c_3}\theta + n_{c_4}\gamma), \\ &\mathrm{E}(S_{2,k}) = u_k[\frac{r(n_{c_1}-1)}{S-r} + \frac{r(n_{c_5} - 4n_{c_1} + 3)}{S-r}\theta - \frac{r(n_{c_5} - 3n_{c_1} + 2)}{S-r}\gamma], \quad \text{and} \\ &\mathrm{E}(S_{3,k}) = u_k(1 - 3\theta + 2\gamma), \end{aligned} \tag{2.16}$$

where $u_k = p_k(1-p_k)(1-2p_k)$ and

$$\begin{aligned} n_{c_2} &= \frac{1}{(r-1)(r-2)}(S\sum_{i=1}^{r}\frac{1}{n_i} - 3r + 2), \\ n_{c_3} &= \frac{1}{(r-1)(r-2)}(Sr - S\sum_{i=1}^{r}\frac{1}{n_i} - 3S + 3r + \frac{2\sum_{i=1}^{r} n_i^2}{S} - 2), \\ n_{c_4} &= \frac{1}{(r-1)(r-2)}(S^2 - 3Sr + 2S\sum_{i=1}^{r}\frac{1}{n_i} - 3\sum_{i=1}^{r} n_i^2 + 9S - 6r \\ &\quad + \frac{2\sum_{i=1}^{r} n_i^3}{S} - \frac{6\sum_{i=1}^{r} n_i^2}{S} + 4), \quad \text{and} \\ n_{c_5} &= \frac{1}{r-1}(\sum_{i=1}^{r} n_i^2 - \frac{\sum_{i=1}^{r} n_i^3}{S}). \end{aligned} \tag{2.17}$$

The expectations defined in (2.16) are 0 if the frequency of the allele A_k is 0.5. So under this situation these expectations are not informative about the parameters θ and γ. It is also important to note that we need at least three populations to have information

about the parameters from our statistics. In principle, from the three equations given in the equation (2.16), we can find three statistics $T_{1,k}$, $T_{2,k}$ and $T_{3,k}$ that are linear combinations of $S_{1,k}$, $S_{2,k}$ and $S_{3,k}$ such that

$$\begin{aligned} \mathrm{E}(T_{1,k}) &= p_k(1-p_k)(1-2p_k), \\ \mathrm{E}(T_{2,k}) &= p_k(1-p_k)(1-2p_k)\theta, \text{ and} \\ \mathrm{E}(T_{3,k}) &= p_k(1-p_k)(1-2p_k)\gamma. \end{aligned} \tag{2.18}$$

After doing some algebra we get

$$\begin{aligned} T_{1,k} &= [\frac{2r(n_{c_5}-4n_{c_1}+3)}{S-r} - \frac{3r(n_{c_5}-3n_{c_1}+2)}{S-r}]S_{1,k} + (-3n_{c_4}-6n_{c_3})S_{2,k} \\ &\quad +[-3n_{c_3}\frac{r(n_{c_5}-3n_{c_1}+2)}{S-r} - n_{c_4}\frac{r(n_{c_5}-4n_{c_1}+3)}{S-r}]S_{3,k}, \\ T_{2,k} &= [-\frac{r(n_{c_5}-3n_{c_1}+2)}{S-r} - \frac{2r(n_{c_1}-1)}{S-r}]S_{1,k} + (2n_{c_2}-n_{c_4})S_{2,k} \\ &\quad +[n_{c_4}\frac{r(n_{c_1}-1)}{S-r} + n_{c_2}\frac{r(n_{c_5}-3n_{c_1}+2)}{S-r}]S_{3,k}, \text{ and} \\ T_{3,k} &= [-\frac{3r(n_{c_1}-1)}{S-r} - \frac{r(n_{c_5}-4n_{c_1}+3)}{S-r}]S_{1,k} + (3n_{c_3}+3n_{c_2})S_{2,k} \\ &\quad +[n_{c_2}\frac{r(n_{c_5}-4n_{c_1}+3)}{S-r} - 3n_{c_3}\frac{r(n_{c_1}-1)}{S-r}]S_{3,k}. \end{aligned} \tag{2.19}$$

When $T_{1,k} \neq 0$ then using ratio estimation theory we get our moment estimator of θ and γ as

$$\hat{\theta}_{M,k} = \frac{T_{2,k}}{T_{1,k}} I(T_{1,k} \neq 0) \text{ and} \tag{2.20}$$

$$\hat{\gamma}_{M,k} = \frac{T_{3,k}}{T_{1,k}} I(T_{1,k} \neq 0). \tag{2.21}$$

The above estimators are based on the frequency data of allele A_k. For each allele frequency data at each locus we have one new estimator of θ and γ. Weir-Cockerham and Robertson-Hill proposed two different methods for combing the information from

different alleles. Here we use both the combining methods to get two sets of estimators. The expectations of the statistics $T_{1,k}$, $T_{2,k}$, $T_{3,k}$ are positive when $\tilde{p}_k < 0.5$ and are negative when $\tilde{p}_k > 0.5$. There is only one allele that has frequency greater than 0.5. If we add $T_{1,k}$ over $\forall k$, then there is a chance that we might end up getting some value very close to 0. For two alleles we always get $\sum_{k=1}^{2} T_{1,k} = 0$. Under these situations the estimators do not work well if we get the final estimator using Weir-Cockerham's weight. So we exclude the allele that has observed frequency greater than 0.5 and work with the rest independent allele frequencies. We combine the estimators corresponding to different alleles and different loci by our modified method and get the final estimators

$$\begin{aligned}
\hat{\theta}_{1,M} &= \frac{\sum_{k=1}^{s} T_{2,k} I(\tilde{p}_{w,k} < 0.5)}{\sum_{k=1}^{s} T_{1,k} I(\tilde{p}_{w,k} < 0.5)}, \\
\hat{\theta}_{2,M} &= \frac{1}{s} \sum_{k=1}^{s} \frac{T_{2,k}}{T_{1,k}} I(T_{1,k} \neq 0), \\
\hat{\gamma}_{1,M} &= \frac{\sum_{k=1}^{s} T_{3,k} I(\tilde{p}_{w,k} < 0.5)}{\sum_{k=1}^{s} T_{1,k} I(\tilde{p}_{w,k} < 0.5)}, \quad \text{and} \\
\hat{\gamma}_{2,M} &= \frac{1}{s} \sum_{k=1}^{s} \frac{T_{3,k}}{T_{1,k}} I(T_{1,k} \neq 0).
\end{aligned} \tag{2.22}$$

If we have data from L independent loci then our final estimators would be

$$\hat{\theta}_{1,M} = \frac{\sum_{l=1}^{L} T_{2,l}}{\sum_{l=1}^{L} T_{1,l}}, \qquad \hat{\theta}_{2,M} = \frac{1}{L} \sum_{l=1}^{L} \hat{\theta}_{2,M,l}, \tag{2.23}$$

$$\hat{\gamma}_{1,M} = \frac{\sum_{l=1}^{L} T_{3,l}}{\sum_{l=1}^{L} T_{1,l}}, \quad \text{and} \quad \hat{\gamma}_{2,M} = \frac{1}{L} \sum_{l=1}^{L} \hat{\gamma}_{2,M,l}, \tag{2.24}$$

where $\theta_{2,M,l}$ and $\gamma_{2,M,l}$ are the estimators of θ and γ respectively based on the l^{th} locus. $T_{1,l}$, $T_{2,l}$ and $T_{3,l}$ are the sum of $T_{1,k}$, $T_{2,k}$ and $T_{3,k}$ respectively over different alleles that has frequency less than 0.5 at the locus l. For equal sample size our statistics reduce to Weir's (1994) statistics. We have found that the expressions for the expectations of the statistics derived by us are not identical to the expressions derived by Weir. For

equal sample sizes case the equation (2.19) reduces to

$$\begin{aligned} T_{1,k} &= S_{1,k} + 3(n-1)S_{2,k} + (n-1)(n-2)S_{3,k}, \\ T_{2,k} &= S_{1,k} + (n-3)S_{2,k} - (n-2)S_{3,k}, \text{ and} \\ T_{3,k} &= S_{1,k} - 3S_{2,k} + 2S_{3,k}. \end{aligned} \tag{2.25}$$

2.5.2 Estimators with Probabilistic Interpretation

In this section we provide estimators of θ and γ using a Probabilistic Interpretation. Here we work with the same set up as in the previous sections. If we choose two alleles from a population then the probability that both the alleles are of the type A_k, depends on the parameter θ. The probability of getting three A_k alleles from a population depends on θ and γ. The above two probabilities depend on the expected frequency of A_k, p_k, as well. We explore these probabilities and find estimators of the parameters θ and γ. Now we define a new set of parameters

$$\pi_{i,j,k} = \mathrm{E}_{sub-pop}[Pr(i \text{ allele(s) from } j \text{ population(s) being of the type } A_k)]. \tag{2.26}$$

We are interested in the parameters $\pi_{1,1,k}$, $\pi_{2,1,k}$, $\pi_{2,2,k}$, $\pi_{3,1,k}$, $\pi_{3,2,k}$ and $\pi_{3,3,k}$. Using the population genetics theory we get

$$\begin{aligned} \pi_{1,1,k} &= p_k, \quad \pi_{2,2,k} = p_k^2, \quad \pi_{3,3,k} = p_k^3, \\ \pi_{2,1,k} &= p_k^2 + (p_k - p_k^2)\theta, \quad \pi_{3,2,k} = p_k^3 + (p_k^2 - p_k^3)\theta, \text{ and} \\ \pi_{3,1,k} &= p_k^3 + 3(p_k^2 - p_k^3)\theta + (p_k - 3p_k^2 + 2p_k^3)\gamma. \end{aligned} \tag{2.27}$$

After doing some algebra using the equations in (2.27) we get

$$\theta = \frac{\pi_{2,1,k} - \pi_{2,2,k}}{\pi_{1,1,k} - \pi_{2,2,k}}, \quad \theta = \frac{\pi_{3,2,k} - \pi_{3,3,k}}{\pi_{2,2,k} - \pi_{3,3,k}} \quad \text{and} \tag{2.28}$$

$$\gamma = \frac{\pi_{3,1,k} - 3\pi_{3,2,k} + 2\pi_{3,3,k}}{\pi_{1,1,k} - 3\pi_{2,2,k} + 2\pi_{3,3,k}}. \tag{2.29}$$

From the equations (2.28) and (2.29), the estimates of the parameters θ and γ based on the k^{th} allele frequency at locus A are

$$\hat{\theta}_{1,P,k} = \frac{\hat{\pi}_{2,1,k} - \hat{\pi}_{2,2,k}}{\hat{\pi}_{1,1,k} - \hat{\pi}_{2,2,k}}, \quad \hat{\theta}_{2,P,k} = \frac{\hat{\pi}_{3,2,k} - \hat{\pi}_{3,3,k}}{\hat{\pi}_{2,2,k} - \hat{\pi}_{3,3,k}} \quad \text{and} \tag{2.30}$$

$$\hat{\gamma}_{P,k} = \frac{\hat{\pi}_{3,1,k} - 3\hat{\pi}_{3,2,k} + 2\hat{\pi}_{3,3,k}}{\hat{\pi}_{1,1,k} - 3\hat{\pi}_{2,2,k} + 2\hat{\pi}_{3,3,k}}. \tag{2.31}$$

Now we have to find the estimates of $\pi_{1,1,k}, \pi_{2,1,k}, \pi_{2,2,k}, \pi_{3,1,k}, \pi_{3,2,k}$ and $\pi_{3,3,k}$. Giving equal weights to each population, the estimates of the probabilities are

$$\begin{aligned}
\hat{\pi}_{1,1,k} &= \frac{1}{r}\sum_{i=1}^{r} \tilde{p}_{k,i} = \tilde{p}_{uw,k}, \\
\hat{\pi}_{2,1,k} &= \frac{1}{r}\sum_{i=1}^{r} \frac{n_i\tilde{p}_{k,i}^2 - \tilde{p}_{k,i}}{n_i - 1}, \\
\hat{\pi}_{2,2,k} &= \frac{(\sum_{i=1}^{r} \tilde{p}_{k,i})^2 - \sum_{i=1}^{r} \tilde{p}_{k,i}^2}{r(r-1)}, \\
\hat{\pi}_{3,1,k} &= \frac{1}{r}\sum_{i=1}^{r} \frac{n_i^2\tilde{p}_{k,i}^3 - 3n_i\tilde{p}_{k,i}^2 + 2\tilde{p}_{k,i}}{(n_i-1)(n_i-2)}, \\
\hat{\pi}_{3,2,k} &= \frac{1}{r(r-1)}\Big[\sum_{i=1}^{r} \frac{n_i\tilde{p}_{k,i}^2 - \tilde{p}_{k,i}}{n_i - 1}\Big] \sum_{i=1}^{r} \tilde{p}_{k,i} - \frac{1}{r(r-1)} \sum_{i=1}^{r} \frac{n_i\tilde{p}_{k,i}^3 - \tilde{p}_{k,i}^2}{n_i - 1}, \quad \text{and} \\
\hat{\pi}_{3,3,k} &= \frac{(\sum_{i=1}^{r} \tilde{p}_{k,i})^3 - 3(\sum_{i=1}^{r} \tilde{p}_{k,i})(\sum_{i=1}^{r} \tilde{p}_{k,i}^2) + 2(\sum_{i=1}^{r} \tilde{p}_{k,i}^3)}{r(r-1)(r-2)}.
\end{aligned} \tag{2.32}$$

In theory the equation (2.29) does not exist when $p_k = 0.5$. When $p_k = 0.5$, then the equation provides $\gamma = \frac{0}{0}$ which does not make any sense. This can be observed from the equation (2.14) which says that the third moment of observed frequency of an allele does not involve γ if the allele frequency is 0.5. So when $p_k = 0.5$, we can not estimate γ from the frequency data of A_k. At this point we have estimators of θ and γ based on a single allele frequency. There are several methods to combine the estimators corresponding different allele frequencies to get a final estimator. We consider two different approaches and get two different sets of estimators. The first approach was

proposed by Weir and Cockerham (1984) and they combined the estimates by taking the ratio of the sum of the numerators of each estimator to the sum of the denominators of each estimator. Suppose there are L independent loci and the l^{th} locus has s_l alleles. Then the final estimates based on the Weir-Cockerham's method are

$$\hat{\theta}_{1,P} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{2,1,k,l} - \hat{\pi}_{2,2,k,l})}{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{1,1,k,l} - \hat{\pi}_{2,2,k,l})}, \tag{2.33}$$

$$\hat{\theta}_{2,P} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{3,2,k,l} - \hat{\pi}_{3,3,k,l})}{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{2,2,k,l} - \hat{\pi}_{3,3,k,l})}, \text{ and} \tag{2.34}$$

$$\hat{\gamma}_{1,P} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{3,1,k,l} - 3\hat{\pi}_{3,2,k,l} + 2\hat{\pi}_{3,3,k,l})I(\tilde{p}_{w,k,l} < 0.5)}{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{1,1,k,l} - 3\hat{\pi}_{2,2,k,l} + 2\hat{\pi}_{3,3,k,l})I(\tilde{p}_{w,k,l} < 0.5)}, \tag{2.35}$$

where $\hat{\pi}_{1,1,k,l}$, $\hat{\pi}_{2,1,k,l}$, $\hat{\pi}_{2,2,k,l}$, $\hat{\pi}_{3,1,k,l}$, $\hat{\pi}_{3,2,k,l}$ and $\hat{\pi}_{3,3,k,l}$ are estimate of $\pi_{1,1,k}$, $\pi_{2,1,k}$, $\pi_{2,2,k}$, $\pi_{3,1,k}$, $\pi_{3,2,k}$ and $\pi_{3,3,k}$ respectively for the l^{th} locus. $\tilde{p}_{w,k,l}$ is the weighted average frequency of the allele A_k at the l^{th} locus.

On the other hand, Robertson and Hill (1984) combined the estimates by taking a weighted average of the ratio estimators over all alleles at different locus. Using this method we get another set of estimators

$$\hat{\theta}_{3,P} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{s_l}\sum_{k=1}^{s_l}\frac{(\hat{\pi}_{2,1,k,l} - \hat{\pi}_{2,2,k,l})}{(\hat{\pi}_{1,1,k,l} - \hat{\pi}_{2,2,k,l})}, \tag{2.36}$$

$$\hat{\theta}_{4,P} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{s_l}\sum_{k=1}^{s_l}\frac{(\hat{\pi}_{3,2,k,l} - \hat{\pi}_{3,3,k,l})}{(\hat{\pi}_{2,2,k,l} - \hat{\pi}_{3,3,k,l})}, \text{ and} \tag{2.37}$$

$$\hat{\gamma}_{2,P} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{s_l}\sum_{k=1}^{s_l}\frac{(\hat{\pi}_{3,1,k,l} - 3\hat{\pi}_{3,2,k,l} + 2\hat{\pi}_{3,3,k,l})}{(\hat{\pi}_{1,1,k,l} - 3\hat{\pi}_{2,2,k,l} + 2\hat{\pi}_{3,3,k,l})}I(\text{denominator} \neq 0), \tag{2.38}$$

where $\hat{\pi}_{1,1,k,l}$, $\hat{\pi}_{2,1,k,l}$, $\hat{\pi}_{2,2,k,l}$, $\hat{\pi}_{3,1,k,l}$, $\hat{\pi}_{3,2,k,l}$ are $\hat{\pi}_{3,3,k,l}$ are estimates of $\pi_{1,1,k}$, $\pi_{2,1,k}$, $\pi_{2,2,k}$, $\pi_{3,1,k}$, $\pi_{3,2,k}$ and $\pi_{3,3,k}$ respectively for the l^{th} locus.

When sample sizes are equal then $\hat{\theta}_{1,P}$ is exactly the same as θ_{WC}, the classical estimator given by Weir and Cockerham (1984). In appendix A we have shown the result algebraically. When sample sizes are not equal then the equality does not hold.

2.6 New Estimators of Population-specific θ and γ

Long and Kittles (2003) and Balding (2003) showed the overall estimates of descent measures fail to describe the important local patterns and amount of genetic variance because of demographic variances among populations. In this section we assume the value of θ and γ in the i^{th} population is θ_i and γ_i respectively. Researchers provided estimators of θ_i using different methods such as MOM, ML, and Bayesian methods. In the next section we describe new estimates of the population-specific θ using a probabilistic interpretation. In literature there is no estimator for the population-specific γ. We propose several estimators of the population-specific γ.

2.6.1 Estimators with Probabilistic Interpretation

In this section we provide estimators of population-specific θ and γ using a probabilistic interpretation. For estimating θ_i and γ_i we define the following parameters

$$\begin{aligned}
\pi_{2,1,k,i} &= \mathrm{E}_{sub-pop}[Pr(\text{Two alleles from } i^{th} \text{ population are of the type } A_k)], \\
\pi_{3,1,k,i} &= \mathrm{E}_{sub-pop}[Pr(\text{Three alleles from } i^{th} \text{ population are of the type } A_k)], \\
\pi_{3,2,k,i} &= \mathrm{E}_{sub-pop}[Pr(\text{Two alleles from } i^{th} \text{ population and one allele from another} \\
&\quad \text{population are of the type } A_k)].
\end{aligned} \tag{2.39}$$

We also need to consider three more parameters, $\pi_{1,1,k}$, $\pi_{2,2,k}$ and $\pi_{3,3,k}$ that are defined in the equation (2.26). The relation of the above parameters with expected allele frequencies and population-specific descent measures can be found from the equation

$$\mathrm{E}(x_{ij,k}x_{i'j',k}x_{i''j'',k}) = \begin{cases} p_k^3 + 3p_k^2(1-p_k)\theta_i + p_k(1-p_k)(1-2p_k)\gamma_i & \text{if } i = i' = i'' \\ p_k^3 + p_k^2(1-p_k)\theta_i & \text{if } i = i' \neq i'' \\ p_k^3 & \text{if } i \neq i' \neq i'', \end{cases} \tag{2.40}$$

and the relations are

$$\begin{aligned}
\pi_{1,1,k} &= p_k, \quad \pi_{2,2,k} = p_k^2, \quad \pi_{3,3,k} = p_k^3, \\
\pi_{2,1,k,i} &= p_k^2 + (p_k - p_k^2)\theta_i, \quad \pi_{3,2,k,i} = p_k^3 + (p_k^2 - p_k^3)\theta_i, \quad \text{and} \\
\pi_{3,1,k,i} &= p_k^3 + 3(p_k^2 - p_k^3)\theta_i + (p_k - 3p_k^2 + 2p_k^3)\gamma_i.
\end{aligned} \tag{2.41}$$

After doing some algebra using the equations in (2.41) we get

$$\theta_i = \frac{\pi_{2,1,k,i} - \pi_{2,2,k}}{\pi_{1,1,k} - \pi_{2,2,k}}, \quad \theta_i = \frac{\pi_{3,2,k,i} - \pi_{3,3,k}}{\pi_{2,2,k} - \pi_{3,3,k}}, \quad \text{and} \tag{2.42}$$

$$\gamma_i = \frac{\pi_{3,1,k,i} - 3\pi_{3,2,k,i} + 2\pi_{3,3,k}}{\pi_{1,1,k} - 3\pi_{2,2,k} + 2\pi_{3,3,k}}. \tag{2.43}$$

From the equations (2.42) and (2.43), the estimates of the parameters θ_i and γ_i based on the frequency data of the allele A_k are

$$\hat{\theta}_{1,P,k,i} = \frac{\hat{\pi}_{2,1,k,i} - \hat{\pi}_{2,2,k}}{\hat{\pi}_{1,1,k} - \hat{\pi}_{2,2,k}}, \quad \hat{\theta}_{2,P,k,i} = \frac{\hat{\pi}_{3,2,k,i} - \hat{\pi}_{3,3,k}}{\hat{\pi}_{2,2,k} - \hat{\pi}_{3,3,k}}, \quad \text{and} \tag{2.44}$$

$$\hat{\gamma}_{P,k,i} = \frac{\hat{\pi}_{3,1,k,i} - 3\hat{\pi}_{3,2,k,i} + 2\hat{\pi}_{3,3,k}}{\hat{\pi}_{1,1,k} - 3\hat{\pi}_{2,2,k} + 2\hat{\pi}_{3,3,k}}. \tag{2.45}$$

Now we have to find the estimates of $\pi_{1,1,k}, \pi_{2,1,k,i}, \pi_{2,2,k}, \pi_{3,1,k,i}, \pi_{3,2,k,i}$ and $\pi_{3,3,k}$. The estimates of $\pi_{1,1,k}, \pi_{2,2,k}$ and $, \pi_{3,3,k}$ are given in the equation (2.32). Giving equal weights to each population, we get the estimates of the other probabilities as

$$\begin{aligned}
\hat{\pi}_{2,1,i,k} &= \frac{n_i\tilde{p}_{k,i}^2 - \tilde{p}_{k,i}}{n_i - 1}, \\
\hat{\pi}_{3,1,i,k} &= \frac{n_i^2\tilde{p}_{k,i}^3 - 3n_i\tilde{p}_{k,i}^2 + 2\tilde{p}_{k,i}}{(n_i - 1)(n_i - 2)}, \quad \text{and} \\
\hat{\pi}_{3,2,i,k} &= \frac{1}{(r-1)}\frac{n_i\tilde{p}_{k,i}^2 - \tilde{p}_{k,i}}{n_i - 1}(\sum_{j=1}^{r}\tilde{p}_{k,j} - \tilde{p}_{k,i}).
\end{aligned} \tag{2.46}$$

In theory the equation (2.29) does not exist when $p_k = 0.5$. When $p_k = 0.5$, then the equation provides $\gamma_i = \frac{0}{0}$ which does not make any sense. So when $p_k = 0.5$, we can not estimate γ_i from the frequency data of the allele A_k. At this point we have estimators of θ_i and γ_i based on a single allele frequency. There are several methods to combine the estimators corresponding different allele frequencies to get a final estimator. Weir and Cockerham (1984) combined the estimates by taking the ratio of the sum of the numerators of each estimator to the sum of the denominators of each estimator. Suppose there are L independent loci and the l^{th} locus has s_l alleles. Then the final estimates based on the Weir-Cockerham's method are

$$\hat{\theta}_{1,P,i} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{2,1,k,i,l} - \hat{\pi}_{2,2,k,l})}{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{1,1,k,l} - \hat{\pi}_{2,2,k,l})}, \tag{2.47}$$

$$\hat{\theta}_{2,P,i} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{3,2,k,i,l} - \hat{\pi}_{3,3,k,l})}{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{2,2,k,l} - \hat{\pi}_{3,3,k,l})}, \text{ and} \tag{2.48}$$

$$\hat{\gamma}_{1,P,i} = \frac{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{3,1,k,i,l} - 3\hat{\pi}_{3,2,k,i,l} + 2\hat{\pi}_{3,3,k,l})I(\tilde{p}_{w,k,l} < 0.5)}{\sum_{l=1}^{L}\sum_{k=1}^{s_l}(\hat{\pi}_{1,1,k,l} - 3\hat{\pi}_{2,2,k,l} + 2\hat{\pi}_{3,3,k,l})I(\tilde{p}_{w,k,l} < 0.5)}. \tag{2.49}$$

On the other hand, Robertson and Hill (1984) combined the estimates by taking weighted average of the ratio estimators over all alleles at different locus. Using Robertson-Hill's method we get another set of estimators

$$\hat{\theta}_{3,P,i} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{s_l}\sum_{k=1}^{s_l}\frac{(\hat{\pi}_{2,1,k,i,l} - \hat{\pi}_{2,2,k,l})}{(\hat{\pi}_{1,1,k,l} - \hat{\pi}_{2,2,k,l})}, \tag{2.50}$$

$$\hat{\theta}_{4,P,i} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{s_l}\sum_{k=1}^{s_l}\frac{(\hat{\pi}_{3,2,k,i,l} - \hat{\pi}_{3,3,k,l})}{(\hat{\pi}_{2,2,k,l} - \hat{\pi}_{3,3,k,l})}, \text{ and} \tag{2.51}$$

$$\hat{\gamma}_{2,P,i} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{s_l}\sum_{k=1}^{s_l}\frac{(\hat{\pi}_{3,1,k,i,l} - 3\hat{\pi}_{3,2,k,i,l} + 2\hat{\pi}_{3,3,k,l})}{(\hat{\pi}_{1,1,k,l} - 3\hat{\pi}_{2,2,k,l} + 2\hat{\pi}_{3,3,k,l})}I(\text{denominator} \neq 0). \tag{2.52}$$

where $\hat{\pi}_{1,1,k,l}$, $\hat{\pi}_{2,1,k,i,l}$, $\hat{\pi}_{2,2,k,l}$, $\hat{\pi}_{3,1,k,i,l}$, $\hat{\pi}_{3,2,k,i,l}$ and $\hat{\pi}_{3,3,k,l}$ are the estimate of $\pi_{1,1,k}$, $\pi_{2,1,i,k}$, $\pi_{2,2,k}$, $\pi_{3,1,i,k}$, $\pi_{3,2,i,k}$ and $\pi_{3,3,k}$ respectively for the l^{th} locus. $\tilde{p}_{w,k,l}$ is the weighted average frequency of the allele A_k at the l^{th} locus.

2.6.2 Method of Moments Estimator

In this section we propose a new estimator of population-specific γ. The estimator is based on the MOM approach. For estimating γ_i we propose a new statistic

$$S_{3,k,i} = \frac{n_i^2}{(n_i-1)(n_i-2)}\tilde{p}_{i,k}(1-\tilde{p}_{i,k})(1-2\tilde{p}_{i,k}), \tag{2.53}$$

and the expectation of the statistic is

$$\mathrm{E}(S_{3,k,i}) = p_k(1-p_k)(1-2p_k)(1-3\theta_i+2\gamma_i). \tag{2.54}$$

Let us assume $\hat{\theta}_{WH,i}$ is the moment estimator of θ_i proposed by Weir and Hill (2002). Since this is a moment estimator, the bias of the estimator is close to zero. So we can assume $\mathrm{E}(\hat{\theta}_{WH,i}) \approx \theta_i$. Using ratio estimate we get

$$\begin{aligned}
& \mathrm{E}\Big[\frac{\sum_{k=1}^{s} S_{3,k,i}}{1-3\sum_{k=1}^{s}\hat{\pi}_{2,2,k}+2\sum_{k=1}^{s}\hat{\pi}_{3,3,k}} + 1.5\hat{\theta}_{WH,i} - 0.5\Big] \\
\approx\ & \frac{\mathrm{E}(\sum_{k=1}^{s} S_{3,k,i})}{2[1-3\mathrm{E}(\pi_{2,2})+2\mathrm{E}(\pi_{3,3})]} + \mathrm{E}(1.5\theta_{WH,i}) - 0.5 \\
\approx\ & \frac{(1-3\sum_{k=1}^{s}p_k^2+2\sum_{k=1}^{s}p_k^3)(1-3\theta_i+2\gamma_i)}{2(1-3\sum_{k=1}^{s}p_k^2+2\sum_{k=1}^{s}p_k^3)} + 1.5\theta_i - 0.5 \\
=\ & \gamma_i.
\end{aligned}$$

So a moment estimator of γ_i based on locus A is

$$\hat{\gamma}_{M,i} = \frac{\sum_{k=1}^{s} S_{3,k,i}}{1-3\sum_{k=1}^{s}\hat{\pi}_{2,2,k}+2\sum_{k=1}^{s}\hat{\pi}_{3,3,k}} + 1.5\hat{\theta}_{WH,i} - 0.5. \tag{2.55}$$

2.7 Bias and Variance of the Estimators

In this section we discuss the sampling properties of different estimators that are given in this chapter. We mainly concentrate on the bias and variance of the estimators. The expressions for bias and variance of the estimators are very complex. Theoretically, we

can find expressions for bias and variance of the estimators which are multi-allelic, and have unequal sample sizes; in practice they are intractable. So we restrict our research to a simple situation. We assume that there are L independent loci and each locus has two different allelic forms. So the estimators of the descent measures are based on one allelic frequency at each locus. Our expressions for bias and variance will be developed based on a locus A and then extended them for a multi-locus situation. The locus A has two alleles, A and a, with expected frequencies p_A and $1-p_A$ respectively. Without loss of generality we work with the frequency of the allele A, p_A. We also assume that each population has the same sample size, n. The descent measures have the same value in different populations.

Our estimators are based on the second and third moments of allele frequencies. The bias and variance of an estimator that is based on second order allele frequencies involve second, third, and fourth order descent measures. If the estimator is based on third order allele frequencies then the bias and variance involve second, third, fourth, fifth, and sixth order descent measures. In the literature we have descriptions about second, third, and fourth order descent measures. Here we define fifth and sixth order decent measures. There are two different types of fifth order descent measure and four different sixth order descent measures. We parameterize all these descent measures as:

$$\begin{aligned}
\eta &= \mathrm{E}_{sub-pop}[Pr(\text{Five random alleles are ibd})], \\
\Delta_{3,2} &= \mathrm{E}_{sub-pop}[Pr(\text{Three and two random allele are ibd})], \\
\tau &= \mathrm{E}_{sub-pop}[Pr(\text{Six random alleles are ibd})], \\
\Delta_{4,2} &= \mathrm{E}_{sub-pop}[Pr(\text{Four and two random alleles are ibd})], \\
\Delta_{3,3} &= \mathrm{E}_{sub-pop}[Pr(\text{Two sets of three random alleles are ibd})], \text{ and} \\
\Delta_{2,2,2} &= \mathrm{E}_{sub-pop}[Pr(\text{Three random pairs of alleles are ibd})].
\end{aligned} \tag{2.56}$$

Let us denote $P^*(\text{an event}) = \mathrm{E}_{sub-pop}[Pr(\text{an event})]$. Now suppose we have five/six alleles of the type A from a population. Then after some tedious algebra which is

skipped here we get the relations

$$\begin{aligned}
P^*(A,A,A,A,A) &= p_A^5 + p_A(1-p_A)(1-2p_k)(1-12p_A+12p_A^2)\eta \\
&\quad +10p_A^2(1-p_A)^2(1-2p_A)\Delta_{3,2} + 15p_A^3(1-p_A)^2\Delta_{2,2} \\
&\quad +5p_A^2(1-p_A)(1-6p_A+6p_A^2)\delta \qquad (2.57) \\
&\quad +10p_A^3(1-p_A)(1-2p_A)\gamma + 10p_A^4(1-10p_A)\theta \text{ and} \\
P^*(A,A,A,A,A,A) &= p_A^6 + p_A(1-p_A)(1-30p_A+135p_A^2-210p_A^3+105p_A^4)\tau \\
&\quad +15p_A^2(1-p_A)^2(1-5p_A+5p_A^2)\Delta_{4,2} \\
&\quad +15p_A^3(1-p_A)^3\Delta_{2,2,2} + 10p_A^2(1-p_A)(1-2p_A)^2\Delta_{3,3} \\
&\quad +6p_k^2(1-p_A)(1-2p_A)(1-12p_A+12p_A^2)\eta \qquad (2.58) \\
&\quad +60p_A^3(1-p_A)^2(1-2p_A)\Delta_{3,2} + 45p_A^4(1-p_A)^2\Delta_{2,2} \\
&\quad +15p_A^3(1-p_A)(1-6p_A+6p_A^2)\delta \\
&\quad +20p_A^4(1-p_A)(1-2p_A)\gamma + 15p_A^5(1-p_A)\theta.
\end{aligned}$$

Now we define the data at locus A as follows:

$$x_{ij,A} = \begin{cases} 1 & \text{if the } j^{th} \text{ allele in } i^{th} \text{ population is } A \\ 0 & \text{otherwise} \end{cases}$$

The population-specific (i^{th} population) and overall frequency of the allele A are

$$\tilde{p}_{i,A} = \frac{1}{n}\sum_{j=1}^{n} x_{ij,k} \text{ and } \tilde{p}_A = \frac{1}{r}\sum_{i=1}^{r}\tilde{p}_{i,A}. \qquad (2.59)$$

The first six raw moments of the frequency of the allele A_k in a population are

$$\begin{aligned}
\mu_1' &= \mathrm{E}(\tilde{p}_{A,i}), \ \ \mu_2' = \mathrm{E}(\tilde{p}_{A,i}^2), \ \ \mu_3' = \mathrm{E}(\tilde{p}_{A,i}^3), \\
\mu_4' &= \mathrm{E}(\tilde{p}_{A,i}^4), \ \ \mu_5' = \mathrm{E}(\tilde{p}_{A,i}^5), \ \text{ and } \ \mu_6' = \mathrm{E}(\tilde{p}_{A,i}^6).
\end{aligned}$$

The first six central moments of the frequency of the allele A_k in a population are

$$\begin{aligned}
\mu_1 &= \mathrm{E}[(\tilde{p}_{A,i} - p_A)] = 0, \\
\mu_2 &= \mathrm{E}[(\tilde{p}_{A,i} - p_A)^2] = \mu_2' - p_A^2, \\
\mu_3 &= \mathrm{E}[(\tilde{p}_{A,i} - p_A)^3] = \mu_3' - 3p_A\mu_2' + 2p_A^3, \\
\mu_4 &= \mathrm{E}[(\tilde{p}_{A,i} - p_A)^4] = \mu_4' - 4p_A\mu_3' + 6p_A^2\mu_2' - 3p_A^4, \\
\mu_5 &= \mathrm{E}[(\tilde{p}_{A,i} - p_A)^5] = \mu_5' - 5p_A\mu_4' + 10p_A^2\mu_3' - 10p_A^3\mu_2' + 4p_A^5, \text{ and} \\
\mu_6 &= \mathrm{E}[(\tilde{p}_{A,i} - p_A)^6] = \mu_6' - 6p_A\mu_5' + 15p_A^2\mu_4' - 20p_A^3\mu_3' + 15p_A^4\mu_2' - 5p_A^6.
\end{aligned} \tag{2.60}$$

Using the results of Li (1996) and equations (2.57), (2.58) and (2.60) we get the expressions for first six raw moments of allele frequencies for a population as

$$\begin{aligned}
\mu_1' &= p_A \\
\mu_2' &= p_A^2 + p_A(1-p_A)\theta', \\
\mu_3' &= p_A^3 + 3p_A^2(1-p_A)\theta' + p_A(1-p_A)(1-2p_A)\gamma', \\
\mu_4' &= p_A^4 + 6p_A^3(1-p_A)\theta' + 4p_A^2(1-p_A)(1-2p_A)\gamma' + 3p_A^2(1-p_A)^2\Delta_{2,2}' \\
&\quad + p_A(1-p_A)(1-6p_A+6p_A^2)\delta', \\
\mu_5' &= \frac{1}{n^4}p_A + \frac{15}{n^3}(1-\frac{1}{n})\Big[p_A^2 + p_A(1-p_A)\theta\Big] + \frac{25}{n^2}(1-\frac{1}{n})(1-\frac{2}{n})\Big[p_A^3 \\
&\quad + 3p_A^2(1-p_A)\theta + p_A(1-p_A)(1-2p_A)\gamma\Big] + \frac{10}{n}(1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})\Big[p_A^4 \\
&\quad + 6p_A^3(1-p_A)\theta + 4p_A^2(1-p_A)(1-2p_A)\gamma + 3p_A^2(1-p_A)^2\Delta_{2,2} \\
&\quad + p_A(1-p_A)(1-6p_A+6p_A^2)\delta\Big] + (1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})(1-\frac{4}{n})\Big[p_A^5 \\
&\quad + p_A(1-p_A)(1-2p_A)(1-12p_A+12p_A^2)\eta + 10p_A^2(1-p_A)^2(1-2p_A)\Delta_{3,2} \\
&\quad + 5p_A^2(1-p_A)(1-6p_A+6p_A^2)\delta + 15p_A^3(1-p_A)^2\Delta_{2,2} \\
&\quad + 10p_A^3(1-p_A)(1-2p_A)\gamma + 10p_A^4(1-p_A)\theta\Big], \text{ and}
\end{aligned}$$

(For continuation see next page)

$$
\begin{aligned}
\mu_6' &= \frac{1}{n^5}p_A + \frac{31}{n^4}(1-\frac{1}{n})\Big[p_A^2 + p_A(1-p_A)\theta\Big] + \frac{90}{n^3}(1-\frac{1}{n})(1-\frac{2}{n})\Big[p_A^3 \\
&+3p_A^2(1-p_A)\theta + p_A(1-p_A)(1-2p_A)\gamma\Big] + \frac{65}{n^2}(1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})\Big[p_A^4 \\
&+6p_A^3(1-p_A)\theta + 4p_A^2(1-p_A)(1-2p_A)\gamma + 3p_A^2(1-p_A)^2\Delta_{2,2} \\
&+p_A(1-p_A)(1-6p_A+6p_A^2)\delta\Big] + \frac{15}{n}(1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})(1-\frac{4}{n})\Big[p_A^5 \\
&+p_A(1-p_A)(1-2p_A)(1-12p_A+12p_A^2)\eta + 10p_A^2(1-p_A)^2(1-2p_A)\Delta_{3,2} \\
&+5p_A^2(1-p_A)(1-6p_A+6p_A^2)\delta + 15p_A^3(1-p_A)^2\Delta_{2,2} \\
&+10p_A^3(1-p_A)(1-2p_A)\gamma + 10p_A^4(1-p_A)\theta\Big] \\
&+(1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})(1-\frac{4}{n})(1-\frac{5}{n})\Big[p_A^6 + 45p_A^4(1-p_A)^2\Delta_{2,2} \\
&+p_A(1-p_A)(1-30p_A+135p_A^2-210p_A^3+105p_A^4)\tau \qquad (2.61) \\
&+15p_A^2(1-p_A)^2(1-5p_A+5p_A^2)\Delta_{4,2} + 10p_k^2(1-p_k)(1-2p_A)^2\Delta_{3,3} \\
&+15p_A^3(1-p_A)^3\Delta_{2,2,2} + 6p_A^2(1-p_A)(1-2p_A)(1-12p_A+12p_A^2)\eta \\
&+60p_A^3(1-p_A)^2(1-2p_A)\Delta_{3,2} + 15p_A^3(1-p_A)(1-6p_A+6p_A^2)\delta \\
&+20p_A^4(1-p_A)(1-2p_A)\gamma + 15p_A^5(1-p_A)\theta\Big],
\end{aligned}
$$

where,

$$
\begin{aligned}
\theta' &= \frac{1}{n} + (1-\frac{1}{n})\theta, \\
\gamma' &= \frac{1}{n^2} + \frac{3}{n}(1-\frac{1}{n})\theta + (1-\frac{1}{n})(1-\frac{2}{n})\gamma, \\
\delta' &= \frac{1}{n^3} + \frac{7}{n^2}(1-\frac{1}{n})\theta + \frac{6}{n}(1-\frac{1}{n})(1-\frac{2}{n})\gamma + (1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})\delta, \quad \text{and} \\
\Delta_{2,2}' &= \frac{1}{n^2} + \frac{2}{n}(1-\frac{1}{n})(1+\frac{1}{n})\theta + \frac{4}{n}(1-\frac{1}{n})(1-\frac{2}{n})\gamma + (1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})\Delta_{2,2}.
\end{aligned}
$$

The first six central moments can be obtained from the equations (2.60) and (2.61). The expressions for the central moments become simple for large sample sizes. When

$n \to \infty$ the first six central moments are

$$\begin{aligned}
\mu_1 &= 0, \\
\mu_2 &= p_A(1-p_A)\theta, \\
\mu_3 &= p_A(1-p_A)(1-2p_A)\gamma, \\
\mu_4 &= 3p_A^2(1-p_A)^2\Delta_{2,2} + p_A(1-p_A)(1-6p_A+6p_A^2)\delta, \\
\mu_5 &= p_A(1-p_A)(1-2p_A)(1-12p_A+12p_A^2)\eta + 10p_A^2(1-p_A)^2(1-2p_A)\Delta_{3,2}, \text{ and} \\
\mu_6 &= p_A(1-p_A)(1-30p_A+135p_A^2-210p_A^3+105p_A^4)\tau + 15p_A^2(1-p_A)^2(1-5p_A \\
&\quad +5p_A^2)\Delta_{4,2} + 10p_A^2(1-p_A)(1-2p_A)^2\Delta_{3,3} + 15p_A^3(1-p_A)^3\Delta_{2,2,2}.
\end{aligned} \tag{2.62}$$

We use Taylor series expansion for finding the variance and covariances of second and third order polynomials of the allele frequencies. Suppose $f(x_1, x_2, \cdots, x_r)$ is a third order polynomial of $x_1, x_2, \cdots, x_r$. The Taylor series coefficients of f are,

$$\begin{aligned}
&g_i = \frac{\partial f(\mathbf{x}')}{\partial x_i'}|_{\mathbf{x}'=\mathbf{a}}, \quad g_{ii} = \frac{\partial^2 f(\mathbf{x}')}{\partial x_i'^2}|_{\mathbf{x}'=\mathbf{a}}, \quad g_{ii'} = \frac{\partial^2 f(\mathbf{x}')}{\partial x_i' \partial x_{i'}'}|_{\mathbf{x}'=\mathbf{a}}, \\
&g_{iii} = \frac{\partial^3 f(\mathbf{x}')}{\partial x_i'^3}|_{\mathbf{x}'=\mathbf{a}}, \quad g_{iii'} = \frac{\partial^3 f(\mathbf{x}')}{\partial x_i'^2 \partial x_{i'}'}|_{\mathbf{x}'=\mathbf{a}}, \text{ and } g_{ii'i''} = \frac{\partial^3 f(\mathbf{x}')}{\partial x_i' \partial x_{i'}' \partial x_{i''}'}|_{\mathbf{x}'=\mathbf{a}}.
\end{aligned} \tag{2.63}$$

The Taylor series expansion of $f(x_1, x_2, \cdots, x_r)$ is

$$\begin{aligned}
f(x_1, x_2, \cdots, x_r) &= f(a_1, a_2, \cdots, a_r) + \sum_{i=1}^{r} g_i(x_i - a_i) + \frac{1}{2}\sum_{i=1}^{r} g_{ii}(x_i - a_i)^2 \\
&\quad + \frac{1}{2}\sum_{i \neq i'=1}^{r}\sum g_{ii'}(x_i - a_i)(x_{i'} - a_{i'}) + \frac{1}{6}\sum_{i=1}^{r} g_{iii}(x_i - a_i)^3 \\
&\quad + \frac{1}{2}\sum_{i \neq i'=1}^{r}\sum g_{iii'}(x_i - a_i)^2(x_{i'} - a_{i'}) \\
&\quad + \frac{1}{6}\sum_{i \neq i' \neq i''=1}^{r}\sum\sum g_{ii'i''}(x_i - a_i)(x_{i'} - a_{i'})(x_{i''} - a_{i''}).
\end{aligned} \tag{2.64}$$

If x_i's are independent random variable with $\mathrm{E}(x_i) = a_i$, then after some tedious calculation we find the variance of $f(x_1, x_2, \cdots, x_r)$. The variance is

$$\begin{aligned}
\mathrm{Var}(f) &= (\frac{1}{36}\sum_{i=1}^{r} g_{iii}^2)\mu_6 + (\frac{1}{6}\sum_{i=1}^{r} g_{ii}g_{iii})\mu_5 + (\frac{1}{4}\sum_{i=1}^{r} g_{ii}^2 + \frac{1}{3}\sum_{i=1}^{r} g_i g_{iii})\mu_4 \\
&+(\sum_{i=1}^{r} g_i g_{ii})\mu_3 + (\sum_{i=1}^{r} g_i^2)\mu_2 + (\frac{1}{4}\sum_{i\neq i'=1}^{r}\sum^{r} g_{iii'}^2 + \frac{1}{6}\sum_{i\neq i'=1}^{r}\sum^{r} g_{iii'}g_{ii'i'})\mu_4\mu_2 \\
&+(\frac{1}{4}\sum_{i\neq i'=1}^{r}\sum^{r} g_{iii'}g_{ii'i'} - \frac{1}{36}\sum_{i=1}^{r} g_{iii}^2)\mu_3^2 + (\frac{1}{4}\sum_{i\neq i'\neq i''=1}^{r}\sum\sum g_{iii'}g_{i'i''i''} \\
&+\frac{1}{6}\sum_{i\neq i'\neq i''=1}^{r}\sum\sum g_{ii'i''}^2)\mu_2^3 + (\sum_{i\neq i'=1}^{r}\sum^{r} g_{ii'}g_{iii'} + \frac{1}{2}\sum_{i\neq i'=1}^{r}\sum^{r} g_{ii}g_{ii'i'} - \frac{1}{6}\sum_{i=1}^{r} g_{ii}g_{iii})\mu_3\mu_2 \\
&+(\frac{1}{2}\sum_{i\neq i'=1}^{r}\sum^{r} g_{ii'}^2 - \frac{1}{4}\sum_{i=1}^{r} g_{ii}^2 + \sum_{i\neq i'=1}^{r}\sum^{r} g_i g_{ii'i'})\mu_2^2.
\end{aligned} \tag{2.65}$$

Let us assume another function $f_1(x_1, x_2, \cdots, x_r)$ with the Taylor series coefficients g_i^1, g_{ii}^1, $g_{ii'}^1$, g_{iii}^1, $g_{iii'}^1$. If $x_1, x_2, \cdots, x_r$ are independent random variables with $\mathrm{E}(x_i) = a_i$ then the covariance between $f(x_1, x_2, \cdots, x_r)$ and $f_1(x_1, x_2, \cdots, x_r)$ is

$$\begin{aligned}
\mathrm{Covar}(f, f_1) &= (\frac{1}{36}\sum_{i=1}^{r} g_{iii}g_{iii}^1)\mu_6 + (\frac{1}{12}\sum_{i=1}^{r} g_{ii}g_{iii}^1 + \frac{1}{12}\sum_{i=1}^{r} g_{iii}g_{ii}^1)\mu_5 + (\frac{1}{4}\sum_{i=1}^{r} g_{ii}g_{ii}^1 \\
&+\frac{1}{6}\sum_{i=1}^{r} g_i g_{iii}^1 + \frac{1}{6}\sum_{i=1}^{r} g_{iii}g_i^1)\mu_4 + (\frac{1}{2}\sum_{i=1}^{r} g_i g_{ii}^1 + \frac{1}{2}\sum_{i=1}^{r} g_{ii}g_i^1)\mu_3 + (\sum_{i=1}^{r} g_i g_i^1)\mu_2 \\
&+(\frac{1}{4}\sum_{i\neq i'}^{r} g_{iii'}g_{iii'}^1 + \frac{1}{12}\sum_{i\neq i'=1}^{r} g_{iii}g_{ii'i'}^1 + \frac{1}{12}\sum_{i\neq i'=1}^{r} g_{ii'i'}g_{iii}^1)\mu_4\mu_2 + (\frac{1}{4}\sum_{i\neq i'}^{r} g_{iii'}g_{ii'i'}^1 \\
&-\frac{1}{36}\sum_{i=1}^{r} g_{iii}g_{iii}^1)\mu_3^2 + (\frac{1}{6}\sum_{i\neq i'\neq i''}^{r} g_{ii'i''}g_{ii'i''}^1 + \frac{1}{4}\sum_{i\neq i'\neq i''}^{r} g_{iii'}g_{i'ii''}^1)\mu_2^3 + (\frac{1}{2}\sum_{i\neq i'=1}^{r} g_{ii'}g_{iii'}^1 \\
&+\frac{1}{2}\sum_{i\neq i'}^{r} g_{iii'}g_{ii'}^1 + \frac{1}{4}\sum_{i\neq i'}^{r} g_{ii}g_{ii'i'}^1 + \frac{1}{4}\sum_{i\neq i'}^{r} g_{ii'i'}g_{ii}^1 - \frac{1}{12}\sum_{i=1}^{r} g_{ii}g_{iii}^1 \\
&-\frac{1}{12}\sum_{i=1}^{r} g_{iii}g_{ii}^1)\mu_3\mu_2 + (\frac{1}{2}\sum_{i\neq i'}^{r} g_{ii'}g_{ii'}^1 - \frac{1}{4}\sum_{i=1}^{r} g_{ii}g_{ii}^1 + \frac{1}{2}\sum_{i\neq i'}^{r} g_i g_{ii'i'}^1 + \frac{1}{2}\sum_{i\neq i'}^{r} g_{ii'i'}g_i^1)\mu_2^2.
\end{aligned} \tag{2.66}$$

For ratio estimators again we use Taylor series expansion to find bias and variance of the estimators. Suppose $\hat{X}$ and $\hat{Y}$ are two statistics with mean μ_X and μ_Y respectively. If the ratio estimator of some parameter ζ is

$$\hat{\zeta} = \frac{\hat{X}}{\hat{Y}}, \tag{2.67}$$

where $\zeta = \frac{\mu_X}{\mu_Y}$, then the bias and variance of the estimator are

$$\text{Bias}(\hat{\zeta}) = -\frac{1}{\mu_Y^2}[\text{Covar}(\hat{X}, \hat{Y}) - \zeta \text{Var}(\hat{Y})] \text{ and} \tag{2.68}$$

$$\text{Var}(\hat{\zeta}) = \frac{1}{\mu_Y^2}[\text{Var}(\hat{X}) - 2\zeta \text{Covar}(\hat{X}, \hat{Y}) + \zeta^2 \text{Var}(\hat{Y})]. \tag{2.69}$$

Now we use equations (2.68) and (2.69) and find expressions for bias and variance of different estimators. In the following sections we find bias and variance of different estimators.

2.7.1 Weir-Cockerham's Estmator

Here we find the bias and variance of Weir-Cockerham's moment estimator of θ. Under our set up, Weir-Cockerham's moment estimator of θ based on locus A is

$$\hat{\theta}_{WC,A} = \frac{MSP_A - MSG_A}{MSP_A + (n-1)MSG_A}, \tag{2.70}$$

where,

$$MSP_A = \frac{n}{r-1}\sum_{i=1}^{r}(\tilde{p}_{i,A} - \tilde{p}_A)^2 \text{ and } MSG_A = \frac{n}{r(n-1)}\sum_{i=1}^{r}\tilde{p}_{i,A}(1-\tilde{p}_{i,A}).$$

Using the equations (2.65) and (2.66) we get

$$\begin{aligned}
&\mathrm{Var}(MSP_A) = \frac{n^2}{r}\Big[\mu_4 - \frac{r-3}{r-1}\mu_2^2\Big],\\
&\mathrm{Var}(MSG_A) = \frac{n^2}{(n-1)^2 r}\Big[\mu_4 - 2(1-2p_A)\mu_3 + (1-2p_A)^2\mu_2 - \mu_2^2\Big], \text{ and}\\
&\mathrm{Covar}(MSP_A, MSG_A) = \frac{n^2}{(n-1)r}\Big[-\mu_4 + (1-2p_A)\mu_3 + \mu_2^2\Big]. \qquad (2.71)
\end{aligned}$$

After some calculations using the equations (2.68) and (2.69), we find the bias and variance of the estimator $\hat{\theta}_{WC,A}$ as

$$\begin{aligned}
\mathrm{Bias}(\hat{\theta}_{WC,A}) &= \frac{1}{p_A^2(1-p_A)^2}\Big[\{2(n-1)\theta - (n-2)\}\mathrm{Covar}(MSP_A, MSG_A)\\
&\quad +(\theta-1)\mathrm{Var}(MSP_A) + \{(n-1)^2\theta + (n-1)\}\mathrm{Var}(MSG_A)\Big] \text{ and}\\
\mathrm{Var}(\hat{\theta}_{WC,A}) &= \frac{1}{p_A^2(1-p_A)^2}\Big[(1-\theta)^2\mathrm{Var}(MSP_A) + (1-\theta+n\theta)^2\mathrm{Var}(MSG_A)\\
&\quad \{-2-2(n-2)\theta + 2(n-1)\theta^2\}\mathrm{Covar}(MSP_A, MSG_A)\Big], \qquad (2.72)
\end{aligned}$$

where $\mathrm{Var}(MSP_A)$, $\mathrm{Var}(MSG_A)$ and $\mathrm{Covar}(MSP_A, MSG_A)$ are defined in the equation (2.71) and the central moments can be obtained from the equations (2.60) and (2.61). The bias and variance of the estimator can be expressed in terms of descent measures and expected frequency of the allele A, but the expressions will be very complicated. We keep the expressions as above. If we assume that the sample sizes are large then we get simpler expressions for bias and variance of $\hat{\theta}_{WC,A}$. Assuming $n \to \infty$ we get

$$\begin{aligned}
\mathrm{Bias}(\hat{\theta}_{WC,A}) &= -\frac{2}{r(r-1)}\theta^2(1-\theta) + \frac{(1-2p_A)^2}{rp_A(1-p_A)}(\theta^2-\gamma) \text{ and} \qquad (2.73)\\
\mathrm{Var}(\hat{\theta}_{WC,A}) &= \frac{1}{r}\Big[3(1-\theta)^2\Delta - \theta^2\{\frac{(r-3)}{r-1}(1-\theta)^2 + \theta(2-\theta)\}\Big]\\
&\quad +\frac{1}{rp_A(1-p_A)}\Big[(-6p_A+6p_A^2)\delta + (1-2p_A)^2(\theta^3 - 2\theta\gamma)\Big]. \qquad (2.74)
\end{aligned}$$

When sample sizes are large Weir and Hill (2002) assumed a normal distribution for the allele frequencies. Under this model, the higher order descent measures are functions of θ which makes the expressions for bias and variance more simple. Assuming a normal distribution we get

$$\begin{aligned} \text{Bias}(\hat{\theta}_{WC,A}) &= -\frac{2}{r(r-1)}\theta^2(1-\theta) + \frac{(1-2p_A)^2}{rp_A(1-p_A)}\theta^2, \text{ and} \quad (2.75) \\ \text{Var}(\hat{\theta}_{WC,A}) &= \frac{2\theta^2(1-\theta)^2}{r-1} - \frac{\theta^3(2-\theta)}{r} + \frac{(1-2p_A)^2\theta^3}{rp_A(1-p_A)}. \quad (2.76) \end{aligned}$$

Note that these expressions are different from the expressions given by Li (1996). She used the theorem given in section 2.2.4 of Serfling (1980) and found $\text{Var}(MSG_A)$ and $\text{Covar}(MSP_A, MSG_A)$ is 0. The theorem given in Serfling (1980) assumes that the indicator variables are independent. In our case, when $\theta > 0$ the indicator variables are not independent, so we can not use Serfling's theorem. We assumed a normal distribution for allele frequencies and direct calculations produce

$$\begin{aligned} \text{Var}(MSG_A) &= \frac{1}{r}\Big[2p_A^2(1-p_A)^2\theta^2 + p_A(1-p_A)(1-2p_A)^2\theta\Big] \text{ and} \quad (2.77) \\ \text{Covar}(MSP_A, MSG_A) &= -\frac{2p_A^2(1-p_A)^2\theta^2}{r}. \quad (2.78) \end{aligned}$$

The above two expressions are zero only when $\theta = 0$. The expression for bias in the equation (2.73) shows that the bias is always negative. But the bias becomes positive when we assume a normal distribution for the allele frequencies. We think the normal distribution assumes $\gamma = 0$ which is not quite correct.

2.7.2 New Moment Estimator of θ

Under our simple set up, new moment estimator of θ based on locus A is

$$\hat{\theta}_{M,A} = \frac{S_{1,A} + (n-3)S_{2,A} - (n-2)S_{3,A}}{S_{1,A} + 3(n-1)S_{2,A} + (n-1)(n-2)S_{3,A}}, \quad (2.79)$$

where,

$$
\begin{aligned}
S_{1,A} &= \frac{n^2 r}{(r-1)(r-2)} \sum_{i=1}^{r} (\tilde{p}_{i,A} - \tilde{p}_A)^3, \\
S_{2,A} &= \frac{n^2}{(r-1)(n-1)} \sum_{i=1}^{r} \tilde{p}_{i,A}(1-\tilde{p}_{i,A})(\tilde{p}_{i,A} - \tilde{p}_A), \quad \text{and} \\
S_{3,k} &= \frac{n^2}{r(n-1)(n-2)} \sum_{i=1}^{r} \tilde{p}_{i,A}(1-\tilde{p}_{i,A})(1-2\tilde{p}_{i,A}).
\end{aligned} \tag{2.80}
$$

Using equations (2.65) and (2.66) we calculate the following variances and covariances

$$
\begin{aligned}
\mathrm{Var}(S_{1,A}) &= \frac{n^4}{r}\big[\mu_6 - \frac{3(2r-5)}{(r-1)}\mu_4\mu_2 - \frac{(r-10)}{(r-1)}\mu_3^2 + \frac{3(3r^2-12r+20)}{(r-1)(r-2)}\mu_2^3\big], \\
\mathrm{Var}(S_{2,A}) &= \frac{n^4}{r(n-1)^2}\big[\mu_6 - 2(1-2p_A)\mu_5 + (1-2p_A)^2\mu_4 - \frac{(2r-3)}{(r-1)}\mu_4\mu_2 \\
&\quad -\frac{(r-2)}{(r-1)}\mu_3^2 + \frac{(r-2)}{(r-1)}\mu_2^3 + \frac{4(r-2)}{(r-1)}(1-2p_A)\mu_3\mu_2 \\
&\quad -(r-3)(1-2p_A)^2\mu_2^2/(r-1)\big], \\
\mathrm{Var}(S_{3,A}) &= \frac{n^4}{r(n-1)^2(n-2)^2}\big[4\mu_6 - 12(1-2p_A)\mu_5 + \{9(1-2p_A)^2 - 4\mu_3^2 \\
&\quad +4(1-6p_A+6p_A^2)\}\mu_4 - 6(1-6p_A+6p_A^2)(1-2p_A)\mu_3 \\
&\quad +(1-6p_A+6p_A^2)^2\mu_2 - 12(1-2p_A)\mu_3\mu_2 - 9(1-2p_A)^2\mu_2^2\big] \\
\mathrm{Covar}(S_{1,A},S_{2,A}) &= \frac{n^4}{(n-1)r}\big[-\mu_6 + (1-2p_A)\mu_5 + \frac{(4r-7)}{(r-1)}\mu_4\mu_2 + \frac{(r-4)}{(r-1)}\mu_3^2 \\
&\quad -\frac{3(r-2)}{(r-1)}\mu_2^3 - \frac{2(2r-5)}{(r-1)}(1-2p_A)\mu_3\mu_2\big], \\
\mathrm{Covar}(S_{1,A},S_{3,A}) &= \frac{n^4}{(n-1)(n-2)r}\big[2\mu_6 - 3(1-2p_A)\mu_5 + (1-6p_A+6p_A^2)\mu_4 \\
&\quad -6\mu_4\mu_2 - 2\mu_3^2 + 12(1-2p_A)\mu_3\mu_2 - 3(1-6p_A+6p_A^2)\mu_2^2\big], \quad \text{and} \\
\mathrm{Covar}(S_{2,A},S_{3,A}) &= \frac{n^4}{(n-1)^2(n-2)r}\big[-2\mu_6 + 5(1-2p_A)\mu_5 - (4-18p_A+18p_A^2)\mu_4 \\
&\quad +(1-2p_A)(1-6p_A+6p_A^2)\mu_3 + 2\mu_4\mu_2 + 2\mu_3^2 - 8(1-2p_A)\mu_3\mu_2 \\
&\quad +(4-18p_A+18p_A^2)\mu_2^2\big].
\end{aligned} \tag{2.81}
$$

After some calculations using the equations (2.68) and (2.69), we find the bias and variance of the estimator $\hat{\theta}_{M,A}$. The bias and variance are

$$\begin{aligned}
\text{Bias}(\hat{\theta}_{M,A}) &= \frac{1}{p_A^2(1-p_A)^2(1-2p_A)^2}\Big[(\theta-1)\text{Var}(S_{1,A}) \\
&\quad +\{9(n-1)^2\theta - 3(n-1)(n-3)\}\text{Var}(S_{2,A}) \\
&\quad +\{(n-1)^2(n-2)^2\theta + (n-1)(n-2)^2\}\text{Var}(S_{3,A}) \\
&\quad +\{6(n-1)\theta - 2(2n-3)\}\text{Covar}(S_{1,A}, S_{2,A}) \\
&\quad +\{2(n-1)(n-2)\theta - (n-2)^2\}\text{Covar}(S_{1,A}, S_{3,A}) \\
&\quad +\{6(n-1)^2(n-2)\theta - (n-1)(n-2)(n-6)\}\text{Covar}(S_{2,A}, S_{3,A})\Big] \quad \text{and}
\end{aligned} \tag{2.82}$$

$$\begin{aligned}
\text{Var}(\hat{\theta}_{M,A}) &= \frac{1}{p_A^2(1-p_A)^2(1-2p_A)^2}\Big[(\theta-1)^2\text{Var}(S_{1,A}) \\
&\quad +(3n\theta - 3\theta - n + 3)^2\text{Var}(S_{2,A}) \\
&\quad +(n-2)^2(n\theta - \theta + 1)\text{Var}(S_{3,A}) + \{6(n-1)^2(n-2)\theta^2 \\
&\quad -2(n-1)(n-2)(n-6)\theta - 2(n-2)(n-3)\}\text{Covar}(S_{2,A}, S_{3,A}) \\
&\quad +\{6(n-1)\theta^2 - 4(2n-3)\theta + 2(n-3)\}\text{Covar}(S_{1,A}, S_{2,A}) \\
&\quad +\{2(n-1)(n-2)\theta^2 - 2(n-2)^2\theta - 2(n-2)\}\text{Covar}(S_{1,A}, S_{3,A})\Big],
\end{aligned} \tag{2.83}$$

where $\text{Var}(S_{1,A}), \text{Var}(S_{2,A}), \text{Var}(S_{3,A}), \text{Covar}(S_{1,A}, S_{2,A}), \text{Covar}(S_{1,A}, S_{3,A})$ and $\text{Covar}(S_{2,A}, S_{3,A})$ are defined in the equations (2.81) and the central moments can be obtained from the equations (2.60) and (2.61). The estimator can be expressed in terms of descent measures and expected frequency of the allele A but the expressions will be very complicated. So we keep the expressions as above. If we assume that the sample sizes are large then we get simpler expressions for bias and variance of $\theta_{M,A}$.

Assuming $n \to \infty$ and after some algebra we get

$$\begin{aligned}
\text{Bias}(\hat{\theta}_{M,A}) &= \frac{1}{p_A^2(1-p_A)^2(1-2p_A)^2}\Big[(\theta-1)\text{Var}(S^*_{1,A}) + (9\theta-3)\text{Var}(S^*_{2,A}) \\
&\quad +\theta\text{Var}(S^*_{3,A}) + (6\theta-4)\text{Covar}(S^*_{1,A}, S^*_{2,A}) + (2\theta-1)\text{Covar}(S^*_{1,A}, S^*_{3,A}) \\
&\quad +(6\theta-1)\text{Covar}(S^*_{2,A}, S^*_{3,A})\Big] \quad \text{and} \qquad (2.84) \\
\text{Var}(\hat{\theta}_{M,A}) &= \frac{1}{p_A^2(1-p_A)^2(1-2p_A)^2}\Big[(\theta-1)^2\text{Var}(S^*_{1,A}) + (3\theta-1)^2\text{Var}(S^*_{2,A}) \\
&\quad +\theta^2\text{Var}(S^*_{3,A}) + (6\theta^2-8\theta+2)\text{Covar}(S^*_{1,A}, S^*_{2,A}) \\
&\quad +(2\theta^2-2\theta)\text{Covar}(S^*_{1,A}, S^*_{3,A}) + (6\theta^2-2\theta)\text{Covar}(S^*_{2,A}, S^*_{3,A})\Big], \qquad (2.85)
\end{aligned}$$

where,

$$\begin{aligned}
\text{Var}(S^*_{1,A}) &= \frac{1}{r}[\mu_6 - \frac{3(2r-5)}{(r-1)}\mu_4\mu_2 - \frac{(r-10)}{(r-1)}\mu_3^2 + \frac{3(3r^2-12r+20)}{(r-1)(r-2)}\mu_2^{3}], \\
\text{Var}(S^*_{2,A}) &= \frac{1}{r}[\mu_6 - 2(1-2p_A)\mu_5 + (1-2p_A)^2\mu_4 - \frac{(2r-3)}{(r-1)}\mu_4\mu_2 \\
&\quad +\frac{(r-2)}{(r-1)}(\mu_2^3 - \mu_3^2 + 4(1-2p_A)\mu_3\mu_2) - \frac{(r-3)}{(r-1)}(1-2p_A)^2\mu_2^2], \\
\text{Var}(S^*_{3,A}) &= \frac{1}{r}[4\mu_6 - 12(1-2p_A)\mu_5 + \{9(1-2p_A)^2 + 4(1-6p_A+6p_A^2)\}\mu_4 \\
&\quad -6(1-6p_A+6p_A^2)(1-2p_k)\mu_3 + (1-6p_A+6p_A^2)^2\mu_2 - 4\mu_3^2 \\
&\quad -12(1-2p_A)\mu_3\mu_2 - 9(1-2p_k)^2\mu_2^2], \\
\text{Covar}(S^*_{1,A}, S^*_{2,A}) &= \frac{1}{r^2}[-\mu_6 + (1-2p_A)\mu_5 + \frac{(4r-7)}{(r-1)}\mu_4\mu_2 + \frac{(r-4)}{(r-1)}\mu_3^2 \\
&\quad -\frac{3(r-2)}{(r-1)}\mu_2^3 - \frac{2(2r-5)}{(r-1)}(1-2p_A)\mu_3\mu_2], \\
\text{Covar}(S^*_{1,A}, S^*_{3,A}) &= \frac{1}{r}[2\mu_6 - 3(1-2p_A)\mu_5 + (1-6p_A+6p_A^2)\mu_4 - 6\mu_4\mu_2 - 2\mu_3^2 \\
&\quad +12(1-2p_A)\mu_3\mu_2 - 3(1-6p_A+6p_A^2)\mu_2^2], \quad \text{and} \qquad (2.86) \\
\text{Covar}(S^*_{2,A}, S^*_{3,A}) &= \frac{1}{r}[-2\mu_6 + 5(1-2p_A)\mu_5 - (4-18p_A+18p_A^2)\mu_4 \\
&\quad +(1-2p_A)(1-6p_A+6p_A^2)\mu_3 + 2\mu_4\mu_2 + 2\mu_3^2 - 8(1-2p_A)\mu_3\mu_2. \\
&\quad +(4-18p_A+18p_A^2)\mu_2^2].
\end{aligned}$$

The first six central moments are defined in equation (2.62). When sample sizes are large Weir and Hill (2002) assumed a normal distribution for the allele frequencies. Under this model higher order descent measures are functions of θ which makes the expressions for bias and variance more simple. Using standard normal theory we have

$$\mu_2 = p_A(1-p_A)\theta,\ \mu_4 = 3p_A^2(1-p_A)^2\theta^2,\ \mu_6 = 15p_A^3(1-p_A)^3\theta^3, \text{ and } \mu_1 = \mu_3 = \mu_5 = 0,$$

which gives $\gamma = \delta = \eta = \tau = \Delta_{3,2} = \Delta_{4,2} = \Delta_{3,3} = 0; \quad \Delta = \theta^2; \quad \Delta_{2,2,2} = \theta^3$. Under the normality assumption the bias and variance of $\hat{\theta}_{M,A}$ will remain same as the expression given in (2.84) and (2.85). But the expressions of $\text{Var}(S^*_{1,A})$, $\text{Var}(S^*_{2,A})$, $\text{Var}(S^*_{3,A})$, $\text{Covar}(S^*_{1,A}, S^*_{2,A})$, $\text{Covar}(S^*_{1,A}, S^*_{3,A})$ and $\text{Covar}(S^*_{2,A}, S^*_{3,A})$ will be the functions of θ and p_A only. The relations are

$$\begin{aligned}
\text{Var}(S^*_{1,A}) &= \frac{6r}{(r-1)(r-2)}p_A^3(1-p_A)^3\theta^3, \\
\text{Var}(S^*_{2,A}) &= \frac{(10r-8)}{r(r-1)}p_A^3(1-p_A)^3\theta^3 + \frac{2}{(r-1)}p_A^2(1-p_A)^2(1-2p_A)^2\theta^2, \\
\text{Var}(S^*_{3,A}) &= \frac{1}{r}[60p_A^3(1-p_A)^3\theta^3 + (30-144p_A+144p_A^2)p_A^2(1-p_A)^2\theta^2 \\
&\quad +(1-6p_A+6p_A^2)^2p_A(1-p_A)\theta], \qquad (2.87) \\
\text{Covar}(S^*_{1,A}, S^*_{2,A}) &= \frac{-6p_A^3(1-p_A)^3\theta^3}{(r-1)}, \\
\text{Covar}(S^*_{1,A}, S^*_{3,A}) &= \frac{12p_A^3(1-p_A)^3\theta^3}{r}, \text{ and} \\
\text{Covar}(S^*_{2,A}, S^*_{3,A}) &= \frac{-24p_A^3(1-p_A)^3\theta^3 - 4(2-9p_A+9p_A^2)p_A^2(1-p_A)^2\theta^2}{r}.
\end{aligned}$$

2.7.3 New Moment Estimator of γ

Under our simple set up, new moment estimators of γ based on locus A is

$$\hat{\gamma}_{M,A} = \frac{S_{1,A} - 3S_{2,A} + 2S_{3,A}}{S_{1,A} + 3(n-1)S_{2,A} + (n-1)(n-2)S_{3,A}}, \qquad (2.88)$$

where $S_{1,A}$, $S_{2,A}$ and $S_{3,A}$ are given in the equation (2.80). The bias and variance of $\hat{\gamma}_{M,A}$ will be very similar to the expressions of bias and variance of $\hat{\theta}_{M,A}$. After some calculations, we find the bias and variance of the estimator $\hat{\gamma}_{M,A}$ as follows

$$\begin{aligned}
\text{Bias}(\hat{\gamma}_{M,A}) &= \frac{1}{p_A^2(1-p_A)^2(1-2p_A)^2}\Big[(\theta-1)\text{Var}(S_{1,A}) + \{9(n-1)^2\theta \\
&\quad +9(n-1)\}\text{Var}(S_{2,A}) + \{(n-1)^2(n-2)^2\theta \\
&\quad -2(n-1)(n-2)\}\text{Var}(S_{3,A}) + \{6(n-1)\theta - 3(n-2)\}\text{Covar}(S_{1,A}, S_{2,A}) \\
&\quad +\{2(n-1)(n-2)\theta - (n^2-3n+4)\}\text{Covar}(S_{1,A}, S_{3,A}) \\
&\quad +\{6(n-1)^2(n-2)\theta + 3(n-1)(n-4)\}\text{Covar}(S_{2,A}, S_{3,A})\Big] \text{ and } \quad (2.89)
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\gamma}_{M,A}) &= \frac{1}{p_A^2(1-p_A)^2(1-2p_A)^2}\Big[(\theta-1)^2\text{Var}(S_{1,A}) + 9(n\theta-\theta^2+1)^2\text{Var}(S_{2,A}) \\
&\quad +\{(n-1)(n-2)\theta+2\}^2\text{Var}(S_{3,A}) \\
&\quad +\{6(n-1)\theta^2 - 6(n-2)\theta - 6\}\text{Covar}(S_{1,A}, S_{2,A}) \\
&\quad +\{2(n-1)(n-2)\theta^2 - 2(n^2-3n+4)\theta + 4\}\text{Covar}(S_{1,A}, S_{3,A}) \quad (2.90) \\
&\quad +\{6(n-1)^2(n-2)\theta^2 + 3(n-1)(n-4)\theta - 12\}\text{Var}(S_{2,A}, S_{3,A})\Big],
\end{aligned}$$

where $\text{Var}(S_{1,A})$, $\text{Var}(S_{2,A})$, $\text{Var}(S_{3,A})$, $\text{Covar}(S_{1,A}, S_{2,A})$, $\text{Covar}(S_{1,A}, S_{3,A})$ and $\text{Covar}(S_{2,A}, S_{3,A})$ are defined in the equation (2.81) and the central moments can be obtained from the equations (2.60) and (2.61). If we assume that the sample sizes are large then we get simpler expressions for bias and variance of $\gamma_{M,A}$. Assuming $n \to \infty$ and after some algebra we get

$$\begin{aligned}
\text{Bias}(\hat{\gamma}_{M,A}) &= \frac{1}{p_A^2(1-p_A)^2(1-2p_A)^2}\Big[(\theta-1)\text{Var}(S^*_{1,A}) + 9\theta\text{Var}(S^*_{2,A}) \\
&\quad +\theta\text{Var}(S^*_{3,A}) + 6\theta\text{Covar}(S^*_{1,A}, S^*_{2,A}) + (2\theta-1)\text{Covar}(S^*_{1,A}, S^*_{3,A}) \\
&\quad +6\theta\text{Covar}(S^*_{2,A}, S^*_{3,A})\Big] \text{ and } \quad (2.91)
\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\hat{\gamma}_{M,A}) &= \frac{1}{p_A^2(1-p_A)^2(1-2p_A)^2}\Big[(\theta-1)^2\mathrm{Var}(S^*_{1,A}) + 9\theta^2\mathrm{Var}(S^*_{2,A}) \\ &\quad +\theta^2\mathrm{Var}(S^*_{3,A}) - 6\theta(1-\theta)\mathrm{Covar}(S^*_{1,A}, S^*_{2,A}) \\ &\quad -2\theta(1-\theta)\mathrm{Covar}(S^*_{1,A}, S^*_{3,A}) + 6\theta^2\mathrm{Covar}(S^*_{2,A}, S^*_{3,A})\Big],\end{aligned} \tag{2.92}$$

where $\mathrm{Var}(S^*_{1,A})$, $\mathrm{Var}(S^*_{2,A})$, $\mathrm{Var}(S^*_{3,A})$, $\mathrm{Covar}(S^*_{1,A}, S_{2,A})$, $\mathrm{Covar}(S^*_{1,A}, S^*_{3,A})$ and $\mathrm{Covar}(S^*_{2,A}, S^*_{3,A})$ are defined in the equation (2.86) and the first six central moments are defined in the equation (2.62). When sample sizes are large we assume a normal distribution for the allele frequencies. Under normality the bias and variance of $\hat{\gamma}_{M,A}$ will remain same as the expression given in the equations (2.91) and (2.92). But the expressions of $\mathrm{Var}(S^*_{1,A})$, $\mathrm{Var}(S^*_{2,A})$, $\mathrm{Var}(S^*_{3,A})$, $\mathrm{Covar}(S^*_{1,A}, S^*_{2,A})$, $\mathrm{Covar}(S^*_{1,A}, S^*_{3,A})$ and $\mathrm{Covar}(S^*_{2,A}, S^*_{3,A})$ are given in the equation (2.87).

If we use Robertson-Hill's method to get the final estimator for more than one independent loci then the bias (variance) of the final estimate is the average of the biases (variances) for locus specific estimates. On the other hand for Weir-Cockerham's method we need to find the variances of the numerator and denominator of the final estimate separately. Then we can find the bias and variance of the final estimator using the Taylor series. Finding variances of numerator and denominator are easy as the loci are independent.

Using the same approach we can find the biases and variances for the estimators obtained by the probabilistic approach. Here we skip the calculation and the algebraic expressions. We can not estimate the bias and variances of these estimators. This is because these expressions contain fourth, fifth and sixth order descent measures and we can not estimate them. But we can estimate the empirical bias and variance of the estimators using the MCMC method.

Chapter 3

Testing Hypotheses about θ

3.1 Introduction

In the previous chapters we have seen the role of descent measures in the study of population structure. We have also seen several methods for estimating the descent measures. Among all the descent measures, the coancestry coefficient (θ) is the parameter which is the most widely used among population geneticists. Almost all the important quantities about a population involve θ in their expressions. When the value of θ is 0, then the expressions for the important population parameters become very simple and easy to interpret. Thus, population geneticists are interested in checking if they may assume the value of θ is zero. The population geneticists formed this problem formally in terms of testing hypotheses about θ and proposed several different methods for testing the hypothesis. The hypothesis is

$$H_0 : \theta = 0 \qquad \text{vs.} \qquad H_1 : \theta > 0 \,.$$

Several testing procedures have been proposed under different assumptions. Some geneticists proposed a testing procedure for θ under a fixed population set up (Roff and Bentzen, 1989; Raymond and Rousset, 1995), while others used a random population model (Li, 1996; Dodds, 1986). The χ^2 test based on an allelic contingency table has been used for testing population differentiation. Raymond and Rousset (1995), Roff

and Bentzen (1989) proposed an exact chi-square test based on the permutation procedure for testing population differentiation. The above two procedures are tests for a fixed population setup. Dodds (1986) proposed a non-parametric bootstrap resampling procedure for testing if θ is equal to zero. He worked with small sample sizes. Li (1996) proposed a large sample test that is based on the Analysis of Variance. Her test procedure assumes two alleles at each locus in the population. The last two methods are proposed under a random population setup.

In this chapter we propose two methods for testing the hypothesis. The first approach is based on a parametric bootstrap method and works better for small sample sizes. On the other hand, the second approach is based on the large sample properties of allele frequencies. Both the approaches are defined for a random population and work for any number of alleles in a particular locus. In Chapter 6, we show that our test procedures are better than existing testing methods in terms of power.

3.2 Review on Testing Procedure

In this section we discuss the existing methods for testing population differentiation. Some testing methods apply to fixed populations while the others apply to random populations. We need to clarify the sampling processes involved in the population structure, and the difference between a fixed population and a random population. The population structure include two sampling processes, (i) statistical sampling and (ii) genetical sampling. Different samples from one population will show different levels of genetic variation. This is because of statistical sampling. If all the populations have descended from a common founder population and remain in random mating within populations, then the genetical sampling is also included. For the fixed population structure, only statistical sampling is involved while in the random population structure, both statistical and genetical sampling are involved. The fixed population model holds when we are interested in one particular population and ignore the prior evolutionary history. On the other hand if we want to know about the evolutionary

history of a population, then we need to work with a random population structure. In a random mating population we are interested in testing $\theta = 0$. This null hypothesis is different from the fixed population model. In a random population we focus on the properties of θ that measures the population differentiation. Weir (1996) concluded that the comparison between different fixed populations can be done by testing equal allele frequencies between populations.

Four different methods are often used for testing population differentiation. The conventional χ^2 test (Nei, 1987) based on the allelic contingency table was criticized for lack of power (Roff and Bentzen, 1989). The exact χ^2 test was proposed by Roff and Bentzen (1989). The significance level of this test procedure can be computed by permutation procedure. The above two methods apply to the fixed population model. Dodds (1986) proposed a distribution free approach for random population model. It was to construct the confidence interval for θ by using bootstrap procedures over loci. Later Li (1996) proposed a parametric test that is based on large sample approximation. She proposed a test statistic that has an asymptotic χ^2 distribution under the null hypothesis. She worked with two alleles per locus.

Goodness of fit χ^2 test

The traditional method of testing population differentiation is a χ^2 test. This test is based on an allelic contingency table under the null hypothesis with equal frequencies between populations (Workman and Niswander, 1970; Nei, 1973). Suppose there are two alleles at a locus, A and a. We assume there are r populations and the i^{th} population has n_i sampled alleles. We also assume that the i^{th} population has $n_{A,i}$ sampled alleles that are of the type A and $\tilde{p}_{A,i} = n_{A,i}/n_i$. The overall frequency of allele A is $\tilde{p}_A$ that satisfies $\tilde{p}_A = \sum_{i=1}^{r} \tilde{p}_{A,i}/r$. From Table 3.1, the goodness of fit test statistic is

$$X^2 \;=\; \sum_{i=1}^{r} \frac{n_i(\tilde{p}_{A,i} - \tilde{p}_A)^2}{\tilde{p}_A(1-\tilde{p}_A)}\,. \tag{3.1}$$

Under the null hypothesis, X^2 should be distributed approximately as a χ^2 random variable with $r-1$ degrees of freedom. However, the χ^2 contingency test is appropriate for large samples. If the expected values within cells are very small, X^2 in the equation (3.1) may be very large and we can not use the usual tabulated value of χ^2 to assess the significance of the observed values. The general solution is to group the rare alleles into one category (Roff and Bentzen, 1989). This approach potentially reduces the power of the test.

Table 3.1: The contingency table for χ^2 test when there are two alleles A and a

	Observed Frequencies		Expected Frequencies		
Population	A	a	A	a	Total
1	$n_1\tilde{p}_{A,1}$	$n_1(1-\tilde{p}_{A,1})$	$n_1\tilde{p}_A$	$n_1(1-\tilde{p}_A)$	n_1
2	$n_2\tilde{p}_{A,2}$	$n_2(1-\tilde{p}_{A,2})$	$n_1\tilde{p}_A$	$n_1(1-\tilde{p}_A)$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$n_i\tilde{p}_{A,i}$	$n_i(1-\tilde{p}_{A,i})$	$n_1\tilde{p}_A$	$n_1(1-\tilde{p}_A)$	n_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$n_r\tilde{p}_{A,r}$	$n_r(1-\tilde{p}_{A,r})$	$n_1\tilde{p}_A$	$n_1(1-\tilde{p}_A)$	n_r
Total			$S\tilde{p}_A$	$S(1-\tilde{p}_A)$	S

Permutation Test

An exact χ^2 test based on a permutation procedure for testing the population differentiation was discussed by Roff and Bentzen (1989) and Raymond and Rousset (1995). The permutation test is based on the fixed population model and is designed to test the equality of allele frequencies between populations. As in Table 3.1 the marginal numbers of the contingency table are fixed. By shuffling all the alleles randomly we can reconstruct contingency tables with the same marginal numbers. So we can generate

the exact probability distribution of X^2 under the null hypothesis. The exact value of the Type I error can be computed by summing up the probabilities of all the contingency tables which have the same or less probability than the observed table. If this exact probability is less than the significant level then we reject the test. Enumerating the entire contingency table with fixed marginal numbers is possible but computationally hard when the total number of tables is large. In this situation instead of enumerating all possible tables, a set of contingency table can be generated by a random permutation. The proportion of permutated tables that has a probability equal or less than the probability of the original table forms an estimate of the significance level.

Bootstrap Resampling

The bootstrap resample is used to test the population differentiation for the random population level. Dodds (1986) proposed this test procedure. He assumed that there is more than one locus (L) in the data set. He constructed $100(1-\alpha)\%$ confidence interval of θ from bootstrap resampling the original data. If the confidence interval does not contain 0, then the hypothesis $\theta = 0$ will be rejected at the α significance level. Bootstrapping can be done either over loci or over populations. The resampling over populations is not recommended as it may break the population structure. Resampling over independent loci address the genetical sampling from the fonder population. Dodds proposed to resample over loci. Each time, a particular locus is sampled with replacement, then the data from all populations at that locus are included in. We repeat the above method L times to get a particular bootstrap sample. In this way get bootstrap samples for B times and estimate θ each time to get a bootstrap confidence interval of θ. In general, when the distribution of estimates of a parameter under the null is not known, then the numerical resampling becomes a powerful tool for inference about the estimates. The bootstrap resampling needs lot of computing power.

A Test from ANOVA

The bootstrap resampling method is computationally intensive and time consuming. Li (1996) proposed a testing method using the Analysis of Variance. She proposed a test statistic with a null distribution using the large sample approximation. In this section we discuss her method. She worked with r independent populations that are descended from a common ancestral population. She assumed that the expected allele frequencies in each population are the same. She also assumed that there are 2 alleles, A and a, with expected frequency p_A and $1 - p_A$ in each population. The frequency data for alleles are available and each population has n sampled alleles, where n is very large. The data can be described as

$$x_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ allele in } i^{th} \text{ population is } A \\ 0 & \text{otherwise.} \end{cases}$$

The frequency of the allele A in the i^{th} population is $\tilde{p}_{A,i}$. This can be found from the data as $\tilde{p}_{A,i} = \sum_{j=1}^{n} x_{ij}/n$. The overall frequency of the allele A is $\tilde{p}_A = \frac{1}{r}\sum_{i=1}^{r} \tilde{p}_{A,i}$ The basic model assumes the absence of any disturbing force. Li (1996) defined the following two statistics in her work:

$$MSP = \frac{1}{r-1}\sum_{i=1}^{r} n(\tilde{p}_{A,i} - \tilde{p}_A)^2, \quad \text{and} \tag{3.2}$$

$$MSG = \frac{1}{r(n-1)}\sum_{i=1}^{r} n\tilde{p}_{A,i}(1 - \tilde{p}_{A,i}). \tag{3.3}$$

She found a test statistic using MSP and MSG and then found the asymptotic distribution of that statistic under null hypothesis. The starting point was to find the asymptotic distribution of MSP and MSG under the null hypothesis. When $\theta = 0$, the alleles in a population are independent which means the random variables x_{ij} are

also independent. When $\theta = 0$, the mean and variance of x_{ij} are

$$\mathrm{E}(x_{ij}) \;=\; p_A \quad \text{and} \quad \mathrm{Var}(x_{ij}) = p_A(1-p_A). \tag{3.4}$$

Using the Central Limit Theorem (CLT) we get

$$\sqrt{n}(\tilde{p}_{A,i} - p_A) \;\xrightarrow{d}\; N(0, p_A(1-p_A)) \quad \text{as } n \to \infty\ . \tag{3.5}$$

Since the populations are independent, the $\tilde{p}_{A,i}$ are independent random variables. So the equation (3.5) gives

$$\mathbf{Z}_n \xrightarrow{d} \mathrm{MVN}_r\Big(\mathbf{0}, p_A(1-p_A)I_r\Big), \tag{3.6}$$

where $\mathbf{Z}_n' = \sqrt{n}(\tilde{p}_{A,1} - p_A, \tilde{p}_{A,2} - p_A, \cdots, \tilde{p}_{A,r} - p_A)$ and I_r is the r-dimensional identity matrix. Now $(I_r - r^{-1}1_r 1_r')$ is an idempotent matrix with trace and rank is equal to $(r-1)$. From the Corollary 1.7 and the Theorem 3.5 of Serfling (1980), Li found

$$\frac{MSP}{p_A(1-p_A)} \;\xrightarrow{d}\; \frac{\chi^2_{r-1}}{r-1} \quad \text{as } n \to \infty. \tag{3.7}$$

By Chebyshev's inequality and some algebra, Li (1996) got

$$MSG \xrightarrow{p} p_A(1-p_A). \tag{3.8}$$

Combining equations (3.7) and (3.8) and appealing to Slutsky's theorem, Li derived

$$\frac{MSP}{MSG} \xrightarrow{d} \frac{\chi^2_{r-1}}{r-1}. \tag{3.9}$$

So the test statistic is $\frac{MSP}{MSG}$ and we reject the hypothesis at α significance level if the value of the test statistic is greater than $100(1-\alpha)^{th}$ quantile of the distribution $\frac{\chi^2_{r-1}}{r-1}$.

3.3 New Testing Procedures

3.3.1 Parametric Bootstrap

In this section we propose a new testing procedure for testing population differentiation under a random population model. This test is developed specifically for small sample sizes. We assume that there are r independent populations that are descended from a common ancestral population. We have data from L (> 1) independent loci. We also assume that the expected allele frequencies in each population are the same. We also assume that there are s_l different allelic forms at the l^{th} locus with expected frequencies $p_1, p_2, \cdots, p_{s_l}$ respectively, in each population. We have n_i sampled alleles at each locus in the i^{th} population. We assume no disturbing force is acting on the loci that we are interested in. The populations are in a random mating system. Now we describe our test procedure based on the parametric bootstrap sampling. First estimate the allele frequencies at each locus from the frequency data. In our case the observed allele frequencies in each locus will be the maximum likelihood estimate of the population allele frequencies in that particular locus. When $\theta = 0$, the sampled alleles in a locus and population has a Multinomial distribution with appropriate index parameter and probability vector. We use this fact and generate a sample for each locus and population (n_i sampled alleles for the i^{th} population) from a Multinomial distribution with the frequencies obtained from the maximum likelihood estimator. In this way we generate one bootstrap sample. We repeat this procedure B times to get B bootstrap samples. Now we find the estimate of θ for each bootstrap sample. All these bootstrap samples are generated under the null hypothesis $H_0 : \theta = 0$. Using these B estimates of θ we can construct an empirical confidence interval for θ when the true value of θ is 0. We reject the hypothesis $\theta = 0$ with the level α if the estimate of θ based on the original data does not belong to the lower $100(1-\alpha)\%$ of the bootstrap estimates. Since both the estimators, the bootstrap and the original have the same negative bias, the biases will cancel with each other.

3.3.2 Large Sample Test

In this section we develop a test statistic for testing the population differentiation under a random population model. This test statistic is based on the large sample properties of random variables and works better for large sample. We assume the same set up as in the previous section. We first develop our test statistic based on a single locus. Then extend it for multi-locus situation. The development of the test statistic will be based on the locus A. The locus A has s different allelic types, $A_1, A_2, \cdots, A_s$. The expected frequencies of these allelic types are $p_1, p_2, \cdots, p_s$. We assume the value of θ is the same in different populations. We have n_i sampled alleles from the i^{th} population and $S = \sum_{i=1}^{r} n_i$. The data is defined as

$$x_{ij,k} = \begin{cases} 1 & \text{if the } j^{th} \text{ allele at locus } A \text{ in } i^{th} \text{ population is } A_k \\ 0 & \text{otherwise.} \end{cases}$$

In our notation, the allele frequencies based on the data are

$$\tilde{p}_{k,i} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij,k} \quad \text{and} \quad \tilde{p}_{w,k} = \frac{1}{S} \sum_{i=1}^{r} n_i \tilde{p}_{k,i}. \tag{3.10}$$

As $\sum_{k=1}^{s} \tilde{p}_{k,i} = 1$, there are $s-1$ independent observed allele frequencies. So if we have information from any $s-1$ allele frequencies, the last allele frequency would not provide any extra information. Without loss of generality we will work with the first $s-1$ allele frequencies. Now define some new quantities:

$$\tilde{\mathbf{p}}_i = \sqrt{n_i} \begin{bmatrix} \tilde{p}_{1,i} \\ \tilde{p}_{2,i} \\ \vdots \\ \tilde{p}_{s-1,i} \end{bmatrix}, \ \mathbf{p}_i = \sqrt{n_i} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{s-1} \end{bmatrix}, \ \tilde{\mathbf{P}} = \begin{bmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \\ \vdots \\ \tilde{\mathbf{p}}_r \end{bmatrix}, \ \mathbf{P} = \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_r \end{bmatrix},$$

$$C = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_{s-1} \\ -p_1p_2 & p_2(1-p_2) & \dots & -p_2p_{s-1} \\ \vdots & \vdots & \cdots & \vdots \\ p_1p_{s-1} & -p_2p_{s-1} & \dots & p_{s-1}(1-p_{s-1}) \end{bmatrix}, \mathbf{V} = \begin{bmatrix} C & (0)_{s-1} & \dots & (0)_{s-1} \\ (0)_{s-1} & C & \dots & (0)_{s-1} \\ \vdots & \vdots & \cdots & \vdots \\ (0)_{s-1} & (0)_{s-1} & \dots & C \end{bmatrix}.$$

Our aim is to define a statistic and find its distribution under the null hypothesis. When the value of θ is 0, then $(n_i\tilde{p}_{1,i}, \cdots, n_i\tilde{p}_{s,i}) \sim$ Multinomial$(n_i; p_1, p_2, \cdots, p_s)$. This fact provides us

$$\mathrm{E}(\tilde{\mathbf{P}}) \;=\; \mathbf{P} \quad \text{and} \quad \mathrm{Var}(\tilde{\mathbf{P}}) = \mathbf{V}. \tag{3.11}$$

When $n_i \to \infty \; \forall i$, the Multinomial distribution converges to a Multivariate Normal distribution with appropriate mean and variance (using CLT). So we get

$$\tilde{\mathbf{P}} - \mathbf{P} \sim MVN(\mathbf{0}, \mathbf{V}). \tag{3.12}$$

Since C is a positive definite matrix, there exists a nonsingular matrix T satisfying $C = TT'$. Now we define some new random variables:

$$\tilde{\mathbf{z}}_i = T^{-1}\tilde{\mathbf{p}}_i, \qquad \mathbf{z}_i = T^{-1}\mathbf{p}_i, \qquad \tilde{\mathbf{Z}} = \begin{bmatrix} \tilde{\mathbf{z}}_1 \\ \tilde{\mathbf{z}}_2 \\ \vdots \\ \tilde{\mathbf{z}}_r \end{bmatrix}, \qquad \text{and} \qquad \mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_r \end{bmatrix}.$$

Using standard theory when $n_i \to \infty$ we get

$$\tilde{\mathbf{Z}} - \mathbf{Z} \sim MVN(\mathbf{0}, I_{r(s-1)}). \tag{3.13}$$

Now we consider $A = (I_r - \frac{1}{r}1_r 1_r') \bigotimes I_{s-1}$. We will show that A is idempotent, symmetric and its rank is $(r-1)(s-1)$. Using the formula $(A1 \bigotimes B1)(A2 \bigotimes B2) = (A1A2 \bigotimes B1B2)$ we get

$$\begin{aligned} A^2 &= ((I_r - \frac{1}{r}1_r 1_r') \bigotimes I_{s-1})((I_r - \frac{1}{r}1_r 1_r') \bigotimes I_{s-1}) \\ &= ((I_r - \frac{1}{r}1_r 1_r')(I_r - \frac{1}{r}1_r 1_r')) \bigotimes (I_{s-1} I_{s-1}) \\ &= (I_r - \frac{1}{r}1_r 1_r') \bigotimes I_{s-1} = A. \end{aligned}$$

So A is idempotent. It is easy to check that A is also symmetric as I_r, I_{s-1} and $1_r 1_r'$ are symmetric. Since A is idempotent,

$$rank(A) = tr(A) = tr((I_r - \frac{1}{r}1_r 1_r') \bigotimes I_{s-1}) = tr(I_r - \frac{1}{r}1_r 1_r') tr(I_{s-1}) = (r-1)(s-1).$$

Now A is an idempotent, symmetric matrix with trace and rank is equal to $(s-1)(r-1)$. From the Corollary 1.7 and the Theorem 3.5 of Serfling (1980) we get

$$\tilde{\mathbf{Z}}' A \tilde{\mathbf{Z}} = \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(\sqrt{n_i}\tilde{p}_{k,i} - 1/r \sum_{i=1}^{r} \sqrt{n_i}\tilde{p}_{k,i})^2}{p_k} \quad \sim \quad \chi^2_{(r-1)(s-1)}. \tag{3.14}$$

If $n_1 = n_2 = \cdots = n_r = n$ then the above equation simplifies to

$$n \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(\tilde{p}_{k,i} - \tilde{p}_{w,k})^2}{p_k} \quad \sim \quad \chi^2_{(r-1)(s-1)}. \tag{3.15}$$

The first equality in the equation (3.14) is shown in Appendix B. It can be noted from the above equation that the left hand side is not a statistic. Because it involves the population allele frequencies. Now we propose a Lemma which will be used to find a test statistic. The proof of the lemma is given in Appendix C.

Lemma : Suppose $\mathbf{X}_n$ and $\mathbf{Y}_n$ are positive random variables of length s. Let us assume that $\sum_{i=1}^{s} X_{n,i} \xrightarrow{d} Z$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{c}(\text{or}, Y_{n,i} \xrightarrow{p} c_i)$, where Z is another random

variable and $\mathbf{c}$ is a constant vector. Then $\sum_{i=1}^{s} X_{n,i}c_i/Y_{n,i} \xrightarrow{d} Z$.

In our case, $\mathbf{X}_n = (X_{n,1}, \cdots, X_{n,s})$ and $\mathbf{Y}_n = \frac{1}{S}\sum_{i=1}^{r}\sqrt{n_i}\tilde{\mathbf{p}}_i$, where

$$X_{n,k} = \sum_{i=1}^{r}(\sqrt{n_i}\tilde{p}_{k,i} - 1/r\sum_{i=1}^{r}\sqrt{n_i}\tilde{p}_{k,i})^2/p_k.$$

Using a large sample property it is easy to show $\frac{1}{S}\sum_{i=1}^{r}\sqrt{n_i}\tilde{\mathbf{p}}_i \xrightarrow{p} \mathbf{p}$ which implies $\mathbf{c} = \mathbf{p}$. Let us define

$$\sum_{k=1}^{s} X_{n,k} = \sum_{k=1}^{s}\frac{\sum_{i=1}^{r}(\sqrt{n_i}\tilde{p}_{k,i} - 1/r\sum_{i=1}^{r}\sqrt{n_i}\tilde{p}_{k,i})^2}{p_k} \xrightarrow{d} \chi^2_{(r-1)(s-1)}. \quad (3.16)$$

So using the Lemma we conclude that $\sum_{k=1}^{s} X_{n,k}c_k/Y_{n,k} = \sum_{k=1}^{s} X_{n,k}p_k/\tilde{p}_{w,k}$ converges in distribution to $\chi^2_{(r-1)(s-1)}$. So we get

$$T = \sum_{i=1}^{r}\sum_{k=1}^{s}\frac{(\sqrt{n_i}\tilde{p}_{k,i} - 1/r\sum_{i=1}^{r}\sqrt{n_i}\tilde{p}_{k,i})^2}{\tilde{p}_{w,k}} \xrightarrow{d} \chi^2_{(r-1)(s-1)}. \quad (3.17)$$

For equal sample sizes our test statistic reduces to

$$T = n\sum_{i=1}^{r}\sum_{k=1}^{s}\frac{(\tilde{p}_{k,i} - \tilde{p}_{w,k})^2}{\tilde{p}_{w,k}}. \quad (3.18)$$

The test statistic distributed as χ^2 with $(r-1)(s-1)$ degrees of freedom under null hypothesis. We reject the hypothesis at α significance level if the test statistic is greater than the $100(1-\alpha)^{th}$ percentile of $\chi^2_{(r-1)(s-1)}$ distribution.

Now we assume there are L independent loci and s_l alleles at the l^{th} locus. Then the final test statistic is $T = \sum_{l=1}^{L} T_l$, where T_l is the single locus test statistic corresponding to the l^{th} locus. The test statistic T has a $\chi^2_{(r-1)\sum_{l=1}^{L}(s_l-1)}$ distribution under the null hypothesis. In this case we reject the null hypothesis if the test statistic is bigger than the $100(1-\alpha)^{th}$ percentile of $\chi^2_{(r-1)\sum_{l=1}^{L}(s_l-1)}$ distribution.

Chapter 4

Two Loci

4.1 Introduction

The probability of identity by descent simultaneously at two or more loci is a generalization of Wright's inbreeding coefficient. The multi-locus identity is a useful parameter in predicting the joint ancestry of pair of loci that is frequently used in mapping studies. The joint ancestry also helps in making inferences about historic population structure from current data (Hernández-Sánchez et al., 2004). The variances and covariances of quantitative traits in a finite population also involve multi-locus descent measures. The contributions to variance in the absence of epistasis depend only on two-locus identities or disequilibria, with epistasis multi-locus terms may be involved.

Weir and Cockerham (1969) extended the inbreeding coefficient concept for two loci to evaluate a measure of identity of descent for alleles at each of two linked loci. The multi-locus inbreeding coefficients depend on population size, inbreeding structure at each single locus and the linkage relationships between loci. The authors first studied sib mating and then established methods for determining a two-locus inbreeding function for any pedigree or mating system of individuals. For two-locus descent measures there are several components of inbreeding and it is necessary to introduce trigametic and quadrigametic measures in addition to the digametic measures. Later, Cockerham and Weir (1973) worked with the behavior of two-locus descent measures. The authors

discussed the use of two-locus descent measures. They also found expressions for digenic descent measures for finite populations. Afterwards, Weir and Cockerham (1974) presented an exact treatment of the behavior of pair of loci for infinite randomly mating, monoecious populations. All the above work were developed with distinct generations and in the absence of any kind of disturbing force such as selection.

The descent measures in conjunction with the frequency of initial population provide exact frequencies for all possible categories of two, three and four genes involving at the most two alleles at each of the two loci. This does not imply that the population has only two alleles per locus. From these measures and frequencies we can deduce various disequilibrium functions, the variance of linkage disequilibrium and related moments. The rates of approach to equilibrium conditions for linkage disequilibrium in monoecious population also depends on two-locus descent measures along with initial gamete frequencies.

In two-locus theory there are two different methods for transmitting gametes in the next generation. The first model is known as "random union of zygotes" and is used by Littler (1973), Watterson (1970) and by many others. The second model is "random union of gametes" and is used by Hill and Robertson (1968) and Karlin and McGregor (1968). All of these first authors have commented on the differences of these two approaches. A general theory of Weir and Cockerham (1974) yields many results for both the models. The qualitative behavior of both the models is identical. When the population size is large both models yield identical result. Here we work with the random union of zygotes model.

In this chapter we first discuss the parametrization of the two-locus descent measure described by Weir and Cockerham (1974). We assume that the loci we are interested in are neutral. We also assume that the populations follow a random mating system. Under this set up we find estimators of different components of the two-locus descent measures.

4.2 Two-Locus Parameters

Weir and Cockerham (1969, 1974) have extended Wright's inbreeding parameters for two loci. In their papers they parameterized the two-locus association in different levels of the population. Here we describe the two-locus inbreeding parameters for the locus A and the locus B. The frequently used two-locus parameters are $\mathbf{F}$, Θ, Γ and Δ. Each of these parameters is a vector of length 8 (Weir and Cockerham, 1974). Suppose a and a' are two alleles at the locus A and b and b' are two alleles at the locus B. Now define a vector of length 8

$$\begin{aligned} X(ab, a'b') &= [X_{11}^{11}(ab, a'b'), X_{11}(ab, a'b'),\ X^{11}(ab, a'b'),\ {}_{11}X(ab, a'b'),\ {}_{1}X_{1}^{1}(ab, a'b'), \\ &\quad X_1(ab, a'b'),\ X^1(ab, a'b'),\ {}_{1}X(ab, a'b')]', \end{aligned} \tag{4.1}$$

where the components are

$$\begin{aligned} X_{11}^{11}(ab, a'b') &= \mathrm{E}_{sub-pop}[Pr(a \equiv a' \equiv b \equiv b')], \\ X^{11}(ab, a'b') &= \mathrm{E}_{sub-pop}[Pr(a \equiv b, a' \equiv b')], \\ {}_{11}X(ab, a'b') &= \mathrm{E}_{sub-pop}[Pr(a \equiv b', a' \equiv b)], \\ X_{11}(ab, a'b') &= \mathrm{E}_{sub-pop}[Pr(a \equiv a', b \equiv b')], \\ {}_{1}X_{1}^{1}(ab, a'b') &= \mathrm{E}_{sub-pop}[\frac{1}{4}(Pr(a \equiv a' \equiv b) + Pr(a \equiv a' \equiv b') \\ &\quad + Pr(a \equiv b \equiv b') + Pr(a' \equiv b \equiv b'))], \\ X_1(ab, a'b') &= \mathrm{E}_{sub-pop}[\frac{1}{2}(Pr(a \equiv a') + Pr(b \equiv b'))], \\ X^1(ab, a'b') &= \mathrm{E}_{sub-pop}[\frac{1}{2}(Pr(a \equiv b) + Pr(a' \equiv b'))], \quad \text{and} \\ {}_{1}X(ab, a'b') &= \mathrm{E}_{sub-pop}[\frac{1}{2}(Pr(a \equiv b') + Pr(a' \equiv b))]. \end{aligned} \tag{4.2}$$

The equivalence relation denoted by $\equiv$ means that equivalent alleles are descended

from alleles on one initial gamete. The two-locus parameters are defined as

$$\begin{aligned}
\mathbf{F} &= X(ab,\, a'b' : ab \text{ and } a'b' \text{ are on two gametes from one individual}),\\
\Theta &= X(ab,\, a'b' : ab \text{ and } a'b' \text{ are on two gametes from two individuals}), \qquad (4.3)\\
\Gamma &= X(ab,\, a'b' : ab,\, a' \text{ and } b' \text{ are on three gametes from three individuals}),\\
\Delta &= X(ab,\, a'b' : a,\, b,\, a' \text{ and } b' \text{ are on four gametes from four individuals}).
\end{aligned}$$

The components of these parameters can be described in notation as we define each component of the X vector above. For example, Θ_{11}^{11} is the first component of the vector Θ while $\mathbf{F}^{11}$ is the second component of the vector $\mathbf{F}$. $\mathbf{F}^{1}$ and Θ^{1} are the average value F and θ respectively over the loci A and B. In a random mating population, $\mathbf{F} = \Theta$. We are interested in estimating the different components of Θ.

4.3 Theoretical Values of the Parameters

In this research our main focus will be on a random mating population. In this section we discuss the behavior of the two-locus descent measures over generations in a random mating population. Under this set up the descent measures $\mathbf{F}$ and Θ are exactly the same. In particular we discuss about the behavior of Θ, more precisely, last five components of Θ. We discuss about the transition equations for Θ_1, Θ^1, ${}_1\Theta$, ${}_1\Theta_1^1$ and Θ_{11}. For discussing the behavior of the above components we need to know the values of ${}_1\Gamma_1^1$, Γ_{11} and Δ_{11}. The two-locus parameters depend on the population size, inbreeding structure at each single locus and the linkage relationships between loci. The value of the two-locus parameters in a particular generation depend on the value of the one-locus, two-locus parameters in the previous generation, effective population size and recombination rate between the loci. In our notation, N is the population size in each generation and ρ is the recombination rate between the two loci we are interested in. We define a new quantity, $\lambda = 1 - 2\rho$. The value assume by the parameters $\mathbf{F}$, Θ, Γ and Δ

in the t^{th} generation are $\mathbf{F}_{(t)}$, $\Theta_{(t)}$, $\Gamma_{(t)}$ and $\Delta_{(t)}$ respectively. The following transitions describe the behavior of last five components of two-locus parameter Θ (Weir and Cockerham, 1974). These equations also describe the change of the theoretical values of ${}_1\Gamma^1_1$, Γ_{11} and Δ_{11} over generations. The equations are

$$\begin{aligned}
\Theta_{1(t+1)} &= \frac{1}{2}Q_2 + (1-\frac{1}{2}Q_2)\Theta_{1(t)},\\
\Theta^1_{(t+1)} &= (1-\rho)\Theta^1_{(t+1)} + \rho\,{}_1\Theta_{(t)},\\
{}_1\Theta_{(t+1)} &= \frac{1}{2}Q_2\,\Theta^1_{(t)} + (1-\frac{1}{2}Q_2){}_1\Theta_{(t)},\\
{}_1\Theta^1_{1(t+1)} &= [\frac{1}{2}+\frac{\lambda}{2}(1-Q_2)]{}_1\Theta^1_{1(t)} + \rho(1-Q_2){}_1\Gamma^1_{1(t)} + Q_2(\frac{1-\rho}{2}\Theta^1_{(t)} + \frac{\rho}{2}\,{}_1\Theta_{(t)}),\\
{}_1\Gamma^1_{1(t+1)} &= (\frac{1}{2}Q_3+\frac{1}{3}Q_{21}){}_1\Theta^1_{1(t)} + (\frac{1}{2}Q_{21}+\frac{1}{3}Q_{111}){}_1\Gamma^1_{1(t)} + \frac{1}{4}Q_3\Theta^1_{(t)}\\
&\quad +\frac{1}{2}(\frac{1}{2}Q_3+\frac{1}{3}Q_{21}){}_1\Theta_{(t)},\\
\Theta_{11(t+1)} &= \Omega_{11}\Theta_{11(t)} + \Omega_{12}\Gamma_{11(t)} + \Omega_{13}\Delta_{11(t)} + \frac{(1-\lambda^2)}{2N}\Theta_{1(t)} + \frac{(1+\lambda^2)}{4N},\\
\Gamma_{11(t+1)} &= \Omega_{21}\Theta_{11(t)} + \Omega_{22}\Gamma_{11(t)} + \Omega_{23}\Delta_{11(t)} + \frac{(2N-1)}{2N^2}\Theta_{1(t)} + \frac{1}{4N^2}, \text{ and}\\
\Delta_{11(t+1)} &= \Omega_{31}\Theta_{11(t)} + \Omega_{32}\Gamma_{11(t)} + \Omega_{33}\Delta_{11(t)} + \frac{(2N-1)}{2N^2}\Theta_{1(t)} + \frac{1}{4N^2},
\end{aligned} \tag{4.4}$$

where, $Q_2 = \frac{1}{N}$, $Q_3 = \frac{1}{N^2}$, $Q_{21} = \frac{3(N-1)}{N^2}$ and $Q_{111} = \frac{(N-1)(N-2)}{N^2}$. Ω_{ij}'s are the elements of the matrix Ω and that is

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{bmatrix} = \begin{bmatrix} \frac{(1+\lambda)^2}{4} - \frac{\lambda}{2N} & \frac{(N-1)(1-\lambda^2)}{2N} & \frac{(N-1)(1-\lambda)^2}{4N} \\ \frac{1+\lambda}{4N} - \frac{\lambda}{4N^2} & \frac{(N-1)[N+1+\lambda(N-2)]}{2N^2} & \frac{(N-1)(2N-3)(1-\lambda)}{4N^2} \\ \frac{2N-1}{4N^3} & \frac{(N-1)(2N-1)}{N^3} & \frac{(N-1)(2N-1)(2N-3)}{4N^3} \end{bmatrix}.$$

Weir and Cockerham (1974) also gave the transition equations for Θ^{11}_{11}, Θ^{11} and ${}_{11}\Theta$. Since we will not discuss these components, we skip the transition equations in this research. If the population size changes over generations, then the above equations approximately hold good if N replaced by N_e, the effective population size.

4.4 Notation

In this section we define our notation properly. We have r independent populations. Each population has evolved from the same ancestral population. We also assume that the ancestral population has infinitely many individuals. The population size at each generation (except the ancestral population) is N. We work with two loci, A and B. The locus A has s_A different alleles, $A_1, A_2, \cdots, A_{s_A}$ and the locus B has s_B different alleles, $B_1, B_2, \cdots, B_{s_B}$. The recombination rate between the locus A and the locus B is ρ and $\lambda = 1$ - 2ρ. The expected gamete frequencies in the present-day populations are the same. There are different types of gamete frequencies and digametic, trigametic and quadrigametic gamete frequencies for the present generation are

$$
\begin{aligned}
p_k &= \text{expected frequency of the allele } A_k \text{ ,} \\
q_l &= \text{expected frequency of the allele } B_l, \\
P_{\cdot\cdot}^{kl} &= \text{expected frequency of the gamete } A_kB_l, \\
P_{\cdot l}^{k\cdot} &= \text{expected frequency of a random pair of gametes that carry allele } A_k \text{ and } B_l, \\
P_{u|\cdot}^{kl} &= \text{expected frequency of a random pair of gametes that carry } A_kB_l \text{ and } A_u, \qquad (4.5) \\
P_{\cdot|v}^{kl} &= \text{expected frequency of a random pair of gametes that carry } A_kB_l \text{ and } B_v, \\
P_{u|\cdot}^{k|l} &= \text{expected frequency of a random triplet of gametes that carry } A_k, A_u \text{ and } B_l, \\
P_{\cdot|v}^{k|l} &= \text{expected frequency of a random triplet of gametes that carry } A_k, B_l \text{ and } B_v, \\
P_{uv}^{kl} &= \text{expected frequency of a random pair of gametes } A_kB_l \text{ and } A_uB_v, \\
P_{u|v}^{kl} &= \text{expected frequency of a random triplet of gametes that carry } A_kB_l, A_u \text{ and } B_v, \\
P_{u|v}^{k|l} &= \text{expected frequency of a random quadruple of gametes that carry } A_k, A_u, B_l, B_v, \\
D_{kl} &= P_{\cdot\cdot}^{kl} - p_kq_l = \text{expected linkage disequilibrium for alleles } A_k \text{ and } B_l \text{ ,}
\end{aligned}
$$

where, $k, u = 1, 2, \cdots, s_A$ and $l, v = 1, 2, \cdots, s_B$. The corresponding parameters (frequencies) in the ancestral population are p_k, q_l, $\mathscr{P}_{\cdot\cdot}^{kl}$, $\mathscr{P}_{\cdot l}^{k\cdot}$, $\mathscr{P}_{u|\cdot}^{kl}$, $\mathscr{P}_{\cdot|v}^{kl}$, $\mathscr{P}_{u|\cdot}^{k|l}$, $\mathscr{P}_{\cdot|v}^{kl}$, $\mathscr{P}_{uv}^{kl}$, $\mathscr{P}_{u|v}^{kl}$, $\mathscr{P}_{u|v}^{k|l}$ and $\mathscr{D}_{kl}$. It is important to note that the expected allele frequencies in the ancestral and present population are the same. This is because the populations

are mating at random. The gamete frequencies in the present generation are related to the ancestral gamete frequencies through the descent measures at the present generation. Weir and Cockerham (1974) found the relationship between the gamete frequencies in the present generation, the ancestral gamete frequencies, and the two-locus descent measures in the present generation. Our moment estimator is based on the following relations:

$$\begin{aligned}
P^{ij}_{\cdot\cdot} &= p_i q_j + \Theta^1 \mathscr{D}_{ij};\ \ P^{i\cdot}_{\cdot j} = p_i q_j +\ {}_1\Theta \mathscr{D}_{ij}, \\
P^{ij}_{i|\cdot} &= p_i^2 q_j + \Theta_1 p_i q_j (1-p_i) + (\Theta^1 +\ {}_1\Theta) p_i \mathscr{D}_{ij} +\ {}_1\Theta^1_1 (1-2p_i)\mathscr{D}_{ij}, \\
P^{ij}_{\cdot|j} &= p_i q_j^2 + \Theta_1 p_i q_j (1-q_j) + (\Theta^1 +\ {}_1\Theta) q_j \mathscr{D}_{ij} +\ {}_1\Theta^1_1 (1-2q_j)\mathscr{D}_{ij}, \qquad (4.6) \\
P^{i|j}_{i|\cdot} &= p_i^2 q_j + \Theta_1 p_i q_j (1-p_i) + 2_1\Theta p_i \mathscr{D}_{ij} +\ {}_1\Gamma^1_1 (1-2p_i)\mathscr{D}_{ij}, \ \text{ and} \\
P^{i|j}_{\cdot|j} &= p_i q_j^2 + \Theta_1 p_i q_j (1-q_j) + 2_1\Theta q_j \mathscr{D}_{ij} +\ {}_1\Gamma^1_1 (1-2q_j)\mathscr{D}_{ij}.
\end{aligned}$$

4.5 Data

We have data from r independent populations. These populations have evolved from the same ancestral population. We have haplotype data for two loci, locus A and locus B. We have n_i haplotype sampled from the i^{th} population. So there are total $\sum_{i=1}^{r} n_i = S$ sampled haplotypes. Let us define the data at the locus A as follows:

$$x_{ij,k} = \begin{cases} 1 & \text{if the } j^{th} \text{ allele at locus } A \text{ in } i^{th} \text{ population is } A_k \\ 0 & \text{otherwise} \end{cases}$$

The data at the locus B can be defined as follows:

$$y_{ij,l} = \begin{cases} 1 & \text{if the } j^{th} \text{ allele at locus } B \text{ in } i^{th} \text{ population is } B_l \\ 0 & \text{otherwise} \end{cases}$$

Since we have information at both the loci for a particular gamete, we can recover the haplotypes from the data at locus A and locus B. Now we define several observed gamete frequencies using the haplotype data. These sample gamete frequencies will be used to define different statistics which will be used to find the estimators of the parameters. The observed gamete frequencies for the i^{th} population are

$$
\begin{aligned}
&\tilde{p}_{i,k} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij,k}\ , \quad \tilde{q}_{i,l} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij,l}\ ,\\
&\tilde{P}_{1,i,kl} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij,k}y_{ij,l}\ , \quad \tilde{P}_{2,i,kl} = \frac{1}{n_i(n_i-1)}\sum_{j\neq j'=1}^{n_i}\sum x_{ij,k}y_{ij',l}\ ,\\
&\tilde{P}_{3,i,kl} = \frac{1}{n_i(n_i-1)}\sum_{j\neq j'=1}^{n_i}\sum x_{ij,k}y_{ij,l}x_{ij',k}\ ,\\
&\tilde{P}_{4,i,kl} = \frac{1}{n_i(n_i-1)}\sum_{j\neq j'=1}^{n_i}\sum x_{ij,k}y_{ij,l}y_{ij',l}\ ,\\
&\tilde{P}_{5,i,kl} = \frac{1}{n_i(n_i-1)(n_i-2)}\sum_{j\neq j'\neq j''=1}^{n_i}\sum\sum x_{ij,k}y_{ij',l}x_{ij'',k}\ ,\\
&\tilde{P}_{6,i,kl} = \frac{1}{n_i(n_i-1)(n_i-2)}\sum_{j\neq j'\neq j''=1}^{n_i}\sum\sum y_{ij,k}y_{ij',l}y_{ij'',l}\ ,\\
&\tilde{P}_{7,i,kl} = \frac{1}{n_i(n_i-1)}\sum_{j\neq j'=1}^{n_i}\sum x_{ij,k}y_{ij,l}x_{ij',k}y_{ij',k}\ ,\\
&\tilde{P}_{8,i,kl} = \frac{1}{n_i(n_i-1)(n_i-2)}\sum_{j\neq j'\neq j''=1}^{n_i}\sum\sum x_{ij,k}y_{ij,l}x_{ij',k}y_{ij'',l}\ , \text{ and}\\
&\tilde{P}_{9,i,kl} = \frac{1}{n_i(n_i-1)(n_i-2)(n_i-3)}\sum_{j\neq j'\neq j''\neq j'''=1}^{n_i}\sum\sum\sum x_{ij,k}y_{ij',l}x_{il'',k}y_{ij''',l}.
\end{aligned}
\tag{4.7}
$$

The overall gamete frequencies are $\tilde{p}_k$, $\tilde{q}_l$, $\tilde{P}_{1,kl}$, $\tilde{P}_{2,kl}$, $\tilde{P}_{3,kl}$, $\tilde{P}_{4,kl}$, $\tilde{P}_{5,kl}$, $\tilde{P}_{5,kl}$, $\tilde{P}_{6,kl}$, $\tilde{P}_{7,kl}$, $\tilde{P}_{8,kl}$, and $\tilde{P}_{9,kl}$. These are all weighted gamete frequencies and the weight for the i^{th} population is the denominator in the observed frequency in the i^{th} population.

Now we define some new quantities that are functions of sample sizes in different

populations and the total number of populations r. These numbers will be used to find the estimators of the parameters. The new quantities are

$$\begin{aligned}
n_{c_1} &= \frac{1}{r-1}(\sum_{i=1}^{r} n_i - \frac{\sum_{i=1}^{r} n_i^2}{\sum_{i=1}^{r} n_i}) = \frac{1}{r-1}(S - \frac{\sum_{i=1}^{r} n_i^2}{S}), \\
n_{c_2} &= \frac{1}{(r-1)(r-2)}(S\sum_{i=1}^{r}\frac{1}{n_i} - 3r + 2), \\
n_{c_3} &= \frac{1}{(r-1)(r-2)}(Sr - S\sum_{i=1}^{r}\frac{1}{n_i} - 3S + 3r + \frac{2\sum_{i=1}^{r} n_i^2}{S} - 2), \qquad (4.8) \\
n_{c_4} &= \frac{1}{(r-1)(r-2)}(S^2 - 3Sr + 2S\sum_{i=1}^{r}\frac{1}{n_i} - 3\sum_{i=1}^{r} n_i^2 + 9S - 6r + \frac{2\sum_{i=1}^{r} n_i^3}{S} \\
&\quad -\frac{6\sum_{i=1}^{r} n_i^2}{S} + 4), \text{ and} \\
n_{c_5} &= \frac{1}{r-1}(\sum_{i=1}^{r} n_i^2 - \frac{\sum_{i=1}^{r} n_i^3}{S}).
\end{aligned}$$

For equal sample sizes i.e. $n_1 = n_2 = \cdots = n_r = n$, the above quantities reduce to

$$n_{c_1} = n,\ n_{c_2} = 1,\ n_{c_3} = n-1,\ n_{c_4} = (n-1)(n-2),\ \text{and}\ n_{c_5} = n^2. \qquad (4.9)$$

4.6 Identifiability Problem

Our main aim is to find the estimators of different components of Θ. The equation (4.6) shows that the expectations of second and third order gamete frequencies involve only Θ^1, ${}_1\Theta$, Θ_1 (or θ), ${}_1\Theta_1^1$ and ${}_1\Gamma_1^1$. There are three independent second order gamete frequencies and two independent third order gamete frequencies that depend on alleles A_k and B_l. It is possible to find the estimates the above five parameters using these five independent gamete frequencies. For fourth order gamete frequencies we have only three new gamete frequencies, but these frequencies will involve 12 new unknown descent measures (Weir and Cockerham, 1974). So it is not possible to estimate these 12

descent measures from three independent frequencies. If we assume that the ancestral population is in linkage equilibrium then we have a completely different situation. In this case, the second and third order gamete frequencies will involve only one descent measure, θ. The fourth order gamete frequencies involve three new descent measures, Θ_{11}, Γ_{11} and Δ_{11}, and these measures can be estimated from the three fourth order gamete frequencies.

In finding the moment estimators of the parameters we use the following strategy. First we use the data to define some statistics that estimate different types of population gamete frequencies in the present-day generation. Now using the relations given in the equation (4.6), the population gamete frequencies in the present generation can be expressed in terms of ancestral gamete frequencies and the two-locus descent measures such as Θ, Γ and Δ. Our estimators of the descent measures are based on these relations.

The second and third order gamete frequencies described by the alleles A_k and B_l in the present generation depend on the frequencies of the alleles A_k and B_l, the linkage disequilibrium of the gamete A_kB_l in the ancestral population, and the two-locus descent measures in the present population. So the present-day gamete frequencies depend on the parameters p_k, q_l, $\mathscr{D}_{kl}$, Θ^1, ${}_1\Theta$, Θ_1 (or θ), ${}_1\Theta_1^1$, and ${}_1\Gamma_1^1$. The dependency of the gamete frequencies in the present generation on the descent measures and the linkage-disequilibrium in the ancestral population is through $\Theta^1\mathscr{D}_{kl}$, ${}_1\Theta\mathscr{D}_{kl}$, θ, ${}_1\Theta_1^1\mathscr{D}_{kl}$, and ${}_1\Gamma_1^1\mathscr{D}_{kl}$. So we cannot estimate the parameters Θ^1, ${}_1\Theta$, ${}_1\Theta_1^1$, and ${}_1\Gamma_1^1$ separately, but we can estimate the compound parameters $\Theta^1\mathscr{D}_{kl}$, ${}_1\Theta\mathscr{D}_{kl}$, ${}_1\Theta_1^1\mathscr{D}_{kl}$, and ${}_1\Gamma_1^1\mathscr{D}_{kl}$. So the parameters Θ^1, ${}_1\Theta$, ${}_1\Theta_1^1$, and ${}_1\Gamma_1^1$ are not identifiable. We can estimate the parameter θ separately and it is identifiable. When the ancestral population is in linkage equilibrium then we are interested in the parameters θ, Θ_{11}, Γ_{11} and Δ_{11}. In this situation, these parameters are estimable.

4.7 Moment Estimator of $\Theta^1\mathscr{D}_{kl}$ and ${}_1\Theta\mathscr{D}_{kl}$

In this section we propose unbiased estimators of $\Theta^1\mathscr{D}_{kl}$ and ${}_1\Theta\mathscr{D}_{kl}$. We propose the following two statistics using observed gamete frequencies to get information about the parameters. These statistics are motivated from Weir and Hill (2002)'s statistics that are defined in the equations (2.3) and (2.4). The statistics are

$$MSP_{kl}^{AB} = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\tilde{p}_{i,k}-\tilde{p}_k)(\tilde{q}_{i,l}-\tilde{q}_l) \text{ and} \tag{4.10}$$

$$MSG_{kl}^{AB} = \frac{1}{\sum_{i=1}^{r}(n_i-1)}\sum_{i=1}^{r} n_i\Big[\tilde{p}_{i,k}(1-\tilde{q}_{i,l})+(1-\tilde{p}_{i,k})\tilde{q}_{i,l}\Big]. \tag{4.11}$$

We use the following equations to find the expectations of the above statistics:

$$\mathrm{E}(x_{ij,k}y_{i'j',l}) = \begin{cases} P_{\cdot\cdot}^{kl} = p_kq_l + \Theta^1\mathscr{D}_{kl} & \text{if } i=i',\, j=j' \\ P_{\cdot|l}^{k|\cdot} = p_kq_l + {}_1\Theta\mathscr{D}_{kl} & \text{if } i=i',\, j\neq j' \\ p_kq_l & \text{if } i\neq i' \end{cases} \tag{4.12}$$

The expectations of the statistics are

$$\mathrm{E}(MSP_{kl}^{AB}) = \Theta^1\mathscr{D}_{kl} + (n_{c_1}-1){}_1\Theta\mathscr{D}_{kl} \text{ and} \tag{4.13}$$

$$\mathrm{E}(MSG_{kl}^{AB}) = \frac{Sp_k+Sq_l-2Sp_kq_l}{S-r} - \frac{2}{S-r}\Big[r\Theta^1\mathscr{D}_{kl}+(S-r)\,{}_1\Theta\mathscr{D}_{kl}\Big]. \tag{4.14}$$

When we have equal sample sizes for different populations, then the right hand side of the equation (4.13) reduces to $(\Theta^1\mathscr{D}_{kl}+(n-1){}_1\Theta\mathscr{D}_{kl})$, while the second term in the right had side of the equation (4.14) is $\frac{2}{n-1}(\Theta^1\mathscr{D}_{kl}+(n-1){}_1\Theta\mathscr{D}_{kl})$. This implies that the two equations (4.13) and (4.14) provide the same information about $\Theta^1\mathscr{D}_{kl}$ and ${}_1\Theta\mathscr{D}_{kl}$ when sample sizes are equal. In order to find estimators of these parameters we need two independent linear equations. So for unequal sample sizes the two statistics will be sufficient to find estimators of the parameters, but for equal sample sizes these two

statistics will not be enough. We propose to replace the second statistic by a statistic which always provides different information about the parameters than MSP_{kl}^{AB}. Here we consider the statistic $S_{4,kl} = \tilde{P}_{1,kl} - \tilde{P}_{2,kl}$. The expectation of the statistic is

$$\mathrm{E}(S_{4,kl}) \;=\; \Theta^1\mathscr{D}_{kl} - {}_1\Theta\mathscr{D}_{kl}. \tag{4.15}$$

After doing some algebra with the equations (4.13) and (4.15) we have

$$\mathrm{E}\Big[MSP_{kl}^{AB} + (n_{c_1}-1)S_{4,kl}\Big] \;=\; n_{c_1}\,\Theta^1\mathscr{D}_{ij} \;\text{ and} \tag{4.16}$$

$$\mathrm{E}\Big[MSP_{kl}^{AB} - S_{4,kl}\Big] \;=\; n_{c_1}\,{}_1\Theta\mathscr{D}_{kl}. \tag{4.17}$$

The above two equations give our moment estimators of the compound parameters

$$\widehat{\Theta^1\mathscr{D}_{kl}} \;=\; \frac{MSP_{kl}^{AB} + (n_{c_1}-1)S_{4,kl}}{n_{c_1}} \;\text{ and} \tag{4.18}$$

$$\widehat{{}_1\Theta\mathscr{D}_{kl}} \;=\; \frac{MSP_{kl}^{AB} - S_{4,kl}}{n_{c_1}}. \tag{4.19}$$

4.8 Moment Estimator Of ${}_1\Theta_1^1\mathscr{D}_{kl}$ and ${}_1\Gamma_1^1\mathscr{D}_{kl}$

We propose four statistics that are based on the third order gamete frequencies to estimate the compound parameters ${}_1\Theta_1^1\mathscr{D}_{kl}$ and ${}_1\Gamma_1^1\mathscr{D}_{kl}$. The statistics are

$$MSP_{kl}^{AAB} \;=\; \frac{S}{(r-1)(r-2)}\sum_{i=1}^{r} n_i(\tilde{p}_{i,k}-\tilde{p}_k)^2(\tilde{q}_{i,l}-\tilde{q}_l), \tag{4.20}$$

$$MSG_{kl}^{AAB} \;=\; \frac{1}{\sum_{i=1}^{r}(n_i-1)}\sum_{i=1}^{r} n_i[\tilde{p}_{i,k}^2(1-\tilde{q}_{i,l}) + \tilde{p}_{i,k}(1-\tilde{q}_{i,l})^2], \tag{4.21}$$

$$MSP_{kl}^{ABB} \;=\; \frac{S}{(r-1)(r-2)}\sum_{i=1}^{r} n_i(\tilde{p}_{i,k}-\tilde{p}_k)(\tilde{q}_{i,l}-\tilde{q}_l)^2, \;\text{ and} \tag{4.22}$$

$$MSG_{kl}^{AAB} \;=\; \frac{1}{\sum_{i=1}^{r}(n_i-1)}\sum_{i=1}^{r} n_i[\tilde{p}_{i,k}(1-\tilde{q}_{i,l})^2 + \tilde{p}_{i,k}(1-\tilde{q}_{i,l})]. \tag{4.23}$$

To find the expectations of the statistics defined above we need to use the following equation:

$$\mathrm{E}(x_{ij,k}x_{i'j',k}y_{i''j'',l}) = \begin{cases} P^{kl}_{\cdot\cdot} & \text{if } i=i'=i'',\ j=j'=j'' \\ P^{k\cdot}_{\cdot|l} & \text{if } i=i'=i'',\ j=j'\neq j'' \\ P^{kl}_{k|\cdot} & \text{if } i=i'=i'',\ j=j''\neq j' \\ P^{k|l}_{k|\cdot} & \text{if } i=i'=i'',\ j\neq j'\neq j'' \\ q_l P^{k|\cdot}_{k|\cdot} & \text{if } i=i'\neq i'',\ j\neq j' \\ p_k q_l & \text{if } i=i'\neq i'',\ j=j' \\ p_i P^{kl}_{\cdot\cdot} & \text{if } i=i''\neq i',\ j=j'' \\ p_k P^{k|\cdot}_{\cdot|l} & \text{if } i=i''\neq i',\ j\neq j'' \\ p_k^2 q_l & \text{if } i\neq i'\neq i'' \end{cases} \tag{4.24}$$

We also need to use $\mathrm{E}(x_{ij,k}y_{i'j',l}y_{i''j'',l})$ which can be found in the same pattern as equation (4.24). The expectations of the statistics are

$$\mathrm{E}(MSP^{AAB}_{kl}) = (1-2p_k)[n_{c_2}\Theta^1\mathscr{D}_{kl} + n_{c_3\ 1}\Theta\mathscr{D}_{kl} + 2n_{c_3\ 1}\Theta^1_1\mathscr{D}_{kl} + n_{c_4\ 1}\Gamma^1_1\mathscr{D}_{kl}],$$

$$\begin{aligned}\mathrm{E}(MSG^{AAB}_{kl}) &= \frac{Sp_k + Sq_l - 2Sp_kq_l}{S-r} + \frac{p_k(1-p_k)(r+(S-r)\Theta_1)}{S-r} \\ &\quad -\frac{2(r\Theta^1\mathscr{D}_{kl} + (S-r)_1\Theta\mathscr{D}_{kl})}{S-r},\end{aligned} \tag{4.25}$$

$$\mathrm{E}(MSP^{ABB}_{kl}) = (1-2q_l)[n_{c_2}\Theta^1\mathscr{D}_{kl} + n_{c_3\ 1}\Theta\mathscr{D}_{kl} + 2n_{c_3\ 1}\Theta^1_1\mathscr{D}_{kl} + n_{c_4\ 1}\Gamma^1_1\mathscr{D}_{kl}], \text{ and}$$

$$\begin{aligned}\mathrm{E}(MSG^{ABB}_{kl}) &= \frac{Sp_l + Sq_l - 2Sp_kq_l}{S-r} + \frac{q_l(1-q_l)(r+(S-r)\Theta_1)}{S-r} \\ &\quad -\frac{2(r\Theta^1\mathscr{D}_{kl} + (S-r)_1\Theta\mathscr{D}_{kl})}{S-r}.\end{aligned}$$

The equation (4.25) shows that $\mathrm{E}(MSG_{kl}^{AAB})$ and $\mathrm{E}(MSG_{kl}^{ABB})$ do not provide any information about ${}_1\Theta_1^1\mathscr{D}_{kl}$ and ${}_1\Gamma_1^1\mathscr{D}_{kl}$. In fact, the statistics MSG_{kl}^{AAB} and MSG_{kl}^{ABB} involve only second order gamete frequencies. We propose the following two statistics which give information about ${}_1\Theta_1^1\mathscr{D}_{kl}$ and ${}_1\Gamma_1^1\mathscr{D}_{kl}$:

$$S_{5,kl} = \tilde{P}_{3,kl} - \tilde{P}_{5,kl} - \frac{1}{r(r-1)} \sum_{i \neq i'=1}^{r} \sum (\tilde{P}_{1,i,kl} - \tilde{P}_{2,i,kl}) \tilde{p}_{i',k} \text{ and} \tag{4.26}$$

$$S_{6,kl} = \tilde{P}_{4,kl} - \tilde{P}_{6,kl} - \frac{1}{r(r-1)} \sum_{i \neq i'=1}^{r} \sum (\tilde{P}_{1,i,kl} - \tilde{P}_{2,i,kl}) \tilde{q}_{i',j} \tag{4.27}$$

The expectations of the statistics are

$$\mathrm{E}(S_{5,kl}) = (1-2p_k)[\,{}_1\Theta_1^1\mathscr{D}_{kl} - {}_1\Gamma_1^1\mathscr{D}_{kl}] \text{ and} \tag{4.28}$$

$$\mathrm{E}(S_{6,kl}) = (1-2q_l)[\,{}_1\Theta_1^1\mathscr{D}_{kl} - {}_1\Gamma_1^1\mathscr{D}_{kl}]. \tag{4.29}$$

These expectations are zero when $p_l = 0.5$ and $q_l = 0.5$. Thus, they do not provide any information about the parameters when $p_k = 0.5$ and $q_l = 0.5$. This is consistent with the population genetics theory. It is easy to check that $\tilde{p}_k$ and $\tilde{q}_l$ are the unbiased estimators of p_k and q_l respectively. After doing some algebra with the equations (4.20)-(4.23) we construct four new statistics for estimating the compound parameters. The four statistics are

$$S_{7,kl} = \frac{MSP_{kl}^{AAB} + n_{c_4} S_{5,kl}}{1-2\tilde{p}_k} - n_{c_2}\widehat{\Theta^1\mathscr{D}_{kl}} - n_{c_3}\widehat{{}_1\Theta\mathscr{D}_{kl}}, \tag{4.30}$$

$$S_{8,kl} = \frac{MSP_{kl}^{AAB} - 2n_{c_3} S_{5,kl}}{1-2\tilde{p}_k} - n_{c_2}\widehat{\Theta^1\mathscr{D}_{kl}} - n_{c_3}\widehat{{}_1\Theta\mathscr{D}_{kl}}, \tag{4.31}$$

$$S_{9,kl} = \frac{MSP_{kl}^{ABB} + n_{c_4} S_{6,kl}}{1-2\tilde{q}_l} - n_{c_2}\widehat{\Theta^1\mathscr{D}_{kl}} - n_{c_3}\widehat{{}_1\Theta\mathscr{D}_{kl}}, \text{ and} \tag{4.32}$$

$$S_{10,kl} = \frac{MSP_{kl}^{ABB} - 2n_{c_3} S_{6,kl}}{1-2\tilde{q}_l} - n_{c_2}\widehat{\Theta^1\mathscr{D}_{kl}} - n_{c_3}\widehat{{}_1\Theta\mathscr{D}_{kl}}. \tag{4.33}$$

The statistics in (4.30) and (4.31) are well defined only when $\tilde{p}_k$ is not equal to 0.5. Similarly, the statistics in (4.32) and (4.33) are not defined when $\tilde{q}_l = 0.5$. The expectations of the statistics are defined only when $p_k \neq 0.5$ or $q_l \neq 0.5$. The ratio estimation theory provides the expectations of the statistics and they are

$$\begin{aligned} \mathrm{E}(S_{7,kl}) &= \mathrm{E}(S_{9,kl}) \approx (2n_{c_3} + n_{c_4})\, {}_1\Theta^1_1\mathscr{D}_{kl}, \text{ and} \quad (4.34) \\ \mathrm{E}(S_{8,kl}) &= \mathrm{E}(S_{10,kl}) \approx (2n_{c_3} + n_{c_4})\, {}_1\Gamma^1_1\mathscr{D}_{kl}. \quad (4.35) \end{aligned}$$

The above discussion demonstrates that when both $\tilde{p}_k$ and $\tilde{q}_l$ are 0.5 then there do not exist any estimators for ${}_1\Theta^1_1\mathscr{D}_{kl}$ and ${}_1\Gamma^1_1\mathscr{D}_{kl}$. But the estimators of these parameters exist if one of the allele frequency is not equal to 0.5 and the estimators are

$$\begin{aligned} \widehat{{}_1\Theta^1_1\mathscr{D}_{kl}} &= \frac{S_{7,kl} I(\tilde{p}_k \neq 0.5) + S_{9,kl} I(\tilde{q}_l \neq 0.5)}{(2n_{c_3} + n_{c_4})\Big[I(\tilde{p}_k \neq 0.5) + I(\tilde{q}_l \neq 0.5)\Big]}, \text{ and} \quad (4.36) \\ \widehat{{}_1\Gamma^1_1\mathscr{D}_{kl}} &= \frac{S_{8,kl} I(\tilde{p}_k \neq 0.5) + S_{10,kl} I(\tilde{q}_l \neq 0.5)}{(2n_{c_3} + n_{c_4})\Big[I(\tilde{p}_k \neq 0.5) + I(\tilde{q}_l \neq 0.5)\Big]}. \quad (4.37) \end{aligned}$$

4.9 Ancestral Population is in Linkage Equilibrium

In this section we assume that the ancestral population is in linkage equilibrium. This assumption reduces lot of computation burden. For example, if the ancestral population is in linkage equilibrium then the different gamete frequencies in the present population do not depend on Θ^1, ${}_1\Theta$, ${}_1\Theta^1_1$ and ${}_1\Gamma^1_1$. In this setup, we are interested in the four parameters θ, Θ_{11}, Γ_{11} and Δ_{11}. These parameters characterize the gamete frequencies in the present generation. Now we propose the moment estimator of these parameters. We have found different estimators of θ in the previous section. To find the estimator of other three parameters we have used the equations for quadrigametic

descent measures. The equations are

$$\begin{aligned} P_{kl}^{kl} &= p_k^2 q_l^2 + \Theta_1 p_k q_l (p_k + q_l - 2p_k q_l) + \Theta_{11} p_k q_l (1-p_k)(1-q_l), \\ P_{k|l}^{kl} &= p_k^2 q_l^2 + \Theta_1 p_k q_l (p_k + q_l - 2p_k q_l) + \Gamma_{11} p_k q_l (1-p_k)(1-q_l), \quad \text{and} \\ P_{k|l}^{k|l} &= p_k^2 q_l^2 + \Theta_1 p_k q_l (p_k + q_l - 2p_k q_l) + \Delta_{11} p_k q_l (1-p_k)(1-q_l). \end{aligned} \tag{4.38}$$

Now we propose four statistics which will be used to estimate the parameters. The statistics are

$$\begin{aligned} R_{1,kl} &= \tilde{P}_{7,kl} - R_{kl}, \\ R_{2,kl} &= \tilde{P}_{8,kl} - R_{kl}, \\ R_{3,kl} &= \tilde{P}_{9,kl} - R_{kl}, \quad \text{and} \\ R_{4,kl} &= \frac{1}{r(r-1)} \sum_{i \neq i'=1}^{r} \sum \tilde{p}_{i,k} \tilde{q}_{i,l} (1 - \tilde{p}_{i',k})(1 - \tilde{q}_{i'l}), \end{aligned} \tag{4.39}$$

where,

$$R_{kl} = \frac{1}{r(r-1)} \sum_{i \neq i'=1}^{r} \sum \Big[\frac{\tilde{q}_{i',l}(\tilde{P}_{3,i,kl} + \tilde{P}_{5,i,kl})}{2} + \frac{\tilde{p}_{i',k}(\tilde{P}_{4,i,kl} + \tilde{P}_{6,i,kl})}{2} - \tilde{p}_{i,k} \tilde{q}_{i,l} \tilde{p}_{i',k} \tilde{q}_{i',l} \Big].$$

Using population genetics theory we have

$$\begin{aligned} \mathrm{E}(R_{kl}) &= p_k^2 q_l^2 + \Theta_1 p_k q_l (p_k + q_l - 2p_k q_l), \\ \mathrm{E}(\tilde{P}_{7,kl}) &= p_k^2 q_l^2 + \Theta_1 p_k q_l (p_k + q_l - 2p_k q_l) + p_k q_l (1-p_k)(1-q_l)\Theta_{11}, \\ \mathrm{E}(\tilde{P}_{8,kl}) &= p_k^2 q_l^2 + \Theta_1 p_k q_l (p_k + q_l - 2p_k q_l) + p_k q_l (1-p_k)(1-q_l)\Gamma_{11}, \\ \mathrm{E}(\tilde{P}_{9,kl}) &= p_k^2 q_l^2 + \Theta_1 p_k q_l (p_k + q_l - 2p_k q_l) + p_k q_l (1-p_k)(1-q_l)\Delta_{11}, \quad \text{and} \\ \mathrm{E}(R_{4,kl}) &= p_k q_l (1-p_k)(1-q_l). \end{aligned} \tag{4.40}$$

Using the relations defined in the equations (4.40) and ratio estimation theory we have

$$\begin{aligned} \mathrm{E}(\frac{R_{1,kl}}{R_{4,kl}}) &\approx \Theta_{11}, \\ \mathrm{E}(\frac{R_{2,kl}}{R_{4,kl}}) &\approx \Gamma_{11}, \text{ and} \\ \mathrm{E}(\frac{R_{3,kl}}{R_{4,kl}}) &\approx \Delta_{11}. \end{aligned} \tag{4.41}$$

From the above relations we propose our moment estimators of the parameters based on the frequency data of the gamete A_kB_l. The estimators are

$$\begin{aligned} \hat{\Theta}_{11} &= \frac{R_{1,kl}}{R_{4,kl}}, \\ \hat{\Gamma}_{11} &= \frac{R_{2,kl}}{R_{4,kl}}, \text{ and} \\ \hat{\Delta}_{11} &= \frac{R_{3,kl}}{R_{4,kl}}. \end{aligned} \tag{4.42}$$

The overall estimates of the parameters are

$$\begin{aligned} \hat{\Theta}_{11} &= \frac{\sum_{k=1}^{s_A}\sum_{l=1}^{s_B} R_{1,kl}}{\sum_{k=1}^{s_A}\sum_{l=1}^{s_B} R_{4,kl}}, \\ \hat{\Gamma}_{11} &= \frac{\sum_{k=1}^{s_A}\sum_{l=1}^{s_B} R_{2,kl}}{\sum_{k=1}^{s_A}\sum_{l=1}^{s_B} R_{4,kl}}, \text{ and} \\ \hat{\Delta}_{11} &= \frac{\sum_{k=1}^{s_A}\sum_{l=1}^{s_B} R_{3,kl}}{\sum_{k=1}^{s_A}\sum_{l=1}^{s_B} R_{4,kl}}. \end{aligned} \tag{4.43}$$

In this chapter we have proposed moment estimators for different components for two-locus inbreeding parameters. Now we need to check the performance of our estimators. We have done this in Chapter 6. In that chapter we simulate data under a pure drift model and a both-way mutation model and calculate the estimates of the parameters to find the performance of our methods in terms of bias and standard error.

Chapter 5

Variance of Heterozygosity

5.1 Introduction

The summary statistics heterozygosity and gene diversity are the basic tools for summarizing the amount of genetic variability in a population. The characteristics of population genetic variation are of key interest in studies of evolution and for commercial and conservational breeding programs that seek to develop and maintain a desirable variance. The genetic variance also determines the extent to which a population can adapt to a changing environment. While it is very convenient to characterize the population variation using descriptive measures, heterozygosity and gene diversity, a number of studies published each year reporting these parameters fail to report the sampling errors of their estimates. It would be better to report the sampling properties of the estimators for increasing the quality of statistical inference. In this chapter we consider several approaches to obtaining the variance of sample heterozygosity and discuss the efficiency of these approaches.

The measure of genetic variation was first described by Marshall and Allard (1970) where they coined the word polymorphic index. Later Nei (1973) and Nei and Roychoudhury (1974) worked on determining the variance of polymorphic index. It was in that work that the name 'gene diversity' was first used for this measure. Their approach focused on a single random mating population so they did not consider the

variation due to differentiation between populations. They worked with genetic distance measures to account for the variation between pairs of population. This variation can be summarized as the total variance of heterozygosity or gene diversity.

The purpose of this research is to advocate the regular inclusion and consideration of the variances of the estimators of heterozygosity and gene diversity. We have noticed that in last five years several papers presented estimates of these statistics, but less than 50% of these papers also reported variances with their point estimates. The studies that gave variances of their estimates typically only reported the variance under the within-population scope.

Weir (1989) and Weir et al. (1990) developed extensive theory for the variances of sample gene diversity and heterozygosity, respectively. These papers emphasized that the appropriate expressions for these variances are dependent upon the scope of inference. The total population scope is used when making inferences about a larger group of populations. In this scope, the evolutionary history of the population has been involved. In contrast, the within-population scope infers about the specific populations sampled and individuals can be regarded as being independent. For both total and within scopes, Weir (1989) and Weir et al. (1990) considered the dependencies between loci. Weir (1989) found an expression for the variance of the sample heterozygosity for different mating systems. He used a linear model approach for estimating the variance assuming the loci effects are fixed. Latter Johnson (2004) assumed the loci effects are random and estimated the variance of heterozygosity using the same linear model approach. Because the genotypes are either 1 or 0 based on heterozygote or homozygote we think the generalized linear model would be a better fit. In this research we find the variance of heterozygosity using a generalized linear model approach. We believe the locus effects are fixed rather than random.

Shete (2003) found a uniformly minimum variance unbiased estimator (UMVUE) of gene diversity by correcting the bias of the classical estimator. Shete suggested a bootstrap procedure for estimating the variance of sample gene diversity for a single population at multiple loci. This variance is equivalent to the within-population vari-

ance. Resampling over the population is not recommended as it breaks the population structure. So we can not use bootstrap method for estimating the variance of heterozygosity and gene diversity in total scope. It is very difficult to obtain an exact expression for the total variance of the UMVUE of sample gene diversity because of the complexity of the expression. However, the UMVUE should give similar values to the classical estimator of gene diversity for large samples.

The data sets used in studies are frequently unbalanced because they are surveys of natural populations. Typically in these data sets different numbers of individuals are observed in different population according to the availability of individuals that can be genotyped during the data collection. The statistical properties of sample variances can not be determined easily for unbalanced data. For balanced data the usual ANOVA procedure gives the estimates of variance components with desirable and known statistical properties such as uniformly best unbiasedness. We can not find a unique set of sum of squares that are uniformly best for unbalanced data set. Scientists (Rao and Kleffe, 1988; Searle et al., 1992) advocated MIVQUE (minimum variance quadratic unbiased estimation), ML (maximum likelihood), REML (restricted maximum likelihood) for analyzing an unbalanced data set. The implementation of these methods is available in many statistical software packages including R and SAS.

Here we address a number of issues such as the need for the consideration of the sampling properties of heterozygosity and gene diversity and the best procedures for obtaining the variances. We also illustrate the difference between within-population and total scope variance by re-analysis of one published data set (Olsen and Schaal, 2001). Throughout it will be demonstrated that failure to account for source of variance in population could strongly affect the quality of inferences that can be made in an analysis. The difference between a linear model and a generalized linear model for estimating the total variance of sample heterozygosity will be discussed.

5.2 Estimation of Variance of Heterozygosity

We define some notation and our setup for analysis here and then develop the theory of sample heterozygosity and gene diversity. We have genotype data at L different loci from r independent populations. The l^{th} locus has s_l different allelic forms. We assume that there are n_{il} sampled genotypes at the l^{th} locus in the i^{th} population. We denote the data using an indicator function. Let x_{ijl} takes the value of 1 if the j^{th} individual in the i^{th} population is a heterozygous at the l^{th} locus and 0 otherwise. The sample heterozygosity is the observed frequency of heterozygote in a data set. The term H_i denotes the proportion of heterozygosity in the i^{th} population. $\tilde{H}_{il}$ and $\tilde{H}_i$ are the sample value of heterozygosity of the i^{th} population at the l^{th} locus and over all L loci respectively. The symbol $\sim$ is used here to distinguish the sample values from the population values of parameters. Using indicator variables we can write

$$\tilde{H}_{il} = \frac{\sum_{j=1}^{n_{il}} x_{ijl}}{n_{il}} \quad \text{and} \quad \tilde{H}_i = \frac{\sum_{l=1}^{L}\sum_{j=1}^{n_{il}} x_{ijl}}{\sum_{l=1}^{L} n_{il}}. \tag{5.1}$$

Gene diversity is the frequency of heterozygote expected for a population with Hardy-Weinberg genotype proportion. Let us assume that the d_i is the gene diversity of the i^{th} population. The sample values for locus-specific and overall gene diversity in the i^{th} population are

$$\tilde{d}_{il} = 1 - \sum_{k=1}^{s_l} \tilde{p}_{il,k}^2 \quad \text{and} \quad \tilde{d}_i = \frac{\sum_{l=1}^{L} n_{il}\tilde{d}_{il}}{\sum_{l=1}^{L} n_{il}}, \tag{5.2}$$

where $\tilde{p}_{il,k}$ is the sample frequency of k^{th} allele at locus l in the i^{th} population.

Heterozygosity and gene diversity are both measures that summarize the amount of genetic variation found in populations. While gene diversity quantifies the variation at the allelic level, heterozygosity summarizes the variation at the genotypic level. The observed heterozygosity is much simpler measure than gene diversity but it fails to capture the true extent of genetic variation in populations with a high amount

of selfing or asexual reproduction which is found in many plant species and simpler organisms. In a selfing population alleles tend to associate within individuals. In this case gene diversity would be more appropriate measure to use than heterozygosity.

The general expression for the variance of $\tilde{H}_i$ for both within and total population scope include the variance of the population estimates at single locus, $\tilde{H}_{il}$, and the covariance between these estimates. The exact relation is

$$\mathrm{Var}(\tilde{H}_i) \;=\; \frac{1}{L^2}\Big[\sum_{l=1}^{L}\mathrm{Var}(\tilde{H}_{il}) + \sum_{l\neq l'=1}^{L}\sum \mathrm{Covar}(\tilde{H}_{il},\tilde{H}_{il'})\Big], \tag{5.3}$$

where the $\mathrm{Var}(\tilde{H}_{il})$ and $\mathrm{Covar}(\tilde{H}_{il},\tilde{H}_{il'})$ are either within or total population scope. For within-population scope we can calculate the value of $\mathrm{Var}(\tilde{H}_i)$ by calculating the quantities in equation (5.3). Shete (2003) suggested to estimate the value of $\mathrm{Var}(\tilde{H}_i)$ through bootstrapping the data from the population of interest. On the other hand by using the sample values of $\mathrm{Var}_W(\tilde{H}_{il})$ and $\mathrm{Covar}_W(\tilde{H}_{il},\tilde{H}_{il'})$ and the relation in the equation (5.3) we get the exact expression

$$\mathrm{Var}_W(\tilde{H}_i) \;=\; \frac{1}{L^2}\Big[\sum_{l=1}^{L}\frac{\tilde{H}_{il}(1-\tilde{H}_{il})}{n_{il}} + \sum_{l\neq l'=1}^{L}\sum \frac{\tilde{H}_{ill'}-\tilde{H}_{il}\tilde{H}_{il'}}{n_{ill'}}\Big], \tag{5.4}$$

where $\tilde{H}_{ill'}$ is the observed proportion of individuals that are heterozygous at locus l and l'. $n_{ill'}$ is the number of individuals that have been genotyped at both the loci l and l' in the i^{th} population.

The estimate of total variance for sample heterozygosity is relatively hard to find. We can express the variance of the heterozygosity in terms of variances and covariances of the data. This means we express the variance of sample heterozygosity in terms of variances and the covariances of the indicator random variables. In particular we need to know $\mathrm{Var}_T(x_{i1l})$, $\mathrm{Covar}_T(x_{i1l},x_{i2l})$, $\mathrm{Covar}_T(x_{i1l},x_{i1l'})$ and $\mathrm{Covar}_T(x_{i1l},x_{i2l'})$. Weir (1989) found expressions for these quantities for different mating systems. We can estimate the variances and covariances using a linear or generalized linear model. We

discuss this in the next two sections. A straightforward algebra shows

$$\begin{aligned}
\mathrm{Var}_T(\tilde{H}_{il}) &= \frac{1}{n_{il}}\mathrm{Var}_T(x_{i1l}) + (1-\frac{1}{n_{il}})\mathrm{Covar}_T(x_{i1l}, x_{i2l}) \text{ and} \\
\mathrm{Var}_T(\tilde{H}_i) &= \frac{1}{(\sum_{l=1}^{L} n_{il})^2}\Big[\sum_{l=1}^{L} n_{il}\mathrm{Var}_T(x_{i1l}) + \sum_{l=1}^{L}(n_{il}^2 - n_{il})\mathrm{Covar}_T(x_{i1l}, x_{i2l}) \qquad (5.5) \\
&\quad +(\sum_{l\neq l'=1}^{L}\sum n_{ill'})\mathrm{Covar}_T(x_{i1l}, x_{i1l'}) + \sum_{l\neq l'=1}^{L}\sum(n_{il}n_{il'} - n_{ill'})\mathrm{Covar}_T(x_{i1l}, x_{i2l'})\Big].
\end{aligned}$$

5.2.1 A Linear Model Approach

Our main aim is to estimate $\mathrm{Var}_T(\tilde{H}_{il})$ and $\mathrm{Var}_T(\tilde{H}_i)$ and for that we need to estimate $\mathrm{Var}_T(x_{i1l})$, $\mathrm{Covar}_T(x_{i1l}, x_{i2l})$, $\mathrm{Covar}_T(x_{i1l}, x_{i1l'})$ and $\mathrm{Covar}_T(x_{i1l}, x_{i2l'})$. Weir et al. (1990) used a linear model to obtain the estimates of the total variance components. The linear model used by Weir et al. (1990) was

$$x_{ijl} = \alpha_i + \beta_{j(i)} + \gamma_l + (\alpha\gamma)_{il} + (\beta\gamma)_{jl(i)}, \qquad (5.6)$$

where α_i, $\beta_{j(i)}$, $(\alpha\gamma)_{il}$ $(\beta\gamma)_{jl(i)}$ are random effects of population, individual, population interacted with locus and individual interacted with locus respectively and γ_l is fixed locus effect. The third interaction (population interacted with locus and individual) term is essentially the error term in the linear model. It is important to note the individuals are nested within a population. The variance components for a total population scope are

$$\begin{aligned}
\alpha_i &\overset{iid}{\sim} N(0, \sigma_p^2) && \Rightarrow \mathrm{Var}_T(\alpha_i) = \sigma_p^2, \\
\beta_{j(i)} &\overset{iid}{\sim} N(0, \sigma_{i(p)}^2) && \Rightarrow \mathrm{Var}_T(\beta_{j(i)}) = \sigma_{i(p)}^2, \qquad (5.7) \\
(\alpha\gamma)_{il} &\overset{iid}{\sim} N(0, \sigma_{pl}^2) && \Rightarrow \mathrm{Var}_T((\alpha\gamma)_{il}) = \sigma_{pl}^2, \text{ and} \\
(\beta\gamma)_{jl(i)} &\overset{iid}{\sim} N(0, \sigma_{il(p)}^2) && \Rightarrow \mathrm{Var}_T((\beta\gamma)_{jl(i)}) = \sigma_{il(p)}^2.
\end{aligned}$$

Johnson (2004) considered a fully random model where she assumed the locus effects are random as well. The linear model in the equation (5.6) remains the same but Johnson's assumption would have the effect of adding the variance component $\gamma_l \overset{iid}{\sim} N(0, \sigma_l^2)$ which gives $\mathrm{Var}_T(\gamma_l) = \sigma_l^2$ to those listed in the equation (5.7). The total variance of the heterozygosity can be expressed in terms of σ_p^2, $\sigma_{i(p)}^2$, σ_l^2, σ_{pl}^2 and $\sigma_{il(p)}^2$. Johnson (2004) found the general expressions for the total variance of heterozygosity in terms of the above variance components but we found that her expressions are incorrect. After some algebra we get the total variance of sample heterozygosity using a fully random model. The total variances are

$$\begin{aligned}
\mathrm{Var}_T(\tilde{H}_i) &= \frac{1}{(\sum_{l=1}^{L} n_{il})^2}\Big[\sigma_p^2\big(\sum_{l=1}^{L} n_{il}^2 + \sum_{l\neq l'=1}^{L}\sum n_{il}n_{il'}\big) + \sigma_{i(p)}^2\big(\sum_{l=1}^{L} n_{il} + \sum_{l\neq l'=1}^{L}\sum n_{ill'}\big) \\
&\quad +(\sigma_l^2 + \sigma_{pl}^2)\big(\sum_{l=1}^{L} n_{il}^2\big) + \sigma_{il(p)}^2\big(\sum_{l=1}^{L} n_{il}\big)\Big] \text{ and} \qquad (5.8) \\
\mathrm{Var}_T(\tilde{H}_{il}) &= \big(\sigma_p^2 + \sigma_l^2 + \sigma_{pl}^2\big) + \frac{1}{n_{il}}\big(\sigma_{i(p)}^2 + \sigma_{il(p)}^2\big). \qquad (5.9)
\end{aligned}$$

For a fixed loci effect, the variance component of loci effect is zero, i.e. $\sigma_l^2 = 0$. To find the above expression we first find the variances and covariances in the equation (5.5) in terms of variance components and then find variance of observed heterozygosity.

5.2.2 A Generalized Linear Mixed Model Approach

Since x_{ijl}'s are 0-1 random variables, normality of the errors generally does not hold. The linear model does not guaranty that the estimates of $\mathrm{E}(x_{ijl})$'s always belong to $[0, 1]$. To resolve these problems we propose a generalized linear model. Our model is divided in two steps and the model is

$$\begin{aligned}
&x_{ijl} \mid p_{ijl} \sim \text{Bernoulli}(p_{ijl}) \text{ and} \qquad (5.10) \\
&g(p_{ijl}) = \alpha_i + \beta_{j(i)} + \gamma_l + (\alpha\gamma)_{il}.
\end{aligned}$$

where g is the link function. It is important to note that we do not have the error term in the second step. In theory any function that has the range [0,1] can be used as a link function. One popular choice for the link function is the distribution function of a random variable. Statisticians have shown that different types of distribution functions are useful for different families. In our the case the family is binomial (to be more specific, bernoulli), and obvious choices for the link function are logit, probit and complementary log-log. In our research we consider all the three different link functions and choose the best link function based on model selection criteria. α_i, $\beta_{j(i)}$, $(\alpha\gamma)_{il}$ $(\beta\gamma)_{ijl}$ are random effects of population, individual, and population interacted with locus respectively. The distribution of these effects are given in the equation (5.7). We treat the locus effect as fixed rather than random which implies γ_l's are also fixed locus effects.

Our aim is to get an estimate of $\mathrm{Var}_T(\tilde{H}_{il})$ and $\mathrm{Var}_T(\tilde{H}_i)$. First we estimate the variance components for different random effects and then use the equation (5.5) for estimating the variance of heterozygosity. For a generalized linear model we can not estimate the quantities $\mathrm{Var}_T(x_{11l})$, $\mathrm{Covar}_T(x_{11l}, x_{12l})$, $\mathrm{Covar}(x_{11l}, x_{11l'})$, and $\mathrm{Covar}(x_{11l}, x_{12l'})$ directly. Using the second degree Taylor series approximation we have

$$\begin{aligned}
\mathrm{E}(p_{ijl}) &= f(\gamma_l) + \frac{1}{2} f''(\gamma_l)\Big[\sigma_p^2 + \sigma_{i(p)}^2 + \sigma_{pl}^2\Big], \\
\mathrm{Covar}(p_{ijl}, p_{ij'l}) &= [f'(\gamma_l)]^2\Big[\sigma_p^2 + \sigma_{pl}^2\Big], \\
\mathrm{Covar}(p_{ijl}, p_{ijl'}) &= f'(\gamma_l) f'(\gamma_{l'})\Big[\sigma_p^2 + \sigma_{i(p)}^2\Big], \quad \text{and} \\
\mathrm{Covar}(p_{ijl}, p_{ij'l'}) &= f'(\gamma_l) f'(\gamma_{l'})\sigma_p^2,
\end{aligned} \tag{5.11}$$

where $f = g^{-1}$. Now using the theory of conditional probability we get

$$\begin{aligned}
\mathrm{Var}(x_{ijl}) &= \mathrm{E}(p_{ijl}) - [\mathrm{E}(p_{ijl})]^2, \\
\mathrm{Covar}(x_{ijl}, x_{ij'l}) &= \mathrm{Covar}(p_{ijl}, p_{ij'l}) = [f'(\gamma_l)]^2\Big[\sigma_p^2 + \sigma_{pl}^2\Big], \\
\mathrm{Covar}(x_{ijl}, x_{ijl'}) &= \mathrm{Covar}(p_{ijl}, p_{ijl'}) = f'(\gamma_l)f'(\gamma_{l'})\Big[\sigma_p^2 + \sigma_{i(p)}^2\Big], \text{ and} \\
\mathrm{Covar}(x_{ijl}, x_{ij'l'}) &= \mathrm{Covar}(p_{ijl}, p_{ij'l'}) = f'(\gamma_l)f'(\gamma_{l'})\sigma_p^2.
\end{aligned} \tag{5.12}$$

From the equation (5.12) we get

$$\begin{aligned}
\mathrm{Var}_T(\tilde{H}_{il}) &= \frac{1}{n_{il}}\Big[\mathrm{E}(p_{ijl}) - [\mathrm{E}(p_{ijl})]^2\Big] + (1 - \frac{1}{n_{il}})[f'(\gamma_l)]^2\Big[\sigma_p^2 + \sigma_{pl}^2\Big] \text{ and} \\
\mathrm{Var}_T(\tilde{H}_i) &= \frac{1}{(\sum_{l=1}^{L} n_{il})^2}\Big[\sum_{l=1}^{L}\mathrm{E}(p_{ijl})\big[1 - \mathrm{E}(p_{ijl})\big] + (\sum_{l=1}^{L}(n_{il}^2 - n_{il})[f'(\gamma_l)]^2\big[\sigma_p^2 + \sigma_{pl}^2\big] \\
&\quad + \big[\sum_{l\neq l'=1}^{L}\sum n_{ill'}\big]f'(\gamma_l)f'(\gamma_{l'})\big[\sigma_p^2 + \sigma_{i(p)}^2\big] \\
&\quad + \sum_{l\neq l'=1}^{L}\sum(n_{il}n_{il'} - n_{ill'})f'(\gamma_l)f'(\gamma_{l'})\sigma_p^2\Big].
\end{aligned}$$

Sometimes f' and f'' may be very complicated and not easy to find. In those cases we use the following algorithm for estimating $\mathrm{Var}_T(p_{ijl})$, $\mathrm{Covar}_T(p_{ijl}, p_{ij'l})$, $\mathrm{Covar}_T(p_{ijl}, p_{ijl'})$ and $\mathrm{Covar}_T(p_{ijl}, p_{ij'l'})$.

Step 1. Estimate $\sigma_\alpha^2, \sigma_\beta^2, \sigma_{\alpha\gamma}^2, \sigma_{\beta\gamma}^2$ and γ_l from the model using an appropriate method

Step 2. Generate $\alpha_i, \beta_{j(i)}, (\alpha\gamma)_{il}, (\beta\gamma)_{jl(i)}$ from the distributions given in the equation (5.7). Use the estimated values for the variances.

Step 3. Calculate p_{ijl} from γ_l, $\alpha_i, \beta_{j(i)}, (\alpha\gamma)_{il}, (\beta\gamma)_{jl(i)}$.

Step 4. Repeat Step1-Step3 B times and store the p_{ijl}'s each time. B is a large number, say 1000.

Step 5. Use B independent p_{ijl}'s and get the empirical estimates of $\mathrm{Var}_T(p_{ijl})$,

$\text{Covar}_T(p_{ijl}, p_{ij'l})$, $\text{Covar}_T(p_{ijl}, p_{ijl'})$, and $\text{Covar}_T(p_{ijl}, p_{ij'l'})$.

Because we are using a random number generator, we will not get identical results if we repeat the method but results will be very similar.

It is necessary to have data from more than one population for estimating the variance component σ_p^2 for both linear or generalized linear models. If we do not have data from multiple populations, it has often been suggested that data from individual independent loci can be used for approximating the genetical sampling. This approach can be written as the average of the variance of the single-locus heterozygosity. The variance of single-locus heterozygosity is

$$\text{Var}_T(\tilde{H}_i) \widehat{=} s^2_{H_i} = \frac{\sum_{i=1}^{L}(\tilde{H}_{il} - \tilde{H}_i)^2}{L(L-1)}. \tag{5.13}$$

By taking the expectation it can be seen that this approximation fails to allow the dependencies in the data that are accounted for in the underlying population model. Weir et al. (1990) found the expectation of the variance of single-locus heterozygosity and it is

$$\text{E}_T(s^2_{H_i}) = \text{Var}_T(\tilde{H}_i) + \frac{1}{L(L-1)}\sum_{l=1}^{L}(H_l - H)^2 - \frac{\sum\sum_{l\neq l'}^{L}\text{Cov}_T(\tilde{H}_{il}, \tilde{H}_{il'})}{nL(L-1)}.$$

This approximation is biased for the variance of sample heterozygosity if the heterozygosity at single-locus varies over loci or the loci are dependent. Because of this the single-locus approximation given in equation (5.13) is not a good approximation of total variance of sample heterozygosity.

5.3 Variance Component Methods

For estimating variance components there are several methods such as ML, REML, ANOVA, MIVQUE. For balanced data sets all the methods are identical but for un-

balanced data sets they differ. The methods REML and ML are based on the full form of the probability distribution of the data. The ML approach estimates the parameter by maximizing the joint likelihood for the model parameters given the observed data. REML attempts to remedy drawbacks of the ML method such as negative estimates of variances. It also maximizes the likelihood for the model parameter to estimate the parameters but with some restriction. In contrast ANOVA requires a less restrictive assumption of the form of the first two moments. The ANOVA estimators are the familiar ones obtained by equating observed and expected mean squares from an analysis of variance. For a balanced data set these estimators have many desirable properties. However, unbalanced data destroy all the properties except unbiasedness. The MIVQUE approach does not depend on the distribution of the data. MIVQUE estimates are unbiased, translation invariant and have a minimum variance. This method is desirable as it does not depend on any distribution. Johnson (2004) compared the different methods and found that the estimates of variance components using different methods are very similar. So we can use any method for estimating the variance components. For the linear model we use REML for estimating the variance components.

In the case of the generalized linear mixed model we always use the log-likelihood to estimate the variance components. But this criterion does not have a closed form expression and must be approximated. The default approximation in R is "PQL" or penalized quasi-likelihood. Alternatives are "Laplace" or "AGQ" indicating the Laplacian and adaptive Gaussian quadrature approximations respectively. The "PQL" method is fastest but least accurate. The "Laplace" method is intermediate in speed and accuracy. The "AGQ" method is the most accurate but can be considerably slower than the others. We have used "Laplace" method to approximate the log-likelihood.

Chapter 6

Simulation Studies

6.1 Introduction

In the previous chapters we have proposed several estimators for overall and population-specific descent measures. We also proposed different testing procedures for testing hypotheses about the coancestry coefficient. A generalized linear mixed model is proposed to find the variance of observed heterozygosity. These expressions have been developed under the random population setup. We need to verify the performance of our methods under different values of the parameters. We also need to check the effect of number of loci, number of individuals in populations, mutation rates, age of the current generation, unbalanced sampling on our methods. For this we need to simulate populations that follow our model with different parameter values. Once we generate different independent populations we can estimate the descent measures. We also can estimate the variance of heterozygosity using the genotypic data from more than one population. Since we know the true values of these parameters we can compare the accuracy of our estimates by evaluating the empirical biases and standard errors of our estimators. Using the simulated data we can find the empirical power of our testing procedures.

We generate data assuming two models: (i) a Pure drift model and (ii) a Both-way mutation model. In the next section we describe these two models and the procedures

to generate data using these two models. We use these data sets and estimate the descent measures. The biases and variances of the estimators involve higher order descent measures which are unknown. So we can not find the theoretical value of the biases and variances of our estimators. We evaluate the performance of our estimators by calculating the empirical bias and variance of the estimators. The empirical power of different tests has also been calculated. We give the results in tabular forms.

6.2 Pure Drift Model

Model

Falconer and Mackay (1996) described the assumptions of an idealized population. We construct different populations using these assumptions. To make our simulation study easy, we modify some of these assumptions. We consider a random population setup, and under this all the present populations are evolved from a single ancestral population. The evolution of one population may or may not depend on the evolution of other populations. In this research we assume that the evolution of one population is independent of the evolution of other populations. So the populations are essentially independent. We consider L independent loci. Different loci may have a different number of allelic forms, but in our simulation study we assume each locus has s different allelic forms. The initial reference population is non-inbred and has infinitely many individuals. This reference population is also in Hardy-Weinberg equilibrium at each locus. We assume a random mating system within each population that also includes self-fertilization in a random amount. The loci we are interested in are neutral. The pure drift model assumes that the evolutionary forces such as mutation, migration and selection are assumed not to occur. The generations are distinct and do not overlap. The population size remains the same over generations and it is N. So in a population there are N individuals or $2N$ alleles. Based on these assumptions the theoretical value

of the descent measures at the t^{th} generation are (see equation (1.11))

$$\theta_t = 1-(1-\frac{1}{2N})^t \text{ and} \tag{6.1}$$

$$\gamma_t = 1-\frac{3}{2}(1-\frac{1}{2N})^t+\frac{1}{2}\Big\{(1-\frac{1}{2N})(1-\frac{2}{2N})\Big\}^t. \tag{6.2}$$

We can not control the values of both the parameters θ and γ in a population, but we certainly can control the value of one parameter. Here we control the value of θ. Given a value of θ, θ_0, we can compute the time t_0 such that the population at the generation t_0 will have θ_0 as the value of θ. The value of γ in the population at t_0 can be computed from the equation (6.2). The value of t_0 can be computed by

$$t_0 = \frac{\log(1-\theta_0)}{\log(1-\frac{1}{2N})}. \tag{6.3}$$

Hence, the number of generations needed for specific θ values can be determined. The simulated allelic data in the final population is used to find the performance of our estimators. The pure drift model is presented diagrammatically in Figure 6.1.

Simulation Procedures

Here we describe the simulation procedure for one locus. The description is based on locus A that has s different allelic forms, namely $A_1, A_2, \cdots, A_s$. The frequency of these alleles in the ancestral population is $p_1, p_2, \cdots, p_s$. Since there are L independent loci, we have to repeat the following procedure L times to get data at L independent loci. The simulation procedure for generating a population with coancestry coefficient θ_0 is

1. Set the frequency of the allele A_k in the reference or ancestral population to p_k for $k = 1, 2, \cdots, s$; where $0 < p_k < 1$ and $\sum_{k=1}^{s} p_k = 1$.
2. Generate one allele in the first generation by drawing a random number, u such that $u \sim uni(0, 1)$.

- If $0 \leq u \leq p_1$, then the allelic type is A_1.
- If $\sum_{i=1}^{k-1} p_i < u \leq \sum_{i=1}^{k} p_i$, then the allelic type is A_k where $k = 2, 3, \cdots, s-1$
- If $\sum_{i=1}^{s-1} p_i < u \leq 1$, then the allelic type is A_s.

3. Repeat the above step $2N$ times to get $2N$ alleles for one population in the first generation. Store the allelic type of the k^{th} allele, $k = 1, 2, \cdots, 2N$. This way we generate the first generation population from the ancestral population.
4. Generate the $(t+1)^{th}$ generation from the previous generation, $t \geq 1$
   - Create one allele at the $(t+1)^{th}$ generation by drawing a random number I uniformly from $\{1, 2, \cdots, 2N\}$.
   - Determine the type of the I^{th} allele in the t^{th} generation. This will be the allelic type of one allele in the population at time $(t+1)$.
   - Repeat the above two steps $2N$ times to generate the whole population at the $(t+1)^{th}$ generation.
5. This simulation is stopped at the generation t_0. The value of t_0 can be calculated from the value of the coancestry coefficient, θ_0 (see the equation (6.3)).
6. We randomly draw samples of n alleles with replacement from this simulated population at the final generation and record the allele frequencies of all the alleles. These sampled alleles will be our data.
7. We follow the same procedure (step 1 to step 6) to construct r independent populations with L independent loci.
8. Now we estimate descent measures, calculate test statistics, estimate the variance of heterozygosity from the simulated data.
9. Repeat procedures 1 to 8 for 1000 times independently. We use these 1000 replications to find empirical bias, variance, and power for different methods.

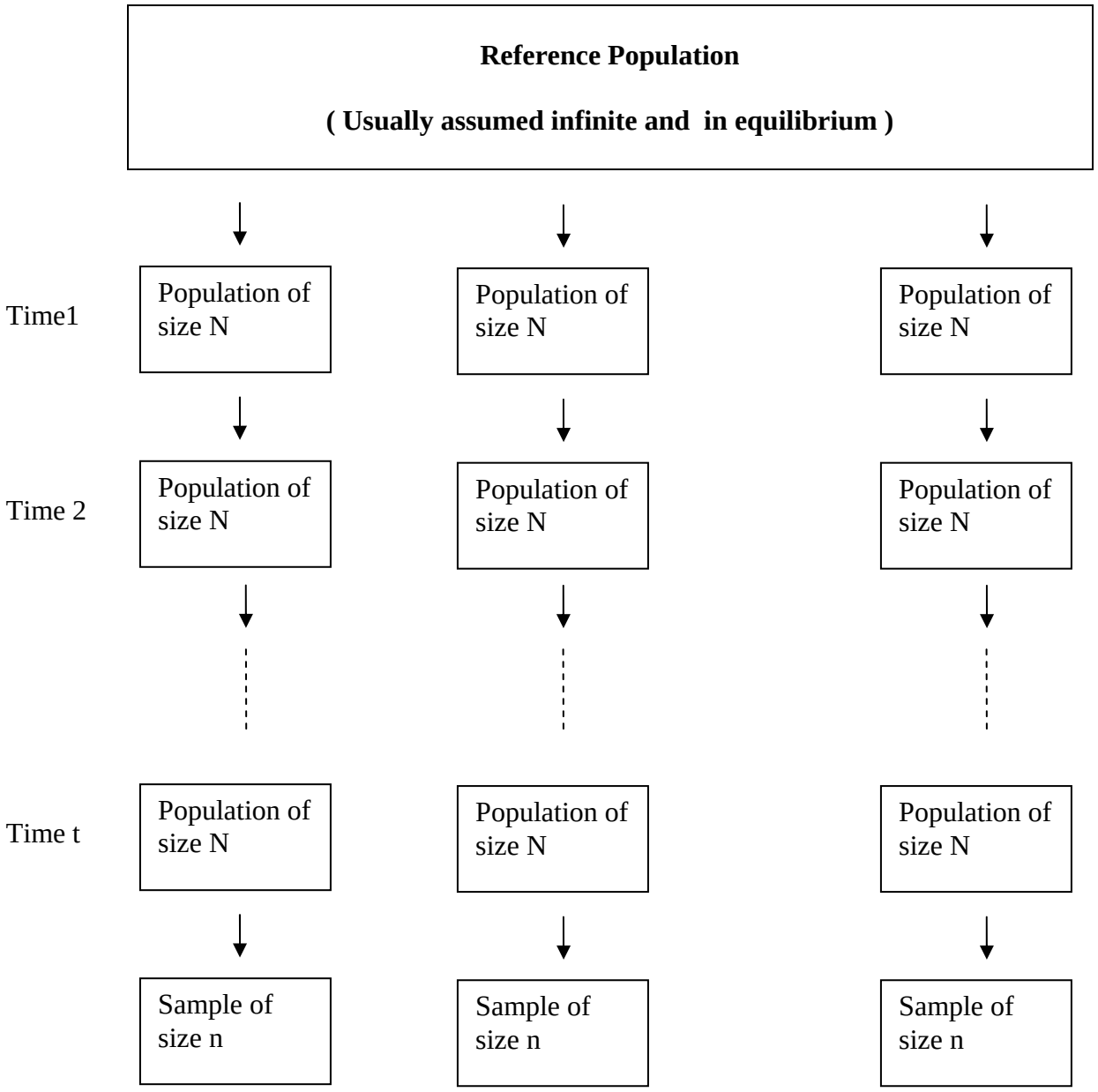

Figure 6.1: The fission and sampling process for the pure drift model and both-way mutation model. This also shows the genetic and statistical sampling involved in genetic data analysis.

6.3 Both-way Mutation Model

Model

The evolutionary process involves forces such as mutation, migration and selection. Therefore it is not realistic to assume a pure drift model. But on the other hand it is extremely complex to include all the evolutionary forces in the model. In this section we include only mutation in the evolutionary process. We assume a both-way mutation model in which any allele can mutate to any other allele. The mutation rate from one allele to another allele does not depend on the forms of the alleles. The rate of mutation is u per generation per allelic form which means that for s alleles at a locus, the total mutation rate per generation is $(s-1)u$. In general, the basic fission process is the same as for the pure drift model, but now the allelic type A_i can mutate to the allelic type A_j with probability u per generation. We can compute the theoretical value of θ and γ given the number of generations and population size using the equation (1.18).

Simulation Procedures

The simulation procedure is similar to the simulation for a pure random drift model. But only "step 4" needs to be modified:

Create the $(t+1)^{th}$ generation from previous generation:

- Create one allele in the $(t+1)^{th}$ generation by drawing a random number I uniformly from $\{1, 2, \cdots, 2N\}$
- Determine the type of the I^{th} allele in generation t. WLOG assumes that the allelic form of this allele is A_i. Now this may be mutated to any other allele with rate u. Now generate a random number, $\eta \sim uni(0,1)$. Create an allele in the $(t+1)^{th}$ generation as follows
  - If $(j-1)u \leq \eta < ju$, then the allelic type is A_j. $j = 1, 2, \cdots, s$;
  - If $su \leq \eta \leq 1$, then the allelic type is A_i.
- Repeat the above two steps $2N$ times to generate the $(t+1)^{th}$ generation

6.4 Results

In this section we discuss the results. This discussion compares the performance our estimators and test statistics with other proposed measures. We have five different topics, (i) different estimators of θ and γ, (ii) different estimators of population-specific θ and γ (iii) estimators for two-locus descent measures, (iv) power of different test statistics for testing $\theta = 0$, and (v) estimators of the variance of observed heterozygosity. In the next five sections we have discussed these five topics separately. For estimation and testing of descent measures we generate 5 independent populations and each population has 500 individuals or 1000 alleles in each generation. For estimating the one-locus descent measures we sample 300, 325, 350, 375 and 400 alleles from 5 populations and for the two-locus descent measures we sample the whole population which means that there are 1000 sampled alleles in each population. We consider different sample sizes for testing hypotheses about the coancestry coefficient. The sample sizes are given in the caption of the tables where we describe the results. For estimating the variance of heterozygosity we generate 10 populations with 150 individuals in each population and at each generation. In this case, we sample 30 alleles from each population. In the caption of the tables we have used various notations and here we describe the notations. L is the number of independent loci and s is the number of different alleles at each locus. $\mathbf{p}$ is a vector which denotes the expected frequencies of different allelic forms. t denotes the age of the current population while ρ is the recombination rate between the locus A and the locus B. $\mathscr{D}_{11}$ is the linkage disequilibrium between the allele A_1 and the allele B_1.

Overall Descent Measures

In this subsection we discuss the performance of the different estimators of θ and γ. There are several estimators available for θ, but there is no estimator for γ. Weir and Cockerham's moment estimator is a well established moment estimator of θ and it has been compared with other estimators. We compare our new moment estimators with

Weir and Cockerham's estimator.

The Table 6.1, Table 6.2, and Table 6.3 show that all the moment estimators have a small bias but a large standard deviation. The information from independent loci increases the accuracy of the estimators by decreasing the standard deviations of the estimators. The information from independent loci also decreases the magnitude of the biases of the estimators. For example, in Table 6.2, when $\theta = 0.101$ the bias of $\hat{\theta}_{WC}$ based on a single-locus and 20 loci are 0.00125 and 0.00013 respectively. Sometimes an increment in the number of loci may cause the change of sign in the bias of an estimator, but it may also decrease the magnitude of the bias. For example, in Table 6.1, when $\theta = 0.101$ the bias of $\hat{\theta}_{1,M}$ based on a single locus is -0.00224. If we take information from 20 independent loci then the bias is 0.00003. Theoretically, the bias of the estimators should be negative but we can not calculate the theoretical value of the biases as they involve unknown higher order descent measures, δ, $\Delta_{2,2}$ etc. In some situations, we have found that an increase in the number of locus may cause an increment in the bias although the increment is very small. For example, in Table 6.3 when $\theta = 0.051$ the biases of $\hat{\theta}_{1,P}$ based on a single locus and 20 loci are 0.00018 and 0.00114. This should not happen in theory and we conclude that one or two outliers in the simulation process cause these discrepancies. In general, the biases of the moment estimators are small for different parameter values and decrease with an increasing number of loci. On the other hand, the information from different loci always reduces the standard deviations of the estimators. For example, in Table 6.2 when $\theta = 0.051$ then the standard deviations of $\hat{\theta}_{4,P}$ based on a single locus and 20 loci are 0.01961 and 0.00439 respectively. So we can control the value of standard deviations of the moment estimators by taking information from different loci. The above facts are true for a pure drift model and a both-way mutation model.

The biases and standard deviations of estimators heavily depend on the number of different allelic forms at each locus. The simulation studies show that the value of biases and standard deviations of different estimators decrease as the number of allelic forms per locus increases. This fact does not depend on the allelic frequencies at a

locus. If we have 2 alleles with frequencies 0.7 and 0.3 at each locus, then the bias and standard deviation of $\hat{\theta}_{3,P}$ based on a single locus are -0.00079 and 0.06432 when the true value of θ is 0.101 (see Table 6.1). Under the same setup with four allelic forms at each locus with equal frequencies, the bias and standard deviation of $\hat{\theta}_{3,P}$ are -0.00070 and 0.03640 (see Table 6.2). If we generate data using a both-way mutation model, then the standard errors of the MOM estimators also decreases with an increasing number of loci.

The performance of the estimators does depend on the population model but the relative performance of the estimators remain similar. In the simulation studies we consider two different models. Under the same setup, Table 6.1 and Table 6.2 show the results of different estimators of θ under a pure drift model while Table 6.3 displays the results for a both-way mutation model. We compare Table 6.1, Table 6.2, and Table 6.3 and find that all the tables show similar results. The results are not exact as the true values of θ are not the same for both the models.

Now we compare different estimators with Weir-Cockerhams's moment estimator. As we discussed earlier, all the estimators have negligible biases and we do not compare our estimators in terms of biases. We compare our estimators in terms of their standard deviations. The newly proposed moment estimators of θ, $\hat{\theta}_{1,M}$ and $\hat{\theta}_{2,M}$ are less efficient than $\hat{\theta}_{WC}$ in terms of standard deviation. In Table 6.1 when $\theta = 0.011$, the standard deviations of $\hat{\theta}_{WC}$ based on a single locus and 20 loci are 0.00927 and 0.00214. Under the same setup as above, the standard deviations of $\hat{\theta}_{1,M}$ are 0.01170 and 0.00267 while the standard deviations of $\hat{\theta}_{2,M}$ are 0.01170 and 0.00266. In Table 6.3 when $\theta = 0.051$, the standard deviations of $\hat{\theta}_{WC}$ based on a single locus and 20 loci are 0.03597 and 0.00834. Under the same setup as above, the standard deviations of $\hat{\theta}_{1,M}$ are 0.03971 and 0.00931 while the standard deviations of $\hat{\theta}_{2,M}$ are 0.03971 and 0.00909. The new proposed estimators $\hat{\theta}_{1,P}$ and $\hat{\theta}_{2,P}$ are very competitive with $\hat{\theta}_{WC}$. Sometimes the first two estimators are more accurate than $\hat{\theta}_{WC}$, but in some other cases we get the opposite result. In Table 6.1 when $\theta = 0.011$, the standard deviations of $\hat{\theta}_{WC}$ based on a single locus and 20 loci are 0.00927 and 0.00214. Under the same setup, the standard

deviations of $\hat{\theta}_{1,P}$ are 0.00927 and 0.00213 while the standard deviations of $\hat{\theta}_{2,P}$ are 0.00927 and 0.00213. In Table 6.2 when $\theta = 0.051$, the standard deviations of $\hat{\theta}_{WC}$ based on a single locus and 20 loci are 0.02101 and 0.00463. Under the same setup, the standard deviations of $\hat{\theta}_{1,P}$ are 0.02091 and 0.00463 while the standard deviations of $\hat{\theta}_{2,P}$ are 0.02117 and 0.00467. In the first example $\hat{\theta}_{1,P}$ and $\hat{\theta}_{2,P}$ are doing well, but in the second example $\hat{\theta}_{WC}$ is more accurate. So in general we cannot prefer some estimators over others. The $\hat{\theta}_{3,P}$ has lesser standard deviation than $\hat{\theta}_{1,M}$, $\hat{\theta}_{2,M}$, $\hat{\theta}_{1,P}$ and $\hat{\theta}_{2,P}$. In most of the cases $\hat{\theta}_{3,P}$ is more accurate than $\hat{\theta}_{WC}$ in terms of variance. But in some cases, the $\hat{\theta}_{3,P}$ has a larger standard error than $\hat{\theta}_{WC}$. For example, in Table 6.3 when $\theta = 0.011$, the standard errors of $\hat{\theta}_{WC}$ based on a single locus and 20 loci are 0.00998 and 0.00206. Under the same setup, the standard deviations of $\hat{\theta}_{3,P}$ are 0.00999 and 0.00205. The $\hat{\theta}_{3,P}$ has more variance than $\hat{\theta}_{WC}$.

Now we compare the performance of $\hat{\theta}_{4,P}$ with $\hat{\theta}_{WC}$. From our simulation studies we have found that the estimator $\hat{\theta}_{4,P}$ is at least as accurate as $\hat{\theta}_{WC}$. In most cases $\hat{\theta}_{4,P}$ has smaller variance than $\hat{\theta}_{WC}$, but sometimes they may have the same variance. In Table 6.2 when $\theta = 0.101$, the standard deviations of $\hat{\theta}_{WC}$ based on a single locus and 20 loci are 0.03848 and 0.00897. Under the same setup the standard deviations of $\hat{\theta}_{4,P}$ are 0.03478 and 0.00796. In both the cases $\hat{\theta}_{4,P}$ has less standard deviations than $\hat{\theta}_{WC}$. On the other hand, in Table 6.3 when $\theta = 0.011$, the standard deviations of $\hat{\theta}_{WC}$ based on a single locus and 20 loci are 0.00998 and 0.00206. Under the same setup, the standard deviations of $\hat{\theta}_{4,P}$ are 0.00998 and 0.00205. In the first case both the estimators have the same standard error while in the last case $\hat{\theta}_{4,P}$ has a lower standard error. So we conclude $\hat{\theta}_{4,P}$ is a better estimator than $\hat{\theta}_{WC}$ in terms of standard deviation.

All the moment estimators have negligible biases. It is not possible to find the best estimator in terms of bias. But we will not worry about the biases as they are negligible for all estimators. The estimators $\hat{\theta}_{1,M}$ and $\hat{\theta}_{2,M}$ are less accurate than $\hat{\theta}_{WC}$ in terms of the variance. The estimators $\hat{\theta}_{1,P}$, $\hat{\theta}_{2,P}$, and $\hat{\theta}_{3,P}$ are very competitive with $\hat{\theta}_{WC}$ in terms of variance. In most of the situations $\hat{\theta}_{1,P}$, $\hat{\theta}_{2,P}$, and $\hat{\theta}_{3,P}$ have smaller variances than $\hat{\theta}_{WC}$ but sometimes $\hat{\theta}_{WC}$ may have a larger variance. On the other hand, $\hat{\theta}_{4,P}$ is

at least as efficient as $\hat{\theta}_{WC}$ in terms of variance. Most of cases $\hat{\theta}_{4,P}$ has smaller variance than $\hat{\theta}_{WC}$ but sometimes they may have the same variance. So we advocate using $\hat{\theta}_{4,P}$ as a moment estimator of θ as this estimator has smaller standard error.

Now we discuss the performance of different estimators of γ. The biases of the estimators are relatively small. If we gather information from different loci then the bias become lesser. For example, in Table 6.4 when the number of alleles per locus is 2 and $\gamma = 0.00017$, the biases of $\hat{\gamma}_{1,M}$ are -0.00009 and $6.4e^{-6}$ respectively for 1 and 20 loci. In the same setup the biases of $\hat{\gamma}_{1,P}$ based on a single locus and 20 loci are -0.00010 and $2.6e^{-6}$ respectively. The standard deviations of the moment estimators are relatively large, but we can decrease the standard deviations by including independent loci. For example, in Table 6.5 when the number of alleles per locus is 2 and $\gamma = 0.01464$, the standard deviations of $\hat{\gamma}_{1,P}$ based on a single locus and 20 loci are 0.14925 and 0.01467 respectively. In the same setup, the biases of $\hat{\gamma}_{2,P}$ are 0.14925 and 0.05753. When the true value of γ is small then all four estimators of γ are equivalent. For example, Table 6.4 shows that when the number of alleles per locus is 4 and $\gamma = .00017$ all the four estimators have equal standard deviations for a different number of loci. On the other hand, Table 6.4 shows when the number of alleles per locus is 2 and $\gamma = .00017$, the standard deviations of different estimators are not exactly equal but they are very close to each other. As the true value of γ increases we find differences in the performance of the estimators. In general, $\hat{\gamma}_{1,P}$ and $\hat{\gamma}_{1,M}$ are more stable and better estimators than the other two for large values of γ. These two estimators have lower standard deviation than the other two estimators. For example, Table 6.4 when $\gamma = 0.01464$ and there are 4 alleles per locus, the standard deviations of $\hat{\gamma}_{1,P}$ based on a single locus and 20 loci are 0.02368 and 0.00592. Under the same setup, the standard deviations of $\hat{\gamma}_{1,M}$ are 0.02373 and 0.00593, standard deviations of $\hat{\gamma}_{2,M}$ are 0.02879 and 0.00944, and standard deviations of $\hat{\gamma}_{2,P}$ are 0.03437 and 0.03132. When the true value of $\gamma = 0.00377$, then the standard deviations of $\hat{\gamma}_{1,M}$ and $\hat{\gamma}_{1,P}$ are also slightly smaller than the other two estimators. Table 6.4 shows that most of the cases the standard deviation of $\hat{\gamma}_{1,P}$ is smaller than the standard deviation of $\hat{\gamma}_{1,M}$ but

in some cases they are the same. So we prefer $\hat{\gamma}_{1,P}$ over $\hat{\gamma}_{1,M}$ as an estimator of γ. In conclusion, we advocate using $\hat{\gamma}_{1,P}$ to estimate γ.

Population-specific Descent Measures

In this section we discuss about the performance of four different estimators of population specific θ. The moment estimators of population-specific descent measures have small biases. Moreover adding information from different independent loci also generally decreases the biases. But the standard deviations of the estimators are relatively large. In fact, the magnitude of the standard deviations the estimators based on a single locus are larger than the estimates. The variances of the estimators decrease as the number of independent loci increases. For example, in Table 6.5 the standard deviations of $\hat{\theta}_{4,P}$ based on a single locus and 20 loci in the first population are 0.17187 and 0.03912 which shows that the variance decreases as number loci increases. Among the four estimators, $\hat{\theta}_{4,P}$ is the most efficient in terms of variance and $\hat{\theta}_{2,P}$ is the least efficient. The other three estimators, $\hat{\theta}_{WC}$, $\hat{\theta}_{1,P}$ and $\hat{\theta}_{3,P}$ are very competitive. We cannot order the last three estimators according to their efficiency in terms of standard error. For example, Table 6.5 shows that the standard deviations of $\hat{\theta}_{4,P}$ based on a single locus and 20 loci in the 2^{nd} population are 0.18636 and 0.04314. Under the same setup, the standard deviations of $\hat{\theta}_{WC}$ are 0.21815 and 0.04859, the standard deviations of $\hat{\theta}_{1,P}$ are 0.21430 and 0.04785, the standard deviations of $\hat{\theta}_{2,P}$ are 0.37568 and 0.08182 and the standard deviations of $\hat{\theta}_{3,P}$ are 0.21430 and 0.04989. This shows that the estimator $\hat{\theta}_{4,P}$ has least variance. Table 6.5 and Table 6.6 show that the estimator $\hat{\theta}_{4,P}$ has the least variance in all the cases. In terms of biases all the estimators are very competitive but the biases are so small we can neglect them. So we suggest using the estimator $\hat{\theta}_{4,P}$ for estimating population specific θ. The biases and variances of the estimators of population-specific θ are larger than the biases and variances of the estimators of overall θ. This is obvious because in estimating population-specific θ heavily depends on a single population although they need more than one population.

The estimator of population-specific γ does not work properly when there are two or three different allelic forms per locus. When there are four or more allelic forms the estimators work well. As with the other moment estimators, the estimators of population-specific γ have less biases but large standard deviations. If the true value of γ is very small then the biases of the estimators are not negligible, but for moderate and large values of γ the biases become smaller. We can reduce the biases by collecting information from independent loci. In general the bias of $\hat{\gamma}_{1,M}$ is smaller than other estimators, but in a few cases the bias of $\hat{\gamma}_{1,P}$ is smaller than $\hat{\gamma}_{1,M}$. In these cases the difference in biases are very small and the biases are negligible. So we conclude that $\hat{\gamma}_{1,M}$ is a better estimator than others in terms of bias. $\hat{\gamma}_{1,M}$ has the least variance among four different estimators. This is true for all populations and different number of loci. The standard deviation of $\hat{\gamma}_{1,M}$ decreases as we increase the number of independent loci. The bias and variance of the estimator of population-specific γ are larger than the bias and variance of the estimator of overall γ. We suggest using $\hat{\gamma}_{1,M}$ for estimating population-specific γ. The supporting results are given in Table 6.7 in tabular form.

Two-locus Descent Measures

In this section we discuss the performance of the estimators of $\Theta^1\mathscr{D}_{kl}$, ${}_1\Theta\mathscr{D}_{kl}$, ${}_1\Theta^1_1\mathscr{D}_{kl}$ and ${}_1\Gamma^1_1\mathscr{D}_{kl}$. We also have found the moment estimators of Θ_{11}, Γ_{11} and Δ_{11} when the ancestral population is in linkage equilibrium but we skip the discussion on them. When we have two alleles in each locus, then $\mathscr{D}_{11}$ can characterize the full linkage structure by the relation $\mathscr{D}_{11}$ = -$\mathscr{D}_{12}$ = -$\mathscr{D}_{21}$ = $\mathscr{D}_{22}$. For two alleles per locus we found the estimators of the composite parameters follow the above relation, for example $\widehat{\Theta^1\mathscr{D}_{11}} = -\widehat{\Theta^1\mathscr{D}_{12}} = -\widehat{\Theta^1\mathscr{D}_{21}} = \widehat{\Theta^1\mathscr{D}_{22}}$. If we have more than two allelic forms in each locus then we do not get the above relations but we have found that the estimate of compound parameters corresponding different $\mathscr{D}_{kl}$ behave similarly. So without loss of generality here we present the behavior of the estimators of $\Theta^1\mathscr{D}_{11}$, ${}_1\Theta\mathscr{D}_{11}$, ${}_1\Theta^1_1\mathscr{D}_{11}$ and ${}_1\Gamma^1_1\mathscr{D}_{11}$ only.

Table 6.8 describes the values of the estimators when $\rho = 0.01, 0.05, 0.10$ and $\mathscr{D}_{11} = 0.16, 0.08$. The estimators of $\Theta^1\mathscr{D}_{11}$, ${}_1\Theta\mathscr{D}_{11}$ and ${}_1\Theta_1^1\mathscr{D}_{11}$ work very well for the setup in Table 6.8. The biases of these estimators are very small compare to the true value of the parameters. The standard deviations are of the same order of the true value of the parameters. For $t = 105$ the standard deviations are comparatively larger than the standard deviations of the estimates when $t = 52$. The estimators of ${}_1\Gamma_1^1\mathscr{D}_{11}$ have biases of the same order as true value and the standard deviations are comparatively large. For example, in Table 6.8 when $t = 52$, $\mathscr{D}_{11} = 0.16$ and $\rho = 0.01$, the estimate is 0.00033 while bias and standard deviation of the estimate are -0.00013 and 0.00384 respectively. In fact, the estimate of ${}_1\Gamma_1^1\mathscr{D}_{11}$ is always negative while the true value of the compound parameter is positive. The reason behind this the true value of the compound parameter is very close to 0 and ratio estimator does not work well in this situation. The value of the parameter increases as t increases and we get positive estimate of the parameter for large generation values.

In the Table 6.9 we present the results under different parameter values when each locus has four different forms of alleles with equal frequencies. The interpretation of these results are more or less the same as the interpretation of the results of Table 6.8. The biases of the estimators are small compare to the true value of the parameters except for one or two cases. The bias of ${}_1\Gamma_1^1\mathscr{D}_{11}$ is relatively larger than the others. In fact, for $t = 11$ the estimates of ${}_1\Gamma_1^1\mathscr{D}_{11}$ are negative while the true value of the parameter is positive. The standard deviations of the moment estimator $\Theta^1\mathscr{D}_{11}$ are of the order of the true value of the parameters. In the other cases the standard deviations are relatively larger. The standard deviations of the estimators increase with time t. We observed that the estimator of ${}_1\Gamma_1^1\mathscr{D}_{11}$ does not work properly.

Testing hypotheses about θ

In this section we compare the empirical power of newly proposed parametric bootstrap test with the non-parametric bootstrap test. The comparison has been done under

small sample sizes. For large sample sizes we compare our newly proposed F test with the test statistic proposed by Li (1996). We have performed the tests at a 5% level. We have found that the empirical level of the parametric bootstrap is very close to 5% all the time. In some situations when there are small number of loci (say, 5), the empirical significance level of non-parametric bootstrap test may well exceed the theoretical significance level. The empirical power of the parametric bootstrap test and large sample test increases with the true value of θ and the sample sizes. The power of both the bootstrap tests increases as the number of loci increases. For example, when $\theta = 0.05070$, the sample sizes are 10, and there are two alleles per loci with equal frequencies, then the power of the parametric bootstrap test are 0.409, 0.611 and 0.838 for 5, 10 and 20 loci. For the above setup the powers of non-parametric bootstrap test are 0.348, 0.516 and 0.787. Table 6.10 shows that the power of the tests also increases with the number of allelic form per locus. For example, when $\theta = 0.01095$ and there are 10 loci with two allelic form per locus with equal frequencies, the power calculations of the parametric bootstrap test are 0.130, 0.332 and 0.553 for 10, 25 and 40 sampled alleles in each population. Under the same setup, when there are four alleles per locus with equal frequencies the power of the parametric bootstrap test are 0.200, 0.595 and 0.892 for 10, 25 and 40 sampled alleles in each population. As we increase the number of sampled alleles in each locus, the power of the non-parametric tests also increase. For example, when $\theta = 0.10062$, two alleles per locus and there are 5 loci the the power of the non-parametric bootstrap test are 0.658, 0.952 and 0.995. Table 6.10 shows that the parametric bootstrap method is better than non-parametric bootstrap method in terms of powers. In most of the cases, parametric bootstrap method has more power than non-parametric bootstrap method. When θ is very large and there are more loci then both the method have a power 1. So we recommend using the parametric bootstrap method when the sample sizes are small.

The power of the two F tests is given in Table 6.11. For a 5% significance level, both the F-test have approximately 5% power when the null hypothesis is true, showing that the tests have a correct size. The power of the F-tests increases when the true value

of θ increases. The power of the F-tests increases when the number of sampled alleles in each population increases. When there are two alleles in each locus then both tests have approximately equal power. In most of the situations they have equal power, but in some cases the test proposed by Li (1996) has a slightly more power than our test procedure but these differences are negligible. The power of the tests does not vary with the expected allele frequencies. So when we have two allelic forms in each locus then the performances of both the tests are very similar to each other. When we have more than two alleles, then Li's F-test does not exist but our F-test works fine. Table 6.11 shows that the power of our F-test increases when the number of alleles per locus increases and this increment is significantly large. For example, when true value of θ is 0.011 and the locus has two alleles with equal frequencies then the power of our test are 0.313, 0.566, and 0.825 for 100, 200 and 500 sampled alleles. For the same setup when there are five alleles with equal frequencies then the powers are 0.671, 0.937 and 0.998. Throughout this simulation studies, we have noticed that the power of our F test is more than Li's test, if we have more allelic forms per loci. So our F-test can be used to get more power when there are more than two alleles per locus. So we suggest to use our method for more than two alleles and for two alleles we can use any one F-statistics.

Variance of heterozygosity

In this part we compare the performance of our newly proposed generalized linear model with the existing linear model approach for estimating the variance of sample heterozygosity. It is very hard to find the true value of the variance of observed heterozygosity in the total-population sense. We generate data with independent loci and each locus has the same allele frequencies. This gives the expectation of the statistic $s^2_{\tilde{H}_i}$ is equal to the total variance of the observed heterozygosity. Using this approach we have an idea about the true value of the variance. The another approach for having an idea about the true value of the total variance is to store the estimates

of the observed heterozygosity for each monte carlo simulation. Then find the sample variance of the observed heterozygosity and get an empirical estimate of the variance of the heterozygosity. The Table 6.12 shows that the two empirical estimates of the variance of the heterozygosity agree with each other, but the true value is unknown to us. The estimate of sample heterozygosity using a generalized linear model is very close to its empirical value and the linear model has a large bias. The magnitude of the standard error of both the estimates is the same. The above two facts suggest that the MSE of the GLMME is smaller than the MSE of LMME. Both the estimates decrease as the true value of the population differentiation increases. The Table 6.12 shows that the estimates of the within-population variance is always smaller than the empirical estimate of the total-population variance of sample heterozygosity. This is because the within variance fails to incorporate the variance due to the population differentiation.

We use a model selection criteria to find which model is a better fit. There are many model selection methods, and we have used Akaike information criterion (AIC), Bayesian information criteria (BIC), and Deviance information criterion (DIC). We have found that in almost all the cases our generalized linear mixed model is a better fit than the linear model. We also have found that in general, the logit link function for generalized mixed model works better than other link functions.

The Akaike information criterion (AIC), developed by Hirotsugu Akaike in 1971 and is grounded in the concept of entropy. The AIC is an operational way of trading off the complexity of an estimated model against how well the model fits the data. In the general case, $AIC = 2k - 2\ln(L)$, where k is the number of parameters, and L is the likelihood function. Given any two estimated models, the model with the lower value of AIC is the one to be preferred. The AIC is a decreasing function of RSS, the goodness of fit, and an increasing function of k.

The Bayesian information criteria (BIC) penalizes free parameters more strongly than does the AIC. In the general case, $BIC = k\ln(n) - 2\ln(L)$, where k is the number of parameters, and L is the likelihood function. It is important to keep in mind that the BIC can be used to compare estimated models only when the numerical values

of the dependent variable are identical for all estimates being compared. The models being compared need not be nested, unlike the case when models are being compared using an F or likelihood ratio test.

The deviance information criterion (DIC) is a hierarchical modeling generalization of the AIC and BIC. It is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation. Like AIC and BIC it is an asymptotic approximation as the sample size becomes large. It is only valid when the posterior distribution is approximately multivariate normal. Define the deviance as $D(\beta) = -2\log(p(y \mid \beta)) + C$, where y is the data, β are the unknown parameters of the model and $p(y \mid \beta)$ is the likelihood function. C is a constant that cancels out in all calculations that compare different models, and which therefore does not need to be known. The expectation $\bar{D} = E_\beta[D(\beta)]$ is a measure of how well the model fits the data; the larger this is, the worse the fit. The effective number of parameters of the model is computed as $p_D = \bar{D} - D(\bar{\beta})$, where $\bar{\beta}$ is the expectation of β. The larger this is, the easier it is for the model to fit the data. The deviance information criterion is calculated as $DIC = p_D + \bar{D}$. The models with a smaller DIC should be preferred to models with larger DIC. Models are penalized both by the value of $\bar{D}$, which favors a good fit, but also by the effective number of parameters p_D. Since $\bar{D}$ will decrease as the number of parameters in a model increases, the p_D term compensates for this effect by favoring models with a smaller number of parameters.

The advantage of DIC over other criteria, for Bayesian model selection, is that the DIC is easily calculated from the samples generated by a MCMC simulation. AIC and BIC require calculating the likelihood at its maximum over β, which is not readily available from the MCMC simulation. But to calculate DIC, simply compute $\bar{D}$ as the average of $D(\beta)$ over the samples of β, and $D(\bar{\beta})$ as the value of D evaluated at the average of the samples of β. Then the DIC follows directly from these approximations.

Table 6.1: Different estimators of θ. Parameters: $s = 2$; $p = (0.7, 0.3)$; $L = 1, 20$; The data is generated with a pure drift model.

		$L = 1$			$L = 20$		
θ	Method	Average	Bias	SD	Average	Bias	SD
.011	$\hat{\theta}_{WC}$	.01038	-.00056	.00927	.01101	.00006	.00214
	$\hat{\theta}_{1,M}$	.01019	-.00076	.01170	.01106	.00011	.00267
	$\hat{\theta}_{2,M}$	.01019	-.00076	.01170	.01097	.00002	.00266
	$\hat{\theta}_{1,P}$	.01036	-.00059	.00927	.01100	.00005	.00213
	$\hat{\theta}_{2,P}$	.01036	-.00059	.00927	.01100	.00005	.00213
	$\hat{\theta}_{3,P}$	.01036	-.00059	.00927	.01097	.00002	.00211
	$\hat{\theta}_{4,P}$	.01037	-.00058	.00926	.01097	.00002	.00211
.051	$\hat{\theta}_{WC}$	.04861	-.00209	.03549	.05031	-.00038	.00784
	$\hat{\theta}_{1,M}$	.04878	-.00191	.03904	.05051	-.00019	.00902
	$\hat{\theta}_{2,M}$	.04878	-.00191	.03904	.04980	-.00089	.00885
	$\hat{\theta}_{1,P}$	.04873	-.00196	.03552	.05036	-.00033	.00782
	$\hat{\theta}_{2,P}$	.04873	-.00196	.03552	.05036	-.00033	.00782
	$\hat{\theta}_{3,P}$	.04873	-.00196	.03552	.04994	-.00076	.00770
	$\hat{\theta}_{4,P}$	.04864	-.00205	.03520	.04992	-.00078	.00764
.101	$\hat{\theta}_{WC}$	.09952	-.00110	.06322	.10083	.00020	.01466
	$\hat{\theta}_{1,M}$	.09839	-.00224	.07052	.10065	.00003	.01695
	$\hat{\theta}_{2,M}$	.09839	-.00224	.07052	.09737	-.00325	.02768
	$\hat{\theta}_{1,P}$	.09983	-.00079	.06342	.10086	.00024	.01460
	$\hat{\theta}_{2,P}$	.09983	-.00079	.06342	.10086	.00024	.01460
	$\hat{\theta}_{3,P}$	.09983	-.00079	.06342	.09924	-.00139	.01416
	$\hat{\theta}_{4,P}$	.09986	-.00077	.06304	.09908	-.00154	.01397

Table 6.2: Different estimators of θ. Parameters: $s = 4$; $p = (0.25, 0.25, 0.25, 0.25)$; $L = 1, 20$; The data is generated with a pure drift model.

		$L = 1$			$L = 20$		
θ	Method	Average	Bias	SD	Average	Bias	SD
.011	$\hat{\theta}_{WC}$	.01073	-.00021	.00544	.01094	-1.8e-6	.00126
	$\hat{\theta}_{1,M}$	.01073	-.00021	.00556	.01095	2.5e-6	.00127
	$\hat{\theta}_{2,M}$	.01064	-.00031	.00550	.01084	-.00010	.00126
	$\hat{\theta}_{1,P}$	.01073	-.00021	.00541	.01094	-6.7e-7	.00126
	$\hat{\theta}_{2,P}$	.01073	-.00022	.00541	.01094	-5.4e-6	.00127
	$\hat{\theta}_{3,P}$	.01071	-.00024	.00539	.01090	-.00004	.00125
	$\hat{\theta}_{4,P}$	.01071	-.00023	.00536	.01090	-.00004	.00124
.051	$\hat{\theta}_{WC}$	.05067	-.00002	.02101	.05023	-.00046	.00463
	$\hat{\theta}_{1,M}$	.05083	.00014	.02171	.05025	-.00045	.00483
	$\hat{\theta}_{2,M}$	.05021	-.00048	.02129	.04943	-.00126	.00468
	$\hat{\theta}_{1,P}$	.05067	-.00002	.02091	.05025	-.00045	.00463
	$\hat{\theta}_{2,P}$	.05050	-.00020	.02117	.05023	-.00047	.00467
	$\hat{\theta}_{3,P}$	.05022	-.00048	.02023	.04963	-.00106	.00449
	$\hat{\theta}_{4,P}$	.05003	-.00066	.01961	.04958	-.00112	.00439
.101	$\hat{\theta}_{WC}$	.10187	.00125	.03848	.10075	.00013	.00897
	$\hat{\theta}_{1,M}$	.10090	.00028	.03945	.10079	.00017	.00965
	$\hat{\theta}_{2,M}$	.09939	-.00123	.03839	.09805	-.00257	.00923
	$\hat{\theta}_{1,P}$	.10198	.00135	.03835	.10073	.00011	.00891
	$\hat{\theta}_{2,P}$	.10272	.00210	.04038	.10070	.00007	.00894
	$\hat{\theta}_{3,P}$	.09992	-.00070	.03640	.09843	-.00219	.00846
	$\hat{\theta}_{4,P}$	.09938	-.00124	.03478	.09796	-.00267	.00796

Table 6.3: Different estimators of θ. Parameters: $s = 2$; $p = (0.7, 0.3)$; $L = 1, 20$; The data is generated with a both-way mutation model.

		$L = 1$			$L = 20$		
θ	Method	Average	Bias	SD	Average	Bias	SD
.011	$\hat{\theta}_{WC}$	.01069	-.00019	.00998	.01095	.00007	.00206
	$\hat{\theta}_{1,M}$	.01054	-.00033	.01214	.01095	.00007	.00254
	$\hat{\theta}_{2,M}$	.01054	-.00033	.01214	.01085	-.00003	.00252
	$\hat{\theta}_{1,P}$	.01073	-.00014	.00999	.01095	.00007	.00206
	$\hat{\theta}_{2,P}$	.01073	-.00014	.00999	.01095	.00007	.00206
	$\hat{\theta}_{3,P}$	.01073	-.00014	.00999	.01092	.00004	.00205
	$\hat{\theta}_{4,P}$	.01073	-.00014	.00998	.01092	.00004	.00205
.051	$\hat{\theta}_{WC}$	.05022	.00006	.03597	.05127	.00111	.00834
	$\hat{\theta}_{1,M}$	.04967	-.00049	.03971	.05117	.00102	.00931
	$\hat{\theta}_{2,M}$	.04967	-.00049	.03971	.05042	.00027	.00909
	$\hat{\theta}_{1,P}$	.05034	.00018	.03590	.05130	.00114	.00833
	$\hat{\theta}_{2,P}$	.05034	.00018	.03590	.05130	.00114	.00833
	$\hat{\theta}_{3,P}$	.05034	.00018	.03590	.05085	.00069	.00819
	$\hat{\theta}_{4,P}$	.05032	.00016	.03570	.05083	.00067	.00812
.101	$\hat{\theta}_{WC}$	.10482	.00424	.06893	.10606	.00548	.01624
	$\hat{\theta}_{1,M}$	.10230	.00171	.08197	.10589	.00531	.01825
	$\hat{\theta}_{2,M}$	.10230	.00171	.08197	.10278	.00219	.02678
	$\hat{\theta}_{1,P}$	.10474	.00415	.06873	.10605	.00547	.01629
	$\hat{\theta}_{2,P}$	.10474	.00415	.06873	.10605	.00547	.01629
	$\hat{\theta}_{3,P}$	.10474	.00415	.06873	.10411	.00353	.01565
	$\hat{\theta}_{4,P}$	.10461	.00403	.06774	.10392	.00333	.01544

Table 6.4: Estimators of γ. Parameters: $L = 1$ and 20; $s = 2$ and 4; $p = (0.7, 0.3)$ and $(0.25, 0.25, 0.25, 0.25)$; The data is generated with a pure drift model.

			$L = 1$			$L = 20$		
allele	γ	Method	Average	Bias	SD	Average	Bias	SD
2	.00017	$\hat{\gamma}_{1,M}$	.00008	-.00009	.00268	.00018	6.4e-6	.00067
		$\hat{\gamma}_{2,M}$	.00008	-.00009	.00268	.00018	5.1e-6	.00067
		$\hat{\gamma}_{1,P}$	.00007	-.00010	.00264	.00018	6.6e-6	.00066
		$\hat{\gamma}_{2,P}$	.00007	-.00010	.00264	.00018	2.6e-6	.00067
	.00377	$\hat{\gamma}_{1,M}$	.00447	.00070	.02295	.00376	-9.0e-6	.00498
		$\hat{\gamma}_{2,M}$	.00447	.00070	.02295	.00356	-.00020	.00537
		$\hat{\gamma}_{1,P}$	.00444	.00067	.02248	.00379	.00002	.00499
		$\hat{\gamma}_{2,P}$	.00444	.00067	.02248	.00357	-.00020	.00537
	.01464	$\hat{\gamma}_{1,M}$	.00918	-.00546	.07136	.01455	-.00009	.01297
		$\hat{\gamma}_{2,M}$	.00918	-.00546	.07136	.00909	-.00555	.06614
		$\hat{\gamma}_{1,P}$	.00948	-.00516	.07074	.01456	-.00008	.01294
		$\hat{\gamma}_{2,P}$	.00948	-.00516	.07074	.01353	-.00111	.03532
4	.00017	$\hat{\gamma}_{1,M}$	.00016	-.00002	.00098	.00019	.00001	.00024
		$\hat{\gamma}_{2,M}$	.00016	-.00002	.00098	.00019	.00001	.00024
		$\hat{\gamma}_{1,P}$	.00016	-.00001	.00098	.00019	.00001	.00024
		$\hat{\gamma}_{2,P}$	.00016	-.00002	.00097	.00019	.00001	.00024
	.00377	$\hat{\gamma}_{1,M}$	.00422	.00045	.00850	.00371	-.00006	.00193
		$\hat{\gamma}_{2,M}$	.00412	.00035	.00846	.00354	-.00022	.00190
		$\hat{\gamma}_{1,P}$	.00419	.00042	.00840	.00370	-.00007	.00193
		$\hat{\gamma}_{2,P}$	.00406	.00030	.00834	.00351	-.00026	.00189
	.01464	$\hat{\gamma}_{1,M}$	.01412	-.00052	.02318	.01456	-.00008	.00571
		$\hat{\gamma}_{2,M}$	.01311	-.00153	.02381	.01309	-.00155	.00583
		$\hat{\gamma}_{1,P}$	.01411	-.00053	.02311	.01455	-.00009	.00570
		$\hat{\gamma}_{2,P}$	.01300	-.00164	.02372	.01301	-.00163	.00577

Table 6.5: Estimators of θ_i. Parameters: $s = 2$; $p = (0.7, 0.3)$; $L = 1$ and 20; The data is generated with a pure drift model.

			$L = 1$			$L = 20$		
Pop	θ_i	Method	Average	Bias	SD	Average	Bias	SD
1	.049	$\hat{\theta}_{WH}$	.03802	-.01078	.20344	.04979	.00100	.04449
		$\hat{\theta}_{1,P}$	.03993	-.00887	.19707	.04971	.00092	.04335
		$\hat{\theta}_{2,P}$	.03451	-.01428	.34796	.05087	.00208	.07485
		$\hat{\theta}_{3,P}$	.03993	-.00887	.19707	.04473	-.00407	.04498
		$\hat{\theta}_{4,P}$	.05016	.00137	.17187	.05470	.00591	.03912
2	.063	$\hat{\theta}_{WH}$	.04939	-.01357	.21815	.06531	.00235	.04859
		$\hat{\theta}_{1,P}$	.05084	-.01212	.21430	.06522	.00226	.04785
		$\hat{\theta}_{2,P}$	.03995	-.02301	.37568	.06745	.00448	.08182
		$\hat{\theta}_{3,P}$	.05084	-.01212	.21430	.06230	-.00066	.04989
		$\hat{\theta}_{4,P}$	.05558	-.00738	.18636	.06700	.00404	.04314
3	.077	$\hat{\theta}_{WH}$	.07604	-.00088	.23051	.07608	-.00084	.05165
		$\hat{\theta}_{1,P}$	.07685	-7.2e-5	.22960	.07597	-.00096	.05139
		$\hat{\theta}_{2,P}$	.08331	.00639	.39388	.07547	-.00145	.08848
		$\hat{\theta}_{3,P}$	.07685	-7.2e-5	.22960	.07523	-.00169	.05368
		$\hat{\theta}_{4,P}$	.07729	.00037	.19787	.07527	-.00166	.04647
4	.091	$\hat{\theta}_{WH}$	.08833	-.00234	.24322	.09183	.00116	.05253
		$\hat{\theta}_{1,P}$	.08832	-.00235	.24482	.09168	.00101	.05284
		$\hat{\theta}_{2,P}$	.08371	-.00696	.41958	.09152	.00085	.08980
		$\hat{\theta}_{3,P}$	.08832	-.00235	.24482	.09257	.00189	.05539
		$\hat{\theta}_{4,P}$	.08187	-.00880	.21001	.08713	-0.0035	.04740
5	.104	$\hat{\theta}_{WH}$	.11188	.00767	.24591	.10138	-.00284	.05671
		$\hat{\theta}_{1,P}$	.11126	.00705	.24950	.10119	-.00303	.05769
		$\hat{\theta}_{2,P}$	.12572	.02151	.42000	.09845	-.00577	.09659
		$\hat{\theta}_{3,P}$	.11126	.00705	.24950	.10357	-.00064	.06014
		$\hat{\theta}_{4,P}$	.10222	-.00199	.21280	.09414	-.01008	.05102

Table 6.6: Estimators of θ_i. Parameters: $s = 4$; $p = (0.25, 0.25, 0.25, 0.25)$; $L = 1, 20$; The data is generated with a both-way mutation model.

			$L = 1$			$L = 20$		
Pop	θ_i	Method	Average	Bias	SD	Average	Bias	SD
1	.049	$\hat{\theta}_{WH}$	.05551	.00357	.04703	.05368	.00174	.01043
		$\hat{\theta}_{1,P}$	.05551	.00357	.04703	.05368	.00174	.01043
		$\hat{\theta}_{2,P}$	.05439	.00245	.06368	.05404	.00211	.01402
		$\hat{\theta}_{3,P}$	.06307	.01113	.04171	.06023	.00829	.00892
		$\hat{\theta}_{4,P}$	.07052	.01858	.03081	.06868	.01674	.00655
2	.063	$\hat{\theta}_{WH}$	.07135	.00626	.05998	.06729	.00220	.01249
		$\hat{\theta}_{1,P}$	.07135	.00626	.05998	.06729	.00220	.01249
		$\hat{\theta}_{2,P}$	.06863	.00353	.07256	.06779	.00270	.01526
		$\hat{\theta}_{3,P}$	.07433	.00923	.04995	.06966	.00457	.01043
		$\hat{\theta}_{4,P}$	.07658	.01149	.03487	.07388	.00879	.00746
3	.077	$\hat{\theta}_{WH}$	.08446	.00662	.06765	.08183	.00399	.01432
		$\hat{\theta}_{1,P}$	.08446	.00662	.06765	.08183	.00399	.01432
		$\hat{\theta}_{2,P}$	.08750	.00966	.08108	.08143	.00359	.01659
		$\hat{\theta}_{3,P}$	.08142	.00358	.05321	.08030	.00246	.01184
		$\hat{\theta}_{4,P}$	.08142	.00358	.03739	.07964	.00180	.00809
4	.091	$\hat{\theta}_{WH}$	.09541	.00522	.07334	.09480	.00461	.01646
		$\hat{\theta}_{1,P}$	.09541	.00522	.07334	.09480	.00461	.01646
		$\hat{\theta}_{2,P}$	.09706	.00687	.08364	.09406	.00387	.01922
		$\hat{\theta}_{3,P}$	.08951	-.00068	.05863	.08934	-.00085	.01283
		$\hat{\theta}_{4,P}$	.08581	-.00438	.04057	.08468	-.00552	.00869
5	.104	$\hat{\theta}_{WH}$	.10570	.00354	.08241	.10815	.00599	.01790
		$\hat{\theta}_{1,P}$	.10570	.00354	.08241	.10815	.00599	.01790
		$\hat{\theta}_{2,P}$	.10612	.00396	.09151	.10839	.00624	.02095
		$\hat{\theta}_{3,P}$	.09728	-.00488	.06505	.09826	-.00390	.01385
		$\hat{\theta}_{4,P}$	.08989	-.01227	.04352	.08978	-.01238	.00930

Table 6.7: Different estimators of γ_i. Parameters: $s = 4$; $p = (0.25, 0.25, 0.25, 0.25)$; $L = 1, 20$; The data is generated with a pure drift model.

			$L = 1$			$L = 20$		
Pop	γ_i	Method	Average	Bias	SD	Average	Bias	SD
1	.0035	$\hat{\gamma}_{1,M}$	.00364	.00015	.01297	.00354	4.8e-5	.00327
		$\hat{\gamma}_{2,M}$	.00723	.00374	.07341	.00859	.00510	.01947
		$\hat{\gamma}_{1,P}$	.00266	-.00084	.05406	.00367	.00018	.01274
		$\hat{\gamma}_{2,P}$	.00720	.00371	.07342	.00858	.00509	.01947
2	.0058	$\hat{\gamma}_{1,M}$	.00642	.00063	.02059	.00569	-.00010	.00455
		$\hat{\gamma}_{2,M}$	.00924	.00344	.07704	.00836	.00257	.02102
		$\hat{\gamma}_{1,P}$	.00774	.00194	.06072	.00606	.00027	.01333
		$\hat{\gamma}_{2,P}$	.00928	.00348	.07705	.00835	.00255	.02103
3	.0086	$\hat{\gamma}_{1,M}$	.00770	-.00091	.02582	.00885	.00024	.00630
		$\hat{\gamma}_{2,M}$	.00691	-.00171	.08256	.00730	-.00132	.02021
		$\hat{\gamma}_{1,P}$	.00678	-.00184	.06400	.00877	.00015	.01497
		$\hat{\gamma}_{2,P}$	.00684	-.00178	.08260	.00730	-.00132	.02021
4	.0119	$\hat{\gamma}_{1,M}$	.01268	.00075	.03902	.01153	-.00039	.00783
		$\hat{\gamma}_{2,M}$	.01027	-.00165	.09240	.00867	-.00326	.02154
		$\hat{\gamma}_{1,P}$	.01279	.00087	.07753	.01163	-.00029	.01685
		$\hat{\gamma}_{2,P}$	.01037	-.00156	.09246	.00867	-.00325	.02154
5	.0157	$\hat{\gamma}_{1,M}$	.01527	-.00042	.03907	.01544	-.00024	.00979
		$\hat{\gamma}_{2,M}$	.00939	-.00630	.09004	.01031	-.00538	.02248
		$\hat{\gamma}_{1,P}$	.01383	-.00186	.07584	.01538	-.00031	.01760
		$\hat{\gamma}_{2,P}$	.00931	-.00638	.09008	.01029	-.00540	.02248

Table 6.8: Estimates of two-locus descent measures. Parameters: $s = 2$; $p = (0.7, 0.3)$; The data is generated with a pure drift model.

			$t = 52$			$t = 106$		
$\mathscr{D}_{11}$	ρ	Method	Average	Bias	SD	Average	Bias	SD
0.16	0.01	$\widehat{\Theta^1\mathscr{D}_{11}}$	.09522	.00026	.01797	.05906	.00337	.02246
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00617	-.00027	.00627	.01008	.00021	.01231
		$\widehat{{}_1\Theta^1_1\mathscr{D}_{11}}$	.00484	-4.7e-5	.00665	.00635	.00036	.01956
		$\widehat{{}_1\Gamma^1_1\mathscr{D}_{11}}$	.00033	-.00013	.00384	.00112	-.00025	.01815
	0.05	$\widehat{\Theta^1\mathscr{D}_{11}}$	.01334	.00201	.01091	.00387	.00202	.01413
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00306	.00019	.00576	.00346	.00050	.01104
		$\widehat{{}_1\Theta^1_1\mathscr{D}_{11}}$	.00089	.00020	.00383	.00074	.00036	.01407
		$\widehat{{}_1\Gamma^1_1\mathscr{D}_{11}}$	.00015	-3.9e-5	.00274	.00059	.00024	.01113
	0.1	$\widehat{\Theta^1\mathscr{D}_{11}}$	.00233	.00137	.00866	.00185	.00048	.01234
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00171	.00017	.00555	.00175	.00025	.01096
		$\widehat{{}_1\Theta^1_1\mathscr{D}_{11}}$	7.4e-5	-3.7e-5	.00314	-.00151	-.00168	.08314
		$\widehat{{}_1\Gamma^1_1\mathscr{D}_{11}}$	2.6e-5	-6.5e-5	.00244	-.00030	-.00047	.04286
0.08	0.01	$\widehat{\Theta^1\mathscr{D}_{11}}$	.04704	-.00044	.01792	.02947	.00162	.02175
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00317	-3.2e-6	.00556	.00484	-9.8e-5	.01135
		$\widehat{{}_1\Theta^1_1\mathscr{D}_{11}}$	.00239	-5.4e-5	.00553	.00388	.00088	.02348
		$\widehat{{}_1\Gamma^1_1\mathscr{D}_{11}}$	.00014	-8.7e-5	.00292	.00071	3.3e-5	.01218
	0.05	$\widehat{\Theta^1\mathscr{D}_{11}}$	.00662	.00096	.01060	.00113	.00020	.01440
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00145	9.5e-6	.00548	.00128	-.00020	.01110
		$\widehat{{}_1\Theta^1_1\mathscr{D}_{11}}$	.00027	-7.2e-5	.00321	.00043	.00024	.01084
		$\widehat{{}_1\Gamma^1_1\mathscr{D}_{11}}$	5.6e-5	-3.8e-5	.00213	.00019	1.4e-5	.00956
	0.1	$\widehat{\Theta^1\mathscr{D}_{11}}$	.00129	.00081	.00875	.00087	.00019	.01245
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00090	.00014	.00542	.00097	.00023	.01068
		$\widehat{{}_1\Theta^1_1\mathscr{D}_{11}}$	.00014	8.2e-5	.00286	.00024	.00016	.00841
		$\widehat{{}_1\Gamma^1_1\mathscr{D}_{11}}$	4.6e-5	6.2e-7	.00222	.00038	.00030	.00891

Table 6.9: Estimates of two-locus descent measures. Parameters: $s = 4$; $p = (0.25, 0.25, 0.25, 0.25)$; The data is generated with a pure drift model.

			$t = 52$			$t = 106$		
$\mathscr{D}_{11}$	ρ	Method	Average	Bias	SD	Average	Bias	SD
.0625	0.01	$\widehat{\Theta^1\mathscr{D}_{11}}$	.03874	.00165	.01729	.02479	.00304	.02049
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00268	.00020	.00492	.00455	.00070	.01022
		$\widehat{{}_1\Theta_1^1\mathscr{D}_{11}}$	.00205	.00014	.00398	.00311	.00077	.00913
		$\widehat{{}_1\Gamma_1^1\mathscr{D}_{11}}$	.00022	4.3e-5	.00191	.00093	.00040	.00667
	0.05	$\widehat{\Theta^1\mathscr{D}_{11}}$	.00507	.00064	.00976	.00068	-4.4e-5	.01274
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00126	.00013	.00492	.00087	-.00029	.00988
		$\widehat{{}_1\Theta_1^1\mathscr{D}_{11}}$	.00031	4.6e-5	.00241	.00190	.00175	.05743
		$\widehat{{}_1\Gamma_1^1\mathscr{D}_{11}}$	7.4e-5	1.4e-7	.00169	.00162	.00148	.04894
	0.1	$\widehat{\Theta^1\mathscr{D}_{11}}$	.00089	.00052	.00779	.00058	4.9e-5	.01080
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00060	-4.2e-7	.00492	.00042	-.00017	.00921
		$\widehat{{}_1\Theta_1^1\mathscr{D}_{11}}$	9.8e-5	5.5e-5	.00205	.00014	7.9e-5	.00486
		$\widehat{{}_1\Gamma_1^1\mathscr{D}_{11}}$	2.2e-5	-1.4e-5	.00172	.00021	.00014	.00454
.03125	0.01	$\widehat{\Theta^1\mathscr{D}_{11}}$	.01834	-.00021	.01614	.01150	.00063	.01966
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00130	5.9e-5	.00512	.00170	-.00023	.00987
		$\widehat{{}_1\Theta_1^1\mathscr{D}_{11}}$	.00111	.00016	.00395	.00121	4.0e-5	.00872
		$\widehat{{}_1\Gamma_1^1\mathscr{D}_{11}}$	.00013	3.8e-5	.00176	.00022	-5.2e-5	.00672
	0.05	$\widehat{\Theta^1\mathscr{D}_{11}}$	.00306	.00085	.00962	.00075	.00038	.01262
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00079	.00023	.00485	.00074	.00016	.00981
		$\widehat{{}_1\Theta_1^1\mathscr{D}_{11}}$	.00014	5.4e-6	.00213	5.8e-5	-1.6e-5	.00672
		$\widehat{{}_1\Gamma_1^1\mathscr{D}_{11}}$	2.0e-5	-1.7e-5	.00148	.00014	7.5e-5	.00711
	0.1	$\widehat{\Theta^1\mathscr{D}_{11}}$	.00023	4.6e-5	.00730	1.9e-5	-.00025	.01058
		$\widehat{{}_1\Theta\mathscr{D}_{11}}$	.00023	-6.9e-5	.00476	.00026	-3.1e-5	.00931
		$\widehat{{}_1\Theta_1^1\mathscr{D}_{11}}$	-4.9e-5	-7.1e-5	.00185	6.7e-5	3.8e-5	.00534
		$\widehat{{}_1\Gamma_1^1\mathscr{D}_{11}}$	-4.4e-5	-6.2e-5	.00147	.00012	8.7e-5	.00517

Table 6.10: The comparison between the empirical powers of newly proposed parametric bootstrap test with the non-parametric bootstrap test. We consider equal allele frequencies and equal sample sizes. The data is generated with a pure drift model.

			$L = 5$		$L = 10$		$L = 20$	
θ	allele	n	NP Boot	P Boot	NP Boot	P Boot	NP Boot	P Boot
0.00		10	0.065	0.057	0.043	0.054	0.041	0.050
	2	25	0.047	0.042	0.045	0.044	0.044	0.049
		40	0.044	0.042	0.043	0.053	0.041	0.048
		10	0.080	0.047	0.057	0.056	0.049	0.049
	4	25	0.082	0.052	0.056	0.047	0.055	0.055
		40	0.062	0.044	0.047	0.049	0.044	0.045
.01095		10	0.114	0.110	0.111	0.130	0.151	0.187
	2	25	0.190	0.224	0.260	0.332	0.430	0.506
		40	0.312	0.365	0.455	0.553	0.713	0.797
		10	0.166	0.139	0.190	0.200	0.278	0.303
	4	25	0.386	0.389	0.540	0.595	0.814	0.842
		40	0.609	0.669	0.861	0.892	0.981	0.985
.05070		10	0.348	0.409	0.516	0.611	0.787	0.838
	2	25	0.750	0.847	0.946	0.973	0.999	0.999
		40	0.911	0.966	0.994	0.999	1	1
		10	0.696	0.714	0.903	0.932	0.997	0.998
	4	25	0.991	0.994	1	1	1	1
		40	1	1	1	1	1	1
.10062		10	0.658	0.780	0.893	0.935	0.993	0.997
	2	25	0.952	0.989	1	1	1	1
		40	0.995	0.999	1	1	1	1
		10	0.970	0.981	1	1	1	1
	4	25	1	1	1	1	1	1
		40	1	1	1	1	1	1

Table 6.11: The comparison between the empirical powers of newly proposed chi square test statistics with Li's test procedure. We consider equal sample sizes for different populations. The data is generated with a pure drift model.

			$n = 100$		$n = 200$		$n = 500$	
θ	allele	frequency	Our Test	Li's Test	Our Test	Li's Test	Our Test	Li's Test
.000	2	equal	0.055	0.055	0.052	0.053	0.044	0.044
	2	0.7 & 0.3	0.053	0.053	0.041	0.041	0.049	0.050
	2	0.9 & 0.1	0.042	0.042	0.053	0.053	0.048	0.048
	3	equal	0.050	NA	0.048	NA	0.058	NA
	4	equal	0.046	NA	0.047	NA	0.052	NA
	5	equal	0.048	NA	0.039	NA	0.056	NA
.011	2	equal	0.313	0.317	0.566	0.568	0.825	0.825
	2	0.7 & 0.3	0.316	0.321	0.539	0.543	0.823	0.823
	2	0.9 & 0.1	0.343	0.348	0.557	0.561	0.837	0.838
	3	equal	0.469	NA	0.774	NA	0.964	NA
	4	equal	0.566	NA	0.895	NA	0.994	NA
	5	equal	0.671	NA	0.937	NA	0.998	NA
.051	2	equal	0.827	0.828	0.937	0.937	0.985	0.985
	2	0.7 & 0.3	0.861	0.862	0.951	0.951	0.990	0.990
	2	0.9 & 0.1	0.861	0.862	0.951	0.951	0.990	0.990
	3	equal	0.968	NA	0.997	NA	1	NA
	4	equal	0.997	NA	1	NA	1	NA
	5	equal	1	NA	1	NA	1	NA
.101	2	equal	0.944	0.944	0.983	0.983	0.995	0.995
	2	0.7 & 0.3	0.951	0.952	0.985	0.985	0.996	0.996
	2	0.9 & 0.1	0.940	0.940	0.990	0.990	0.999	0.999
	3	equal	0.997	NA	1	NA	1	NA
	4	equal	1	NA	1	NA	1	NA
	5	equal	1	NA	1	NA	1	NA

Table 6.12: Relationship between several different expressions for the variance of heterozygosity ($\tilde{H}_i$). The terms given are heterozygosity, within and total-population standard deviation of observed heterozygosity, single-locus and empirical approximation of standard deviation of heterozygosity. The data is generated from 10 populations at 5 independent loci using a Pure drift model.

						$SD_T(\tilde{H}_i)$			
						linear		glm+logit	
frequency	θ	$\tilde{H}_i$	$SD_W(\tilde{H}_i)$	$\sqrt{s^2_{\tilde{H}_i}}$	empirical	mean	sd	mean	sd
0.5 & 0.5	.000	0.497	0.040	0.041	0.038	0.189	0.010	0.093	0.017
	.052	0.472	0.040	0.043	0.041	0.182	0.012	0.074	0.018
	.101	0.447	0.040	0.047	0.046	0.174	0.013	0.060	0.015
	.151	0.424	0.039	0.055	0.053	0.167	0.015	0.051	0.014
0.6 & 0.4	.000	0.477	0.040	0.043	0.039	0.186	0.010	0.092	0.017
	.052	0.454	0.040	0.044	0.043	0.179	0.012	0.074	0.018
	.101	0.431	0.039	0.053	0.051	0.171	0.014	0.059	0.015
	.151	0.408	0.038	0.059	0.058	0.164	0.016	0.051	0.014
0.7 & 0.3	.000	0.419	0.039	0.044	0.040	0.176	0.010	0.090	0.016
	.052	0.399	0.039	0.054	0.051	0.170	0.013	0.072	0.017
	.101	0.378	0.037	0.065	0.060	0.163	0.015	0.058	0.015
	.151	0.359	0.036	0.069	0.068	0.157	0.016	0.049	0.014
0.8 & 0.2	.000	0.318	0.037	0.044	0.039	0.163	0.010	0.086	0.015
	.052	0.305	0.036	0.060	0.056	0.159	0.013	0.068	0.016
	.101	0.285	0.034	0.072	0.068	0.151	0.015	0.054	0.014
	.151	0.274	0.034	0.076	0.075	0.147	0.016	0.045	0.013

6.5 Application on HapMap Data

In this section we analyze the HapMap data to characterize the human genome. We used our estimators to estimate the descent measures in different human populations. We measure genome-wide descent measure and show the heterogeneity among genome regions.

The International HapMap Project is an organization whose goal is to develop a haplotype map of the human genome (the HapMap), which will describe the common patterns of human genetic variation. The project is a collaboration among researchers at academic centers, non-profit biomedical research groups and private companies in Canada, China, Japan, Nigeria, the United Kingdom, and the United States. The HapMap is expected to be a key resource for researchers to use to find genes affecting health, disease and responses to drugs and environmental factors. The information produced by the project is freely available to researchers around the world.

The International HapMap Project officially started with a meeting on October 27 to 29, 2002, and was expected to take about three years. It comprises two phases and the complete data for Phase I was published on October 27, 2005. Completion of the HapMap will enable future work. The Japanese teams will study 300,000 people to identify haplotypes that match 47 diseases, and the British will attempt to genotype patients with diabetes, bipolar disorder, rheumatoid arthritis, cardiovascular disease and other common diseases.

Most of the common haplotypes occur in all human populations. However, their frequencies differ among populations. Therefore, data from several populations are needed to choose tag SNPs. In the HapMap project scientists collected SNP data from different populations. Pilot studies have found sufficient differences in haplotype frequencies among population samples from Nigeria (Yoruba), Japan, China and the United States (residents with ancestry from Northern and Western Europe) to warrant developing the HapMap with large-scale analysis of haplotypes in these populations. The HapMap developed from information obtained from these populations should be

useful for all populations in the world. Specifically, the DNA samples for the HapMap will come from a total of 270 people. The groups consist of the Yoruba people in Ibadan, Nigeria (30 adult-and-both-parents trios), Japanese in Tokyo (45 unrelated individuals), Han Chinese in Beijing (45 unrelated individuals) and the U.S. residents of northern and western European ancestry (30 trios). These numbers of samples will allow the project to find almost all haplotypes with frequencies of 5% or higher. For the HapMap project, researchers sequenced all 22 chromosomes human autosomes. In the following we provide the length and the number of SNPs for each chromosome.

Results

We compute values of descent measures using only those SNPs that were found to be segregating in all population samples. We have used $\hat{\theta}_{4,P}$ and $\hat{\gamma}_{1,P}$ for estimating overall θ and γ. For estimating population-specific θ we have used $\hat{\theta}_{4,P,i}$. We also use this data set to find if the coancestry coefficient of human populations is strictly positive. Our estimates are calculated for all markers separately and also for all markers in all the 5-Mb windows centered on each SNP in the autosomal genome. The numbers of markers used are shown in Table 6.13.

A genome-wide survey of descent measures shows that there is substantial variation, even among SNPs that are very close to each other. The estimate of overall θ based on single-locus marker values from four samples has a distribution very much like the χ^2 distribution with two or three degrees of freedom. The extreme noisiness in single-locus estimates is demonstrated in Table 6.14, where the standard deviations of the values for each chromosome are seen to be about the same size as the means. The variation is even higher for the population-specific values. The noisiness of single-locus estimates can be reduced by combining data from several adjacent markers. We have chosen to use 5-Mb windows to clarify the graphical presentations. The distribution of these (approximately) 1000-locus values is close to a normal distribution. Table 6.14 shows that the chromosomal standard deviations have dropped substantially. Even

for the relatively large window size of 5 Mb there is substantial variation along each chromosome, suggesting that values of descent measures are genome region-specific.

Table 6.13: Chromosome lengths and numbers of markers segregating in all populations

Chromosome	Length(Mb)	No. markers	Chromosome	Length(Mb)	No. markers
1	246.02	46,170	2	243.36	54,649
3	199.16	39,741	4	191.64	35,988
5	180.75	35,649	6	170.67	40,993
7	158.41	26,444	8	146.29	46,834
9	136.31	36,513	10	134.89	29,488
11	134.29	26,767	12	131.96	25,156
13	96.17	22,427	14	87.05	17,520
15	81.78	15,430	16	89.88	14,111
17	81.70	14,317	18	76.11	24,697
19	63.58	10,355	20	63.58	12,115
21	36.95	12,639	22	34.76	11,353

Because the usual values of descent measures are averages over populations, they may obscure signatures of past evolutionary events such as selective sweeps; so, we have also estimated population-specific values using our newly proposed estimators. These values show much more variation, and the very large standard deviations shown in Table 6.14 indicate that single-locus values are not reliable. The 5-Mb window values, however, have coefficients of variation that are always < 0.5. The overall estimate of γ based on a single-locus is extremely variable over the genome. The variability of the estimate of γ reduces if we consider estimates based on a 5 Mb window and the standard deviation is about the same size as the means. In the previous section we have seen that the population-specific γ is stable only when there are at least four allelic forms per locus. Since our SNP data has two alleles per locus, we do not estimate the

population-specific γ.

The correlation of pairs of single-locus statistics reflects the linkage disequilibrium between those pairs (Weir et al., 2005). Specifically, the correlation for single-locus within-population inbreeding coefficients is given by r^2, the squared correlation of allele frequencies at those loci. There is a similar relationship for single-locus θ values and within population r^2 values. Attention must be paid to the inherent variation in descent measures values if they are to be used to detect selection. Because the standard deviations differ among chromosomes, a case could be made for using genome-wide standard deviations to identify exceptional values which means population-specific values differ from each other exceptionally. There are many more regions with population differences than there are regions with values different from the mean.

We also have implemented the testing procedures on the HapMap data set. We found that all four methods rejected the null hypothesis which means the coancestry coefficient is strictly positive for human populations. This is expected as in the simulation study we have seen that when the true value of θ is close to 0.10 then the power of all the tests is one.

6.6 Application on Another Published Data set

The analysis of an empirical data set was employed to study the effect of different methods and models of interest for estimating the variance of heterozygosity. We reanalyzed one previously published data set (Olsen and Schaal, 2001) using the software code written in R.

The data set was collected by Olsen and Schaal (2001), who genotyped 5 microsatellite loci in 27 populations of the plants *Manihot esculenta* ssp *flabellifolia*. There are 157 individuals in the sample data set. These plants are found in very small populations of less than 15 individuals. The microsatellites were located in multiple introns of a 962-base-pair sequence of the Glyceraldehyde 3- phosphate dehydrogenase gene. These populations were further pooled according to geographic relationships into five

Table 6.14: Estimates of population-specific and overall θ and overall γ based on single-locus and 5-Mb window for HapMap data

Chr	CEU(θ)	HCB(θ)	JPT(θ)	YRI(θ)	All(θ)	All(γ)
chr1	.09(.31, .04)	.16(.25, .04)	.16(.26, .05)	.07(.36, .04)	.12(.11, .02)	.04(.12, .02)
chr2	.10(.31, .04)	.17(.25, .05)	.17(.25, .04)	.07(.36, .05)	.13(.11, .02)	.04(.12, .02)
chr3	.09(.31, .04)	.17(.25, .04)	.17(.25, .03)	.06(.35, .03)	.12(.11, .02)	.04(.12, .02)
chr4	.10(.30, .05)	.15(.24, .04)	.16(.25, .04)	.06(.36, .04)	.12(.11, .02)	.04(.12, .03)
chr5	.09(.31, .03)	.14(.25, .04)	.15(.25, .04)	.08(.34, .03)	.12(.11, .02)	.03(.11, .02)
chr6	.09(.31, .05)	.14(.24, .03)	.14(.25, .03)	.08(.35, .03)	.11(.11, .01)	.03(.11, .02)
chr7	.08(.32, .05)	.15(.24, .03)	.16(.25, .04)	.07(.34, .04)	.12(.11, .02)	.04(.11, .02)
chr8	.10(.29, .05)	.15(.24, .04)	.15(.24, .03)	.09(.35, .05)	.12(.11, .02)	.04(.12, .02)
chr9	.08(.30, .04)	.16(.24, .03)	.15(.24, .03)	.07(.35, .04)	.12(.10, .02)	.03(.11, .02)
chr10	.10(.31, .05)	.15(.24, .03)	.15(.25, .02)	.08(.34, .04)	.12(.11, .01)	.04(.12, .03)
chr11	.09(.30, .03)	.14(.24, .02)	.13(.24, .02)	.09(.34, .03)	.11(.10, .01)	.03(.11, .02)
chr12	.09(.32, .05)	.16(.25, .02)	.15(.26, .03)	.08(.35, .03)	.12(.11, .01)	.04(.12, .02)
chr13	.09(.30, .03)	.15(.24, .04)	.14(.24, .04)	.07(.35, .04)	.11(.10, .02)	.03(.10, .02)
chr14	.10(.32, .04)	.14(.24, .02)	.14(.25, .02)	.09(.35, .03)	.12(.11, .01)	.03(.11, .02)
chr15	.12(.32, .05)	.16(.24, .05)	.16(.25, .04)	.08(.36, .04)	.13(.11, .02)	.04(.12, .02)
chr16	.09(.31, .02)	.14(.25, .03)	.15(.24, .03)	.08(.35, .01)	.12(.11, .01)	.04(.12, .02)
chr17	.10(.30, .04)	.15(.24, .04)	.16(.25, .04)	.09(.34, .02)	.13(.11, .02)	.05(.13, .04)
chr18	.09(.30, .03)	.15(.23, .03)	.15(.25, .04)	.05(.34, .04)	.11(.10, .01)	.03(.10, .02)
chr19	.10(.31, .02)	.13(.24, .02)	.15(.25, .03)	.07(.36, .03)	.11(.10, .01)	.04(.11, .01)
chr20	.09(.30, .04)	.14(.24, .03)	.14(.24, .03)	.09(.34, .03)	.11(.10, .02)	.03(.11, .02)
chr21	.09(.29, .03)	.14(.23, .02)	.13(.24, .03)	.09(.34, .03)	.11(.10, .01)	.03(.11, .02)
chr22	.08(.30, .05)	.14(.24, .03)	.15(.24, .03)	.10(.34, .03)	.12(.11, .02)	.03(.12, .01)

groups in order to study the effects of increasing the departure from balanced sampling on different methods. The five pooled groups were as follows: Tocantins included the populations Axixa, Luzinopolis, Miranorte and Duere. The group Goia included the populations Campos Belos, Campinorte, Rialma, Corumba, Neropolis, Goias Velho, Ipora and Caiaponia. Mato Grosso was composed of Nova Xavatina, Serra Petrovina, Santa Elvira, Sao Vincente, Lambari dOeste, Pontes e Lacerda-A and -B. The group Randonia included Vilhena, Pimenta Bueno, Jaru, Ariquemes, Teotonio, Taquaras. Finally, the group Acre was composed of the Rio Branco and Sena Madureira populations.

Results

We advocate the regular inclusion of the variance of estimators in statistical analyses, particularly for the estimators of heterozygosity. It is important to include estimates of variance in analyses of this type, because the variances can be quite large and we can get more completely summarizes the data of interest. Examination of the point estimates $\tilde{H}_i$ in Table 6.16 illustrates the benefits of such a summarization. For this data set different plant populations have high levels of genetic variation which is conveyed by the point estimates. However, the estimates of heterozygosity range a great deal across both loci and populations, with underlying $\tilde{H}_{il}$ estimates ranging from (0.0; 0.72). This range is best summarized by including the variance of sample estimates with the point estimates for this data.

We use the general term "single-locus approximation" for the estimator to mean the variance of single-locus estimates as in the equation (5.13). This is a frequently used approach for the small proportion of studies that do give variance estimates of heterozygosity. The approximation will be very similar to the total variance of sample heterozygosity only in the cases where the heterozygosities can be reasonably modeled as having the same expected values and having no dependencies between loci. Weir (1989) noted that the composite linkage disequilibrium coefficients can be used as an

indicator of non-independence between gene diversity estimates at different loci. In our empirical data (Olsen and Schaal, 2001) the different loci are in linkage disequilibrium. Johnson (2004) performed a series of tests and found that for this data 18 out of 154 tests for composite linkage disequilibrium were significant at the 5% level, when 8 would be expected to be significant by chance. The presence of disequilibrium is consistent with the microsatellite markers being located within a 962-base-pair sequence (Olsen and Schaal, 2001). For this data set we have found a great disparity between the variance of heterozygosity as estimated by the single-locus approximations and those obtained with variance component methods or exact expressions (Table 6.16). This supports the idea of testing for composite linkage disequilibrium as an indicator of covariances between sample heterozygosities.

The sources of variation in observed values of heterozygosity are population, individual and different interaction. We believe that the loci contributes to the variation of heterozygosity by adding different fixed effects for different loci. On the other hand Johnson (2004) treated the loci effects as random effects. These two approaches will produce different results and the difference will depend on the variation of the locus effects. In the Table 3.2, Johnson (2004) found the total variances of the heterozygosity for the same data set using both random and mixed models and found that there is not much difference in the result. So it is appropriate to assume the loci effects are fixed. Since our data set is unbalanced we can use different variance component methods for analyzing the data set. But Johnson (2004) showed that within a model the results do not vary a lot for different component methods. So without loss of generality we find the estimates using a REML method.

We mentioned earlier that Johnson (2004) derived the expressions for the variance of heterozygosity incorrectly. We found the correct expressions for the variance of heterozygosity. Table 6.15 and Table 6.16 show that our expression produces a larger estimate than Johnson (2004)'s estimate. But we also think the linear model approach is not correct for indicator random variables. Here we propose a generalized linear model with mixed effects. Now the biggest problem with the generalized linear model

is choosing the link function. In particular any cumulative density function can serve as a link function. Three popular choices of link function for Bernoulli random variables are logit, probit and complementary log-log. We have used all three link functions and found that they give different estimate of the variance of heterozygosity. To find the best linear function we have used model selection criteria such as AIC, BIC, log-likelihood, or deviance. The results show that for the microsatellite data set (Olsen and Schaal, 2001) the logit link function is the best among three link functions. For the pooled data set all three link functions work well and they also produce almost the same result. For this particular data set we can use any one of the three link functions. So the link function depends on the data set. Again the model selection criteria shows that the generalized linear mixed model with logit link function is better than the linear model for the actual data as well as the pooled data. So we suggest using our generalized linear model rather than a linear mixed model for estimating the variance of heterozygosity. The glm estimate for variance is always larger than the estimate found using a linear model.

Table 6.15: Relationships between different expressions for the variances of $\tilde{H}_i$ for the pooled data obtained from Olsen data set

					$SD_T(\tilde{H}_i)$			
Population	n_i	$\tilde{H}_i$	$SD_W(\tilde{H}_i)$	$\sqrt{s^2_{\tilde{H}_i}}$	lm	logit	probit	cloglog
Toc.	26	0.377	0.051	0.037	0.1044	0.0971	0.0980	0.0918
Goiás	48	0.258	0.032	0.075	0.0998	0.0920	0.0930	0.0865
Mato	35	0.303	0.035	0.049	0.1018	0.0943	0.0952	0.0889
Rond.	36	0.411	0.041	0.032	0.1016	0.0940	0.0950	0.0886
Acre	12	0.550	0.053	0.077	0.1154	0.1090	0.1096	0.1039

Table 6.16: Relationships between different expressions for the variances of $\tilde{H}_i$ for the Olsen data set

					$SD_T(\tilde{H}_i)$			
Population	n_i	$\tilde{H}_i$	$SD_W(\tilde{H}_i)$	$\sqrt{s^2_{\tilde{H}_i}}$	lm	logit	probit	cloglog
Axixá	6	0.433	0.087	0.145	0.168	0.190	0.241	0.175
Luzinópolis	6	0.300	0.078	0.133	0.168	0.190	0.241	0.175
Miranorte	8	0.550	0.059	0.085	0.163	0.187	0.240	0.170
Dueré	6	0.167	0.119	0.053	0.168	0.190	0.241	0.175
C. Belos	6	0.267	0.077	0.113	0.168	0.190	0.241	0.175
Campinorte	6	0.067	0.038	0.067	0.168	0.190	0.241	0.175
Rialma	6	0.233	0.073	0.145	0.168	0.190	0.241	0.175
Corumbá	6	0.233	0.030	0.163	0.168	0.190	0.241	0.175
Nerópolis	6	0.233	0.099	0.085	0.168	0.190	0.241	0.175
Goiás Velho	6	0.367	0.110	0.082	0.168	0.190	0.241	0.175
Iporá	6	0.233	0.087	0.113	0.168	0.190	0.241	0.175
Caiapônia	6	0.433	0.087	0.145	0.168	0.190	0.241	0.175
N.Xavatina	6	0.167	0.056	0.091	0.168	0.190	0.241	0.175
S. Petrovina	5	0.000	0.000	0.000	0.172	0.193	0.242	0.178
Sta. Elvira	2	0.400	0.000	0.245	0.202	0.217	0.252	0.204
S. Vincente	4	0.300	0.087	0.146	0.177	0.197	0.244	0.182
L. d'Oeste	6	0.500	0.041	0.158	0.168	0.190	0.241	0.175
P. Lacerda-A	6	0.300	0.062	0.111	0.168	0.190	0.241	0.175
P. Lacerda-B	6	0.467	0.038	0.082	0.168	0.190	0.241	0.175
Vilhena	6	0.433	0.087	0.145	0.168	0.190	0.241	0.175
P. Bueno	6	0.700	0.078	0.062	0.168	0.190	0.241	0.175
Jarú	6	0.133	0.038	0.133	0.168	0.190	0.241	0.175
Ariquemes	6	0.267	0.077	0.113	0.168	0.190	0.241	0.175
Teotónio	6	0.533	0.038	0.111	0.168	0.190	0.241	0.175
Taquaras	6	0.400	0.067	0.113	0.168	0.190	0.241	0.175
Rio Branco	6	0.567	0.073	0.085	0.168	0.190	0.241	0.175
S. Madureira	6	0.533	0.077	0.200	0.168	0.190	0.241	0.175

Chapter 7

Discussion

This research has found different estimators for overall and population-specific descent measures. This study has explored the sampling properties of all the estimators using simulation studies. In particular, we have found the analytical expressions for the biases and standard errors of the moment estimators of overall θ and γ. The simulation study shows that the biases of the moment estimators of descent measures are relatively small in magnitude, and negative in direction. This result is consistent with the theoretical results that we obtained in this research using numerical approximations. Li (1996) also found similar results using a normal approximation. The biases and variances of the moment estimators of θ and γ increase as the differentiation levels increase in a total population. The biases of the moment estimators were found to be unaffected by the number of loci sampled, the amount of linkage between loci, and unbalanced sampling. The biases of the moment estimators are negligible although the sampling variances may be quite large. The sampling variances of the moment estimators increase as the true value of descent measures increases, but is not affected by unbalanced sampling. The sampling variances of the moment estimators decrease strongly as a result of increasing the number of loci sampled and increased number of alleles per locus. But still the variances remain fairly large on the whole due to variance from genetic sampling occurring in populations that cannot be reduced by sampling design. Increasing the number of loci sampled has a stronger effect on reducing the sampling variance of a moment estimator than increasing the number of

individuals sampled. The estimate of γ based on an allele which has frequency close to 0.5 is not stable. The estimate has huge standard error although the bias is not that large. The reason behind this is when the true allele frequency is 0.5, then the third moment of that particular allele does not provide any information about γ. One must be cautious about using the moment estimator of descent measures based on loci with very low polymorphism. The estimators are more robust to polymorphism problems if we increase the number of polymorphic loci. In our research we always assume that the populations are independent, but in general this is not the case. For dependent populations we cannot estimate the population-specific descent measures separately, but we can estimate some particular function of descent measures.

The maximum likelihood estimator of overall and population-specific θ based on a normal distribution gives undesirable estimates for both iterative and non-iterative approaches. For population-specific θ, the MLE fails consistently to converge to estimate values within the possible parameter range. This is true for both of the two proposed iteration methods for the MLE (Weir and Hill, 2002). Possibly future work with different numerical optimization procedures might be able to solve this problem of convergence and provide better estimates of population-specific θ. On the other hand, for overall θ the numerical optimization for the MLE converges but does not produce any reasonable estimates. In summary, the unpredictable behavior of the iterative and non-iterative MLE suggests that the moment estimator is a much better choice to be used in analyses. These estimators of θ are not recommended for general use in analyses. If we use a normal distribution then we do not need to estimate the higher order descent measures as they are functions of θ. In the first chapter we have shown that these approximations of descent measures do not work for a random pure drift model and a both-way mutation model. Since the alleles in a population are equally correlated, we cannot use the central limit theorem to get the asymptotic distribution of the allele frequencies. We simulated data and found that the moments of the allele frequencies differ from a normal distribution moments. We think this is the reason behind the unreasonable estimates produced by MLE method. There are two fourth

order descent measures and they appear in the fourth order moment of the allele frequencies. Since there is only one independent sample moment, it is not possible to estimate both the fourth order descent measures.

The normal theory approach assumes the distributional form of the data but the MOM assumes only the first two moments of the allele frequencies. Because of this reason scientists prefer MOM over MLE based on normal distribution. There are several other methods which only assume the first two moments and estimate descent measures. Here we consider Quasi-Likelihood, extended Quasi-Likelihood, and Pseudo Likelihood methods that have the above property. The performance of these methods are yet to be evaluated. In the following paragraphs we discuss how to estimate descent measures using these methods.

We outline the process of developing an estimator of θ based on a particular allele at a particular locus. Let us assume the allele is A. The expected frequency of the allele is p. We assume that there are r independent populations. The total sampled alleles and the count of the allele A in the i^{th} population are n_i and Y_i respectively.

Quasi-likelihood was proposed by Wedderburn (1974) and is based on the first two moments of the data. It does not assume the distributional form of the data. We have independent observations $z_1, \cdots, z_r$ with $\mathrm{E}(z_i) = \mu_i$ and $\mathrm{Var}(z_i) = V_i(\mu_i)$ where V_i is some known function. Now assume μ_i is a known function of $\beta_1, \cdots, \beta_p$ and the β's are the parameters of our interest. Then the quasi-likelihood function is (Wedderburn, 1974)

$$Q(\mu, Z) = \sum_{i=1}^{r} Q(\mu_i, z_i) = \sum_{i=1}^{r} \int_{z_i}^{\mu_i} \frac{z_i - t}{V_i(t)} dt. \tag{7.1}$$

We get the estimates of the parameters $\beta_1, \cdots, \beta_r$ by maximizing the equation (7.1) with respect to the parameters.

In our case the data are Y_i/n_i's. So $z_i = Y_i/n_i$, $\mu_i = p$, $\mathrm{Var}(z_i) = [1 + (n_i - 1)\theta_i]/n_i$.

So the quasi-likelihood is

$$Q(\mu, Z) = \sum_{i=1}^{r} Q(\mu_i, z_i) = \sum_{i=1}^{r} \int_{z_i}^{p} \frac{n_i(z_i - t)}{t(t-1)[1 + (n_i - 1)\theta_i]} dt. \tag{7.2}$$

We can estimate p and θ_i by solving the following equations simultaneously

$$U_1(p, \theta_i) = \sum_{i=1}^{r} \frac{n_i(z_i - p)}{p(p-1)[1 + (n_i - 1)\theta_i]} \text{ and} \tag{7.3}$$

$$U_2(p, \theta_i) = \sum_{i=1}^{r} \frac{n_i(z_i - p)^2}{p(p-1)[1 + (n_i - 1)\theta_i]}. \tag{7.4}$$

To get an estimate of the overall value of θ, we replace θ_i by θ in the equations (7.3) and (7.4). Then we solve them in terms of θ and p.

For the common descent measures model, ignoring constant terms, the extended quasi-likelihood function of Nelder and Pregibon (1987) is

$$l_Q = \frac{1}{2} \sum_{i=1}^{r} \Big\{ \log[1 + (n_i - 1)\theta] + \frac{D_i(Y_i, p)}{[1 + (n_i - 1)\theta]} \Big\}, \tag{7.5}$$

where D_i is the binomial deviance function for the i^{th} population. The form of D_i is

$$D_i(Y_i, p) = 2\Big[Y_i \log(\frac{Y_i}{n_i p}) + (n_i - Y_i) \log(\frac{n_i - Y_i}{n_i - n_i p}) \Big]. \tag{7.6}$$

Differentiation of l_Q with respect to p and θ leads to the pair of the estimating equations

$$\sum_{i=1}^{r} \frac{Y_i - n_i p}{p(1-p)\phi_i} = 0 \text{ and} \tag{7.7}$$

$$\sum_{i=1}^{r} (n_i - 1)\Big(\frac{D_i - \phi_i}{\phi_i^2} \Big) = 0, \tag{7.8}$$

where $\phi_i = 1 + (n_i - 1)\theta$. For a given value of θ, the solution of the equation (7.7) is simply a weighted average of allele frequencies over different populations. The estimate

of θ based on the extended quasi-likelihood method, θ_{EQL} is the simultaneous solution of the equations. The equation (7.7) is an unbiased estimating equation, but equation (7.8) is not because in general $E(D_i) \neq \phi_i$. The pseudo-likelihood method replaces D_i in the equation (7.8) with

$$X_i^2 = \frac{Y_i - n_i p}{n_i p(1-p)}, \tag{7.9}$$

and leads to the pseudo-likelihood estimator θ_{PL}. Unlike θ_{EQL}, θ_{PL} results from an unbiased estimating equation because $E(X_i^2) = \phi_i$.

We also can estimate γ using the above approaches but for that we need to consider the second order moments of the allele frequencies. This is because γ appears in the variance of the second order allele frequencies. The variance of second order allele frequencies will also have δ and Δ in the expression.

In one part of our research we clarified the problem of testing hypotheses about the coancestry coefficient (θ). We worked with a random population setup. We proposed a parametric bootstrap procedure for small sample sizes and a chi square test for large sample sizes. Under a random population model, bootstrap resampling over loci is the only way to get information about the evolution. Dodds (1986) used a non-parametric bootstrap method for testing purpose for small sample sizes. We suggested a parametric bootstrap testing method for the same setup. Simulation studies show that our testing procedure has higher power than Dodds's method. Li (1996) used the central limit theorem for approximating the distribution of allele frequencies as a normal distribution and proposed a chi square test. Her test is based on the frequency of one particular allele and it loses information when there are more than two alleles per locus. We resolved this problem and found a test procedure which includes all the allele frequencies. As expected, our test statistic has higher power than Li's test procedure for more than two alleles per locus although for two alleles both the test procedures have almost similar power. Weir and Cockerham (1984) have developed higher levels for analysis of variance for estimating higher order F-statistics. The methods for making

inferences regarding F-statistics for a hierarchial population setup will be a topic for future study.

We have found a moment estimator for different components of two-locus descent measures. We also have shown that these measures are not identifiable. The compound parameter, the product of linkage disequilibrium between loci, and the two-locus descent measures are estimable. We derived the sampling properties of these estimators using simulation studies and found that they have negligible biases and large sampling errors. Since the value of two-locus descent measures depend on the value of the recombination rate between two loci, these measures are loci specific. There is a scope to use other methods for estimating the descent measures. One can put some prior on linkage disequilibrium and integrate it to get an estimate of two-locus descent measures. These may be problems for future research.

We have advocated the practice of adding sample variances with the point estimates by illustration with analysis of an empirical data set, and by attempting to clarify the underlying statistical theory motivating the methods and models for obtaining variance estimates. This illustration has shown that the range of heterozygosity and gene diversity estimates can be large and it has demonstrated that this can be well summarized by including associated variances with point estimators. In the analysis of a real data set, we have found a wide range for the estimates of heterozygosity. This should encourage investigators estimating heterozygosity to consider the sampling properties of their estimators in order to increase the quality of their inferences. Generally however, if data from multiple populations is available, one must avoid the approximation of the total variance of sample heterozygosity by the single-locus approximation. Instead, it is best to use the generalized mixed model or generalized random model variances, which will be more reliable than the single-locus approximation because they more generally account for all sources of variation due to evolutionary history.

The magnitude of the total variance of the sample gene diversity is a result of the scope of inferences to be made, the number of loci and individuals per population sampled, mating systems, and the distribution of alleles across the populations of

interest. Differences in these factors can result in differences in the sampling variance of gene diversity from equal sized samples of different species.

Some scientists treat the locus effects as fixed while others assume that they are random. If the variance due to loci is very high then the random effect model can produce a different result than mixed effect model. Otherwise these models give similar results. Johnson (2004) analyzed the empirical data set that is presented here and found that the random effects model produces similar results as the mixed effects model. The variance component for loci may be more likely to be large for loci that are not in linkage disequilibrium. But in theory we would like to think that the effects of the loci are fixed.

It has been shown that the variance component methods studied here can produce different estimates for unbalanced data sets. The ANOVA method performs differently than REML, ML and MIVQUE for unbalanced data. The maximum likelihood based methods appear to be robust to the effect of small sample sizes and the assumption of normality made by these approaches. The REML method is perhaps to be preferred because it accounts for fixed effects with respect to degrees of freedom and has guaranteed minimum variance properties, it does not allow negative variance estimates. The ML approach does not guarantee minimum variance, while, MIVQUE and ANOVA allow negative estimates of variance. In general, all the methods provide similar kind of estimates for variance for the cases where the populations were of relatively similar sizes.

The bootstrap resampling approaches have been used to determine the sampling properties of gene diversity and heterozygosity for a fixed population (Shete, 2003). The bootstrap procedure should not be applied to a random population model because the resampling would disrupt associations between genes in individuals. Weir (1989) and Shete (2003) have shown that the bias of the gene diversity is small, particularly relative to the large sampling variance of the estimator. However, Shete has also determined the form of a uniform minimum variance unbiased estimator (UMVUE) of gene diversity for a fixed population model by correcting the bias.

Expressions for the total variance of sample gene diversity and heterozygosity must account for both within and between population variation for the case of more than one population sampled. Weir (1989) and Weir et al. (1990) determined the total variance of sample gene diversity and heterozygosity, respectively, using exact expressions involving genotype frequencies and descent measures under a variety of different evolutionary models. For a mixed-mating or random mating systems, this variance results mostly from associations of alleles between individuals. Because of this, the total variance of sample heterozygosity is minimized most efficiently by sampling more individuals, rather than increasing the number of loci sampled. In contrast, increasing the number of loci sampled, rather than the number of individuals sampled, has the strongest minimization of variance for unlinked loci in populations at migration-drift equilibrium.

The sources of variation in their observed values of gene diversity and heterozygosity are similar. Due to the complexity of the exact expression, the total variance of gene diversity is hard to find. In this situation one can approximate $\mathrm{Var}_T(\tilde{d}_i)$ by $\mathrm{Var}_T(\tilde{H}_i)$. If this approximation was found to be reasonable, the total variance estimates could be obtained with the use of variance components methods. It would then have well-studied sampling properties due to the extensive statistical theory developed with these statistical methods. Unfortunately, the relationship between the variances of sample heterozygosity and gene diversity appears to be complex and one is not well approximated by the other. Simulation studies show that as heterozygosity decreases relative to a given level of gene diversity, the associated variance of sample heterozygosity increases, while the variance of sample gene diversity is instead decreasing.

Bibliography

Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficents, *Theoretical Population Biology* **63**: 221–230.

Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, *Genetica* **96**: 3–12.

Cockerham, C. C. (1969). Variance of gene frequencies, *Evolution* **23**: 72–84.

Cockerham, C. C. (1971). Higher order probability functions of identity of alleles by descent, *Genetics* **69**: 235–246.

Cockerham, C. C. (1973). Analyses of gene frequencies, *Genetics* **74**: 679–700.

Cockerham, C. C. and Weir, B. S. (1973). Descent measures for two loci with some applications, *Theoretical Population Biology* **4**: 300–330.

Dodds, K. G. (1986). *Resampling methods in Genetics and the effect of Family Structure in Genetic Data*, PhD thesis, North Carolina State University.

Ewens, W. J. (1979). *Mathematical Population Genetics*, Springer-Verlag, Berlin-Heidelberg-New York.

Falconer, D. and Mackay, T. F. (1996). *Introduction to Quantative Genetics*, Longman, Essex, England.

Gillois, M. (1966). Relation d'identite en génétique. I. Postulats et axiomes Mendéliens. II. Corrélation génétique dans le cas de dominance, *Ann. Inst. Henri Poincaré* **Sec. B II**: 349–352.

Graham, J., Curran, J. and Weir, B. S. (2000). Conditional genotypic probabilities for microsatellite loci, *Genetics* **155**: 1973–1980.

Hernández-Sánchez, J., Haley, C. S. and Woolliams, J. A. (2004). On the prediction of simultaneous inbreeding coefficients at multiple loci, *Genetical Research* **83**: 113–120.

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations, *Theor. and Appl. Genet.* **38**: 226–231.

Holsinger, K. E. (1999). Analysis of genetic diversity in geographically structured populations: a Bayesian perspective, *Hereditas* **130**: 245–255.

Holsinger, K. E. and Wallace, L. E. (2004). Bayesian apporaches for the analysis of population genetic structure: an example from *Plantanthera leucophaea*, *Molicular Ecology* **13**: 887–894.

Johnson, A. M. (2004). *Estimation and Sampling Properties of Gene Diversity, Heterozygosity and* F_{ST}, PhD thesis, North Carolina State University.

Karlin, S. and McGregor, J. (1968). Rates and probabilities of fixation for two locus random mating finite populations without selection, *Genetics* **58**: 141–159.

Lange, K. (1995). Applications of the Dirichlet distribution to forensic match probabilities, *Genetica* **96**: 107–117.

Li, Y.-J. (1996). *Characterizing the structure of genetics population*, PhD thesis, North Carolina State University.

Littler, R. A. (1973). Linkagae disequilibrium in two-locus, finite, random mating models without selection or mutation, *Theoretical Population Biology* **4**: 259–275.

Long, J. C. (1986). The allelic correlation structure of Gainj- and Kalman-speaking people. I. The estimation and incorporation of Wright's F-statistics, *Genetics* **112**: 629–647.

Long, J. C. and Kittles, R. A. (2003). Human genetic diversity and the nonexistence of biological races, *Human Biology* **75**: 449–471.

Lynch, M. (1988). Estimation of relatedness by DNA fingerprinting, *Molecular Biology and Evolution* **5**: 584–599.

Malécot, G. (1948). *Les mathématiques*, de l'hérédité, Masson, Paris.

Marshall, D. R. and Allard, R. W. (1970). Isozyme polymorphisms in natural populations of *Avena fatua* and *A. barbata.*, *Heredity* **25**: 373–382.

Maruyama, T. (1977). Stochastic problems in population genetics, *Lecture Notes in Biometrica, Vol 17*, Springer, Berlin.

Nei, M. (1973). Ananlysis of gene diversity in subdivided population, *Proc. Natl. Acad. Sci. U.S.A.* **70**: 3321–3323.

Nei, M. (1987). *Molecular Evolutionary Genetics*, Columbia University Press, New York.

Nei, M. and Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance, *Genetics* **76**: 379–390.

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function, *Biometrika* **74**: 221–232.

Nicholson, G. and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data, *J. R. Statist. Soc.* **B 64**: 1–21.

Olsen, K. M. and Schaal, B. A. (2001). Microsatellite variation in cassava (*Manihot esculenta*, Euphorbiaceae) and its wild relatives: further evidence for a southern Amazonian origin of domestication, *American Journal of Botany* **88**: 131–142.

Rao, C. R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications*, North-Holland, Amsterdam.

Raufaste, N. and Bonhomme, F. (2000). Properties of bias and variance of two multiallelic estimators of F_{ST}, *Theoretical Population Biology* **57**: 285–296.

Raymond, M. and Rousset, F. (1995). An exact test for population differentiation, *Evolution* **49**: 1280–1283.

Ritland, K. (1987). Definition and estimation of higher-order gene fixation indices, *Genetics* **117**: 783–793.

Robertson, A. (1952). The effect of inbreeding on the variation due to recessive genes, *Genetics* **37**: 189–207.

Robertson, A. and Hill, W. G. (1984). Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients, *Genetics* **107**: 703–718.

Roff, D. A. and Bentzen, P. (1989). The statistical analysis of mitochondrial DNA poymorphisms: χ^2 and the probelem of small sample sizes, *Molecular Biology and Evolution* **6**: 539–545.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, Wiley, New York.

Serfling, R. J. (1980). *Approximation Theorem of Mathematical Statistics*, Jhon Wiley, New York.

Shete, S. (2003). Uniformly minimum variance unbiased estimation of gene diversity, *Journal of Heredity* **94**: 421–424.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies, *Genetics* **139**: 457–462.

Smouse, P. E. and Willams, R. C. (1982). Multivariate analysis of HLA-disease associations, *Biometrics* **38**: 757–768.

Watterson, G. A. (1970). The effect of linkage in a finite random-mating population, *Theoretical Population Biology* **1**: 72–87; Errata, *Theoretical Population Biology* 3: 117 (1972).

Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method, *Biometrika* **61**: 439–447.

Weir, B. S. (1989). *Sampling properties of gene diversity, Ch. 2 in Brown, A. H. D., Clegg, M. T. Kahler, A. L. and B. S. Weir, eds. Plant Population Genetics, Breeding and Genetic Resources*, Sinauer, Sunderland,MA.

Weir, B. S. (1994). The effects of inbreeding on forensic calculations, *Annual Review of Genetics* **28**: 597–621.

Weir, B. S. (1996). *Genetic Data Analysis II*, Mass.:Sinauer, Sunderland.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. and Hill, W. G. (2005). Measures of human population structure show heterogeneity among genomic regions, *Genome Research* **15**: 1468–1476.

Weir, B. S. and Cockerham, C. C. (1969). Group inbreeding with two linked loci, *Genetics* **63**: 711–742.

Weir, B. S. and Cockerham, C. C. (1974). Behavior of pairs of loci in finite monoecious populations, *Theoretical Population Biology* **6**: 323–354.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure, *Evolution* **38**: 1358–1370.

Weir, B. S. and Hill, W. G. (2002). Estimating F-statistics, *Annual Review of Genetics* **36**: 721–750.

Weir, B. S., Reynolds, J. and Dodds, K. G. (1990). The variance of sample heterozygosity, *Theoretical Population Biology* **37**: 235–253.

Workman, P. L. and Niswander, J. D. (1970). Population studies on Southwestern Indian tribes. II. Local genetic differentiation in the Papago, *American Journal Human Genetics* **22**: 24–49.

Wright, S. (1921). Systems of mating. i. the biometric relations between partent and offspring, *Genetics* **6**: 111–178.

Wright, S. (1931). Evolution in Mendelian populations, *Genetics* **16**: 97–159.

Wright, S. (1951). The genetical structure of populations, *Annals of Human Genetics* **15**: 323–354.

Yang, R. C. (1998). Estimating hierarchial F-statistics, *Evolution* **52**: 950–956.

APPENDICES

Appendix A

The Relations between the Moment and the Probabilistic Estimators

When the sample sizes are equal, we get the following simplifications:

$$
\begin{aligned}
MSP_k &= \frac{n}{r-1}\sum_{i=1}^{r}(\tilde{p}_{i,k}-\tilde{p}_{w,k})^2 = \frac{n}{r-1}\Big[\sum_{i=1}^{r}\tilde{p}_{i,k}^2 - r\tilde{p}_{w,k}^2\Big] \\
MSG_k &= \frac{n}{n(r-1)}\sum_{i=1}^{r}\tilde{p}_{i,k}(1-\tilde{p}_{i,k}) = \frac{n}{r(n-1)}\Big[r\tilde{p}_{w,k} - \sum_{i=1}^{r}\tilde{p}_{i,k}^2\Big] \\
\hat{\pi}_{1,1,k} &= \frac{1}{r}\sum_{i=1}^{r}\tilde{p}_{k,i} = \tilde{p}_{uw,k} = \tilde{p}_{w,k} \\
\hat{\pi}_{2,1,k} &= \frac{1}{r}\sum_{i=1}^{r}\frac{n\tilde{p}_{k,i}^2 - \tilde{p}_{k,i}}{n-1} = \frac{n}{r(n-1)}\sum_{i=1}^{r}\tilde{p}_{k,i}^2 - \frac{1}{n-1}\tilde{p}_{w,k} \\
\hat{\pi}_{2,2,k} &= \frac{(\sum_{i=1}^{r}\tilde{p}_{k,i})^2 - \sum_{i=1}^{r}\tilde{p}_{k,i}^2}{r(r-1)} = \frac{r}{r-1}\tilde{p}_{w,k}^2 - \frac{1}{r(r-1)}\sum_{i=1}^{r}\tilde{p}_{k,i}^2.
\end{aligned}
\tag{A.1}
$$

Some algebra with the above expression give

$$\begin{aligned} MSP_k - MSG_k &= \frac{n(nr-1)}{r(r-1)(n-1)} \sum_{i=1}^{r} \tilde{p}_{k,i}^2 - \frac{nr}{r-1} \tilde{p}_{w,k}^2 - \frac{n}{n-1} \tilde{p}_{w,k}, \\ MSP_k + (n_{c_1} - 1) MSG_k &= \frac{n}{r(r-1)} \sum_{i=1}^{r} \tilde{p}_{k,i}^2 - \frac{nr}{r-1} \tilde{p}_{w,k}^2 + n\tilde{p}_{w,k}, \\ MSP_k - MSG_k &= n(\pi_{2,1,k} - \pi_{2,2,k}), \text{ and} \\ MSP_k + (n_{c_1} - 1) MSG_k &= n(\pi_{1,1,k} - \pi_{2,2,k}). \end{aligned} \tag{A.2}$$

Using the equations (A.1) and (A.2) we get

$$\hat{\theta}_{WC,k} = \frac{MSP_k - MSG_k}{MSP_k + (n_{c_1} - 1) MSG_k} = \frac{n(\pi_{2,1,k} - \pi_{2,2,k})}{n(\pi_{1,1,k} - \pi_{2,2,k})} = \hat{\theta}_{1,P,k}. \tag{A.3}$$

This shows that the moment estimator of θ proposed by Weir and Cockerham and the newly proposed probabilistic estimator of θ based on a single allele are the same. So the final estimators also be the same irrespective of the combining methods. Using the similar kind of algebra we also can make inferences about the third order moment estimators of θ and moment estimators of γ. Some calculations provide use the equalities

$$\hat{\theta}_{1,M} = \hat{\theta}_{3,P},\ \hat{\theta}_{2,M} = \hat{\theta}_{4,P},\ \hat{\gamma}_{1,M} = \hat{\gamma}_{1,P},\ \text{and } \hat{\gamma}_{1,M} = \hat{\gamma}_{2,P}. \tag{A.4}$$

For unequal sample sizes these equalities do not hold. Simulation studies show that there is a small difference between these estimators. We also have shown through simulation that the estimators based on the probabilistic approach is better than the moment estimators in general.

Appendix B

Derive the Simpler Form of a Test Statistic

$$
\begin{aligned}
\tilde{\mathbf{Z}}' A \tilde{\mathbf{Z}} &= (\tilde{\mathbf{Z}}_1', \tilde{\mathbf{Z}}_2', \cdots, \tilde{\mathbf{Z}}_r')(I_r - \frac{1}{r} 1_r 1_r') \bigotimes I_{s-1} (\tilde{\mathbf{Z}}_1', \tilde{\mathbf{Z}}_2', \cdots, \tilde{\mathbf{Z}}_r')' \\
&= (\tilde{\mathbf{Z}}_1', \cdots, \tilde{\mathbf{Z}}_r')(\tilde{\mathbf{Z}}_1', \cdots, \tilde{\mathbf{Z}}_r')' - (\tilde{\mathbf{Z}}_1', \cdots, \tilde{\mathbf{Z}}_r')(\frac{1}{r} 1_r 1_r') \bigotimes I_{s-1} (\tilde{\mathbf{Z}}_1', \cdots, \tilde{\mathbf{Z}}_r')' \\
&= \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i' \tilde{\mathbf{Z}}_i - (\tilde{\mathbf{Z}}_1', \tilde{\mathbf{Z}}_2', \cdots, \tilde{\mathbf{Z}}_r')(\frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i', \frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i', \cdots, \frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i')' \\
&= \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i' \tilde{\mathbf{Z}}_i - r(\frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i)' (\frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i) \\
&= \sum_{i=1}^{r} (\tilde{\mathbf{Z}}_i - \frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i)' (\tilde{\mathbf{Z}}_i - \frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{Z}}_i) \\
&= \sum_{i=1}^{r} (T^{-1} \tilde{\mathbf{P}}_i - \frac{1}{r} \sum_{i=1}^{r} T^{-1} \tilde{\mathbf{P}}_i)' (T^{-1} \tilde{\mathbf{P}}_i - \frac{1}{r} \sum_{i=1}^{r} T^{-1} \tilde{\mathbf{P}}_i) \\
&= \sum_{i=1}^{r} (\tilde{\mathbf{P}}_i - \frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{P}}_i)' C^{-1} (\tilde{\mathbf{P}}_i - \frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{P}}_i) \\
&= \sum_{i=1}^{r} (\tilde{\mathbf{P}}_i - \frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{P}}_i)' \begin{bmatrix} \frac{1}{p_1} + \frac{1}{p_s} & \frac{1}{p_s} & \cdots \\ \frac{1}{p_s} & \frac{1}{p_2} + \frac{1}{p_s} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} (\tilde{\mathbf{P}}_i - \frac{1}{r} \sum_{i=1}^{r} \tilde{\mathbf{P}}_i) \\
&= \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(\sqrt{n_i} \tilde{p}_{k,i} - 1/r \sum_{i=1}^{r} \sqrt{n_i} \tilde{p}_{k,i})^2}{p_k} \qquad \text{(B.1)}
\end{aligned}
$$

Appendix C

Proof of the Lemma

We have $Y_{n,i} \xrightarrow{p} c_i$ and $Pr(Y_{n,i} = 0) = 0$. Using these two facts we get $c_i/Y_{n,i} \xrightarrow{p} 1$ for $i = 1, 2, ..., s$. So for $\forall \epsilon > 0$, there exists a $\delta(\epsilon) > 0$ which depends on the value of ϵ such that (true for $i = 1, 2, \cdots s$)

$$Pr\Big[c_i/Y_{n,i} > 1 + \delta(\epsilon)\Big] < \frac{\epsilon}{s+1} \text{ and } Pr\Big[c_i/Y_{n,i} < 1 - \delta(\epsilon)\Big] < \frac{\epsilon}{s+1}. \tag{C.1}$$

Using the equation (C.1) we get the following two inequalities:

$$\begin{aligned} Pr\Big[\max_{1\le i\le s} c_i/Y_{n,i} < 1 - \delta(\epsilon)\Big] &\le Pr\Big[c_1/Y_{n,1} < 1 - \delta(\epsilon)\Big] < \frac{\epsilon}{s+1} \\ Pr\Big[\max_{1\le i\le s} c_i/Y_{n,i} > 1 + \delta(\epsilon)\Big] &= Pr\Big[\bigcup_{i=1}^{s}\big\{c_i/Y_{n,i} > 1 + \delta(\epsilon)\big\}\Big] \\ &\le \sum_{i=1}^{s} Pr\Big[c_i/Y_{n,i} > 1 + \delta(\epsilon)\Big] < \frac{s\epsilon}{s+1} \end{aligned} \tag{C.2}$$

From the equation (C.2) we get, $\forall \epsilon > 0$, there exists a positive $\delta(\epsilon)$ such that

$$Pr\Big[\mid \max_{1\le i\le s} c_i/Y_{n,i} - 1 \mid > \delta(\epsilon)\Big] < \frac{\epsilon}{s+1} + \frac{s\epsilon}{s+1} = \epsilon. \tag{C.3}$$

The equation (C.3) concludes that

$$\max_{1\le i\le s} c_i/Y_{n,i} \xrightarrow{p} 1. \tag{C.4}$$

Similarly we can show

$$\min_{1\leq i\leq s} c_i/Y_{n,i} \xrightarrow{p} 1. \tag{C.5}$$

Now by assumption we have $\sum_{i=1}^{s} X_i \xrightarrow{d} Z$. Using Slutksy's theorem we get

$$(\max_{1\leq i\leq s} c_i/Y_{n,i}) \sum_{i=1}^{s} X_i \xrightarrow{d} Z \text{ and } (\min_{1\leq i\leq s} c_i/Y_{n,i}) \sum_{i=1}^{s} X_i \xrightarrow{d} Z. \tag{C.6}$$

Since each X_i and $c_i/Y_{n,i}$ are positive, we get the following inequalities:

$$(\min_{1\leq i\leq s} Y_{n,i}/c_i) \sum_{i=1}^{s} X_{i,n} \leq \sum_{i=1}^{s} X_{i,n}c_i/Y_{n,i} \leq (\max_{1\leq i\leq s} c_i/Y_{n,i}) \sum_{i=1}^{s} X_{i,n} \tag{C.7}$$

Choose x from real line and from the equation (C.7) we get

$$\begin{aligned} Pr\Big[(\min_{1\leq i\leq s} Y_{n,i}/c_i) \sum_{i=1}^{s} X_{i,n} \leq x\Big] &\leq Pr\Big[\sum_{i=1}^{s} X_{i,n}c_i/Y_{n,i} \leq x\Big] \\ &\leq Pr\Big[(\max_{1\leq i\leq s} c_i/Y_{n,i}) \sum_{i=1}^{s} X_{i,n} \leq x\Big]. \end{aligned} \tag{C.8}$$

The point x is a continuous point of Z as Z is a continuous random variable. Now taking limits on both sides of the equation (C.8), we get

$$\begin{aligned} & Pr[Z \leq x] \leq \lim_{n\to\infty} Pr\Big[\sum_{i=1}^{s} X_{i,n}c_i/Y_{n,i} \leq x\Big] \leq Pr[Z \leq x], \\ \Rightarrow \quad & \lim_{n\to\infty} Pr\Big[\sum_{i=1}^{s} X_{i,n}c_i/Y_{n,i} \leq x\Big] = Pr[Z \leq x]. \end{aligned} \tag{C.9}$$

From the equation (C.9) we get

$$\sum_{i=1}^{s} X_{i,n}c_i/Y_{n,i} \xrightarrow{d} Z.$$