ABSTRACT

Title of Document:	BAYESIAN ESTIMATION OF THE INBREEDING COEFFICIENT FOR SINGLE NUCLEOTIDE POLYMORPHISM USING COMPLEX SURVEY DATA Zhenyi Xue, Doctor of Philosophy, 2015
Directed By:	Professor Partha Lahiri Associate Professor Yan Li Joint Program in Survey Methodology

In genome-wide association studies (GWAS), single nucleotide polymorphism (SNP) is often used as a genetic marker to study gene-disease association. Some large scale health sample surveys have recently started collecting genetic data. There is now growing interest in developing statistical procedures using genetic survey data. This calls for innovative statistical methods that incorporate both genetic and statistical sampling.

Under simple random sampling, the traditional estimator of the inbreeding coefficient is given by 1 - (number of observed heterozygotes) / (number of expected heterozygotes). Genetic data quality control reports published by the National Health and Nutrition Examination Survey (NHANES) and the Health and Retirement Study (HRS) use this simple estimator, which serves as a reasonable quality control tool to identify problems such as genotyping error. There is, however, a need to improve on

this estimator by considering different features of the complex survey design. The main goal of this dissertation is to fill in this important research gap. First, a design-based estimator and its associated jackknife standard error estimator are proposed. Secondly, a hierarchical Bayesian methodology is developed using the effective sample size and genotype count. Lastly, a Bayesian pseudo-empirical likelihood estimator is proposed using the expected number of heterozygotes in the estimating equation as a constraint when maximizing the pseudo-empirical likelihood. One of the advantages of the proposed Bayesian methodology is that the prior distribution can be used to restrict the parameter space induced by the general inbreeding model.

The proposed estimators are evaluated using Monte Carlo simulation studies. Moreover, the proposed estimates of the inbreeding coefficients of SNPs from APOC1 and BDNF genes are compared using the genetic data from the 2006 Health and Retirement Study.

BAYESIAN ESTIMATION OF THE INBREEDING COEFFICIENT FOR SINGE NUCLEOTIDE POLYMORPHISM USING COMPLEX SURVEY DATA

By

Zhenyi Xue

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Frank Alt
Professor Partha Lahiri, Chair
Professor Yan Li, Co-Chair
Professor Paul Smith
Professor Jing Zhang

Acknowledgements

First of all, I want to express my sincere gratitude to my academic advisor, Professor Partha Lahiri. Throughout many years of my part-time graduate study, he continuously guided me in my academic research through his insightful and inspiring thinking. I am also very thankful to my academic co-advisor, Professor Yan Li, for her dedicated support and guidance in my dissertation research. I greatly appreciate the enormous amount of time both of them spent with me, listening patiently and then advising using their profound knowledge.

My special thanks go to Professor Paul Smith for his consistent encouragement and support during my entire PhD studies in the Department of Mathematics. I also want to extend my gratitude to Professor Frank Alt and Professor Jing Zhang for their time to read over my work and for their constructive comments and help on this dissertation. I also thank Professor Eric Slud, Professor Abram Kagan, and Professor Benjamin Kedem for their valuable teachings.

I also thank the following in our Mathematics Department and Joint Program in Survey Methodology: Alverda McCoy, Haydee Hidalgo, Celeste Regalado, and Holly Rollins for their administrative help; Dr. Daniel Bonnery and Dr. Neung Soo Ha for their help and research discussion; and other fellow students in the Mathematics Department for their helpful discussions, encouragement, and friendship.

To Ms. Rebecca Torguson, Dr. Ron Waksman at MedStar Health Research Institute, Mrs. Paula Johnson, Dr. Julie Berube and my senior colleagues at Becton, Dickinson

and Company, I want to extend my gratitude for giving me the flexibility that allowed me to manage my work and study simultaneously during my part-time PhD studies.

Last but not the least, I would like to thank my parents Guiyu Liang and Aofu Xue, my wife Chunhong, my daughter Hannah and my son Alex, for their love, patience and support.

Table of Contents

List of Tables

List of Figures

Glossary of Technical Terms

Allele	A variant of similar DNA sequence located at a given locus
Chromosome	A packaged and organized structure containing most of the DNA of a living organism
dbGap	The Database of Genotypes and Phenotypes
dbSNP	The Database of Single Nucleotide Polymorphisms
Diploid cell	A cell that has two homologous copies of each chromosome
DNA	Deoxyribonucleic acid, a molecular that carries most of genetic information of living organism or viruses
F_{IS}	The correlation between genes within individual relative to the genes from subpopulation
F_{IT}	The correlation between genes within individual relative to the genes from population
F_{ST}	The correlation between genes within subpopulation relative to the genes from total population
Gamete	A cell that fuses with another cell during fertilization (conception) in organisms that sexually reproduce
Gene	A locus (or region) of DNA that encodes a functional RNA or protein product, and is the molecular unit of heredity
Gene Drift	The change in the frequency of a gene variant (allele) in a population due to random sampling of organisms
Gene Flow	Transfer of alleles or genes from one population to another
Genotype	A part of DNA sequence of a cell, and therefore of an organism or individual, which determines a specific characteristic (phenotype) of that cell/organism/individual
GWAS	Genome-wide association studies
Heterozygote	Organism's genotype contains different alleles of a gene
Homozygote	Organism's genotype contains same allele pair of a gene

HWE	Hardy-Weinberg Equilibrium, or Hardy-Weinberg Principal, named after Godfrey H. Hardy and Wilhelm Weinberg, states that genetic variation including allele and genotype frequency in a population will remain constant from one generation to the next generation in the absence of disturbing influences
IBD	Identity by descent, both alleles from two individuals come from the same allele copy in a common ancestor
Inbreeding Coefficient (*f*)	or Fixation Index F_{IS}, first introduce by Wright (1921), measures the expected percentage of homozygosity based on a known pedigree (a fully documented genealogy for a fixed system of breeding)
Locus	(plural) Loci, specific location or position of a gene, DNA sequence, on a chromosome
MAF	Minor allele frequency, the lowest allele frequency at a locus in a population
Mendel's Law	Allele pairs separate randomly, or segregate, from each other during the production of gametes: egg and sperm. Because allele pairs separate during gamete production, a sperm or egg carries only one allele for each inherited trait. When sperm and egg unite at fertilization, each contributes its allele, restoring the paired condition in the offspring.
Mutation	A permanent change of the nucleotide sequence of the genome of an organism, virus, or extrachromosomal DNA or other genetic elements
Panmictic	A panmictic population is one where all individuals are potential partners, e.g., no mating restrictions
Panmixia	or Panmixis, means random mating
Pedigree	A fully documented genealogy for a fixed system of breeding
Phenotype	The composite of an organism's observable characteristics or traits
SNP	Single nucleotide polymorphism, a DNA sequence variation occurring when a single nucleotide — adenine (*A*), thymine (*T*), cytosine (*C*), or guanine (*G*) — in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual)

*some references from https://en.wikipedia.org

Chapter 1: Introduction

1.1 Single Nucleotide Polymorphism (SNP)

The human genome contains a complete set of genetic information, which is encoded as DNA sequences within twenty-three chromosome pairs. Most human cells have two versions of each chromosome, one inherited from the father and the other inherited from the mother. Therefore, at each specific location (locus) along the chromosome, there are two versions of the DNA sequence. A variant of the DNA sequence at a given locus is called an allele. In any particular diploid organism like human being, the genotype for each gene comprises of a pair of alleles present at that locus, which are the same in homozygotes and different in heterozygotes. Single nucleotide polymorphism (SNP, pronounced as "snip") is a DNA sequence variation occurring when a single nucleotide — adenine (*A*), thymine (*T*), cytosine (*C*), or guanine (*G*) — in the genome (or other shared sequence) differs among members of a species (or between paired chromosomes in an individual). For example, a two-allele SNP may contain two sequenced DNA fragments *ATTCCG* and *ATTTCG*, which only differ in the fourth nucleotide, changing from *C* to *T*. In this case, we say that there are two alleles: *C* and *T*. Most commonly, these variations are often found in DNA non-coding regions between genes. Nevertheless, they may still have an effect on health and development. These are called linked (or indicative) SNPs. Other variations are within a gene or in a regulatory region of a gene. They can affect how human develop disease and/or respond to pathogens, drugs and vaccines, etc. Thus they are called causative SNPs. For example, a single base mutation in the APOE (apolipoprotein E) gene is associated with a higher risk for Alzheimer's disease (Wolf, Caselli, Reiman, & Valla, 2013). These causative SNPs, therefore, can act as biological markers for early diagnosis. In practice, SNPs are currently used to screen genes that may be associated with disease. They can also be used to study the inheritance of disease genes within families and possible gene-disease associations for more complex diseases such as diabetes, heart disease and cancer, etc. For those cases, usually a group of SNPs work in coordination to manifest a disease condition.

Not all single nucleotide changes are called SNPs. Rare polymorphisms that cause disease with high probability are often referred to as mutations. To be classified as a SNP, two or more alleles

must each be present in at least one percent (1%) of the population. The lowest allele frequency at a locus in a population is called a minor allele frequency (MAF). There are variations for the SNP frequency among human populations. A SNP that is common in one ethnic group or geographic region may be much rarer than in the other. In the National Center for Biotechnology Information (NCBI)'s Database of Single Nucleotide Polymorphisms (*dbSNP*), there are more than 10 million SNPs in the human genome (Wheeler, 2007). These SNPs occur throughout the 3-billion-nucleotide human genome with approximately one in every 300 nucleotide base pairs.

SNPs have been widely used in genome-wide association studies (GWAS) as gene markers related to diseases or important traits. For example, a case-control study is often used to compare the distribution of SNPs between one healthy control group and another with disease. In order to rule out the effect of population subdivision, inbreeding or other evolutionary influence on the gene disease association and to confirm the independence within a population of an individual's alleles at a locus, a test of Hardy-Weinberg Equilibrium is often performed for the study sample before any further screening of SNPs.

1.2 Hardy-Weinberg Equilibrium (HWE)

If the union of gametes to produce the next generation is random, it can be shown that allele frequencies are constant from generation to generation. The Hardy-Weinberg Equilibrium (HWE), named after Godfrey H. Hardy and Wilhelm Weinberg, states that genetic variation including allele and genotype frequency in a population will remain constant from one generation to the next in the absence of disturbing influences. This principle is an ideal condition since one or more influences such as nonrandom mating, mutation, natural selection, gene flow (migration), and gene drift, are always present in a real population.

Non-random mating, mostly caused by inbreeding, is one of the major sources of deviation from HWE by introducing more homozygotes into the population. Mutation, a change of the nucleotide sequence of the genome, introduces possible new alleles and changes allele frequencies in a population. The mutation probability is typically very small, an order of 10^{-6} or 10^{-5}. Its influence on the population genetics alone is limited (Nei, 1987). Natural selection affects the allele frequencies and sometimes leads to the loss of all alleles except the favored one. Migration links

two or more populations together. This results in the transfer of alleles or genes from a more homogeneous population to the other. Migration into or out of a population may be responsible for a change in allele frequencies within such population. Sometimes, immigration introduces new alleles to the established gene pool of the existing population. Gene drift changes allele frequencies due to random sampling in a small population (Masel, 2011). Genes are inherited from parents. Therefore, the alleles in the offspring can be considered as a sample of those in the parents. In a small population, chance plays a great role in deciding each individual's survival and the ability to reproduce. This may cause certain gene variants to disappear completely. Gene flow transfers alleles or genes from one population to the other. Furthermore, in genetic studies, sampling stratification and small sample size of the study may also statistically alter HWE from observed data. Lastly, even with today's advanced technology in DNA sequencing, SNPs with genotype error are often found to have substantial excesses or deficiencies of heterozygosity which ultimately results in outlying departure from HWE (Weir B. , 2010).

Throughout this dissertation, a single digit subscript is used to represent the allele property and a double digit subscript is for the genotype property unless otherwise specified. To be consistent with the genetics literature, the term frequency is used instead of proportion. Suppose at a single locus from a large monoecious diploid population, there are k alleles $A_1, A_2, \ldots, A_k$ with allele frequencies $p_1, p_2, \ldots, p_k$ ($\sum_{i=1}^{k} p_i = 1$). This diploid organism is homozygous at this locus when its cells contain the same allele pair of a gene, denoted as A_iA_i or (A_i, A_i), i=1,…,k. The organism's genotype is a homozygote and can be represented as A_{ii} (i=1,…,k). This diploid organism is heterozygous at this locus when its cells contain different alleles of a gene as A_iA_j, or (A_i, A_j), i, j=1,…,k, $i \neq j$. The organism's genotype is a heterozygote and can be denoted as A_{ij} (i=1,…,k; j=i+1,…, k). Each organism produces the same number of male and female gametes and has two alleles at this locus. The copies of a gene separate so that each gamete receives only one allele. Each parent randomly contributes a single allele copy to the offspring. With current widely used gene sequence data, we cannot differentiate the allele pairs (A_i, A_j) and (A_j, A_i) in the offspring and both pairs contain the same genetic information, thus the same genotype A_{ij} ($j \geq i$). Therefore, these k alleles result in a total of $k(k+1)/2$ distinct genotypes $A_{11}, A_{12}, \ldots, A_{1k}, A_{22}, \ldots, A_{2k}, \ldots, A_{kk}$ with corresponding genotype frequencies $p_{11}, p_{12}, \ldots, p_{1k}, p_{22}, \ldots, p_{2k}, \ldots, p_{kk}$ ($\sum_{i=1}^{k} \sum_{j=i}^{k} p_{ij} = 1$).

In a sample of size n, there are $2n$ alleles at a single locus. Let n_{ij} be the number of genotype A_{ij}, $i = 1, \dots, k; j = i, \dots, k$. Let $n = \sum_{i=1}^{k} \sum_{j=i}^{k} n_{ij}$ be the total sample size, and n_l be the number of allele A_l (l=1, …, k) in the sample. Table 1 shows the genotype counts and allele counts from a single locus with k alleles. In order to count the allele A_l, we need to count all genotypes containing A_l (both vertically and horizontally in Table 1), $n_l = \sum_{i=1}^{l} n_{il} + \sum_{j=l}^{k} n_{lj}$ and $\sum_{l=1}^{k} n_l = 2n$. The last equation is true because each subject in the sample has two alleles at the same locus. From Table 1, the sample estimator of genotype frequency and allele frequency are defined as follows:

$$\begin{aligned} &\hat{p}_{ij} = \frac{n_{ij}}{n}, i = 1, \dots, k; j = i, .., k; \text{and } \hat{p}_i = \frac{n_i}{2n}, i = 1, \dots, k; \\ &\hat{p}_l = \frac{n_l}{2n} = \frac{\sum_{i=1}^{l} n_{il} + \sum_{j=l}^{k} n_{lj}}{2n} = \frac{1}{2}\left(\sum_{i=1}^{l} \hat{p}_{il} + \sum_{j=l}^{k} \hat{p}_{lj}\right) \\ &\qquad = \hat{p}_{ll} + \frac{1}{2}\left(\sum_{i=1}^{l-1} \hat{p}_{il} + \sum_{j=l+1}^{k} \hat{p}_{lj}\right) \end{aligned} \tag{1}$$

Under random mating, organisms produce offspring by the random union of gametes. As defined earlier, $p_{ii} = P(A_{ii}) = P(A_i A_i) = P(A_i)P(A_i)$, where the last equality is due to independence of receiving one copy of an allele from parents. Similarly, $p_{ij} = P(A_{ij}) = P(A_i A_j \cup A_j A_i) = P(A_i A_j) + P(A_j A_i) = P(A_i)P(A_j) + P(A_j)P(A_i) = 2P(A_i)P(A_j)$, where the third equality sign is due to mutually exclusive events and the fourth equality sign is due to independence of receiving one copy of an allele from parents. Therefore, the relationship between genotype frequency p_{ij} and allele frequency p_i can be written as

$$\begin{aligned} &p_{ii} = {p_i}^2, i = 1, \dots, k; \\ &p_{ij} = 2p_i p_j, i = 1, \dots, k; j = i + 1, \dots, k. \end{aligned} \tag{2}$$

From above relationship, it can be easily verified that both total genotype frequencies and total allele frequencies equal one, e.g., $\sum_{i=1}^{k} \sum_{j=i}^{k} p_{ij} = \sum_{i=1}^{k} {p_i}^2 + \sum_{i=1}^{k} \sum_{j=i+1}^{k} 2p_i p_j = (\sum_{i=1}^{k} p_i)^2 = 1$. These frequencies (2) define the Hardy-Weinberg Equilibrium. The genotype frequencies of the next generation depend only on the allele frequencies of the current population. Denoting ${p_l}^{(t+1)}$ to represent the allele A_l frequency at generation t+1, under HWE we have:

$$p_l^{(t+1)} = p_{ll}^{(t+1)} + \frac{1}{2}\left(\sum_{i=1}^{l-1} p_{il}^{(t+1)} + \sum_{j=l+1}^{k} p_{lj}^{(t+1)}\right)$$

$$= \left(p_l^{(t)}\right)^2 + p_l^{(t)}\left(\sum_{i=1}^{l-1} p_i^{(t)} + \sum_{j=l+1}^{k} p_j^{(t)}\right)$$

$$= p_l^{(t)} \sum_{i=1}^{k} p_i^{(t)}$$

$$= p_l^{(t)}.$$

Table 1. Genotype Count from a Single Locus with k Alleles

♂ / ♀	A_1	A_2	A_3	$\ldots A_l \ldots$	A_k	*Total*
A_1	n_{11}	n_{12}	n_{13}	$\ldots n_{1l} \ldots$	n_{1k}	$\sum_{j=1}^{k} n_{1j}$
A_2	n_{21}	n_{22}	n_{23}	$\ldots n_{2l} \ldots$	n_{2k}	$\sum_{j=2}^{k} n_{2j}$
A_3	n_{31}	n_{32}	n_{33}	$\ldots n_{3l} \ldots$	n_{3k}	$\sum_{j=3}^{k} n_{3j}$
…	…	…	…	…	…	
A_l	n_{l1}	n_{l2}	n_{l3}	n_{ll}	n_{lk}	$\sum_{j=l}^{k} n_{lj}$
…	…	…	…	…	…	
A_k	n_{k1}	n_{k2}	n_{k3}	$\ldots n_{kl} \ldots$	n_{kk}	$\sum_{j=k}^{k} n_{kj}$
Total	$\sum_{i=1}^{k} n_{i1}$	$\sum_{i=2}^{k} n_{i2}$	$\sum_{i=3}^{k} n_{i3}$	$\sum_{i=l}^{k} n_{il}$	$\sum_{i=k}^{k} n_{ik}$	n

Note: The greyed upper triangle is the observed genotype count; the lower triangle is a reflection of the observed count under slightly different notation such that $n_{ji} = n_{ij}, j > i$.

Thus, the allele frequencies remain constant from generation to generation. The first equality sign can be similarly derived as the second equation in (1).

1.3 Inbreeding Coefficient (f)

Among various reasons for the deviation from HWE, our research is motivated by the fact that inbreeding causes the production of offspring from the mating of genetically related individuals. Populations usually do not constitute a single panmictic unit without any mating restriction. They are rather further divided into smaller regions geographically. The inbreeding coefficient *f*, or fixation index F_{IS}, first introduced by Wright (1921), measures the expected percentage of homozygosity based on a known pedigree (a fully documented genealogy for a fixed system of breeding). In this dissertation, we follow the genetics literature by using *f* to denote the inbreeding coefficient. This should be differentiated from the traditional use of *f* in statistics as a density function. The inbreeding coefficient is also defined as the probability that two homologous alleles are identity by descent (IBD), which is, both coming from the same allele copy in a common ancestor. Pedigrees are, however, usually not available for individuals within the study population. Instead, the inbreeding coefficient is an expected value (or a statistical value) derived from individual's pedigree. It is not directly measureable by looking at the individual's genome. In another words, its definition exists even if such individual's genome does not actually contains such gene.

When studying a population, *f* measures the degree of deviation from random pairing of genes, that is, the association between pairs of uniting gametes. A positive *f* means excessive homozygotes, while negative value means excessive heterozygotes in the system. Although an inbreeding coefficient can be parameterized as genotype specific, such as $f_{ij}, i = 1, \dots, k; j = i, \dots, k$, it is often assumed in practice that all pairs of allele frequencies are equally influenced and, therefore, have a common *f*. Under this general inbreeding model, the genotype frequency p_{ij} is a function of the allele frequency p_i, p_j, and the inbreeding coefficient *f*. Cavalli-Sforza and Bodmer (1971) explained that the population can be separated into one fully inbred part with relative proportion *f* and the other bred at random with relative proportion 1 - *f*. For the inbred population, the individual homozygotes A_{ii} genotype frequency by descent therefore is *f* times the allele A_i frequency $p_i, \sum_{i=1}^{K} f p_i = f$. For the bred at random population, the genotype frequencies are (1 –

f) times the genotype frequencies under HWE in equation (2). When *f*=1, there is a fully inbred population and *f*=0 means individuals mate completely at random. Therefore, the total genotype frequencies are the sum of these two fractions.

$$p_{ii} = (1-f){p_i}^2 + fp_i = {p_i}^2 + p_i(1-p_i)f, i = 1, \dots, k.$$
$$p_{ij} = 2(1-f)p_i p_j, i = 1, \dots, k; j = i+1, \dots, k. \quad (3)$$

From the above model, the total probability $\sum_{i=1}^{k}\sum_{j=i}^{k} p_{ij}$ is still one, as demonstrated below:

$$\sum_{i=1}^{k}\sum_{j=i}^{k} p_{ij} = \sum_{i=1}^{k}\sum_{j=i+1}^{k} p_{ij} + \sum_{i=1}^{k} p_{ii}$$
$$= \sum_{i=1}^{k}\sum_{j=i+1}^{k} p_{ij} + \sum_{i=1}^{k}[(1-f){p_i}^2 + fp_i]$$
$$= \sum_{i=1}^{k}\sum_{j=i+1}^{k} p_{ij} + (1-f)\sum_{i=1}^{k} {p_i}^2 + f = 1.$$

Simplifying the last expression above to solve for *f* yields:

$$f = 1 - \frac{\sum_{i=1}^{k}\sum_{j=i+1}^{k} p_{ij}}{1-\sum_{i=1}^{k} {p_i}^2} = 1 - \frac{\sum_{i=1}^{k}\sum_{j=i+1}^{k} p_{ij}}{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} p_i p_j} = 1 - \frac{H_O}{H_E}. \quad (4)$$

In the above equation, $H_O = \sum_{i=1, j>i}^{k} p_{ij}$ is the heterozygotes frequency and $H_E = 2\sum_{i=1}^{k}\sum_{j=i+1}^{k} p_i p_j = 1 - \sum_{i=1}^{k} p_i^2$ is the expected heterozygotes frequency with observed allele frequency under HWE. Therefore, *f* can be expressed as one minus the ratio of the observed proportion of heterozygotes to the expected proportion of heterozygotes under HWE. When $k = 2$, $f = 0$ is the sufficient and necessary condition for HWE, in which case $p_{ij} = 2p_i p_j;\ i = 1, j = 2$. However, when $k > 2$, $f = 0$ is only a sufficient condition for $p_{ij} = 2p_i p_j;\ i = 1, \dots, k; j = i+1, \dots, k$ since it only implies $\sum_{i=1}^{k}\sum_{j=i+1}^{k} p_{ij} = 2\sum_{i=1}^{k}\sum_{j=i+1}^{k} p_i p_j$. For a simple illustration, considering a simple bi-allelic case *k*=2 with alleles *A* and *a*, the alleles frequency are $p_1 = p$ and $p_2 = 1-p$. So the inbreeding coefficient can be expressed as

$$f = 1 - \frac{p_{12}}{2p(1-p)} \quad (5)$$

To understand why the inbreeding coefficient is a measure of correlation, suppose we draw two gametes (G_1 and G_2) at random from this population with allele *A* frequency *p*. Assign indicator variable *Y*=1 if a gamete is of allele *A* and *Y*=0 if it is of allele type *a*. Therefore we see immediately that over replicates of this sampling process, $E(y) = 1 \times p + 0 \times (1-p) = p, E(y^2) = p$, $Var(y) = E(y^2) - (E(y))^2 = p - p^2 = p(1-p)$ and $Cov(y_1, y_2) = E(y_1 y_2) - E(y_1)E(y_2) = p - \frac{p_{12}}{2} - p^2 = p(1-p) - p_{12}/2$ as shown in Table 2. The correlation between two homologous genes in uniting gametes can be expressed as $\frac{Cov(y_1,y_2)}{\sqrt{Var(y_1)Var(y_2)}} = \frac{p(1-p)-\frac{p_{12}}{2}}{p(1-p)} = 1 - \frac{p_{12}}{2p(1-p)}$, where p_{12} is the observed heterozygote frequency (H_O) and $2p(1-p)$ is the expected heterozygosity (H_E) under HWE (Balding, Bishop, & Cannings, 2007). This is exactly the same as was shown in (5).

Table 2. Pattern of Union of Gametes in the Total Population

		Gamete 2		
		A	*a*	Total
Gamete 1	*A*	$p - p_{12}/2$	$p_{12}/2$	*p*
	a	$p_{12}/2$	$1 - p - p_{12}/2$	1-*p*
	Total	*p*	1-*p*	1

Notice that all genotype frequencies range from 0 to 1 under the above inbreeding model. There is a natural constraint for *f*:

$$0 \le p_{ii} = (1-f){p_i}^2 + f p_i \le 1 \Rightarrow \frac{-p_i}{1-p_i} \le f \le 1 + \frac{1}{p_i},$$

$$0 \le p_{ij} = 2(1-f)p_i p_j \le 1 \Rightarrow f \le 1.$$

Therefore, conditional on the allele frequencies, we have a constraint for *f*:

$$\frac{-p_{min}}{1-p_{min}} \le f \le 1, \text{where } p_{min} = min(p_1, \dots, p_k).$$

Although the inbreeding coefficient *f* is considered as a measure of correlation, its range is not [-1,1] as it is for the product moment correlation coefficient. The lower limit is dependent on the

distribution of actual population allele frequency. For k=2, Figure 1 shows that the parameter space for the inbreeding coefficient f is conditional on the allele frequency p. When p=0.5, the range of f is fully [-1, 1]. The range slowly decreases to [0, 1] when the allele frequency moves toward either end (0 or 1). For the special case when the MAF is at the cutoff for the definition of SNP, e.g., p=0.01, the parameter space for f is [-0.0101, 1].

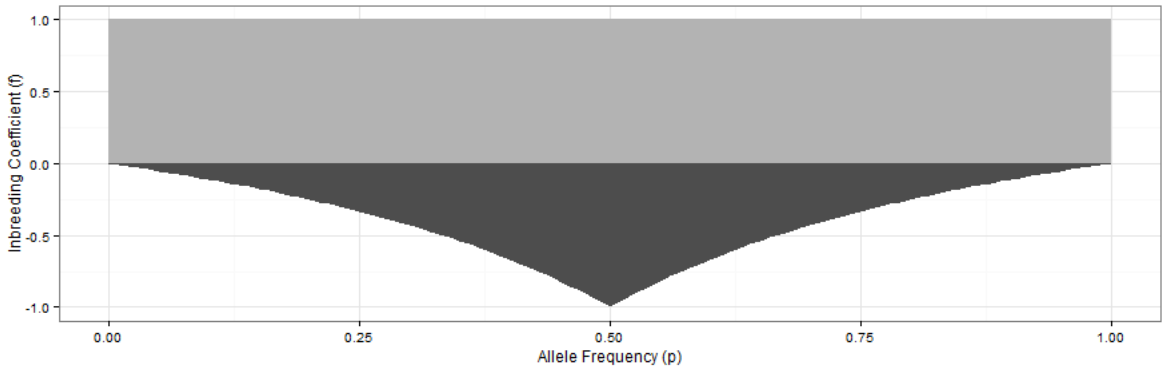


Figure 1. Parameter Space for the Inbreeding Coefficient (f) Conditional on the Allele Frequency (p) Under the General Inbreeding Model. Dark gray area indicates negative f, while light gray area indicates positive f.

When the inbreeding coefficient is estimated from a single sample, it measures the deviation from random pairing of the genes. Sampling from a whole population may be considered as a random sample from an infinitely large pool of zygotes (Curie-Cohen, 1982). Suppose a sample of size n subjects are taken from the target population, and we count the number of subjects with each genotype. By substituting sample values in equation (4), a Direct Sample Estimator (DSE) $f^{(s)}$ can be derived as a function of observed and expected number of heterozygotes under HWE (Li & Horvitz, 1953):

$$f^{(s)} = 1 - \frac{2n\sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}}{\sum_{i=1}^{k}\sum_{j=i+1}^{k} n_i n_j}. \quad (6)$$

The estimator compares this sample with an infinitely large population with no mutation, migration, natural selection and with random mating except for a fixed number of inbred mating.

The above general inbreeding model reduces to HWE when f=0. Therefore, the HWE model is nested under this general model. Instead of a chi-square test based on the expected and the observed counts, the testing of HWE can be constructed as the testing of the null hypothesis H_0: f=0 versus the alternative hypothesis H_1: $f \neq 0$.

In GWAS, researchers are more interested in testing the HWE than in estimating the inbreeding coefficient f. Genes that deviate from HWE are, therefore, screened out in the gene-disease association studies. However, a statistical test based on the frequentist p-value does not always have a direct interpretability to geneticists. The test does not directly measure the effect size f, i.e., the magnitude of deviation. A relatively large f with a non-significant p-value does not prove that a population is in HWE. On the other hand, when the sample size increases toward infinity, a highly significant p-value does not necessarily warrant a final conclusion of departure from HWE when the estimated f is often only barely worth mentioning. For studies with multiple SNPs screening, a simple Bonferroni correction is often performed to adjust the final p-values of tests (Weir, Hill, & Cardon, 2004). It is not clear whether the estimated f with its associated confidence interval should also be adjusted by the multiple estimations in a similar way. For human genetics, there is another challenge because the amount of inbreeding usually is extremely small. Table 7.3 of Cavalli-Sforza & Bodmer (1971) gives estimates of the inbreeding coefficients for a wide range of populations. The average of the inbreeding coefficients in human populations is generally less than 0.0001; an inbreeding coefficient larger than 0.01 is considered exceptional.

1.4 Multinomial Model for the Genotype Distribution

For a total sample size of n subjects with observed numbers of genotypes n_{ij} ($\sum_{i=1}^{k}\sum_{j=i}^{k} n_{ij} = n$) and genotype frequency p_{ij}, i=1,…, k, $j \geq i$, the $k(k+1)/2$ dimensional vector $\boldsymbol{n} = (n_{11}, n_{12}, \dots, n_{1k}, n_{22}, \dots, n_{2k}, \dots, n_{kk})$ follows a multinomial distribution with parameter $\tilde{\boldsymbol{p}} = (p_{11}, p_{12}, \dots, p_{1k}, p_{22}, \dots, p_{2k}, \dots, p_{kk})$:

$$P(\boldsymbol{n}|\tilde{\boldsymbol{p}}) = \frac{n!}{\prod_{i=1}^{k}\prod_{j=i}^{k} n_{ij}!}\prod_{i=1}^{k}\prod_{j=i}^{k} {p_{ij}}^{n_{ij}}.$$

We use $\tilde{\boldsymbol{p}}$ temporarily here for the genotype frequencies, to distinguish it from the notation for the allele frequencies. The marginal distribution of n_{ij} is a binomial with parameter p_{ij}:

$$P(n_{ij}|p_{ij}) = \frac{n!}{n_{ij}!\,(n-n_{ij})!} {p_{ij}}^{n_{ij}} (1-p_{ij})^{n-n_{ij}}.$$

Therefore, the moment estimators of the genotype frequency and allele frequency can be estimated as sample proportions: $\hat{p}_{ij} = \frac{n_{ij}}{n}$ and $\hat{p}_i = \frac{2n_{ii}+(n-n_{ii})}{2n} = \frac{2n_{ii}+\sum_{j:j\neq i} n_{ij}}{2n}$, respectively. They are both unbiased estimators since

$$E(\hat{p}_{ij}) = E\left(\frac{n_{ij}}{n}\right) = p_{ij}, \text{and } E(\hat{p}_i) = E\left(\frac{2n_{ii}+\sum_{j:j\neq i} n_{ij}}{2n}\right) = p_{ii} + \frac{1}{2}\sum\nolimits_{j:j\neq i} p_{ij} = p_i.$$

However, the direct sample estimator of the inbreeding coefficient,

$$f^{(s)} = 1 - \frac{2n\sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}}{\sum_{i=1}^{k}\sum_{j=i+1}^{k} n_i n_j} = 1 - \frac{\sum_{i=1}^{k}\sum_{j=i+1}^{k} \hat{p}_{ij}}{1-\sum_{i=1}^{k} {\hat{p}_i}^2},$$

is consistent but biased (Gorroochurn & Hodge, 2006). Using the delta method, Cohen (1982) showed that the bias is of the order of n^{-1}.

The multinomial model can be further re-parameterized as:

$$P(\boldsymbol{n}|\boldsymbol{p}, f) = \frac{n!}{\prod_{i=1}^{k}\prod_{j=i}^{k} n_{ij}!} \prod\nolimits_{i=1}^{k} ((1-f){p_i}^2 + f p_i)^{n_{ii}} \prod\nolimits_{i=1}^{k}\prod\nolimits_{j=i+1}^{k} (2(1-f)p_i p_j)^{n_{ij}},$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$ is a $k \times 1$ dimensional vector with $p_k = 1 - \sum_{i=1}^{k-1} p_i$. Therefore, the log-likelihood is

$$\begin{aligned} l(\boldsymbol{p}, f) &= \log\frac{n!}{\prod_{i=1}^{k}\prod_{j=i}^{k} n_{ij}!} + \sum_{i=1}^{k} n_{ii}\log\left((1-f){p_i}^2 + f p_i\right) \\ &\quad + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log\left(2(1-f)p_i p_j\right) \\ &= \log\frac{n!}{\prod_{i=1}^{k}\prod_{j=i}^{k} n_{ij}!} + \sum_{i=1}^{k} n_{ii}\log\left((1-f)p_i + f\right) + \sum_{i=1}^{k} n_{ii}\log(p_i) \\ &\quad + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log 2 + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log(1-f) \\ &\quad + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log(p_i) + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log(p_j) \end{aligned}$$

$$= C + \sum_{i=1}^{k} n_{ii} \log\big((1-f)p_i + f\big) + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log(1-f) + \sum_{i=1}^{k}\sum_{j=i}^{k} n_{ij}\log(p_i) + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log(p_j)$$

$$= C + \sum_{i=1}^{k} n_{ii} \log\big((1-f)p_i + f\big) + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log(1-f) + \sum_{i=1}^{k} (n_i - n_{ii})\log(p_i)$$

$$= C + \sum_{i=1}^{k-1} n_{ii} \log\big((1-f)p_i + f\big) + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log(1-f) + \sum_{i=1}^{k-1} (n_i - n_{ii}) \log(p_i) + n_{kk}\log\left((1-f)(1-\sum_{i=1}^{k-1} p_i) + f\right) + (n_k - n_{kk})\log(1-\sum_{i=1}^{k-1} p_i),$$

where $C = \log\frac{n!}{\prod_{i=1}^{k}\prod_{j=i}^{k} n_{ij}!} + \sum_{i=1}^{k}\sum_{j=i+1}^{k} n_{ij}\log 2$ and $n_i = \sum_{l=i}^{k} n_{il} + \sum_{m=1}^{i} n_{mi}$.

1.4.1 Existing Frequentist Estimator

Differentiating the above log-likelihood with respect to p_i and f and setting these equal to 0 yields:

$$\frac{\partial l}{\partial p_i} = \frac{n_{ii}(1-f)}{(1-f)p_i + f} + \frac{n_i - n_{ii}}{p_i} - \frac{n_{kk}(1-f)}{(1-f)p_k + f} - \frac{n_k - n_{kk}}{p_k} = 0,$$

$$\frac{\partial l}{\partial f} = \sum_{i=1}^{k} \frac{n_{ii}(1-p_i)}{(1-f)p_i + f} - \sum_{i=1,j>i}^{k} \frac{n_{ij}}{1-f} = 0.$$

By solving the above equations numerically using the Newton's method (Monohan, 2011), the estimators of $\hat{p}_i$ and $\hat{f}$ can be derived. The Hessian matrix, including the following second partial derivatives of the log-likelihood, can also be estimated

$$\frac{\partial^2 l}{\partial {p_i}^2} = -\frac{n_{ii}(1-f)^2}{\big((1-f)p_i + f\big)^2} - \frac{n_i - n_{ii}}{{p_i}^2} + \frac{n_{kk}(1-f)^2}{\big((1-f)p_k + f\big)^2} + \frac{n_k - n_{kk}}{{p_k}^2},$$

and

$$\frac{\partial^2 l}{\partial f^2} = -\sum_{i=1}^{k} \frac{n_{ii}(1-p_i)^2}{\big((1-f)p_i + f\big)^2} + \sum_{i=1,j>i}^{k} \frac{n_{ij}}{(1-f)^2}.$$

The variance of the estimator can be calculated as $var(\hat{f}) = -E(\frac{\partial^2 l}{\partial f^2})$ and $var(\hat{p}_i) = -E(\frac{\partial^2 l}{\partial {p_i}^2})$ using the information matrix.

When k=2, let p be the population proportion of allele A, and q be the proportion of allele a, where $p + q$ =1. Under the general inbreeding model, the genotype frequencies can be reparametrized as

$$\begin{cases} p_{11} = p^2 + pqf = (1-f)p^2 + fp, \\ p_{12} = 2pq(1-f), \\ p_{22} = q^2 + pqf = (1-f)q^2 + fq. \end{cases}$$

The log-likelihood is given by

$$l(p,f) = log\left\{\frac{n!}{n_{11}!\,n_{12}!\,n_{22}!}\left((1-f)p^2 + fp\right)^{n_{11}}(2(1-f)pq)^{n_{12}}\left((1-f)q^2 + fq\right)^{n_{22}}\right\}$$

$$= log\frac{n!}{n_{11}!\,n_{12}!\,n_{22}!} + n_{11}\,log\left((1-f)p^2 + fp\right) + n_{12}\,log(2(1-f)pq) + n_{22}log\left((1-f)q^2 + fq\right).$$

Letting q = 1 – p and taking the first derivatives with respective to p and f and setting these equal to 0, we have

$$\begin{cases} \frac{\partial l}{\partial p} = \frac{n_{11}(2p(1-f)+f)}{(1-f)p^2 + fp} + \frac{n_{12}(2(1-f)(1-2p))}{2(1-f)pq} + \frac{n_{22}((2p-2)(1-f)-f)}{(1-f)q^2 + fq} = 0, \\ \frac{\partial l}{\partial f} = \frac{n_{11}(p-p^2)}{(1-f)p^2 + fp} + \frac{n_{12}(-2pq)}{2(1-f)pq} + \frac{n_{22}(q-q^2)}{(1-f)q^2 + fq} = 0. \end{cases}$$

In order to simplify the presentation, let $r = f/(1-f)$; then $f = r/(r+1)$. Substituting f into the above equations gives

$$\begin{cases} \frac{n_{11}(2p+r)}{p(p+r)} + \frac{n_{12}(1-2p)}{pq} + \frac{-n_{22}(2q+r)}{q(q+r)} = 0, \\ \frac{n_{11}q}{p+r} - n_{12} + \frac{n_{22}p}{q+r} = 0. \end{cases}$$

Multiplying by $(1-2p)/pq$ in the second equation and adding these two equations together to eliminate the middle term yields

$$\frac{n_{11}(2p+r)}{p(p+r)} + \frac{n_{11}q}{p+r} * \frac{1-2p}{pq} = \frac{n_{22}(2q+r)}{q(q+r)} - \frac{n_{22}p}{q+r} * \frac{1-2p}{pq} \Leftrightarrow$$

$$\frac{n_{11}(2pq+qr+q-2pq)}{pq(p+r)}=\frac{n_{22}(2pq+pr-p+2p^2)}{pq(q+r)}\Leftrightarrow$$

$$\frac{n_{11}(q+qr)}{p+r}=\frac{n_{22}(p+pr)}{q+r}\Leftrightarrow$$

$$\frac{n_{11}q}{p+r}=\frac{n_{22}p}{q+r}.$$

Combining above result with earlier equation $\frac{n_{11}q}{p+r}-n_{12}+\frac{n_{22}p}{q+r}=0$ gives

$$\begin{cases}\frac{n_{11}q}{p+r}=\frac{n_{12}}{2},\\ \frac{n_{22}p}{q+r}=\frac{n_{12}}{2}.\end{cases}$$

Rewriting the above equation only with respect to *p* and *r* gives

$$\begin{cases}n_{11}(1-p)=\frac{n_{12}}{2}p+\frac{n_{12}}{2}r,\\ n_{22}p=\frac{n_{12}}{2}(1-p)+\frac{n_{12}}{2}r.\end{cases}$$

Solving the above equation for *p*, we have estimator $\hat{p}=\frac{2n_{11}+n_{12}}{2(n_{11}+n_{12}+n_{22})}$. This estimator is also valid when *f*=1 ($n_{12}=0$ in such case). Furthermore, solving the above equation for *r*, we have $\hat{r}=\frac{2n_{11}(1-\hat{p})}{n_{12}}-\hat{p}$. Therefore,

$$\hat{f}=\frac{\hat{r}}{1+\hat{r}}=1-\frac{1}{1+\hat{r}}=1-\frac{n_{12}}{2n_{11}(1-\hat{p})-n_{12}\hat{p}+n_{12}}$$

$$=1-\frac{n_{12}}{(2n_{11}+n_{12})(1-\hat{p})}=1-\frac{2n*n_{12}}{(2n_{11}+n_{12})(2n_{22}+n_{12})},$$

which is a special case of $f^{(s)}=1-\frac{2n\sum_{i=1}^{k}\sum_{j=i+1}^{k}n_{ij}}{\sum_{i=1}^{k}\sum_{j=i+1}^{k}n_i n_j}$ when *k*=2. Further simplification yields

$$\hat{f}=1-\frac{2n*n_{12}}{(2n_{11}+n_{12})(2n_{22}+n_{12})}=1-\frac{n_{12}}{2n\hat{p}(1-\hat{p})}, \quad (7)$$

where n_{12} is the observed number of heterozygotes, and $2n\hat{p}(1-\hat{p})$ is the expected number of heterozygotes under HWE (*f*=0). Therefore, $\hat{f}$ can be treated as a measure from the deviation in heterozygosity from the expected under random mating.

The Hessian matrix can be calculated by taking the second partial derivatives of the log-likelihood. However, even for *k*=2, many terms are involved to simply the second derivatves. Without showing that at least one second partial derivative is negative and that the Jacobian of the second partial derivatives is positive, the solution derived in (7) may not be the MLE.

From equation (7), it is obvious that $\hat{f} \leq 1$. The equality sign holds when there is no observed heterozygotes, e.g., $n_{12} = 0$. When the population allele *A* frequency *p* is known, from above equation $n_{11}(1-p) = \frac{n_{12}}{2}p + \frac{n_{12}}{2}r$, we have $\hat{f} = 1 - \frac{n_{12}}{(2n_{11}+n_{12})(1-p)} = \frac{2n_{11}-(2n_{11}+n_{12})p}{(2n_{11}+n_{12})(1-p)} = \frac{-p+\frac{2n_{11}}{2n_{11}+n_{12}}}{1-p} > \frac{-p}{1-p}$. Similarly, from equation $n_{22}p = \frac{n_{12}}{2}(1-p) + \frac{n_{12}}{2}r$ and conditional on population allele *a* frequency *q*, we have $\hat{f} = \frac{-q+\frac{2n_{22}}{2n_{22}+n_{12}}}{1-q} > \frac{-q}{1-q}$. Therefore, for this special case of *k*=2, conditional on the allele frequency *p* (or equivalently, *q*), the estimator $\hat{f}$ satisfies the constraint $max(\frac{-p}{1-p}, \frac{-q}{1-q}) = \frac{-min(p,q)}{1-min(p,q)} \leq \hat{f} \leq 1$. However, in most genetic studies, the population allele frequency is unknown and needs to be estimated from the data. The estimator $\hat{p}$ is included in the calculation of $\hat{f}$. Therefore, $\hat{f}$ may be outside of the parameter space induced from the inbreeding model. When the allele frequency *p* is close to 0.5, this is not likely to be a serious issue since the parameter space for *f* is in the [-1, 1] interval. However, when the minor allele frequency (MAF) approaches zero, the lower limit of *f* shrinks toward zero (Figure 1). The estimator may be seriously outside of parameter space, especially when the sample size is small. When the sample size goes to infinity, the consistent estimator $\hat{p}$ converges to the parameter *p*. Under this large sample size condition, asymptotically the estimator $\hat{f}$ resides within its parameter space for the reason we just outlined earlier.

Using the property of binomial distribution and Taylor expansion, Cohen (1982) showed that

$$E(\hat{f}) = E\left(1 - \frac{n_{12}}{2\hat{p}\hat{q}n}\right) = 1 - E\left(\frac{n_{12}}{2\hat{p}\hat{q}n}\right)$$

$$= 1 - \frac{2(1-f)pqn}{2npq - pq(f+1)} + \frac{2npq(p-q)^2(1-f)}{\left(2npq - pq(f+1)\right)^2} - \frac{2npq(p-q)^2(1+f) * 2(1-f)pqn}{\left(2npq - pq(f+1)\right)^3}$$

$$+ O\left(\frac{1}{n}\right)$$

$$= f + O\left(\frac{1}{n}\right).$$

Thus, $\hat{f}$ has a bias of the order of n^{-1}. Moreover, Cohen (1982) showed that the asymptotic variance of the estimator $\hat{f}$ is given by

$$Var(\hat{f}) = Var\left(1 - \frac{n_{12}}{2\hat{p}\hat{q}n}\right) = Var\left(\frac{n_{12}}{2\hat{p}\hat{q}n}\right)$$

$$= \frac{2(1-f)pqn[1-2(1-f)pq]}{\left(2npq - pq(f+1)\right)^2}$$
$$+ \frac{(2(1-f)pqn)^2 2npq(p-q)^2(1+f) - 4(1-f)pqn\left(2npq - pq(f+1)\right)2npq(p-q)^2(1-f)}{\left(2npq - pq(f+1)\right)^4}$$
$$+ O\left(\frac{1}{n^2}\right)$$

$$= \frac{n^3(1-f)(2pq)^2[2pq - (1-f)(2pq)^2 - (1-f)^2 2pq(p-q)^2]}{\left(2npq - pq(f+1)\right)^4} + O\left(\frac{1}{n^2}\right)$$

$$\approx \frac{n^3(1-f)(2pq)^2 2pq[1 - 2pq(1-f) - (p-q)^2(1-f)^2]}{(2npq)^4}$$

$$= \frac{(1-f)[1 - 2pq(1-f) - (p-q)^2(1-f)^2]}{2npq}. \quad (8)$$

The asymptotic sample variance of the estimator $\hat{f}$ is dependent on p and f. Since $pq = p(1-p) = -\left(p - \frac{1}{2}\right)^2 + \frac{1}{4}$, pq is an increase function in $0 < p \leq \frac{1}{2}$ and reaches its maximum value at p=q=1/2. At the same time, $(p-q)^2$ reaches its minimum value of zero at p=q=1/2. Several observations for the asymptotic variance estimator can be made from equation (8):

1.) The asymptotic variance can be further simplified as $Var(\hat{f}) = \frac{f(2-f)}{2npq} + \frac{(f-1)(2f-1)}{n}$.

When p=1/2, the variance of $\hat{f}$ towards $\frac{(1-f)(1+f)}{n}$ and it reaches its minimal value;

2.) When allele frequency is very small close to zero, the variance of $\hat{f}$ towards $\frac{(1-f)-(1-f)^3}{2npq}$. Therefore, small increase of f may have large increase in its variance when the allele frequency is small (Cohen, 1982).

3.) When the inbreeding coefficient is f=0, the variance of $\hat{f}$ becomes $\frac{1-2pq-(p-q)^2}{2npq} = \frac{1-(p+q)^2+2pq}{2npq} = \frac{1}{n}$. It does not depend on the allele frequency, but only depends on the sample size (Cohen, 1982). Under the null hypothesis of f=0, a Z-test statistic can be calculated as $Z = \frac{\hat{f}-0}{1/\sqrt{n}} = \sqrt{n}\,\hat{f}$. The Z-test of HWE is strongly related to the sample size and observed inbreeding coefficient. For example, when the observed inbreeding coefficient is as moderate as 0.196, the test will reach a statistical significance at 0.05 level when the sample size is 100 or larger.

4.) For the estimator of sample variance of the estimated inbreeding coefficient $\widehat{Var}(\hat{f})$, an estimator of $\hat{p}$ and $\hat{f}$ can be used in the equation (8) instead.

Except for the above simple bi-allelic model (k=2), generally the estimator $\hat{f}$ is not equal to the direct sample estimator $f^{(s)}$ for k≥3. This is not surprising since homozygotes with alleles that are identity by decent (IBD) only provide partial information than the alleles that are not IBD under the influence of inbreeding (Gorroochurn & Hodge, 2006). Although the solution of f cannot be explicitly written when k≥3, it is proven to be biased when k=2 (Li & Horvitz, 1953). Unfortunately, all eight estimators for f proposed by Li & Horvitz (1953) including the above two ($\hat{f}, f^{(s)}$), are biased estimators, even though all of them are essentially identical when k=2.

1.4.2 Bayesian Method for HWE and Estimation

In parallel with the recent increased interest in testing HWE among human genetics studies, Bayesian methods have found enormous applications in GWAS where there are large amount of single nucleotide polymorphism (SNPs) examined from the same or different genes. The rationale for controlling of the family-wise error rate is not obvious in a genome-wide context when it is practically not reasonable to expect all nulls to be true. Therefore, traditional frequentist methods such as Bonferroni correction to control the family level type I error may pose a major risk that true genetic association might be discarded. Even if we accept the concept that type I error need

be controlled, the choice of a specific threshold for all SNPs in a study is difficult (Stephens 2009). In Health and Retirement Study genotype Quality Control report (University of Washington, 2012), a *p*-value cutoff of 0.0001 is recommended for testing HWE. However, other literature propose a more strict cutoff as low as 5×10^{-7} (Pongpanich, Sullivan, & Tzeng, 2010). Bayesian method, on the other hand, can provide measure of evidence that can be interpreted directly and even be compared among different SNPs without worrying about the interpretation of the *p*-value and a fixed error probability α resulted in multiple testing. The main advantage of Bayesian method is its flexibility to incorporate non-sampling information into analysis. For example, under HWE assumption, dominant or co-dominant genetic model as well as prior belief of allele frequency distributions can be considered altogether into the estimation of the allele frequency from nominal measure of the phenotype (Gunel & Wearden, 1995). Using the multinomial distribution, under both HWE and general model, conjugate Dirichlet priors are often applied to the allele or genotype frequencies to derive the Bayes Factor for model selection (Wakefield, 2010). Conssoni etc. (2008) discussed various compatible priors, including Kullback-Leibler Conjugate Approximation (KLCA) and Jeffreys' Conditioning (JC) priors, for the testing of HWE. Later they further introduced a class of objective intrinsic priors for the same testing of HWE model and carried out a sensitivity analysis for the prior (Consonni, Morenob, & Venturinic, 2011).

When there are potential genotype error associated with the gene polymorphism, small sample size, or the existence of population stratification in the sampling design, the study cohort may not be exactly in HWE. On the other hand, a statistical significant test with point estimate close to zero does not likely to provide convincing evidence of departure from HWE. As an alternative to the hypothesis testing, estimation provides more useful information in understanding of the magnitude of departure from HWE. The Bayesian method uses a prior distribution for the effect-size parameter while not categorizes it into zero and non-zero, and then focus on estimating the parameter instead of testing whether it is equal to zero. With the introduction of Markov Chain Monte Carlo (MCMC), the posterior density is available so that the uncertainty of the estimation could be displayed visually. Bayesian credible interval could be constructed from the posterior distribution. The posterior probability of parameter *f* within interval suggested by the National Research Council (1996) therefore could be calculated. This helps scientists to incorporate

background information and reduce data dependence when investigating specific research question.

One of the other advantages of the Bayesian estimation method is that it naturally incorporates prior information about parameter *f* learned from previous researches. For example, studies from isozyme loci and pedigrees suggest that the inbreeding coefficient *f* is small, usually less than 0.03 and often less than 0.01 (Chakraborty & Jin, 1992; Doeder, Escobar, Kadane, & Balazs, 1998). Under the multinomial model with conjugate beta prior, the prior is weighted against small or large departure from HWE. A step prior, which gives equal weight to both small and large departures from HWE, could be constructed as the conditional distribution of $\varphi(f|p)$ or the joint distribution $\varphi(f,p)$ uniformly distributed over the range of *f*|*p* (Shoemaker, Painter, & Weir, 1998). Using dominant phenotype data from seven different populations of an endangered orchid specie, Holsinger applied a hierarchical binomial model with non-informative beta priors for the parameters *f* an F_{ST}. The posterior distribution for the parameters were approximated by a beta distribution with parameters estimated by its posterior mean and variance from MCMC (Holsinger & Wallace, 2004). The prior beta distribution can be chosen such that the median fits the plausible value under our knowledge for the parameter. Such prior places positive probability on a range of values that more than adequately covers all possible values for the parameter. When genotype data is available from different populations, a hierarchical multinomial model can, therefore, be constructed without further consideration of the genetic model in data analysis.

Based on the multinomial likelihood, a hierarchical model was built to incorporate the hypo-parameters for the genotype frequency (Shoemaker, Painter, & Weir, 1998). At the first level, the genotype count $\boldsymbol{n} = (n_{11}, n_{12}, n_{22})$ follows a multinomial distribution:

$$Pr(\boldsymbol{n}|\tilde{\boldsymbol{p}}) = Pr\big((n_{11}, n_{12}, n_{22})\big|(p_{11}, p_{12}, p_{22})\big) = \frac{n!}{n_{11}!\, n_{12}!\, n_{22}!} {p_{11}}^{n_{11}} {p_{12}}^{n_{12}} {p_{22}}^{n_{22}}.$$

The genotype frequency $\tilde{\boldsymbol{p}} = (p_{11}, p_{12}, p_{22})$ can be re-parameterized as the following:

$$\begin{cases} p_{11} = p^2 + pqf \\ p_{12} = 2pq(1-f). \\ p_{22} = q^2 + pqf \end{cases}$$

Under this multinomial model, the likelihood function is given by

$$l(p,f) = Pr(\boldsymbol{n}|p,f)$$

$$= \frac{n!}{n_{11}!\,n_{12}!\,n_{22}!}((1-f)p^2+fp)^{n_{11}}(2(1-f)pq)^{n_{12}}((1-f)q^2+fq)^{n_{22}}.$$

At the second level, a uniform prior $\varphi(p)$ for allele frequency p is applied. Conditionally on p, further assume a uniform prior distribution $\varphi(f|p)$ for f with constrained parameter space on $\frac{-min(p,1-p)}{1-min(p,1-p)} \le f \le 1$ (Wakefield, 2010). Therefore, the joint prior distribution is a non-informative $\varphi(p,f) = \varphi(p)\varphi(f|p) = 1, 0 < p < 1, \frac{-min(p,1-p)}{1-min(p,1-p)} \le f \le 1$. The posterior distribution is given by

$$Pr(p,f|\boldsymbol{n}) = \frac{Pr(\boldsymbol{n}|p,f)\varphi_p(p)\varphi_{f|p}(f|p)}{\int Pr(\boldsymbol{n}|p,f)\varphi_p(p)\varphi_{f|p}(f|p)dfdp}.$$

It is not difficult to verify that denominator is a finite quantity as shown below:

$$\int Pr(\boldsymbol{n}|p,f)\varphi_p(p)\varphi_{f|p}(f|p)dfdp$$

$$= \int_0^1 \int_{\frac{-p_{min}}{1-p_{min}}}^1 l(p,f)\varphi(p,f)dfdp$$

$$= \int_0^1 \int_{\frac{-min(p,q)}{1-min(p,q)}}^1 \frac{n!}{n_{11}!\,n_{12}!\,n_{22}!}\left((1-f)p^2+fp\right)^{n_{11}}(2(1-f)pq)^{n_{12}}\left((1-f)q^2+fq\right)^{n_{22}}dfdp$$

$$< \int_0^1 \int_{\frac{-min(p,q)}{1-min(p,q)}}^1 C(n)dfdp$$

$$= C(n)\int_0^1 \left(1+\frac{min(p,q)}{1-min(p,q)}\right)dp$$

$$< 2C(n),$$

where $C(n) = \frac{n!}{n_{11}!n_{12}!n_{22}!}$, a constant depended on data. The first inequality is due to the fact that all the genotype frequencies are equal to or less than one. Therefore, the posterior distribution is proper and has a closed form when *k*=2. The posterior mean of *f* can be obtained as

$$\int_0^1 \int_{\frac{-\min(p,q)}{1-\min(p,q)}}^1 f\, Pr(p,f|\boldsymbol{n})dfdp$$

$$= \frac{\int_0^1 \int_{\frac{-\min(p,q)}{1-\min(p,q)}}^1 f\, Pr(\boldsymbol{n}|p,f)\varphi_p(p)\varphi_{f|p}(f|p)dfdp}{\int_0^1 \int_{\frac{-\min(p,q)}{1-\min(p,q)}}^1 Pr(\boldsymbol{n}|p,f)\varphi_p(p)\varphi_{f|p}(f|p)dfdp}$$

$$= \frac{\int_0^1 \int_{\frac{-\min(p,q)}{1-\min(p,q)}}^1 f\frac{n!}{n_{11}!\,n_{12}!\,n_{22}!}\left((1-f)p^2+fp\right)^{n_{11}}(2(1-f)pq)^{n_{12}}\left((1-f)q^2+fq\right)^{n_{22}}dfdp}{\int_0^1 \int_{\frac{-\min(p,q)}{1-\min(p,q)}}^1 \frac{n!}{n_{11}!\,n_{12}!\,n_{22}!}\left((1-f)p^2+fp\right)^{n_{11}}(2(1-f)pq)^{n_{12}}\left((1-f)q^2+fq\right)^{n_{22}}dfdp}$$

$$= \frac{\int_0^1 \int_{\frac{-\min(p,q)}{1-\min(p,q)}}^1 f\left((1-f)p^2+fp\right)^{n_{11}}(2(1-f)pq)^{n_{12}}\left((1-f)q^2+fq\right)^{n_{22}}dfdp}{\int_0^1 \int_{\frac{-\min(p,q)}{1-\min(p,q)}}^1 \left((1-f)p^2+fp\right)^{n_{11}}(2(1-f)pq)^{n_{12}}\left((1-f)q^2+fq\right)^{n_{22}}dfdp},$$

which is not a standard integral function, but can be obtained using the numerical integration method. Markov Chain Monte Carlo (MCMC) could also be used to derive the posterior mean and variance. In the following simulation study, we choose to use MCMC from OPENBUGS to derive the posterior mean and its 95% credible interval for the estimator. The model for this MCMC is listed in below.

Level 1: $\boldsymbol{n}\sim Multinomial(n,\tilde{\boldsymbol{p}})$, $\boldsymbol{n}=(n_{11},n_{12},n_{22})$, $\tilde{\boldsymbol{p}}=(p_{11},p_{12},p_{22})$, $\begin{cases} p_{11}=p^2+pqf \\ p_{12}=2pq(1-f) \\ p_{22}=q^2+pqf \end{cases}$

Level 2: $p\sim Uniform[0,\ 1]$;

$f|p\sim Uniform[\frac{-\min(p,1-p)}{1-\min(p,1-p)},1]$; Since $\frac{-x}{1-x}=1-\frac{1}{1-x}$ is a decreasing function in x, this is same as $f|p\sim Uniform[\max(\frac{-p}{1-p},\frac{-(1-p)}{p}),1]$;

```
model
{
   q[1] <- (1 - f) * p * p + f * p
   q[2] <- 2 * (1 - f) * p * (1 - p)
   q[3] <- (1 - f) * (1 - p) * (1 - p) + f * (1 - p)
   y[1:3] ~ dmulti(q[], n)
```

```
    p ~ dunif(0, 1)
    f <- w * (1 - f.min) + f.min
    f.min <- max(-p/(1 - p), -(1 - p)/p)
    w ~ dunif(0, 1)
    n <- sum(y[])
}
```

After MCMC, standard diagnostic plot can be used to check the simulation history, posterior distribution of the parameter, autocorrelation function, and Brooks-Gelman-Rubin convergence statistic (Brooks & Gelman, 1998). Figure 2 shows examples of these plots when the observed genotype data is 3, 32, and 165 (simulated counts for a sample size of 200, the allele frequency p=0.1 and the inbreeding coefficient f=0.05). The skewed posterior distribution of f indicates that the confidence interval of the DSE under asymptotical normality assumption may not be satisfactory.

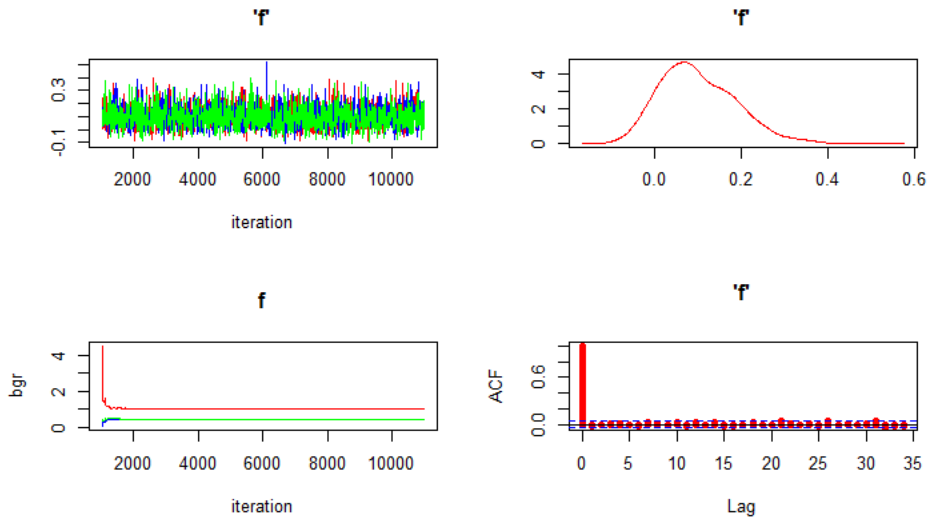


Figure 2. MCMC Diagnostic Plots (History, Posterior Distribution, Autocorrelation and BGR Statistic).

Current Bayesian application in the estimation of inbreeding coefficient is restricted mainly to the simple random sampling from either single population or independent populations (Wakefield, 2010). When sampling design includes unequal selection probabilities or clustering, Bayesian analysis needs to consider the sampling structure so that effective sample size or possible correlations due to clustering may be incorporated in the hierarchical model. However, the final obtained analysis dataset may not contain detailed sample design information. Even when all the

variables for sampling design are available, exact hierarchical modeling may result in a substantial number of hyper-parameters in the model.

1.5 Genetic Data from Complex Survey

In the past few years, population based national surveys, such as the National Health and Nutrition Examination Survey (NHANES) and Health and Retirement Survey (HRS), started collecting a subsample of their participants' DNA specimens. These surveys usually employ a complex sampling design, which often involves stratification, clustering and multi-stage sampling. In order to produce reliable statistics for minority represented subgroup of survey participants, these surveys typically over-sample such subpopulation. Therefore, within a cluster of the population, individuals are often related to one another, due to same ethnicity, some shared ancestry or finite within cluster sample size. On the other hand, complex sampling structure introduces correlations among genes from different clusters. This may introduce a local level inbreeding, which results in an excessive homozygotes and affects our interpretation of the gene polymorphism if it is not appropriately considered in analysis. Statistical tests based on allelic distributions or levels of variability often depend on sample drawn from sub-divided population, and therefore, can be significant in absence of departing from HWE (Nielsen, 2001).

For sample designs with unequal probabilities of inclusion, a generalized weighted quadratic test statistic is used in design-based approaches to conduct hypothesis testing of HWE. Under certain regulatory conditions, this test statistic converges to linear combinations of *i.i.d* chi-square random variables (Li & Graubard, 2009). Similarly, testing HWE is equivalent to testing the null hypothesis of disequilibrium coefficient equals to zero. Under the null hypothesis, the test statistic can be proven to be a chi-square random variable corrected by a simple design correction factor which is a function of design effect estimates of the genotype frequency (Moonesinghe, et al., 2010). In population-based household survey such as NHANES, genetic related individuals are sampled from the same household. To accommodate the correlation induced by genetic relatedness among individuals within a household, as well as the correlation due to multi-stage clustering sampling, an estimation equation under an inbreeding model was used to derive a quasi-score test statistic. Under the null hypothesis, this test statistic converges to a chi-square random variable (Li, Y.; Graubard, B.I., 2011).

The extension of testing HWE for data coming from complex survey provides great opportunities for scientists to screen genes to identify potential ones for gene-disease association study. However, the lack of robust estimator limits the interpretation of these findings in a scientific content. Researchers may be interested in a moderate estimate of the departure from HWE even when the data suggests otherwise. On the other hand, statistical significance from above test with a negligible f does not give researchers much confidence to screen out these potential important genes. Furthermore, although Bayesian method is increasingly used in the testing of HWE or the estimation of the inbreeding coefficient (Shoemaker, Painter, & Weir, 1998; Wakefield, 2010; Consonni, Morenob, & Venturinic, 2011; Holsinger & Wallace, 2004; Ayres & Balding, 1998), we have not found such application to the genetic data from complex sampling design thus far.

When the inbreeding coefficient is estimated from a sample with unequal selection probabilities, it still measures the deviation from random pairing of the genes. Since researchers often devote their time to use the genotype data in the context of population survey, proper estimation of the inbreeding coefficient for the study population may provide an initial screening tool for the potential genes. Extending the Bayesian methods to the genetic data from complex survey may build a bridge between the hypothesis testing and the estimation by using the posterior distribution. A hypothesis testing could be formulized through widely used Bayes Factor or through the credible interval in the Bayesian frame work.

1.5.1 National Health and Nutrition Examination Survey (NHANES)

The National Health and Nutrition Examination Survey (NHANES) is a program designed to assess the health and nutritional status of adults and children in the United States by interviews and physical examinations. The sample for the survey is selected to represent the U.S. population of all ages.

NHANES is a complex sample survey. The sample weights reflect the unequal probabilities of selection, non-response adjustments and adjustments to independent population controls. Primary Sampling Units (PSUs) are generally single counties. In order to meet a minimum population size, small counties are sometimes combined as one PSU. Within a sampled PSU, clusters of households are selected. Each person in a selected household is screened for demographic characteristics. One

or more persons per household are selected for the final sample. Data are sometimes collected on subsamples of the full design for NHANES. Therefore, each subsample involves another stage of selection and separate sample weights that account for that stage of selection and additional non-response, etc. In order to produce reliable statistics, NHANES over-samples persons 60 and older, African Americans, and Hispanics (Ezzati, Massey, Waksberg, Chu, & Maurer, 1992).

NHANES collects data for chronic conditions, previously undiagnosed conditions, as well as those known to and reported by respondents in the survey. The studied disease or medical conditions includes anemia, cardiovascular disease, diabetes, environmental exposures, eye diseases, hearing loss, infectious diseases, kidney disease, nutrition, obesity, oral health, osteoporosis, physical fitness and physical functioning, reproductive history and sexual behavior, respiratory disease (asthma, chronic bronchitis, emphysema), sexually transmitted diseases, vision, etc. NHANES also collects risk factors, including person's lifestyle, constitution, heredity, environment, smoking, alcohol consumption, sexual practices, drug use, physical fitness and activity, weight, and dietary intake, etc.

Starting from NHANES III, DNA specimens have been collected from participant whose age is 20 or more. The genetic data adds to the extensive amount of information collected for the purpose of describing the health of the population. Genotype data is obtained from DNA samples extracted from cell lines (NHANES III) or blood (NHANES 1999-2002) and can be analyzed along with survey data such as NHANES 1999-2000 or NHANES 2001-2002 or all four years combined (NHANES 1999-2002). The quality control of the genetic data is performed by assessing the deviation from Hardy-Weinberg Equilibrium for each of the three main race/ethnicity groups (non-Hispanic white, non-Hispanic black, and Mexican-American) and by assessing the discordant pairs between duplicated samples and the complete set (NHANES, 2010).

Unlike mutation, the genetic data collected in NHANES is mainly for common disease. Each polymorphism contributes only a small amount towards overall disease risk. Most often, a participant's life style, characteristics such as body mass index, and/or social-economic status may be more important risk factors for disease prediction. Nevertheless, NHANES is the first US survey with a national probability sampling design to make unbiased estimates about population genetics and associations between disease and genetic variants (Chang, Lindegren, Butler, & et.al.,

2009). Linkage of the NHANES III or NHANES 1999-2002 phenotype data with the genetic information provides an opportunity to investigate the association of a wide variety of health factors with regard to genetic variation. From the release of NHANES genetic data component, public health scientists have been using the data for screening of genetic marker, studying genetic variation among US population, pre-dispositioning to chronic disease, and identifying of risk factors (Fesinmeyer & et.al, 2013; Zhang & et.al, 2013; Yang & et.al, 2010; Crawford & et.al, 2006; Steinberg & et.al, 2001).

1.5.2 Health and Retirement Survey (HRS)

The Health and Retirement Study (HRS), conducted by the University of Michigan, is a longitudinal panel study of representative samples of Americans over 50 every two years. The target population includes all adults in the contiguous United States born during the years 1931 - 1941 who reside in households. In addition to the core sample, HRS cohort is supplemented by over sampling African Americans, Hispanics as well as respondents who are residents of the state of Florida. The core cohort was augmented in following years to include additional AHEAD cohort (The Study of Assets and Health Dynamics Among the Oldest Old, born before 1924), CODA cohort (Children of Depression, born 1924-30), War Baby (born 1942-47), Early Baby Boomers (born 1948-53), and Mid Boomers (born 1954–59). The current sample includes over 26,000 persons in 17,000 households. The study collects information about participants' income, work, assets, pension plans, health insurance, disability, physical health and functioning, cognitive functioning, and health care expenditures. It is supplemented with administrative linkages to Medicare claims files providing diagnostic and medication utilization information, to the National Death Index, and to the Social Security (Health and Retirement Study Survey Design, 2008).

The HRS sample is selected under a four-stage area probability sampling design. Each multi-stage component of the HRS area probability sample is consistent with the general sample design framework and sampling procedures of the SRC National Sample (Heeringa, Connor, & Darrah, 1986). First, the primary stage of sampling units (PSUs) involve probability proportional to size (PPS) selection of U.S. Metropolitan Statistical Areas (MSAs) and non-MSA counties. The National Sample PSUs are assigned to 84 explicit strata based on MSA/non-MSA status, PSU size and geographic location. Sixteen of the 84 strata contain only a single self-representing (SR) PSU,

each of which is included with certainty in the primary stage of sample selection. The remaining 68 nonself-representing (NSR) strata contain more than one PSU and one PSU is sampled with PPS. To reduce the between-PSU component of sample variation, a controlled selection with "one-per-stratum" design allocation is used to select PSUs. The full 84 strata are 2/3 partitioned into sampling error computation strata. Despite the expected improvement in sample precision, the expected consequence of collapsing design strata into sampling error computation strata is the overestimation of the true sampling error. With the exception of New York, Los Angeles and Chicago MSAs, which have two sampling error computation strata collapsed from similar NSR design strata, each SR design stratum is represented by one sampling error computation stratum. In 1992 survey, there were a total of 61 sampling error computation strata, including 16 self-representing MSA PSUs and a stratified subsampling of 45 of the 68 nonself-representing PSUs (HRS 1992 (Wave 1) Documentation, 1992).

Secondly, sampling of area segments (SSUs) is conducted within each selected PSU. In order to estimate the sampling error using the Balanced Repeated Replication (BRR) method or the approximate Taylor Series method, the half sample units are created by dividing sample cases into random halves based on SSU number order to preserve the stratification and second stage clustering properties within each computation stratum. The half samples to each NSR computation stratum has a one-to-one correspondence to sample design NSR PSUs.

Thirdly, a systematic selection of the housing units (HUs) is conducted from a complete list of all HUs that are physically located within the bounds of the selected SSU. The household financial unit must contain at least one age-eligible member from the cohort. This includes 1) a single unmarried age-eligible person; 2) a married couple in which both persons are age-eligible; or 3) a married couple in which only one spouse is age-eligible.

The fourth and last stage is the selection of an age-eligible person within a sampled HU. If selected age-eligible person has a spouse, the spouse is automatically selected even if he or she is not age-eligible.

In 2006 and 2008, HRS genotyped 12,507 respondents who provided saliva DNA samples and signed consent forms. Additional 6,000 more samples were expected to be added with the inclusion

of a substantial expansion of the minority samples. The genotyping was performed by the NIH Center for Inherited Disease Research using the Illumina Human Omni-2.5 Quad beadchip (Weir D. R., 2012). This technology has capability to cover more than 2.5 million SNPs. These restricted genotype data and a limited set of phenotype measures are deposited in the NIH GWAS repository *dbGap* (http://www.ncbi.nlm.nih.gov/gap) for approved researchers to use. Researchers may link the genotype data with the HRS main survey data by applying the access of the HRS-*dbGaP* Cross-Reference File (http://hrsonline.isr.umich.edu/gwas). In addition, *dbGap* provides genotypes for exonic DNA variants on approximate 16,000 samples. Furthermore, it maintains the imputation of genotypes for approximately 21 million DNA variants from the 1000 Genomes Project (http://www.1000genomes.org).

1.6 Empirical Likelihood

The Empirical Likelihood (EL), first introduced by Owen (1988), is a robust semiparametric alternative to the classical likelihood approach. Assume y_i, i=1,…, n, is a random variable from a distribution $P(\theta)$ with parameter θ. Instead of formalizing the likelihood through the density function as $l(\theta|\boldsymbol{y}) = P(\boldsymbol{y}|\theta)$, the empirical likelihood is defined as $l_{EL}(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} \pi_i$ or $\sum_{i=1}^{n} \log(\pi_i)$, $\boldsymbol{\pi}$ is in the set $\{\boldsymbol{\pi} \in [0,1]^n, \pi_i \in [0,1],\ \sum_{i=1}^{n} \pi_i = 1,\ \text{and}\ \sum_{i=1}^{n} \pi_i\, h(y_i|\theta) = 0\}$. The last constraint is an estimation equation such that $E[h(Y,\theta)] = 0$. Without this last constraint, the empirical likelihood has a nonparametric empirical cumulative distribution. Extensive discussion and application of this topic can be found in Owen (2001).

The concept of empirical likelihood in survey sampling was first used by Hartley & Rao (1968) through their "scale load" approach. To obtain an empirical likelihood under general sampling design with unequal probabilities of selection and clustering, Chen & Sitter (1999) proposed a pseudo-empirical likelihood function. Wu & Rao (2006), Rao & Wu (2010), Yang & Qin & Qin (2011) extended this method to stratified complex sampling for the estimation of confidence intervals. The proposed method has been proved to have design-based frequentist asymptotical property. The formal application of Bayesian approach to the empirical likelihood started with the discussion in Monahan & Boos (1992). They proposed that empirical likelihood is appropriate for the Bayesian inference if the associated posterior credible intervals has correct specified coverage. Using simulation approach, Lazar (2003) showed that empirical likelihood has the correct posterior

coverage for moderate sample size. Using higher order asymptotics, Fang & Mukerjee first theoretically investigated the coverage probability of posterior credible interval for empirical type likelihood admitting a probability-matching prior (2005; 2006). Chang & Mukerjee (2008) later extended to a general class of empirical-type likelihoods for the population mean and studied its existence of a confidence interval that has approximately correct posterior as well as frequentist coverage for any given prior. Rao & Wu (2010) applied the Bayesian empirical likelihood to estimate the posterior mean and coverage interval in a finite population sampling setting. Chaudhuri & Ghosh (2011) considered Bayesian empirical likelihood in the context of small area estimation, which handles discrete and continuous data in a unified manner and does not require a parametric likelihood or any linearity assumptions.

In this dissertation, the genotype data is discrete. The inbreeding coefficient is a function of the genotype frequencies which can be considered as a correlation measure. Extension of the Bayesian empirical likelihood approach to the estimation of a function of population proportions such as the inbreeding coefficient f is one of our main interests.

1.7 Discussion and Overview of Dissertation

In this chapter, we have presented a broad overview of the estimation methods of the inbreeding coefficient, including their advantages and disadvantages. We have introduced two national level complex surveys that collect genotypic data. Lastly, we have briefly reviewed empirical likelihood with the Bayesian approach and its current application in the complex survey. Such discussion motivates our research in the application of these methods to the population genetics based on national level complex surveys. The more and more genetic data coming from these types of surveys make us believe that a wide range of applications of the Bayesian methods including a nonparametric approach could benefit future research. As an initial gene screening tool for GWAS focusing on the measured magnitude of the deviation from HWE, we believe that it is necessary to take the complex survey sampling design into consideration.

This dissertation is organized as follows. In Chapter 2, we first illustrate the maximum empirical likelihood estimator and the parametric Bayesian estimator of the inbreeding coefficient. We further propose the nonparametric Bayesian estimator with the empirical likelihood under a simple

random sampling design. We compare those estimators in simulation studies. In Chapter 3, we extend above methods to estimate the finite population inbreeding coefficient based on samples with unequal selection probabilities. We have developed the Bayesian pseudo-empirical likelihood estimator with the sampling weight. The proposed estimator is compared with the parametric Bayesian estimator using the effective sample size, and with the design-based estimator in simulation studies. In Chapter 4, we further extend the proposed methods to family level genetic data which includes within family genetic correlation. In Chapter 5, we explore the analysis approach incorporating between subject correlations for clustered genetic data with population subdivision. In Chapter 6, we apply the proposed Bayesian methods to genotype data from the 2006 Health and Retirement Study. Finally in Chapter 7, we provide a summary of this dissertation and discuss directions for future research.

Chapter 2: Estimation of the Inbreeding Coefficient under Simple Random Sampling

In this chapter, to lay out the framework for our research, we first focus on the estimation of the inbreeding coefficient under simple random sampling design. The existing frequentist estimator and parametric Bayesian estimator from posterior distribution is compared. In order to simplify our illustration, we used a single locus with two alleles (k=2) throughout discussion. As a benchmark for comparison, the frequentist Direct Sample Estimator (DSE, 6) or the estimator derived in equation (7) by setting the first derivative of the log likelihood to zero is used. A Bayesian model described in Section 1.4.2 is used as our parametric Bayesian estimator.

2.1 Maximum Empirical Likelihood Estimator (MELE)

Let Y be the number of allele A for each subject in the population. We define Y=2 if the observed genotype is AA, Y=1 if the genotype is Aa, and Y=0 if the genotype is aa. Under the general inbreeding model, the distribution of Y is

$$P(Y=y)=\begin{cases}(1-p)^2+p(1-p)f & y=0\\ 2(1-f)p(1-p) & if \quad y=1.\\ p^2+p(1-p)f & y=2\end{cases}$$

Let $y_1, y_2, \dots, y_n$ be a random sample from the distribution of Y. Further, define a binary indicator variable $V=Y(2-Y)$ for the homozygotes AA and aa (Y=0 or 2, therefore V=0) or heterozygote Aa (Y=1, therefore V=1), $v_i=y_i(2-y_i)$. The indicator variable V has the following distribution:

$$P(V=v)=\begin{cases}1-2(1-f)p(1-p) & if \quad v=0\\ 2(1-f)p(1-p) & v=1\end{cases}.$$

It should be recognized that V can be derived from Y, but not the reverse. Based on their distributions, $E(Y)=2p$ and $E(V)=2(1-f)p(1-p)$. Therefore, a nonparametric way to analyze the data is to maximize the empirical log likelihood $l_{EL}(f)=\sum_{i=1}^{n}\log(\pi_i)$ with respect to the constraints $\pi_i>0$, $\sum_{i=1}^{n}\pi_i=1$, $\sum_{i=1}^{n}\pi_i y_i=2p$, and $\sum_{i=1}^{n}\pi_i v_i=2(1-f)p(1-p)$. The objective function $\sum_{i=1}^{n}\log(\pi_i)$ is a strictly concave function on a convex set of weight vectors.

Therefore a unique global maximum exists (Owen, 2001). Similar as below case with a single estimating equation constraint in the next section, we use the Lagrange multiplier method by first setting up $G = \sum_{i=1}^{n} \log(\pi_i) - n\lambda_1 \sum_{i=1}^{n} \pi_i \big(v_i - 2(1-f)p(1-p)\big) - n\lambda_2 \sum_{i=1}^{n} \pi_i (y_i - 2p) - \gamma(\sum_{i=1}^{n} \pi_i - 1)$. If there is only one constraint for $\sum_{i=1}^{n} \pi_i = 1$, we know that empirical likelihood $\sum_{i=1}^{n} \log(\pi_i) = \log(\prod_{i=1}^{n} \pi_i)$ is maximized at $\hat{\pi}_i = 1/n$ (Owen, 2001). At this maximized empirical likelihood, we have $\sum_{i=1}^{n} \hat{\pi}_i \, y_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ and $\sum_{i=1}^{n} \hat{\pi}_i \, v_i = \frac{1}{n}\sum_{i=1}^{n} v_i = \bar{v}$. Therefore, with additional constraints based on estimating equations for the means $E(Y - 2p) = 0$ and $E(V - 2(1-f)p(1-p)) = 0$, we expect to see that the MELEs are similar to DSEs for both $\hat{p}$ and $\hat{f}$. Furthermore, if we use DSE $\hat{p}$ to replace p and derive a profiled likelihood, the MELE is expected to remain the same as DSE for $\hat{f}$. Such dimension reduction can significantly decrease the calculation time.

As Table 3 shown, genotype frequency may be rare when both the allele frequency *p* and the inbreeding coefficient *f* are small. With small sample size and conditioning on *p*, these parameter settings are more likely to produce estimators that are outside of the parameter space for *f*. To illustrate that MELE has a similar performance as DSE, a simulation study (250 runs) was conducted for the allele frequency p=0.1 and the inbreeding coefficient f=0.05 under a sample size of 50. To derive MELE, an EM algorithm implemented in the *R*/***emplik*** package (Zhou, 2005) is used to maximize the empirical log likelihood function with above constraints. Figure 3 shows that DSE and MELE have similar values in each simulation iteration so that the data pairs stay along the diagonal line. Furthermore, MELE has the same problem as DSE. At the sample size of 50, 14.4% of the simulated estimators are outside of the lower limit of the parameter space for *f* under p=0.1 and thus $-p/(1-p)$ = -0.1111. There are still 0.4% of the simulated estimators that are outside of the range when the sample size increases to 200. The simulation is in line with what we discussed earlier that DSE $f^{(s)}$ is a biased (in the order of n^{-1}) but consistent estimator. The estimator moves toward true value *f* when sample size increases toward infinity.

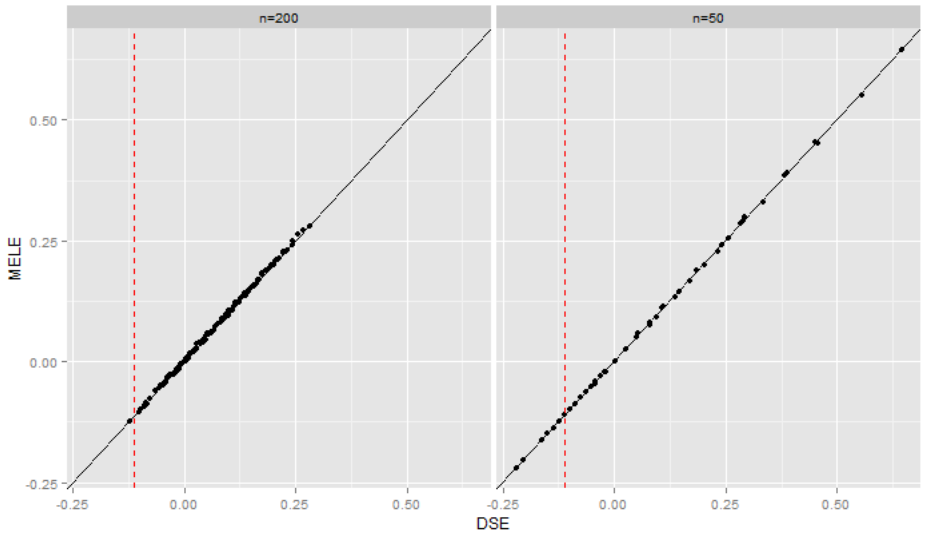


Figure 3. Simulation Results of DSE vs. MELE (p=0.1, f=0.05, sample size of 50 or 200). Red dash line is the lower limit of the parameter space for f when p=0.1.

2.2 Bayesian Pseudo-Empirical Likelihood Estimator (BPELE)

The above computation burden to derive MELE is not small. In order to search estimators of p and f so that they maximize the empirical likelihood with two constraints based on estimating functions, we need first calculate the empirical likelihood for each pair of p and f. Since we are not interested in the unknown population nuisance parameter p, a profile likelihood method can be used to replace p with its MLE $\hat{p} = \frac{\bar{Y}}{2}$. As we discussed in Chapter 1, the estimator $\hat{p}$ of the allele frequency is unbiased. A general estimation equation therefore can be formulized as $E[V - 2(1-f)\hat{p}(1-\hat{p})] = 0$. We want to maximize the empirical likelihood $l_{EL}(f) = \sum_{i=1}^{n} \log(\pi_i)$ subject to $\pi_i > 0$, $\sum_{i=1}^{n} \pi_i = 1$ and $\sum_{i=1}^{n} \pi_i(v_i - 2(1-f)\hat{p}(1-\hat{p})) = 0$. A Lagrange multiplier method can be used to find $\hat{l}_{EL}(\hat{f})$. First, we set up

$$G = \sum_{i=1}^{n} \log(\pi_i) - n\lambda \sum_{i=1}^{n} \pi_i\big(v_i - 2(1-f)\hat{p}(1-\hat{p})\big) - \gamma\Big(\sum_{i=1}^{n} \pi_i - 1\Big).$$

Taking the partial derivative with respect to π_i, this equals 0 yields:

$$\frac{1}{\pi_i} - n\lambda\big(v_i - 2(1-f)\hat{p}(1-\hat{p})\big) - \gamma = 0.$$

Summing the above equation over *i*=1 to *n*, it yields $n - n\lambda \sum_{i=1}^{n} \pi_i \big(v_i - 2(1-f)\hat{p}(1-\hat{p})\big) = \gamma \sum_{i=1}^{n} \pi_i = \gamma, \gamma = n$. Solving this equation for π_i, it yields $\pi_i = \frac{1}{n\lambda\left(v_i - 2(1-f)\hat{p}(1-\hat{p})\right)+n}$. To find the solution of $\lambda = \lambda(f)$, it needs solve the equation $\sum_{i=1}^{n} \frac{v_i - 2(1-f)\hat{p}(1-\hat{p})}{1+\lambda\left(v_i - 2(1-f)\hat{p}(1-\hat{p})\right)} = 0$. This usually can be done through the Newton algorithm. The empirical likelihood is then derived as

$$l_{EL}(f) = \sum_{i=1}^{n} \log(\pi_i) = -n\log(n) - \sum_{i=1}^{n} \log\Big(1 + \lambda(f)\big(v_i - 2(1-f)\hat{p}(1-\hat{p})\big)\Big).$$

As we showed in previous section, MELE is expect to be approximately the same as DSE $f^{(s)}$. When the sample size is small and the allele frequency is small, it is likely to produce estimators that lie outside of the parameter space. Borrowing the same idea from the parametric Bayesian analysis, Bayesian approach can also be applied to the empirical likelihood setting. With our inbreeding model, we assume a uniform prior $\varphi(f) \propto 1$ on the parameter space $[\frac{-\min(\hat{p},1-\hat{p})}{1-\min(\hat{p},1-\hat{p})}, 1]$. The posterior distribution therefore is

$$\varphi(f|y) = c(y) \exp\Big[-\sum_{i=1}^{n} \log\big\{1 + \lambda(f)\big(v_i - 2(1-f)\hat{p}(1-\hat{p})\big)\big\}\Big],$$

where $c(y)$ is a normalizing constant such that $\int \varphi(f|y)\, df = 1$. Since the uniform prior is proper and the empirical likelihood is bounded, the posterior distribution is therefore proper (Chaudhuri & Ghosh, 2011). We can write the posterior distribution as

$$\varphi(f|y) = Pr(f|y_1, \dots, y_n) \propto \varphi(f) l_{EL}(f) = \exp\big(\log\big(\varphi(f)\big) + \log\big(l_{EL}(f)\big)\big).$$

Without loss of generality, let f_0 maximizes the prior $\varphi(f)$ and $\hat{f}_n$ maximizes the empirical likelihood $l_{EL}(f)$. We expand the prior and the empirical likelihood by the Taylor linearization up to the second derivative. To simplify the notation, we ignore the remainder term after the second derivatives.

$$\log\big(\pi(f)\big) = \log\big(\varphi(f_0)\big) + (f - f_0)\varphi'(f_0) - \frac{1}{2}(f - f_0)^2\varphi''(f_0)$$

$$= \log\big(\varphi(f_0)\big) - \frac{1}{2}(f - f_0)^2\varphi''(f_0)$$

$$\log(l_{EL}(f)) = \log\left(l_{EL}(\hat{f}_n)\right) + (f-\hat{f}_n)l_{EL}{}'(\hat{f}_n) - \frac{1}{2}(f-\hat{f}_n)^2 l_{EL}{}''(\hat{f}_n)$$

$$= log\left(l_{EL}(\hat{f}_n)\right) - \frac{1}{2}(f-\hat{f}_n)^2 l_{EL}{}''(\hat{f}_n).$$

Substituting above equations into $\varphi(f|y)$, the posterior distribution therefore is

$$\varphi(f|y) = \exp\left\{\log(\varphi(f_0)) - \frac{1}{2}(f-f_0)^2\pi''(f_0) + \log\left(l_{EL}(\hat{f}_n)\right) - \frac{1}{2}(f-\hat{f}_n)^2 l_{EL}{}''(\hat{f}_n)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}(f-f_0)^2\varphi''(f_0) - \frac{1}{2}(f-\hat{f}_n)^2 l_{EL}{}''(\hat{f}_n)\right\}$$

$$= \exp\left\{-\frac{1}{2}\left(f - \frac{f_0\varphi''(f_0) + \hat{f}_n l_{EL}{}''(\hat{f}_n)}{\varphi''(f_0) + l_{EL}{}''(\hat{f}_n)}\right)^2 \left(\varphi''(f_0) + l_{EL}{}''(\hat{f}_n)\right)\right\}.$$

Under general regularity condition, as $n \to \infty$, the posterior distribution $\varphi(f|y)$ converges to a normal distribution with mean $\frac{f_0\varphi''(f_0)+\hat{f}_n l_{EL}{}''(\hat{f}_n)}{\varphi''(f_0)+l_{EL}{}''(\hat{f}_n)}$ and variance $\varphi''(f_0) + l_{EL}{}''(\hat{f}_n)$ (Owen, 2001). In our case, we select a non-informative uniform prior $\varphi(f) \propto 1$, therefore $\varphi''(f_0) = 0$. The posterior converges to a normal distribution with mean $\hat{f}_n$ and variance $l_{EL}{}''(\hat{f}_n)$ as $n \to \infty$.

2.3 Simulation Study

A simulation study is conducted for a single locus with two alleles (*A*, *a*) to compare the frequentist estimator and the Bayesian estimators of *f*. At each one of *R*=1000 iterations (The rationale of the choice of 1000 simulation runs is evaluated below), a sample size of *n*=50, 100 or 200 genotype records are generated randomly from the general multinomial inbreeding model with parameter *p* and *f*:

$$Pr(\boldsymbol{n}|p,f) = \frac{n!}{\prod_{i=1,j\geq i}^{2} n_{ij}!}\prod\nolimits_{i=1,j\geq i}^{2} {p_{ij}}^{n_{ij}}$$

$$= \frac{n!}{n_{11}!\,n_{12}!\,n_{22}!}[p(p+f-fp)]^{n_{11}}[2(1-f)p(1-p)]^{n_{12}}[(1-p)(1-p+fp)]^{n_{22}}.$$

The parameters for the simulation are the allele *A* frequency of *p*=0.1, 0.3, or 0.5 and the inbreeding coefficient of *f*=-0.1, 0, 0.05, 0.1, or 0.2. Under those parameter setting, Table 3 shows the expected

genotype frequencies. When the allele frequency is small (p=0.1), the expected genotype *AA* frequency is around 0.01~0.03 range depending on the value of f. If the sample size is small, the likelihood to have samples without any observed genotype *AA* is high. When this happens, the sample estimator of $p_{11} = p^2 + p(1-p)f$ tends to get zero, which results in the estimator of f approaches lower bound of parameter space, -p/(1-p).

Table 3. Expected Genotype Frequency under Allele Frequency p and Inbreeding Coefficient f

p	f	Expected Genotype Frequency		
		p(aa)	p(Aa)	p(AA)
0.1	-0.1	0.801	0.198	0.001
	0	0.81	0.18	0.01
	0.05	0.8145	0.171	0.0145
	0.1	0.819	0.162	0.019
	0.2	0.828	0.144	0.028
0.3	-0.1	0.469	0.462	0.069
	0	0.49	0.42	0.09
	0.05	0.5005	0.399	0.1005
	0.1	0.511	0.378	0.111
	0.2	0.532	0.336	0.132
0.5	-0.1	0.225	0.55	0.225
	0	0.25	0.5	0.25
	0.05	0.2625	0.475	0.2625
	0.1	0.275	0.45	0.275
	0.2	0.3	0.4	0.3

For the realized data, estimators of the inbreeding coefficient are calculated by both the frequentist and the Bayesian methods, respectively. The bias, variance and mean square error (MSE) of the estimator $\hat{f}$ are calculated as

$$\text{Bias}_f(\hat{f}) = E_f(\hat{f}) - f = \frac{1}{R}\sum_{r=1}^{R}(\hat{f}^{(r)} - f);$$

$$\text{Var}(\hat{f}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{f}^{(r)} - \bar{\hat{f}})^2;$$

$$\text{MSE}(\hat{f}) = E[(\hat{f} - f)^2] = \frac{1}{R}\sum_{r=1}^{R}\left((\hat{f}^{(r)} - \bar{\hat{f}}) + (\bar{\hat{f}} - f)\right)^2 = \text{Bias}_f(\hat{f})^2 + Var(\hat{f}),$$

where $\hat{f}^{(r)}$ is the corresponding estimator at the *r*-th simulation run and $\bar{\hat{f}} = \frac{1}{R}\sum_{r=1}^{R}\hat{f}^{(r)}$. The ratio of the root mean square error (RRMSE) may also be calculated to compare Bayesian estimators with DSE. Furthermore, 95% confidence interval $[f_{0.025}^{\mathrm{DSE}(r)}, f_{0.975}^{\mathrm{DSE}(r)}]$ for the frequentist estimators is calculated as $\hat{f}^{(r)} \pm 1.96 \times stderr(\hat{f}^{(r)})$. For Bayesian estimators, 95% credible intervals $[f_{0.025}^{\mathrm{PBE}(r)}, f_{0.975}^{\mathrm{PBE}(r)}]$ and $[f_{0.025}^{\mathrm{BPELE}(r)}, f_{0.975}^{\mathrm{BPELE}(r)}]$ are calculated from the posterior distribution. The interval length is calculated as the difference between the upper bound and the lower bound of each confidence (or credible) interval for the estimator. The average interval length therefore is calculated and compared among all three methods. The coverage probability, defined as the proportion of simulations that the confidence intervals or the credible intervals contain the true value (simulation parameter), is estimated as

$$\text{coverage probabity} = 100\% \times \frac{1}{R}\sum_{r=1}^{R} I^{r},$$

where $I^{r} = \begin{cases} 1, & \text{if the confidence or credible interval covers } f \\ 0, & \text{otherwise} \end{cases}, r = 1, \ldots, R.$

2.3.1 Monte Carlo Simulation Error

To evaluate the impact of the number of simulation runs to the Monte Carlo Simulation Standard Error (MCSE), we perform *R*=5,000 simulation runs for each inbreeding coefficient *f*=-0.1, 0, 0.05, 0.1 or 0.2 under a moderate allele frequency *p*=0.3. At each simulation, a sample size of 50 genotypes are generated from the general inbreeding model. Since $f^{(s)}$ or $\hat{f}$ is a consistent estimator, we choose the sample size of 50 to evaluate the performance of the proposed estimators under small sample size situation. Let $\hat{f}^{(r)}$ be the Monte Carlo estimator of *f* from a simulation with *r* replicates. At each cumulative simulation run *r* (=1, …, 5000), the Monte Carlo Simulation Standard Error, $\mathrm{MCSE}(\hat{f}^{(r)}) = \sqrt{\mathrm{Var}(\hat{f}^{(r)})}$, is calculated as if the simulations are stopped at that iteration (Koehler, Brown, & Haneuse, 2009). The results are plotted against the simulation iterations in Figure 4 (*f*=0.05, other parameter settings not shown since all have the similar pattern). Similar to Wu & Rao's (2006) simulation studies specifically for the empirical likelihood and many other studies discussed in the review literature (Koehler, Brown, & Haneuse, 2009), Figure 4 suggests that a total of 1000 simulation iterations provide a reasonable compromise between

computational burden and the control of simulation error. For simulation runs at 5000, 2000, 1000 and 500, the corresponding MCSEs for BPELE are 0.0018, 0.0029, 0.0042 and 0.0059 respectively; the corresponding MCSEs for DSE are 0.0020, 0.0032, 0.0047 and 0.0066 respectively. The MCSEs for BPE are similar to the ones for BPELE at all simulation iterations. Considering both the simulation time and MCSE, we decide to use a total of 1000 runs in most of our simulation studies in this dissertation.

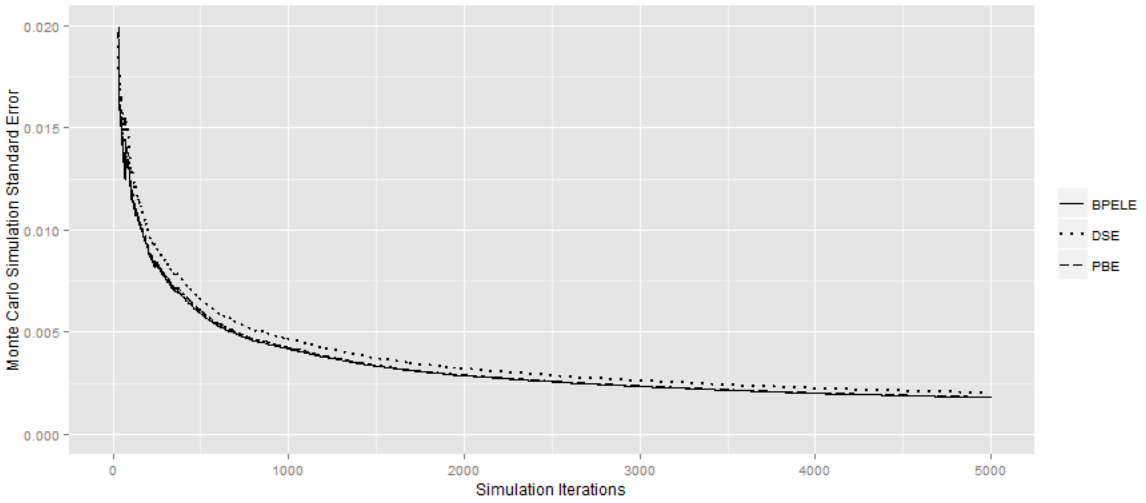


Figure 4. Monte Carlo Simulation Standard Error at Accumulative Simulation Iterations. BPELE=Bayesian Pseudo-Empirical Likelihood Estimator, DSE=Direct Sample Estimator, PBE=Parametric Bayesian Estimator. Allele frequency is p=0.3 and the inbreeding coefficient is f=0.05 in the simulation.

2.3.2 Compare Frequentist and Bayesian Estimators

Generally as expected, when sample size increases, the bias, the MSE as well as the average confidence interval length decrease for both the DSE and the Bayesian estimators. As Cohen (1982) shown previously that the DSE is a consistent estimator, the simulations show that the bias diminishes to zero as the sample size increase to infinity. Table 4, Table 5, and Table 6 present the simulation results in comparing the DSE and two Bayesian estimators at the sample size of 50, 100 or 200, respectively. The parametric Bayesian method tends to have a positive bias, while the frequentist method often has a negative bias. The Bayesian empirical likelihood method has a positive bias when the allele frequency is small (p=0.1), while shows a negative bias when the allele frequency is moderate (p=0.3, 0.5). When the allele frequency is small (p=0.1), the Bayesian

methods have larger bias especially when the inbreeding coefficient is small (f=-0.1, 0, 0.05). A similar observation can be made for the MSE. When the allele frequency or the inbreeding coefficient gets bigger, both Bayesian methods have better performance in terms of the MSE. In all simulations except the first case (p=0.1, f=-0.1), the coverage of the Bayesian empirical likelihood method is able to maintain the nominal level and tends to show better value than the frequentist method and the parametric Bayesian method. This advantage becomes more obvious when the allele frequency is small. When the allele frequency is small (p=0.1), the DSE does not maintain the nominal coverage probability of 95%.

As previously specified, in the multinomial model with re-parameterized genotype frequencies, the inbreeding coefficient f has a natural constraint $\frac{-min(p,1-p)}{1-min(p,1-p)} \leq f \leq 1$ resulted from the law of total probability. This constraint is embedded in the Bayesian method we use. In practice, the allele frequency is often an unknown parameter and need be estimated itself from the data. Therefore, it is practically not feasible to check whether the estimator is within the model specified range or not. This can be easily checked in our simulation studies since the parameter inputs are all known to generate the data. Simulation studies show that a large percentage of the DSEs are outside of the constraint when the allele frequency is small (p=0.1). The problem gets worse if the actual inbreeding coefficient is also small. For example, when the sample size is 50, p=0.1 and f=0, 17.4% of all DSEs are outside of the parameter space of [-0.1111, 1] (Figure 5, Panel A). In human genomic data analysis, the minor allele frequency and the inbreeding coefficient are often small. Ignoring parameter constraint, therefore, results in the DSE resides outside of the parameter space induced by the model. On the other hand, the Bayesian estimator naturally satisfies the f constraint through a proper prior distribution. When the allele frequency is moderate or large (p=0.3, 0.5), all of the DSEs are within the constraint induced by the model.

The simulation also shows that both Bayesian methods have coverage problem when the allele frequency is small (p=0.1) and the inbreeding coefficient is close to the lower boundary (f=-0.1). It should be recognized that under this parameter setting, the expected genotype frequencies are (0.801, 0.198, 0.001) and the lower bound of the expected inbreeding coefficient is -0.11. Given the small sample size of 50~200 in our simulation, the simulated number of genotype AA count maybe sparse. This resulted in a non-stable parametric Bayesian estimator. At the same time, the

empirical likelihood calculation is not stable at the boundary of the parameter space, thus it results in not optimal coverage for the estimator when the true value is close to the boundary. Since DSE has problem of being outside of parameter space, we argue that this situation need be carefully evaluated since none of three methods produce satisfactory solution when considering all four criteria including the coverage, the bias, the MSE, and the parameter space.

Figure 5 plots the kernel densities for all three estimators from those 1000 simulations. When the allele frequency is large (p=0.5), all three density functions overlap with each other. These data is not plotted in this dissertation to simplify the output. Under all three sample sizes, the BPELE always has a single mode that is close to the true parameter input; while both the DSE and the PBE may have two modes in some simulation settings. In all three panels, the DSE is closer to the parameter input when both the allele frequency and the inbreeding coefficient are small (p=0.1, f=-0.1). However, it is also clear that a large proportion of the estimators are outside of the parameter's lower boundary.

In summary, when the allele frequency is moderate or large (p=0.3, 0.5), the PBE has smaller bias and better MSE. It also has larger coverage probability and shorter interval length than the frequentist DSE. The BPELE is comparable to the DSE. When the allele frequency is small (p=0.1), the Bayesian estimators have larger coverage probability and smaller MSE in the exchange of larger bias as well as longer interval length. On the other hand, the DSE is often outside of the model specified parameter range for f.

Table 4. Comparison of Direct Sample Estimator, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Simple Random Sampling, n=50)

p	f	Direct Sample Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0014	0.0033	93.4	0.1622	38.3	0.1416	0.0253	77	0.5067	0.2065	0.0463	55.2	0.5583
	0	-0.0066	0.0187	40.6	0.3621	17.4	0.1132	0.0241	96.2	0.5558	0.1458	0.0266	100	0.606
	0.05	-0.0046	0.0267	52.8	0.4507	10.6	0.1017	0.0252	93.7	0.5759	0.1192	0.0215	99.9	0.6296
	0.1	-0.0102	0.0312	62.4	0.5191	7.8	0.0836	0.0236	95.6	0.5902	0.0896	0.0167	99.5	0.6489
	0.2	-0.0298	0.0402	72.2	0.6037	2.2	0.0478	0.0215	97	0.6168	0.0382	0.0129	100	0.6849
0.3	-0.1	-0.0026	0.0173	93.2	0.5055	0	0.031	0.0152	96	0.4888	0.0186	0.0123	99.1	0.5599
	0	-0.0134	0.0199	91.8	0.5362	0	0.0116	0.0164	95.1	0.51	-0.0073	0.015	98.5	0.5854
	0.05	-0.0109	0.0207	92.1	0.5481	0	0.0096	0.0171	95.2	0.5172	-0.0105	0.0165	98	0.5931
	0.1	-0.0007	0.0217	92.6	0.5574	0	0.0146	0.0183	95.7	0.5229	-0.0061	0.018	97.8	0.5974
	0.2	-0.0145	0.0211	94.2	0.5645	0	-0.0063	0.0176	96.7	0.5279	-0.0266	0.0189	98	0.5973
0.5	-0.1	-0.0106	0.0214	93	0.5431	0	0.007	0.0191	96.4	0.5171	-0.0066	0.0197	96	0.5329
	0	-0.0102	0.0208	93.2	0.548	0	0.0019	0.0183	95.1	0.5199	-0.0102	0.0192	94.9	0.5364
	0.05	-0.0104	0.0197	94.5	0.5485	0	-0.0014	0.0173	95.5	0.52	-0.0124	0.0182	95.9	0.5367
	0.1	-0.0046	0.0201	94.5	0.5469	0	0.0011	0.0177	96.1	0.5188	-0.0087	0.0186	95.6	0.5352
	0.2	-0.0051	0.019	94.8	0.5399	0	-0.0053	0.017	96	0.514	-0.0131	0.0177	95.2	0.5274

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 5. Comparison of Direct Sample Estimator, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Simple Random Sampling, n=100)

p	f	Direct Sample Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0004	0.0016	94.4	0.1241	36.8	0.0816	0.0088	84.6	0.3286	0.1522	0.0248	54.6	0.4271
	0	-0.0066	0.0098	62.1	0.3113	10.7	0.06	0.0108	95.8	0.3919	0.0924	0.0116	99.9	0.4797
	0.05	-0.0052	0.0132	75	0.378	3.1	0.0546	0.012	95.9	0.4215	0.0707	0.0092	99.9	0.5093
	0.1	-0.0063	0.0157	83.7	0.4364	1.6	0.0452	0.0129	96.2	0.443	0.0474	0.0079	100	0.5342
	0.2	-0.0135	0.02	88.3	0.5083	0.8	0.0248	0.0144	95.8	0.4777	0.0081	0.009	99.4	0.5769
0.3	-0.1	-0.002	0.0095	92.4	0.3653	0	0.0159	0.0088	93.6	0.3575	0.0021	0.008	98.3	0.4363
	0	-0.0042	0.0096	93.9	0.3865	0	0.0086	0.0087	95.6	0.3745	-0.0062	0.0088	98	0.4472
	0.05	-0.0064	0.01	93.8	0.3935	0	0.0041	0.0091	95.1	0.3802	-0.0103	0.0095	97.7	0.4473
	0.1	-0.0107	0.0102	94.3	0.3982	0	-0.0024	0.0091	96.2	0.3847	-0.0159	0.0098	98	0.4456
	0.2	-0.0047	0.0109	94.3	0.4018	0	-0.0014	0.01	95.4	0.3873	-0.0123	0.0106	96.9	0.4337
0.5	-0.1	-0.0063	0.0099	94.6	0.3873	0	0.0024	0.0093	94.8	0.3778	-0.0043	0.0095	94.9	0.3834
	0	0.0018	0.01	94.3	0.39	0	0.0071	0.0094	94.6	0.38	0.0016	0.0096	94.7	0.3856
	0.05	-0.0034	0.0099	95	0.3897	0	0.0004	0.0093	94.6	0.38	-0.0044	0.0095	95.2	0.3851
	0.1	-0.0038	0.0108	92.8	0.3883	0	-0.0014	0.0101	93.5	0.3787	-0.0058	0.0104	93.3	0.3837
	0.2	-0.0112	0.0092	94.2	0.3837	0	-0.0114	0.0087	95.1	0.3746	-0.0151	0.009	94.3	0.379

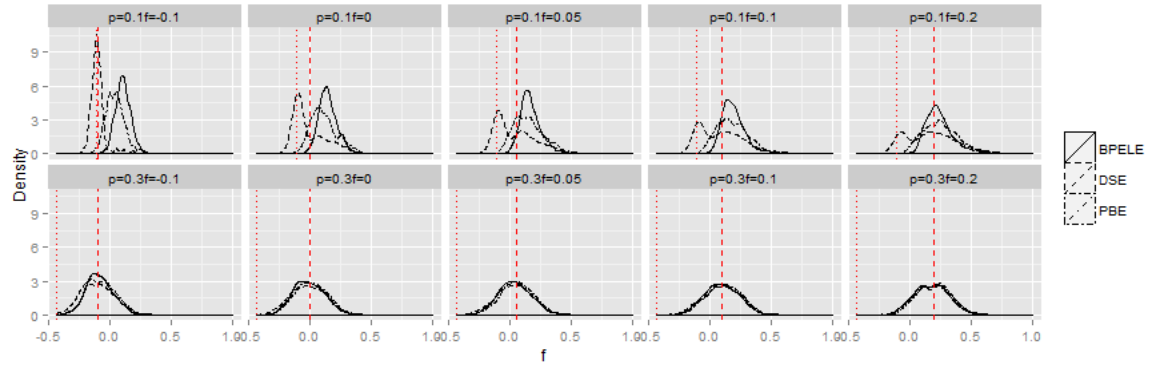
* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 6. Comparison of Direct Sample Estimator, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Simple Random Sampling, n=200)

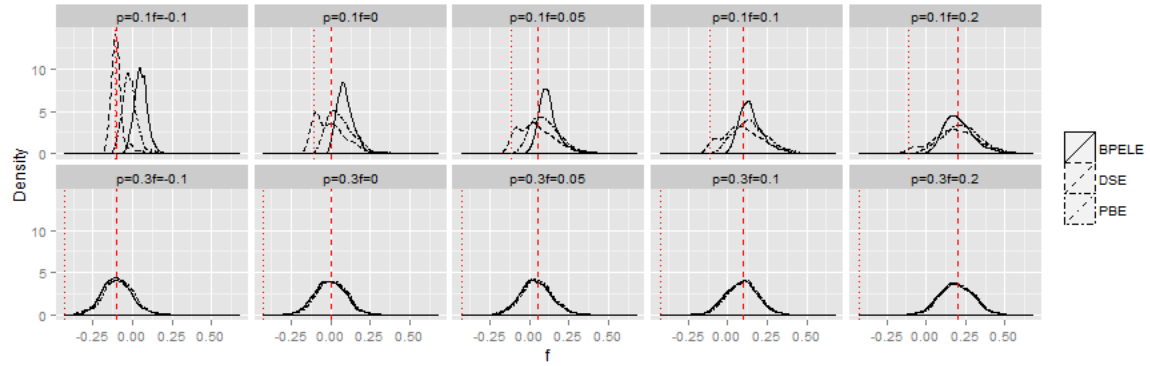
p	f	Direct Sample Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0012	0.0008	95.2	0.0939	37.4	0.0432	0.0028	89.3	0.1989	0.1081	0.0123	61.4	0.3176
	0	-0.0029	0.0053	83.3	0.249	3.7	0.0326	0.0055	95.6	0.2751	0.0544	0.0048	99.9	0.3792
	0.05	-0.0054	0.0064	88.5	0.2958	0.6	0.026	0.0059	96.4	0.3023	0.0326	0.0036	100	0.4077
	0.1	-0.0095	0.0083	88.5	0.3293	0.3	0.0181	0.0071	94.7	0.3243	0.0137	0.0042	99.9	0.4326
	0.2	0.0003	0.0101	92.6	0.3804	0	0.0186	0.0088	94.8	0.3605	-0.0018	0.007	99.6	0.4734
0.3	-0.1	-0.002	0.0047	93.8	0.2609	0	0.0069	0.0045	94.8	0.2576	-0.0025	0.0045	98	0.3244
	0	0.0029	0.0051	94	0.2758	0	0.0091	0.0049	95	0.271	0.001	0.005	97.8	0.323
	0.05	-0.0039	0.0053	94.8	0.2798	0	0.0013	0.0051	95.1	0.2747	-0.0062	0.0053	98	0.321
	0.1	-0.0015	0.0052	94.4	0.2833	0	0.0024	0.0049	95	0.2783	-0.0045	0.0051	97	0.3177
	0.2	-0.0026	0.0058	92.9	0.2858	0	-0.001	0.0056	93.7	0.2806	-0.0065	0.0058	95.7	0.3107
0.5	-0.1	-0.0012	0.0054	93.8	0.2749	0	0.0029	0.0052	94.4	0.2713	-0.0002	0.0052	94	0.2734
	0	-0.0028	0.005	94.9	0.2765	0	0.0001	0.0049	94.9	0.2729	-0.0028	0.0049	95.2	0.2749
	0.05	-0.0055	0.005	95	0.2763	0	-0.0034	0.0049	95.1	0.2727	-0.006	0.0049	95.1	0.2746
	0.1	0.0002	0.0052	93.1	0.2752	0	0.0013	0.005	93.7	0.2718	-0.0009	0.0051	93.6	0.2735
	0.2	-0.0043	0.005	94.6	0.2713	0	-0.0044	0.0049	94.4	0.2676	-0.0062	0.0049	94.6	0.2697

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

A. *n*=50



B. *n*=100



C. *n*=200

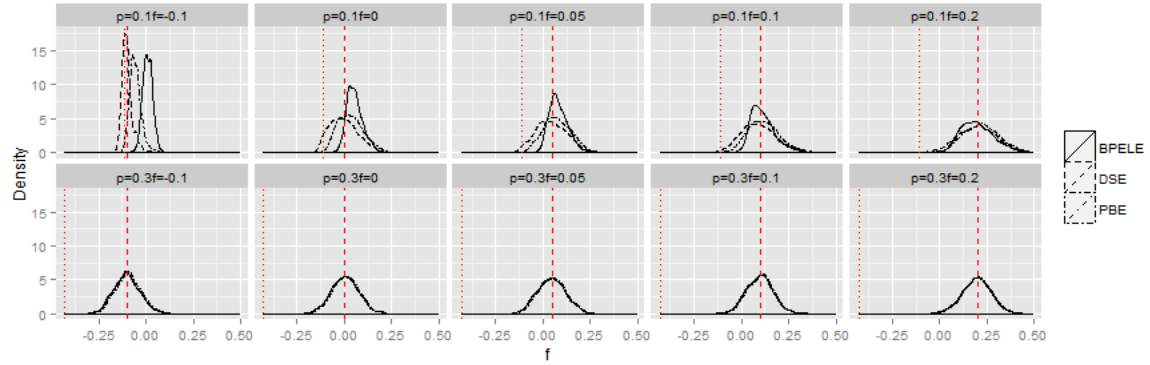


Figure 5 Density Plot of Simulated Estimators under Different Settings (Simple Random Sampling). BPELE=Bayesian Pseudo-Empirical Likelihood Estimator, DSE=Direct Sample Estimator, PBE=Parametric Bayesian Estimator. Parameter p is the allele frequency and f is the inbreeding coefficient in the simulation. Panel A, B and C are for sample size 50, 100 and 200 respectively. Red dashed vertical line is the true parameter f; red dotted vertical line is the lower bound of the parameter space for f conditional on allele frequency p.

2.3.3 Use of Prior Knowledge in Analysis

One of the advantages of Bayesian method is to incorporate prior knowledge of the parameter of interest into modeling. With much advanced genetic sequencing technology in recent years, genomic data approaches exascale into exabyte range. Simply ignoring these accumulative information is a pitiful waste of data. For example, for the important SNPs of interest, the allele frequencies in the population are often well studied. Quite often, such information is also available for a subgroup of the population with certain characteristics. Furthermore, the official definition of the SNP requires that the MAF is equal to or greater than 1%. Using a non-informative uniform distribution on [0, 1] as a prior distribution for the allele frequency is practically convenient but not scientifically well justified.

To investigate the impact of priors to the estimation of the inbreeding coefficient, we performed a simulation study as shown in Table 7. Beside the uniform priors on the whole parameter space that we used in previous simulations, we included a more restricted parameter range for the inbreeding coefficient. For the parametric Bayesian method, instead of a uniform prior for the allele frequency p, a *beta*(0.5, 0.5) is used to give more weights for the small allele frequencies. Conditioning on the p, a uniform prior for f in the range of (0, 0.5) is also studied to justify the situation when we have learned from previous knowledge of the parameter. In all these cases, using previous information produces estimators with less MSE, better coverage, and much less average coverage interval length. Under studied allele frequency p=0.1 and 0.3, compared with a non-informative prior as most literatures do, using previous information for the inbreeding coefficient results in estimators that are more concentrated around the true value as shown in Figure 6. This trend is most obvious for the Bayesian pseudo-empirical likelihood estimator (BPELE2) with a uniform prior (0, 0.5) for f under p=0.3.

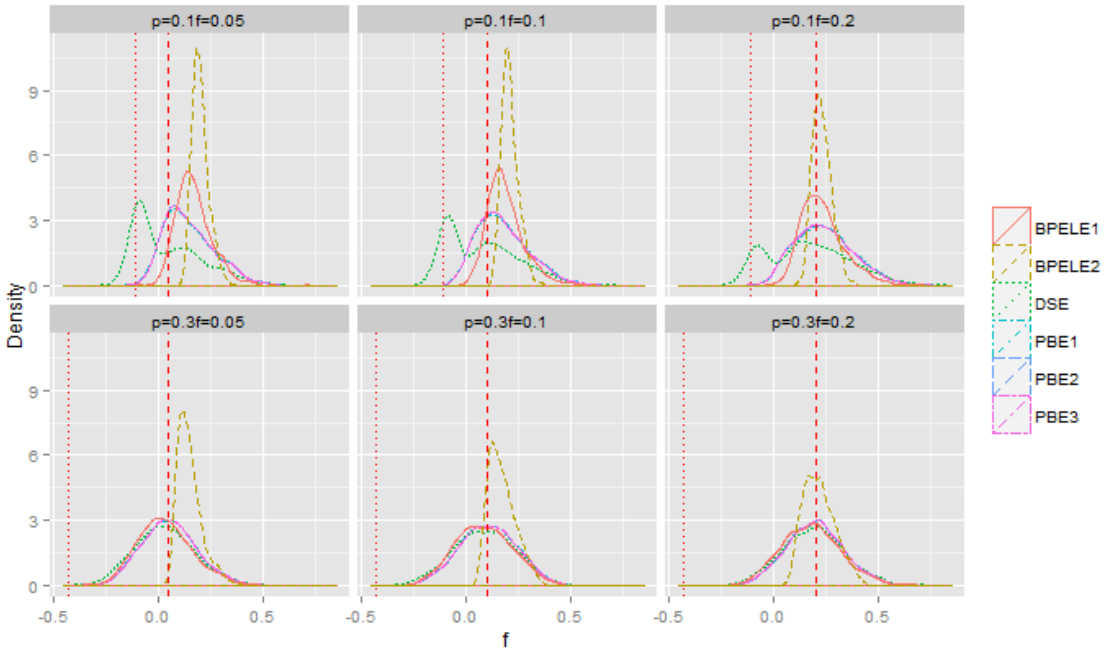


Figure 6. Density Plot of Simulated Estimators under Different Priors (Simple Random Sampling). BPELE1=Bayesian Pseudo-Empirical Likelihood Estimator with uniform prior (0,1) for f; BPELE2=Bayesian Pseudo-Empirical Likelihood Estimator with uniform prior (0,0.5) for f; DSE=Direct Sample Estimator; PBE1=Parametric Bayesian Estimator with uniform prior (0,1) for p and conditional uniform prior (-p/(1-p),1) for f; PBE2=Parametric Bayesian Estimator with beta prior (0.5,0.5) for p and conditional uniform prior (-p/(1-p),1) for f; PBE3=Parametric Bayesian Estimator with beta prior (0.5,0.5) for p and uniform prior (0,0.5) for f.

Table 7. Comparison of Priors in the Bayesian Estimation (Simple Random Sampling, n=50)

		DSE	PBE1	PBE2	PBE3	BPELE1	BPELE2
	Bias	-0.0061	0.0993	0.0978	0.1318	0.1196	0.1505
p=0.1	MSE	0.0266	0.0248	0.0241	0.021	0.0217	0.0242
f=0.05	CP	51.9	95.2	95.3	94.5	99.8	99.7
	AL	0.4455	0.5738	0.566	0.4175	0.6297	0.445
	Bias	-0.0183	0.0782	0.0759	0.0962	0.0885	0.1098
p=0.1	MSE	0.0297	0.0217	0.0208	0.0132	0.0155	0.0136
f=0.1	CP	61	97.6	97.5	98.6	99.7	100
	AL	0.504	0.5887	0.5805	0.4251	0.6478	0.4501
	Bias	-0.0304	0.0449	0.0414	0.0296	0.0354	0.0313
p=0.1	MSE	0.0382	0.0215	0.0206	0.0057	0.0124	0.003
f=0.2	CP	75	97.7	98.1	99.8	99.5	100
	AL	0.6141	0.6159	0.6091	0.4306	0.684	0.4554
	Bias	-0.0185	0.0033	0.0037	0.0939	-0.0177	0.0985
p=0.3	MSE	0.0213	0.0175	0.0174	0.0128	0.0173	0.013
f=0.05	CP	91.3	95.5	95.3	95	97.8	97.3
	AL	0.547	0.5155	0.5138	0.3363	0.593	0.3508
	Bias	-0.0107	0.006	0.0061	0.0698	-0.0153	0.0718
p=0.3	MSE	0.0218	0.0181	0.0179	0.0098	0.0181	0.0092
f=0.1	CP	92.6	95.7	96	97.9	97.4	99.2
	AL	0.5563	0.5224	0.5205	0.3614	0.5973	0.3744
	Bias	-0.0082	-0.0002	-0.0007	0.0209	-0.0204	0.0188
p=0.3	MSE	0.0222	0.0184	0.0183	0.0068	0.0199	0.0058
f=0.2	CP	92.4	94.8	95.1	98.4	96.6	99.2
	AL	0.563	0.5256	0.525	0.3885	0.5944	0.4014

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length
DSE=Direct Sample Estimator;
PBE1=Parametric Bayesian Estimator with uniform prior (0,1) for p and conditional uniform prior (-p/(1-p),1) for f;
PBE2=Parametric Bayesian Estimator with beta prior (0.5,0.5) for p and conditional uniform prior (-p/(1-p),1) for f;
PBE3=Parametric Bayesian Estimator with beta prior (0.5,0.5) for p and uniform prior (0,0.5) for f.
BPELE1=Bayesian Pseudo-Empirical Likelihood Estimator with uniform prior (0,1) for f;
BPELE2=Bayesian Pseudo-Empirical Likelihood Estimator with uniform prior (0,0.5) for f;

Chapter 3: Estimation of the Inbreeding Coefficient under Unequal Probability Sampling

The analysis of genetic data need not only consider the variation associated with statistical sampling, but also take the evolutionary genetic sampling into account. The former can be controlled by increasing the number of sampled units within populations and/or the number of sampled populations. However, the latter is an intrinsic property of the stochastic process (evolution) and therefore cannot be controlled by the sample size. Even if we could take the sampled populations back to a previous time point and run the evolution process under the exact same biological conditions, the genotype frequencies in the new evolved populations would differ from what has been already observed. Increasing the number of sampled individuals or the number of sampled populations will not decrease the variation associated with this evolutionary process. Holsinger & Weir (2009) suggested to characterize such genetic sampling by *F*-statistics. With the collection of genetic data from complex survey, it poses more challenges to data analysis because of its non-simple random sampling design. In both the NHANES and the HRS genetic data quality control report, the test of HWE is conducted as if the data is from a simple random sampling design (NHANES, 2015; Weir D. R., 2012). Furthermore, the estimated inbreeding coefficient from the HRS study is also under a simple random sampling assumption (Weir D. R., 2012). We believe that ignoring the statistical sampling in analysis may introduce some bias to the estimation of the inbreeding coefficient for the study population.

In this chapter, we are trying to estimate the target population inbreeding coefficient, however based on samples with unequal selection probabilities. Previously developed multinomial model under simple random sampling need be extended to such sampling design with unequal probability of selection of units.

3.1 Direct Design-Based Estimator

The design-based and model-assisted frequentist approach have been widely used in survey practice due to its capability to handle complex survey sampling design. The estimators of finite population parameters are proved to be design consistent for large samples. Let's again look at a

biallelic case with alleles *A* and *a* (*k*=2). Suppose a finite population of size *N*, let y_i (*i*=1, 2, …, *N*) be the number of allele *A* for each subject in the population. Individual with genotype *aa* has y_i=0; individual with genotype *Aa* has y_i=1; and individual with genotype *AA* has y_i=2. The population total of each genotype count can be calculated from indicator function as

$$N_0 = \sum_{i=1}^{N} I_{[y_i=0]}, \qquad N_1 = \sum_{i=1}^{N} I_{[y_i=1]}, \qquad N_2 = \sum_{i=1}^{N} I_{[y_i=2]},$$

where $I_{[Y=y]}$ is an indicator function with value of 1 if *Y*=*y* and 0 otherwise. If this finite population genotype is assumed to be selected from a super population with general inbreeding model described in (3), then the finite population inbreeding coefficient can be derived as

$$f = 1 - \frac{N_1/N}{2 * \frac{2N_2 + N_1}{2N} * \frac{2N_0 + N_1}{2N}}.$$

It can also be expressed as 1 - (number of observed heterozygotes) / (number of expected heterozygotes) at the finite population level. Finite population parameter *f* can be viewed as a function of population's genotype frequencies N_0/N, N_1/N, and N_2/N. Similarly as we discussed in Chapter 1, this can be interpreted as a correlation coefficient between two homologous genes in uniting gametes from this finite population.

Under unequal probability sampling of *n* subjects from this finite population, each sampled subject has a design weight d_i, which is the inverse of sampling unit inclusion probability with possible nonresponse and post stratification adjustments. Therefore, Horvitz–Thompson type estimators of the population total and three genotype count totals are

$$\widehat{N} = \sum_{i=1}^{n} d_i, \widehat{N_0} = \sum_{i=1}^{n} d_i I_{[y_i=0]}, \widehat{N_1} = \sum_{i=1}^{n} d_i I_{[y_i=1]}, \text{and } \widehat{N_2} = \sum_{i=2}^{n} d_i I_{[y_i=2]},$$

respectively.

The design-based direct survey estimator for the finite population inbreeding coefficient therefore can be expressed as a function of the estimators of the population genotype frequencies, e.g.,

$$\hat{f} = 1 - \frac{\widehat{N}_1/\widehat{N}}{2 * \frac{2\widehat{N}_2 + \widehat{N}_1}{2\widehat{N}} * \frac{2\widehat{N}_0 + \widehat{N}_1}{2\widehat{N}}}.$$

The analytical form of its standard error using the Taylor linearization is not trivial and is outside of the scope of this dissertation. Instead, the variance can be easily derived by using the resampling technique (Rust & Rao, 1996). We use jackknife method to estimate the variance of the survey deign-based estimator $\hat{f}$.

3.2 Parametric Bayesian Estimator with Survey Weight

As we see from Chapter 2, the Bayesian method can restrict the estimator to its parameter space through the prior distribution. With the introduction of MCMC, the Bayesian method is increasingly used for multi-level modeling. However, the full Bayesian model under the complex sampling design is not feasible to include all the sampling design information (Gelman, 2007). On the other hand, not all survey sampling design information is generally included in the public released data in order to prevent disclosure of participants' social, economic and personal health information. Typical national surveys such as the NHANES and the HRS include sampling design information including the analysis stratum, the primary sampling units (PSUs) as well as the final adjusted sampling weight. Rao (2011) had an extensive review of the application of the Bayesian method in the complex survey practice, focusing on the descriptive finite population parameters. Among all three challenges (appropriate likelihood, proper prior and posterior) to the Bayesian method, formalizing the likelihood function is the biggest hurdle to the survey statistician. For multipurpose complex survey, the parametric Bayesian method with distribution assumptions cannot be easily generalized due to the difficulties in validating its assumptions. For this dissertation, we are interested in screening SNPs through the inbreeding coefficients to measure their magnitude of deviation from the HWE.

Let $\widetilde{d_i} = \frac{d_i}{\sum_{i=1}^{n} d_i}$ be the normalized design weight so that $\sum_{i=1}^{n} \widetilde{d_i} = 1$. Assume that the weights $\widetilde{d_i}$ are independent of y_i, i=1, …, n, therefore they can be considered as fixed numbers given the sample. Follow Skinner (1989) and Ghosh & Maiti (2004), the weighted likelihood is

$$l_d(p,f) = \prod_{i=1}^{n} Pr^{\widetilde{d_i}}(y_i|p,f) = C(d,y) p_0^{\sum_{i=1}^{n} \tilde{d}_i I_{[y_i=0]}} p_1^{\sum_{i=1}^{n} \tilde{d}_i I_{[y_i=1]}} p_2^{\sum_{i=1}^{n} \tilde{d}_i I_{[y_i=2]}},$$

where $C(d, y)$ is a constant depending on the data and the sampling weights. It can be seen from the above likelihood that it still follows a multinomial distribution with the parameter $\boldsymbol{p} = (p_0, p_1, p_2)$ and the normalized weighted counts for genotypes. We use p_0, p_1, p_2 to represent the previous genotype frequency p_{22}, p_{12}, p_{11} when *y*=0, 1, or 2, respectively.

$\boldsymbol{n}^*|\boldsymbol{p} \sim Multinomial(\boldsymbol{n}^*, \boldsymbol{p})$, or,

$$Pr(\boldsymbol{n}^*|\boldsymbol{p}) = Pr\big((n_0{}^*, n_1{}^*, n_2{}^*)\big|(p_0, p_1, p_2)\big) = \frac{n^*!}{n_0{}^*!\, n_1{}^*!\, n_2{}^*!} {p_0}^{n_0{}^*} {p_1}^{n_1{}^*} {p_2}^{n_2{}^*}.$$

In our Bayesian framework, we use 2*p* to denote a finite population parameter $\bar{Y} = \frac{1}{N}\sum_i^N y_i$. The design-based variance is calculated as $\mathrm{Var}_{\mathrm{design}}(\hat{p}) = \frac{1}{4}\mathrm{Var}_{\mathrm{design}}(\hat{\bar{Y}})$. Under simple random sampling without replacement, the variance is $\mathrm{Var}_{\mathrm{srswor}}(\hat{p}) = \frac{1}{4}\mathrm{Var}_{\mathrm{srswor}}(\hat{\bar{Y}})$. The design effect is $\mathrm{Deff} = \mathrm{Var}_{\mathrm{design}}(\hat{p})/\mathrm{Var}_{\mathrm{srswor}}(\hat{p})$ and therefore the effective sample size is $n^* = n/\mathrm{Deff}$. We incorporate the complex sampling design through the effective sample size n^* and effective genotype count $\boldsymbol{n}^* = (n_0{}^*, n_1{}^*, n_2{}^*)$ in our Bayesian multinomial model, in which

$$n_0{}^* = n^* \sum_{i=1}^{n} \tilde{d}_i \, I_{[y_i=0]},$$
$$n_1{}^* = n^* \sum_{i=1}^{n} \tilde{d}_i \, I_{[y_i=1]},$$
$$n_2{}^* = n^* \sum_{i=1}^{n} \tilde{d}_i \, I_{[y_i=2]}.$$

To be included in the multinomial model, all above effective sample size and effective genotype counts are rounded to their nearest integers. The Bayesian multinomial model using the above effective sample size and effective genotype counts can therefore be specified as follows:

Level 1: $\boldsymbol{n}^* \sim \mathrm{Multinomial}(n^*, \boldsymbol{p})$, $\boldsymbol{n}^* = (n_0{}^*, n_1{}^*, n_2{}^*)$, $\boldsymbol{p} = (p_0, p_1, p_2)$, $n^* = n_0{}^* + n_1{}^* + n_2{}^*$,

$$\begin{cases} p_0 = q^2 + pqf, \\ p_1 = 2pq(1-f), \\ p_2 = p^2 + pqf. \end{cases}$$

Level 2: $p \sim \mathrm{Uniform}[0,\ 1]$

$$f|p \sim \mathrm{Uniform}\left[\frac{-\min(p,1-p)}{1-\min(p,1-p)}, 1\right]$$

To model the categorical observations, Jiang & Lahiri proposed generalized linear mix model in the small area estimation context (2006). The survey design may be informative and need be considered in the model (Chen & Lahiri, 2012). Our using of effective sample size in the hierarchical Bayesian model with complex survey data is further inspired by a similar usage in the Bayesian method with pseudo-empirical likelihood (Rao & Wu, 2010) that we will discuss in the next section. Such application of effective sample size is not new for survey practitioners. A recent example can be found in the context of small area estimation from Chen, Wakefield and Lumely (2014). To take the sampling design into account, other alternative method is to model the distribution of the weights for the non-sampled units in the population and simultaneously include them as predictors in a nonparametric Gaussian process regression (Si, Pillai, & Gelman, 2015).

3.3 Bayesian Pseudo-Empirical Likelihood Estimator

The Empirical likelihood (EL), introduced by Owen (1988), is a powerful nonparametric method for statistical inference and recently gets more and more attention. Although in this dissertation the proposed method is mainly for estimating the inbreeding coefficient from survey sample, it can be a useful tool to handle auxiliary variables such as environmental information in the gene-disease association study. For example, disease such as cancer is often age dependent. In the national level complex survey, the population average of age is known from the Census. By incorporating age into the Bayesian model with the empirical likelihood, an age-adjusted genotypic effect can be estimated from complex survey sample. The proposed method is also flexible in terms of incorporating the prior information. For information available on the parameter of interest, either from independent researches or from related measures within the same study, it can be incorporated into the analysis through proper usage of the prior distribution.

Considering the finite population a random sample from an infinite superpopulation, Chen & Sitter (1999) first proposed a pseudo empirical log-likelihood for a general sampling design as $\sum_{i\in s} d_i \log(\pi_i)$, a Horvitz-Thompson estimator of the "census" log-likelihood $l_N(\pi) = \sum_{i=1}^{N} \log(\pi_i)$. Wu and Rao (2006) proposed the pseudo-empirical likelihood in the form of $l_{PEL} = n\sum_{i\in s} \widetilde{d}_i \log(\pi_i)$, where $\widetilde{d}_i = \frac{d_i}{\sum_{i\in s} d_i}$ are the normalized design weights. If the design weights d_i are all equal, then $n\sum_{i\in s} \widetilde{d}_i \log(\pi_i)$ is reduced to the original definition of the empirical likelihood

$\sum_{i\in s}\log(\pi_i)$. The sample size *n* in the above empirical likelihood could also be replaced by the effective sample size, which is defined as sample size *n* divided by the design effect. By using the effective sample size, this method combines the survey design information and Bayesian models in the estimation of population mean (Rao & Wu, 2010). Maximizing those proposed pseudo empirical likelihood functions with the same constraints on those π_i's are essentially the same.

We need to maximize the likelihood l_{PEL} with the following restrictions on the total probability and unbiased estimating functions:

$$\pi_i > 0, \sum_{i\in s}\pi_i = 1, \sum_{i\in s}\pi_i[y_i(2-y_i) - 2(1-f)p(1-p)] = 0.$$

Although it is possible to include another constraint based on estimating equation $E(Y-2p)=0$ by using Lagrange method as we discussed in Section 2.3, the benefit is significantly impacted by the computation burden imposed by the calculation of the Lagrange parameter λ. Lagrange method is used to derive the $\hat{\pi}_i$ to maximize the pseudo empirical log-likelihood by first setting

$$G = n\sum_{i\in s}\tilde{d}_i\log(\pi_i) - n\lambda\sum_{i\in s}\pi_i[y_i(2-y_i) - 2(1-f)p(1-p)] - \gamma(\sum_{i\in s}\pi_i - 1).$$

Taking partial derivative of G with respect to π_i and setting to 0, we have:

$$\frac{\partial G}{\partial \pi_i} = n\frac{\tilde{d}_i}{\pi_i} - n\lambda[y_i(2-y_i) - 2(1-f)p(1-p)] - \gamma = 0.$$

The above equation gives solution

$$\hat{\pi}_i = \frac{\tilde{d}_i}{1+\lambda[y_i(2-y_i) - 2(1-f)p(1-p)]},$$

where λ need be solved by equation

$$\sum_{i\in s}\frac{\tilde{d}_i[y_i(2-y_i) - 2(1-f)p(1-p)]}{1+\lambda[y_i(2-y_i) - 2(1-f)p(1-p)]} = 0.$$

Unlike method developed in Qin & Lawless (1994) where the constraint based on the second moment is a known function of the first moment parameter, constraint in our case depends on the unknown population nuisance parameter *p* as well as the parameter *f* we are interested in.

Therefore, it is not a smooth function. Sitter & Wu (2002) suggested that the estimation based on the second-order population function such as quadratic function requires knowledge of the second-order inclusion probability induced from a synthetic finite population as well as the knowledge of the second-order auxiliary variable information. Without an obvious known auxiliary variable which is closely related to the genotype, to improve the efficiency of computation, an unbiased sample estimator of the allele frequency $\hat{p}$ is used to replace p. To simplify our notation, we define $V=Y(Y\text{-}2)$, an indicator variable for the heterozygote. Let $\theta = 2(1-f)\hat{p}(1-\hat{p})$. Therefore, we have $v_i = y_i(2-y_i)$ and $u_i = y_i(2-y_i) - 2(1-f)\hat{p}(1-\hat{p}) = v_i - \theta$. We want to maximize the profile pseudo-empirical log-likelihood with new constraints as:

$$\pi_i > 0, \sum_{i\in s} \pi_i = 1, \sum_{i\in s} \pi_i u_i = 0.$$

Using the Lagrange method by taking the partial derivative of $G = n\sum_{i\in s} \widetilde{d}_i \log(\pi_i) - n\lambda \sum_{i\in s} \pi_i u_i - \gamma(\sum_{i\in s} \pi_i - 1)$ with respect to π_i and setting this to 0, we have:

$$\frac{\partial G}{\partial \pi_i} = n\frac{\widetilde{d}_i}{\pi_i} - n\lambda u_i - \gamma = 0.$$

The above equation gives the solution

$$\hat{\pi}_i = \frac{\widetilde{d}_i}{1+\lambda u_i}\ .$$

The empirical likelihood can be estimated as

$$l_{PEL} = n\sum_{i\in s} \widetilde{d}_i \log(\frac{\widetilde{d}_i}{1+\lambda u_i}) = n\sum_{i\in s} \widetilde{d}_i \log(\widetilde{d}_i) - n\sum_{i\in s} \widetilde{d}_i \log(1+\lambda u_i),$$

where $\lambda = \lambda(f)$ is the solution to the equation $\sum_{i\in s} \frac{\widetilde{d}_i u_i}{1+\lambda u_i} = 0$. Chen, Sitter & Wu (2002) proposed an algorithm based on the modified Newton's method and it is theoretically guaranteed to be convergent.

Multiplying $1+\lambda u_i - \lambda u_i$ to $\frac{\widetilde{d}_i u_i}{1+\lambda u_i}$, it can be verified that

$$\sum_{i\in s}\frac{\widetilde{d}_i u_i}{1+\lambda u_i}=\sum_{i\in s}\frac{\widetilde{d}_i u_i(1+\lambda u_i-\lambda u_i)}{1+\lambda u_i}=\sum_{i\in s}\frac{-\lambda\widetilde{d}_i {u_i}^2+\widetilde{d}_i u_i(1+\lambda u_i)}{1+\lambda u_i}$$

$$=-\lambda\sum_{i\in s}\frac{\widetilde{d}_i {u_i}^2}{1+\lambda u_i}+\sum_{i\in s}\widetilde{d}_i u_i=0.$$

Therefore $\lambda\sum_{i\in s}\frac{\widetilde{d}_i {u_i}^2}{1+\lambda u_i}=\sum_{i\in s}\widetilde{d}_i u_i$.

Let $U^*=max_{i\in s}u_i$, which is bounded since y_i is the count for the allele A and y_i(2- y_i) is an indicator for the heterozygote, we have $\frac{|\lambda|}{1+|\lambda|U^*}\sum_{i\in s}\widetilde{d}_i {u_i}^2\leq|\sum_{i\in s}\widetilde{d}_i u_i|$. Under the regulatory assumption that the Horvitz-Thompson type estimator is asymptotically normally distributed, we have $\sum_{i\in s}d_i\,v_i=\overline{V}+O_\pi(n^{-1/2})$, where $\overline{V}$ is the finite population heterozygote frequency. Furthermore, assuming the sampling design $\pi(s)$ satisfies $\frac{\sum_{i\in s}d_i}{N}=1+O_\pi(n^{-1/2})$, we have

$$\sum_{i\in s}\tilde{d}_i\,v_i=\frac{\sum_{i\in s}d_i\,v_i}{\sum_{i\in s}d_i\,/N}=\overline{V}+O_\pi\left(n^{-\frac{1}{2}}\right).$$

Since $\sum_{i\in s}\widetilde{d}_i {u_i}^2=\sum_{i\in s}\widetilde{d}_i(v_i-\theta)^2$ is a Hajek estimator of S_v^2, $\sum_{i\in s}\widetilde{d}_i {u_i}^2=S_v^2+O(1)$. It can be derived that $\lambda=\frac{\sum_{i\in s}\widetilde{d}_i u_i}{\sum_{i\in s}\widetilde{d}_i {u_i}^2}+O_\pi(n^{-1/2})$.

Using the Taylor series expansion of log(1+x) at $x=\lambda u_i=\lambda(v_i-\theta)$ up to the second order, suppose we have a non-informative but proper uniform prior $\varphi(f)\propto 1$, the posterior therefor is

$$\varphi(f|y)=e^{l_{PEL}}\times\varphi(f)$$

$$=\exp\{\log[\varphi(f)]+n\sum_{i\in s}\widetilde{d}_i\log(\hat{\pi}_i)\}$$

$$\propto\ \exp\{-n\sum_{i\in s}\widetilde{d}_i\log(1+\lambda u_i)\}$$

$$=\ \exp\{-n\sum_{i\in s}\widetilde{d}_i(\lambda u_i-(\lambda u_i)^2/2)\}$$

$$=\exp\{-n\lambda[\sum_{i\in s}\widetilde{d}_i(v_i-\theta)-\frac{\sum_{i\in s}\widetilde{d}_i(v_i-\theta)}{\sum_{i\in s}\widetilde{d}_i(v_i-\theta)^2}\times\frac{\sum_{i\in s}\widetilde{d}_i(v_i-\theta)^2}{2}]\}$$

$$= \exp\left\{-n\frac{\sum_{i\in s}\tilde{d}_i(v_i-\theta)}{\sum_{i\in s}\tilde{d}_i(v_i-\theta)^2}\times\frac{\sum_{i\in s}\tilde{d}_i(v_i-\theta)}{2}\right\}$$

$$= \exp\left\{-\frac{n\left(\sum_{i\in s}\tilde{d}_i v_i-\theta\right)^2}{2\sum_{i\in s}\tilde{d}_i(v_i-\theta)^2}\right\}$$

$$= \exp\left\{-\frac{[f-\left(1-\frac{\sum_{i\in s}\tilde{d}_i v_i}{2\hat{p}(1-\hat{p})}\right)]^2}{2S_v^2/n}\right\}$$

since$E_\pi\left(\sum_{i\in s}\tilde{d}_i v_i\right)\doteq\bar{V}$, $E_\pi\{\sum_{i\in s}\tilde{d}_i(v_i-\theta)^2\}\doteq S_v^2$ and $\frac{S_v^2}{n}=V_\pi(\sum_{i\in s}\tilde{d}_i v_i)$.Therefore similarly as concluded in Rao & Wu (2010), the Bayesian posterior pseudo-empirical maximum likelihood estimator of the inbreeding coefficient is asymptotically normal.

Unlike a non-informative improper prior for the location parameter in Rao & Wu's (2010) method, in our case the non-informative prior $\varphi(f)$ is a uniform distribution. The empirical likelihood is bounded above by one. It can be verified that $\int e^{l_{PEL}}\varphi(f)df\leq\int\varphi(f)df=1$, which implies that the posterior is proper (Chaudhuri & Ghosh, 2011). For more complicate improper priors, Monahan & Boos (1992) discussed that the empirical likelihood is proper if the resulting credible intervals have correct coverages for all absolutely continuous prior. Lazar (2003) first studied this property in a simulation study. Fang & Mukerjee (2005; 2006) studied the coverage of the posterior credible intervals using the empirical likelihood and proved that the posterior one-sided credible interval have $o(m^{-1})$ margin of error. In Chaudhuri & Ghosh (2011)'s simulation study, the coverage probabilities were compared to their nominal values to justify the use of the empirical likelihood in the Bayesian inference.

Due to the nonparametric nature of the empirical likelihood, the posterior distribution is not in a closed form therefore is not tractable. Approximate Bayesian Computation (ABC) method can be used for the Bayesian inference when the posterior is not tractable but the distribution for the prior and the likelihood function can be both simulated. However, the empirical likelihood makes it difficult for computation based on a complex statistical model. To bypass the explicit computation

of the posterior distribution, the importance sampling method can be performed with obvious two reasons. First, the posterior empirical likelihood is convergent and therefore proper for the Bayesian inference as the number of observations increase to infinity. Second, without requirement of simulating replications of data having many tuning parameters from the sampling model, it substantially reduces the computation time. Using the empirical likelihood as the basis for the Bayesian inference (Lazar, 2003), we can calculate a finite expectation of the form $E_{\varphi(f|y)}\big(u(F)\big) = \int u(f)\varphi(f|y)df$ with respect to the posterior distribution as

$$\int u(f)\frac{\varphi(f|y)}{\varphi(f)}\varphi(f)df = \int u(f)\frac{e^{l_{PEL}} \times \varphi(f)}{\varphi(f)}\varphi(f)df = \int \{u(f)w(f)\}\varphi(f)df$$
$$= E_{\varphi(f)}\{u(F)w(F)\}.$$

To keep the notation consistent, here *F* represents the random variable for inbreeding coefficient *f* and $w(\cdot) = \frac{\varphi(\cdot|y)}{\varphi(\cdot)}$ is the weighting function. If we are interested in estimating the posterior mean, then $u(F) = F$.

A basic sampling algorithm could be developed as following two steps to generate an importance sample $\left(f_i, e^{l_{PEL}(f_i|\boldsymbol{y})}\right), i = 1, \dots, M$:

Step 1: for *i*=1 to M, sampling f_i from $\varphi(f)$;

Step 2: use Lagrange method and Newton algorithm to calculate $e^{l_{PEL}(f_i|\boldsymbol{y})}$, which is a natural distance and importance weight (Mengersen, Pudlo, & Robert, 2013).

Bayesian inference therefore can be generated using the estimated posterior distribution based on the importance sampling. With the above prove that the posterior empirical likelihood is proper, we used a grid search method to find the posterior mean as well as 95% credible interval from the posterior distribution, instead of a rigorous simulation. In order to derive the posterior distribution $\varphi(f|y)$, we need find the normalizing constant function $c(y)$ such that $\int \varphi(f|y)\,df = \int c(y)e^{l_{PEL}}\varphi(f)\,df = 1$. Under our concavity assumption for the posterior, the grid search method is applied to calculate this normalizing factor. The theoretical parameter support for the posterior density is $\frac{-p_{min}}{1-p_{min}} \leq f \leq 1$, $p_{min} = min(\hat{p}, 1-\hat{p})$. Computation of the pseudo posterior

density near the endpoint of the parameter support is problematic since the non-existence of the solution at those endpoints. A non-zero tolerance ϵ (for example, 0.00001) is used such that the practical support (c_1, c_2) is used with $\frac{-p_{min}}{1-p_{min}} < c_1 < \hat{f}_{PEL} < c_2 < 1$, $h(c_1) < \epsilon$ and $h(c_2) < \epsilon$, where

$$h(f) = e^{l_{PEL}(f|y)} = \exp\{n \sum_{i \in s} \tilde{d}_i \log(\hat{\pi}_i(f))\} = \exp\{n \sum_{i \in s} \tilde{d}_i \log(1 + \lambda(f) u_i)\}.$$

First, we divide the un-normalized posterior into m grids, each with width w, starting from c_1, which is a value that is at the far left of parameter space. We then evaluate the un-normalized posterior at each grid point at c_1+w, c_1+2w, …, $c_2=c_1+mw$, where the last grid point is at the far right of the practical support. At each grid point m, f_m, $\lambda(f_m)$, $\hat{\pi}_m$and the height at un-normalized posterior $e^{l_{PEL}}{}_m$ can be calculated. The normalizing constant function $c(y) = \{\int_{c_1}^{c_2} h(f) df\}^{-1}$ is therefore estimated as the inverse of the area under the un-normalized posterior, which is the sum of the areas of rectangles calculated as the width w times the height $e^{l_{PEL}}{}_m$ at the un-normalized posterior ordinates. Finally, the posterior mean and 95% credible interval $[f_{0.025}^{\text{BPELE}}, f_{0.975}^{\text{BPELE}}]$ can be derived from the estimated posterior distribution.

3.4 Simulation Study

A simulation study for a single locus with two alleles (A, a) is conducted to evaluate the performance of the above three estimators. Let a finite population be of size N= 4,800 individuals consisting of two stratum with sub-population size ratio as 5:1. The genotype records are generated independently from the multinomial model with genotype frequencies of $P(aa) = (1-p)^2 + p(1-p)f$, $P(Aa) = 2(1-f)p(1-p)$, and $P(AA) = p^2 + p(1-p)f$.

The parameters for the simulation are the allele A frequency p=0.1, 0.3, or 0.5 and the inbreeding coefficient f=-0.1, 0, 0.05, 0.1, or 0.2. For the realized data, estimators of the inbreeding coefficient are calculated by methods described above. The bias and mean square error (MSE) of the estimator are calculated similarly as in Chapter 2, except that the parameter is based on the finite population of size N=4,800 instead of the super population parameter input.

3.4.1 Stratified Simple Random Sampling

At each one of 1000 iterations in our first simulation design, a sample size of twenty (20), forty (40) or eighty (80) samples were randomly selected from each stratum. Therefore, the ratio of the sample selection probability is 1:5, an inverse of stratum population size. The weights from this stratified simple random sampling design is non-informative.

Table 8, Table 9, and Table 10 report the simulation results in comparing design-based estimator and two Bayesian estimators for the bias, MSE, coverage probability (CP), out of range probability (ORP), and average length (AL). Similarly as the results from simple random sampling in Chapter 2, the parametric Bayesian method tends to have positive bias, while the frequentist method often has negative bias. The Bayesian empirical likelihood method has a positive bias when the allele frequency is small (p=0.1), while shows negative bias when the allele frequency is moderate (p=0,3, 0.5). When the allele frequency is small (p=0.1), the Bayesian methods have larger bias especially when the inbreeding coefficient is also small (f=-0.1, 0, 0.05). When the allele frequency is moderate (p=0.3, 0.5), the Bayesian methods tend to have smaller bias especially for the parametric Bayesian estimator. On average from all simulation scenario, the Bayesian empirical likelihood method has the smallest MSE, followed by the parametric Bayesian method. A similar observation can be made for the coverage probability. In all three simulations with different sample sizes, the design-based estimator does not maintain the nominal coverage probability. When the allele frequency is large (p=0.5), all three methods cannot maintain the nominal coverage probability. In fact, the genotype frequencies are similar in this situation for all inbreeding coefficient values as show in Table 3. When the allele frequency is small (p=0.1) and the inbreeding coefficient is not near the border of the parameter space, the Bayesian pseudo-empirical likelihood estimator has much better MSE and coverage than both the parametric Bayesian estimator and design-based estimator. This is expected as the design-based method is not stable when the data is skewed so the asymptotic assumption may not hold well.

Similarly as shown in Chapter 2, the design-based estimator may produce an estimator outside of the parameter space, especially when the allele frequency is small (p=0.1) and the inbreeding coefficient is also small (f=-0.1, 0). For example, when sample size is 40 per stratum and the true inbreeding coefficient is 0, 14.1% of the simulated design-based estimators have the values lower

than -0.1111, which is the lower boundary of the parameter space for f under p=0.1 (Table 9, Figure 7 Panel B).

Table 8. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Stratified Simple Random Sampling，n=20 per Strata)

p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0045	0.0042	85.9	0.1874	45.5	0.1604	0.0328	76.3	0.5693	0.2286	0.0588	48.2	0.6095
	0	-0.0244	0.027	43.8	0.4758	25.4	0.107	0.0274	94.7	0.5978	0.1565	0.034	99.3	0.6522
	0.05	-0.0204	0.0414	50.5	0.6315	17.4	0.0936	0.0311	94.6	0.6175	0.1323	0.0304	98.8	0.6765
	0.1	-0.0184	0.054	56.3	0.6956	12.9	0.0817	0.0353	94	0.6274	0.11	0.028	98.9	0.6945
	0.2	-0.028	0.0679	69.8	0.8677	4.8	0.0438	0.0351	95.2	0.6525	0.0541	0.0228	98.5	0.7311
0.3	-0.1	-0.0161	0.0317	86.3	0.5913	3.1	0.0277	0.0275	89.2	0.5306	0.0225	0.0216	96.1	0.5901
	0	-0.015	0.0336	88.8	0.6332	1.2	0.0192	0.0282	89.7	0.5495	0.0043	0.0239	95.3	0.6197
	0.05	-0.0242	0.0359	88.9	0.6442	0.5	0.0055	0.0287	90.2	0.5518	-0.0093	0.0252	96.1	0.6263
	0.1	-0.0157	0.0368	88.5	0.6549	0	0.0078	0.0291	89.9	0.5639	-0.0096	0.0269	95.3	0.6362
	0.2	-0.0153	0.0384	88.9	0.6653	0	-0.0025	0.0306	89.2	0.5663	-0.0222	0.0302	93.9	0.6425
0.5	-0.1	-0.0112	0.0339	90.4	0.631	0	0.0116	0.0296	90	0.5606	-0.0054	0.0303	90.9	0.5894
	0	-0.013	0.0341	91	0.6365	0	0.0027	0.0296	90.2	0.5662	-0.0121	0.0307	91.8	0.595
	0.05	-0.0264	0.038	88.1	0.6366	0	-0.0134	0.0326	87.6	0.5654	-0.0277	0.0345	88.6	0.5948
	0.1	-0.0157	0.0353	90.8	0.6355	0	-0.0069	0.0307	90.9	0.5686	-0.0197	0.032	91.3	0.5939
	0.2	-0.0134	0.0379	88.6	0.6275	0	-0.0108	0.0334	87.8	0.563	-0.0224	0.0347	89.6	0.5875

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 9. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Stratified Simple Random Sampling，n=40 per Strata)

p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0028	0.0027	89.2	0.144	43.8	0.093	0.0121	80.9	0.3747	0.1678	0.0312	51.4	0.4696
	0	-0.0097	0.0175	62.4	0.4074	14.1	0.0605	0.0169	92.3	0.4285	0.1036	0.0166	99.4	0.5266
	0.05	-0.015	0.0223	75	0.4965	7.5	0.0464	0.0187	93.8	0.4506	0.076	0.0133	99.8	0.5516
	0.1	-0.0163	0.0285	77.4	0.5412	5.1	0.039	0.0216	93.7	0.4742	0.0557	0.0132	98.8	0.5731
	0.2	-0.0246	0.0388	85.2	0.6286	1.6	0.015	0.0269	92.7	0.506	0.0102	0.0163	98.9	0.6112
0.3	-0.1	-0.0094	0.0168	88.4	0.419	0.3	0.0111	0.0153	89	0.3911	0.0014	0.0133	96.8	0.4665
	0	-0.0076	0.0176	89	0.4422	0	0.0089	0.016	89.3	0.4076	-0.0053	0.0149	95.9	0.484
	0.05	-0.0076	0.0181	89.5	0.4505	0	0.0063	0.0162	90.5	0.4133	-0.0086	0.0158	95.8	0.4896
	0.1	-0.0071	0.0185	89.7	0.4567	0	0.0042	0.0164	90.5	0.4203	-0.0113	0.0167	95.2	0.4906
	0.2	-0.005	0.0191	90.3	0.4616	0	0.0002	0.017	89.9	0.4244	-0.0137	0.0179	93.8	0.4845
0.5	-0.1	-0.0029	0.0172	90.6	0.4415	0	0.0078	0.0161	90	0.4131	-0.0003	0.0163	90.6	0.4265
	0	-0.0047	0.0169	91.5	0.4442	0	0.0031	0.0158	91.7	0.4182	-0.0043	0.0161	91.9	0.4293
	0.05	-0.0042	0.0172	90.4	0.444	0	0.0015	0.0159	90.3	0.419	-0.0052	0.0164	90.8	0.4292
	0.1	-0.005	0.0178	89.3	0.443	0	-0.001	0.0165	89.7	0.4187	-0.0071	0.017	89.8	0.428
	0.2	-0.0039	0.0176	88.4	0.4372	0	-0.004	0.0164	88.3	0.4142	-0.0085	0.0168	89.1	0.4225

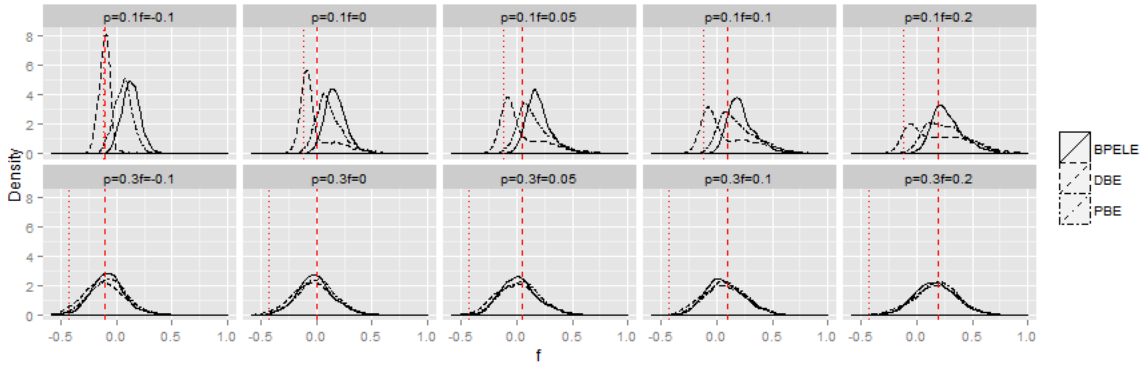
* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 10 Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Stratified Simple Random Sampling，n=80 per Strata)

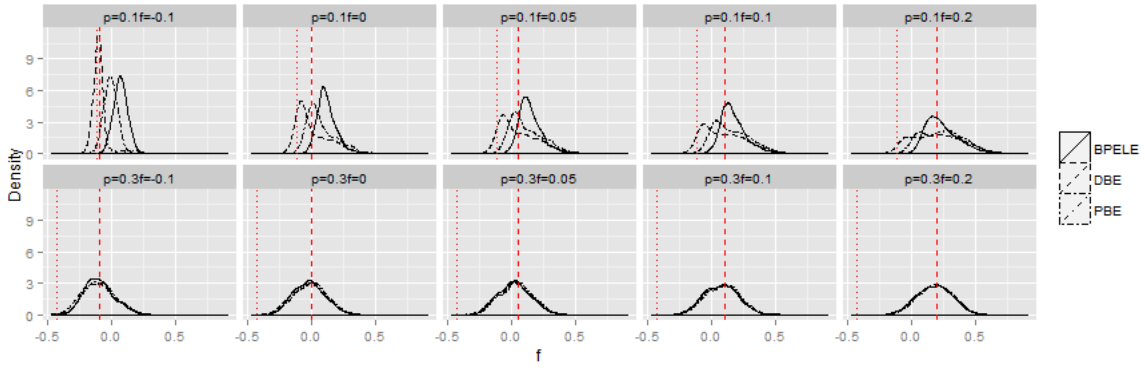
p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0003	0.0014	90.5	0.1111	39.7	0.0529	0.0044	84.7	0.2333	0.1222	0.0161	61.2	0.3531
	0	-0.0062	0.0089	82.2	0.3121	5.2	0.0347	0.0089	93	0.3078	0.0606	0.0069	99.9	0.4175
	0.05	-0.003	0.0121	87.5	0.3755	1.9	0.0309	0.0114	91.3	0.3371	0.0404	0.0068	99.6	0.4481
	0.1	-0.0086	0.0143	87.4	0.4053	0.9	0.0227	0.0125	90.9	0.361	0.0219	0.0072	98.8	0.4696
	0.2	-0.0077	0.0178	90.1	0.4574	0.1	0.0123	0.0148	90.2	0.3951	-0.0057	0.0113	98.4	0.5081
0.3	-0.1	-0.0057	0.0074	91	0.2957	0	0.0054	0.0071	91.6	0.2859	-0.005	0.0069	96.7	0.3567
	0	-0.0007	0.0086	90.5	0.3112	0	0.0079	0.0082	89.8	0.2991	-0.0025	0.0082	95.3	0.3609
	0.05	-0.0058	0.0086	91.1	0.3163	0	0.0017	0.0081	90.3	0.302	-0.0083	0.0083	95.7	0.3613
	0.1	-0.006	0.0092	89.9	0.3205	0	-0.0005	0.0086	90.8	0.3065	-0.0094	0.009	94.6	0.3575
	0.2	-0.0054	0.01	89.3	0.3237	0	-0.0031	0.0093	89.2	0.3092	-0.0102	0.0098	91.7	0.3481
0.5	-0.1	0.0002	0.0094	88.7	0.3101	0	0.0055	0.0091	88.2	0.2993	0.0016	0.0092	89.1	0.3047
	0	-0.0045	0.008	91.2	0.312	0	-0.0008	0.0078	91.9	0.3022	-0.0043	0.0078	91.4	0.3067
	0.05	-0.0071	0.0091	90.4	0.3117	0	-0.0045	0.0088	89.7	0.3022	-0.0075	0.0089	90.6	0.3065
	0.1	-0.0068	0.0082	90.9	0.3112	0	-0.0051	0.0079	90.3	0.3028	-0.0078	0.008	91.1	0.306
	0.2	-0.0033	0.0082	90.7	0.3071	0	-0.0035	0.008	90.7	0.2992	-0.0056	0.008	91.2	0.302

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

A. *n*=20



B. *n*=40



C. *n*=80

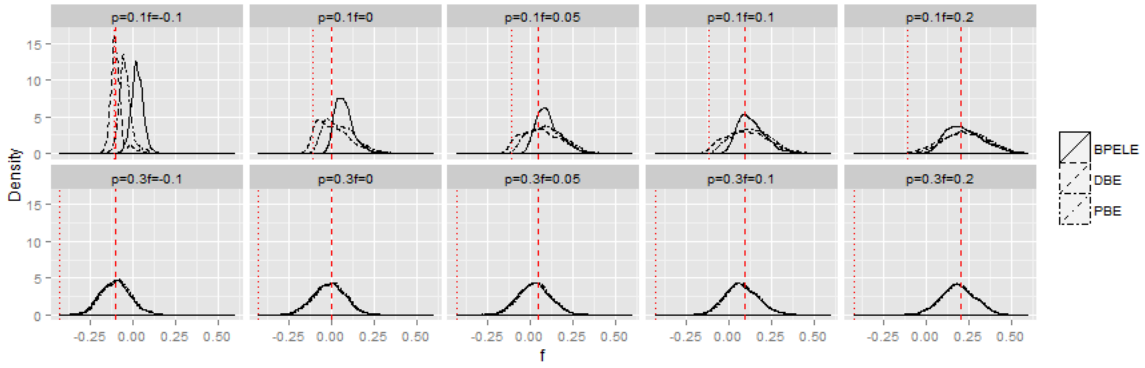


Figure 7. Density Plot of Simulated Estimators under Different Settings (Stratified Simple Random Sampling). BPELE=Bayesian Pseudo-Empirical Likelihood Estimator, DBE=Design-Based Estimator, PBE=Parametric Bayesian Estimator. Parameter p is the allele frequency and f is the inbreeding coefficient in the simulation. Panel A, B and C are for stratified sample size 20, 40 and 80 respectively. Red dashed vertical line is the true parameter f; red dotted vertical line is the lower bound of the parameter space for f conditional on allele frequency p.

3.4.2 Proportional to Population Size Sampling

For our second simulation design, we used proportional to size (PPS) sampling. To generate a size variable *X* for each individual in the population of size *N*=4,800, the number of the minor allele *A* count plus a log normal distributed random error (mean 0, standard deviation 0.2) is used. A total of 20, 40 or 80 samples are randomly selected from the population using the PPS sampling. Therefore, the sample selection probability is the sample size times the relative proportion of the size measure. The sampling weight for analysis is the inverse of the sample selection probability, $wt_i = 1/(n \times \frac{x_i}{\sum_{i=1}^{N} x_i})$, where *i* is the index of each individual in the population and *n* is the sample size. The weights from this PPS sampling design is correlated with the genotype and, therefore, is informative.

Table 11, Table 12, and Table 13 report the simulation results in comparing the design-based estimator and two Bayesian estimators for the bias, MSE, coverage probability (CP), out of range probability (ORP), and average length (AL). For the design-based estimator, the PPS sampling generates higher coverage probability than the stratified simple random sampling under large allele frequency (p=0.5); while the trend is opposite for smaller allele frequency (p=0.1, 0.3). This is likely due to the sample selection probability is correlated with the genetic data. Individuals with genotype *AA* are more likely to be selected in the PPS sampling than individuals with genotype *Aa*. Individuals with genotype *aa* have the lowest probability to be selected. Both Bayesian estimators maintain the nominal coverage probability; while the design-based estimator has lower coverage in all simulations. The design-based estimator also shows higher average length of the 95% confidence interval.

Comparing with the simple random sampling, the design-based estimator from the PPS sampling is more likely to produce result outside of the parameter space. For example, when the sample size is 40 and the inbreeding coefficient is f=0, 25.9% of the simulated design-based estimators have the values lower than -0.1111, which is the lower boundary of the parameter space for f under p=0.1 (Table 12, Figure 8 Panel B). It is about ten percent more than the simulation result from the simple random sampling.

Overall, the comparison result among three methods has similar pattern as that under the stratified simple random sampling design. The Bayesian method is recommended since it maintains nominal coverage probability and it has smaller MSE and average length of the 95% credible interval.

Table 11. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators

(Proportional to Size Sampling, n=20)

p	*f*	***Design-Based Estimator***					***Parametric Bayesian Estimator***				***Bayesian Pseudo-Empirical Likelihood Estimator***			
		Bias	***MSE***	***CP***	***AL***	***ORP***	***Bias***	***MSE***	***CP***	***AL***	***Bias***	***MSE***	***CP***	***AL***
0.1	-0.1	-0.0092	0.0068	81.7	0.2106	40	0.2719	0.0894	83.2	0.8494	0.2807	0.0896	39.3	0.718
	0	-0.0452	0.0424	11.5	0.4112	27.6	0.2071	0.0603	98.1	0.8812	0.2029	0.0554	94.5	0.7383
	0.05	-0.0727	0.0616	13	0.4653	25.9	0.1731	0.0503	98.4	0.8862	0.1658	0.044	94.6	0.7485
	0.1	-0.0929	0.0767	15.7	0.525	22.3	0.142	0.0401	98	0.8965	0.1292	0.0339	93.9	0.7512
	0.2	-0.1319	0.113	23	0.6843	14.1	0.0694	0.0272	99.5	0.9238	0.0406	0.0235	91	0.7512
0.3	-0.1	-0.0239	0.0591	64.2	0.7286	7.9	0.0648	0.0412	94.7	0.7411	0.0741	0.0323	96.9	0.7296
	0	-0.0516	0.0666	62.8	0.7735	5	0.0278	0.0406	94.9	0.7555	0.0323	0.0309	98.2	0.7474
	0.05	-0.0522	0.0703	66.8	0.8204	3.3	0.0181	0.0421	94.6	0.7662	0.0169	0.0315	97.6	0.7618
	0.1	-0.0433	0.0737	70.2	0.863	2	0.0171	0.0431	95.3	0.7777	0.0108	0.0341	97.1	0.7744
	0.2	-0.0446	0.0843	77	0.9328	0.8	-0.0065	0.0481	94.4	0.7876	-0.0168	0.041	96.3	0.7871
0.5	-0.1	-0.0248	0.056	90.8	0.8715	0	0.0268	0.0411	93	0.7608	-0.0058	0.0417	95.3	0.7824
	0	-0.0371	0.055	91.1	0.8879	0	0.0023	0.0391	94.7	0.7701	-0.0266	0.0412	95.6	0.7922
	0.05	-0.0423	0.0584	91.2	0.8953	0	-0.0096	0.0412	94.2	0.7723	-0.0381	0.0445	94.6	0.7963
	0.1	-0.032	0.0565	91.5	0.9029	0	-0.0065	0.0403	95.4	0.7742	-0.0342	0.0441	95.1	0.7967
	0.2	-0.0254	0.0554	92.6	0.9051	0	-0.0161	0.0402	95.5	0.7692	-0.0382	0.0434	95.4	0.7938

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 12. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Proportional to Size Sampling, n=40)

p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0037	0.0051	84.7	0.1533	43.3	0.1914	0.0466	78.2	0.6373	0.2282	0.058	46.9	0.6059
	0	-0.0326	0.0315	20.5	0.3509	25.9	0.1375	0.034	96.1	0.6776	0.1577	0.0342	98.6	0.6468
	0.05	-0.042	0.0452	27.6	0.4477	20.2	0.1167	0.0322	97	0.6951	0.1291	0.0286	98.9	0.666
	0.1	-0.0502	0.0549	34.6	0.5464	15.1	0.0961	0.0297	98.4	0.7095	0.1024	0.024	98.7	0.6829
	0.2	-0.072	0.0804	45.7	0.7269	6.9	0.046	0.0281	98.2	0.7372	0.042	0.0212	97.8	0.7103
0.3	-0.1	-0.0201	0.0292	79.7	0.5429	1.8	0.0307	0.0232	93.6	0.56	0.0225	0.0175	98	0.5861
	0	-0.0232	0.034	83.6	0.5912	0.8	0.019	0.0257	92.7	0.5841	0.0037	0.0209	96.6	0.6108
	0.05	-0.023	0.0357	84.8	0.6165	0.2	0.0135	0.0264	92.9	0.5957	-0.0039	0.0228	96.8	0.6215
	0.1	-0.0248	0.0368	86.9	0.6397	0	0.0048	0.0265	93	0.6058	-0.013	0.0243	96.7	0.6286
	0.2	-0.0267	0.0388	87.8	0.6824	0	-0.0078	0.0277	93.2	0.6209	-0.0286	0.0284	95	0.6367
0.5	-0.1	-0.0119	0.0269	92.3	0.6136	0	0.0129	0.0233	94.7	0.58	-0.0057	0.0238	94.3	0.5893
	0	-0.0161	0.0283	92.5	0.6236	0	0.001	0.0241	94	0.5862	-0.0146	0.0252	94.3	0.5941
	0.05	-0.0146	0.0281	93	0.628	0	-0.0017	0.0238	94.4	0.5892	-0.016	0.0252	94	0.5948
	0.1	-0.0151	0.0285	93.2	0.6299	0	-0.0048	0.0246	94.5	0.5892	-0.0188	0.0258	93.7	0.5938
	0.2	-0.0177	0.027	94.8	0.6299	0	-0.0147	0.023	95.6	0.5862	-0.0261	0.0247	95.1	0.5884

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

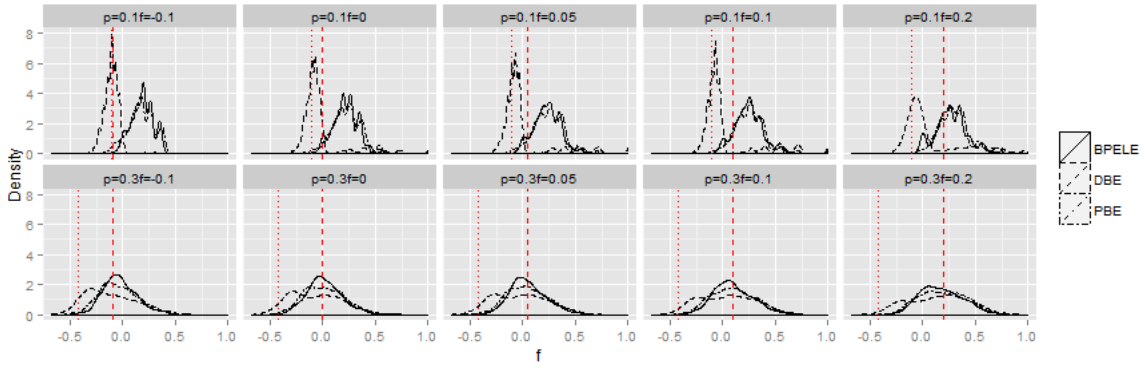
Table 13. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators

(Proportional to Size Sampling, n=80)

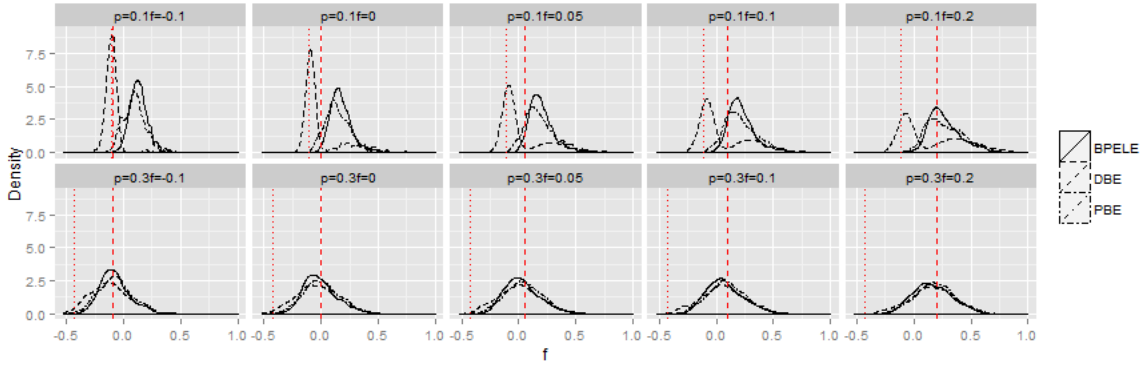
p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0025	0.0026	88.8	0.111	43.9	0.1153	0.0174	79.5	0.4331	0.1707	0.0317	50.4	0.4705
	0	-0.0145	0.0199	37.7	0.3138	18.7	0.0798	0.0177	95	0.497	0.1063	0.0167	99.5	0.5258
	0.05	-0.0112	0.028	50.6	0.4155	12.3	0.0714	0.0209	94.6	0.5266	0.0847	0.0156	98.6	0.553
	0.1	-0.0191	0.0351	56.5	0.4857	8	0.057	0.0224	95.6	0.5488	0.0635	0.0155	98	0.5717
	0.2	-0.0388	0.0458	70.1	0.6119	3.7	0.0171	0.0244	95.7	0.5832	0.0099	0.0166	98.2	0.604
0.3	-0.1	-0.0087	0.0151	86.3	0.393	0.2	0.0168	0.0133	92.4	0.4168	0.0033	0.0113	97	0.4658
	0	-0.0145	0.0166	87.9	0.424	0	0.0054	0.0142	92.6	0.437	-0.0105	0.0134	96.1	0.4809
	0.05	-0.0109	0.0163	90.2	0.4409	0	0.0063	0.0138	94.6	0.4489	-0.0102	0.0135	96.8	0.487
	0.1	-0.0144	0.0173	90	0.4567	0	-0.0004	0.0149	94.6	0.4575	-0.0176	0.0155	96	0.489
	0.2	-0.0251	0.0188	90.8	0.4782	0	-0.0175	0.016	93.5	0.4685	-0.0319	0.0173	93.8	0.4846
0.5	-0.1	-0.0049	0.0134	93.3	0.4324	0	0.0067	0.0124	95	0.4289	-0.0022	0.0127	94.9	0.4265
	0	-0.0024	0.0128	94.5	0.4395	0	0.0054	0.0118	96	0.433	-0.0021	0.0122	95.3	0.4295
	0.05	-0.007	0.0132	94.2	0.4412	0	-0.001	0.0122	94.9	0.4327	-0.008	0.0126	94.8	0.4293
	0.1	-0.0081	0.0132	93.8	0.4421	0	-0.004	0.0121	95.4	0.4334	-0.0102	0.0126	94.3	0.4283
	0.2	-0.015	0.0138	94.1	0.4412	0	-0.014	0.0128	95.4	0.4295	-0.0193	0.0133	94.3	0.4235

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

A. *n*=20



B. *n*=40



C. *n*=80

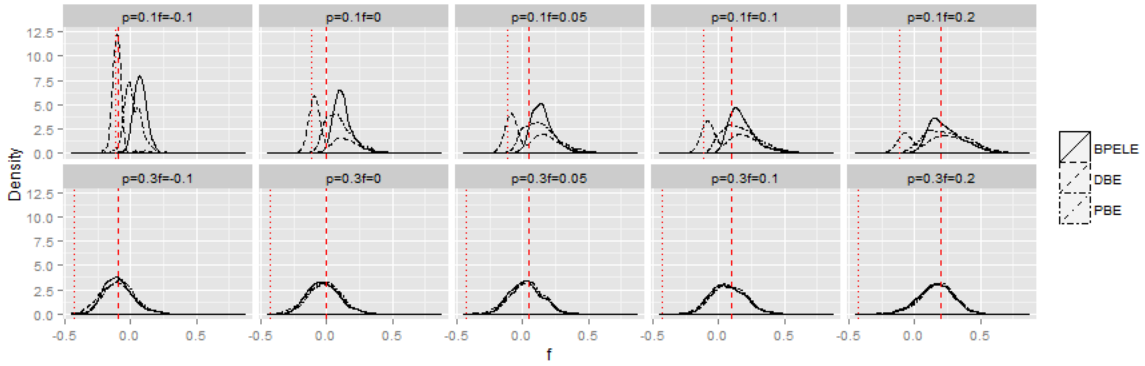


Figure 8. Density Plot of Simulated Estimators under Different Settings (Proportional to Population Size Sampling). BPELE=Bayesian Pseudo-Empirical Likelihood Estimator, DBE=Design-Based Estimator, PBE=Parametric Bayesian Estimator. Parameter p is the allele frequency and f is the inbreeding coefficient in the simulation. Panel A, B and C are for stratum sample size 20, 40 and 80 respectively. Red dashed vertical line is the true parameter f; red dotted vertical line is the lower bound of the parameter space for f conditional on allele frequency p.

3.4.3 Model Misspecification

In our above simulation studies, population genotype records were generated independently from the same multinomial model with parameters p and f. In the real world, this data simulation mechanism may be modified by factors such as population stratification or other natural selection process. To investigate the consequence of model misspecification, we simulated a population with two stratums, the genotype of individual within each stratum is simulated through different multinomial models: one with the allele frequency p=0.1 and the other with p=0.3. We studied same range of the inbreeding coefficient as -0.1, 0, 0.5, 0.1 or 0.2. At each one of 1000 iterations in our first simulation study, a sample size of twenty (20) was randomly selected from each stratum. In the second simulation study, similarly as in Section 3.4.2, a sample size of twenty (20) was randomly selected from the population using the PPS sampling.

Table 14 shows the results under the stratified simple random sampling. Although the design-based estimator shows smallest bias, it fails in all other criteria in terms of the MSE, coverage and average length. On the other hand, only BPELE maintains nominal coverage probability. It also has the smallest MSE and relative smaller average length than the design-based estimator. Table 15 shows a similar pattern when comparing estimators under the PPS sampling. Under the PPS sampling, the coverage probability for the design-based estimator gets improved than the simple random sampling. However, the MSE and average interval length are still less optimal when comparing to the Bayesian estimators.

Two observations worth mention from the two simulation studies. First, when there is population subdivision that results in different allele frequencies among subpopulations, ignoring the population subdivision may increase the variance thus a larger confidence interval for the design-based estimator. Such problem is more serious in the PPS sampling design when the weights are correlated with the genetic data. Second, the Bayesian methods show more resistance to the model misspecification under both sampling designs, especially for the Bayesian pseudo-empirical likelihood estimator. Ignoring the population subdivision while using the average allele frequency for the whole population does not have a large negative impact on the Bayesian estimators in terms of the MSE, coverage probability and average interval length.

Table 14. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Non-Homogenous Population, Stratified Simple Random Sampling，n=20 per Strata)

f	***Design-Based Estimator***				***Parametric Bayesian Estimator***				***Bayesian Pseudo-Empirical Likelihood Estimator***			
	Bias	***MSE***	***CP***	***AL***	***Bias***	***MSE***	***CP***	***AL***	***Bias***	***MSE***	***CP***	***AL***
-0.1	0.0063	0.0342	74.6	0.5611	0.0612	0.0299	90.8	0.5632	0.0599	0.0229	96.1	0.5874
0	0.0179	0.0416	82.5	0.6568	0.0567	0.0348	89.2	0.587	0.0445	0.0276	94.5	0.6226
0.05	0.0041	0.0421	85.6	0.6809	0.0374	0.0336	90.1	0.5913	0.0233	0.0276	94.5	0.6311
0.1	0.0042	0.0432	85.8	0.7007	0.0314	0.0339	90.4	0.6001	0.0155	0.0288	94.2	0.6395
0.2	0.0004	0.0458	87.6	0.7368	0.0146	0.0349	89.5	0.6061	-0.0033	0.0324	92.8	0.6489

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 15. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Non-Homogenous Population, Proportional to Size Sampling，n=20)

f	***Design-Based Estimator***				***Parametric Bayesian Estimator***				***Bayesian Pseudo-Empirical Likelihood Estimator***			
	Bias	***MSE***	***CP***	***AL***	***Bias***	***MSE***	***CP***	***AL***	***Bias***	***MSE***	***CP***	***AL***
-0.1	0.022	0.0363	92.9	0.8428	0.1031	0.0342	94.4	0.7144	0.0975	0.0292	99.5	0.7552
0	0.033	0.0425	89.6	0.8912	0.0855	0.0357	94.6	0.7201	0.0711	0.0273	99.2	0.7982
0.05	0.0204	0.0418	91.8	0.903	0.0648	0.0329	96.6	0.7228	0.0466	0.0247	99	0.8108
0.1	0.0214	0.0429	93.3	0.9144	0.0573	0.0326	96.7	0.7224	0.0344	0.0249	98.9	0.8218
0.2	0.0185	0.0425	94	0.9139	0.0367	0.0305	97.1	0.7222	0.0075	0.0247	99.2	0.8374

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Chapter 4: Estimation of the Inbreeding Coefficient with Family Level Correlation

National level surveys such as the NHANES sometimes collect family level genetic samples. Among these surveys, the genetic correlation is induced by offspring's genetic inheritance from their biological parents within the household. Besides the differential population weights introduced by the sampling design, within household correlations further inflate the variance of estimates for the genetic traits. Methods developed by She (2009), Li (2011), and Li (2013) were the first frequentist approaches to test HWE considering within household genetic correlations among family members. For this chapter, we are trying to extend the Bayesian method to estimate the inbreeding coefficient when the survey collects family level genetic data.

Inspired by the NHANES survey design, let's consider a case with genetic data collected from families with both parents and one child (2P1O). As shown in Table 17, for a diallelic locus (i.e., *A* and *a*) of autosomal genes, there are a total of ten (*G*=10) possible genotype combinations for a sampled family. The mother and father's genotype frequencies are independent following the general inbreeding model. For genes at normal chromosome, child will get one copy of gene from each parent. Therefore for genes located at the autosomal chromosome, switching mother and father's genotypes generally will not affect the gene flow to the next generation. In another words, two parents' genotypes: genotype *AA* for the mother and *aa* for the father, or, genotype *aa* for the mother and *AA* for the father, have their child carrying the same genotype *Aa*. For a diallelic locus, there are six different types of genotype distribution considering the mother and father together: *AA-AA*， *AA-Aa*, *AA-aa*, *Aa-Aa*, *Aa-aa*, or *aa-aa*. Conditional on the parents' genotypes, the probability of child's genotype follows the Mendel's law (Table 16).

Let $\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10})$ be the joint probability vector of each parents-child genotypes combination. It can be derived as the product of the parent genotype probability and the child genotype probability conditional on his/her parents. Table 17 also lists the number of the allele A ($numA_g$) and the number of the heterozygotes *Aa* ($numH_g$) for the *g*-th family for *g*=1,…,10.

Table 16. Child Genotype Probability Conditional on the Parents' Genotypes (Mendel's Law)

Parents' Genotypes	Child Genotype		
	AA	***Aa***	***aa***
AA-AA	1	-	-
AA-Aa	1/2	1/2	-
AA-aa	-	1	-
Aa-Aa	1/4	1/2	1/4
Aa-aa	-	1/2	1/2
aa-aa	-	-	1

Table 17. Family Genotype Distribution for Both Parents One Child Triads (2P1O)

Parent Genotype	Child Genotype	Parent Genotype Probability	Conditional Child Genotype Probability (Mendel's Law)	Joint Probability (q_g)	# Allele A ($numA_g$)	# Heterozygotes ($numH_g$)
AA - AA	*AA*	p_{11}^2	1	$q_1 = p_{11}^2$	6	0
AA - Aa	*AA*	$2p_{11}p_{12}$	½	$q_2 = p_{11}p_{12}$	5	1
AA - Aa	*Aa*	$2p_{11}p_{12}$	½	$q_3 = p_{11}p_{12}$	4	2
AA - aa	*Aa*	$2p_{11}p_{22}$	1	$q_4 = 2p_{11}p_{22}$	3	1
Aa - Aa	*AA*	$p_{12}p_{12}$	¼	$q_5 = p_{12}^2/4$	4	2
Aa - Aa	*Aa*	$p_{12}p_{12}$	½	$q_6 = p_{12}^2/2$	3	3
Aa - Aa	*aa*	$p_{12}p_{12}$	¼	$q_7 = p_{12}^2/4$	2	2
Aa - aa	*Aa*	$2p_{12}p_{22}$	½	$q_8 = p_{12}p_{22}$	2	2
Aa - aa	*aa*	$2p_{12}p_{22}$	½	$q_9 = p_{12}p_{22}$	1	1
aa - aa	*aa*	p_{22}^2	1	$q_{10} = p_{22}^2$	0	0

Similar as the individual genotype counts that follow a multinomial distribution, the family genotype counts also follow a multinomial distribution with probability vector $\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10})$. It is easy to verify that the total joint probability of the family genotype is

$$\sum_{g=1}^{10} q_g = p_{11}^2 + 2p_{11}p_{12} + 2p_{11}p_{22} + 2p_{12}p_{22} + p_{12}^2 + p_{22}^2 = (p_{11}+p_{12} + p_{22})^2 = 1.$$

Furthermore, the expected number of the allele *A* count and the expected number of the heterozygote *Aa* count for the family are, respectively,

$$E(\#Allele\ A) = \sum_{g=1}^{10} q_g numA_g$$

$$= 6p_{11}^2 + 9p_{11}p_{12} + 6p_{11}p_{22} + 3p_{12}p_{22} + 3p_{12}^2$$

$$= 6p_{11}(p_{11}+p_{12} + p_{22}) + 3p_{12}(p_{11}+p_{12} + p_{22})$$

$$= 6p^2 + 6p(1-p)f + 6(1-f)p(1-p) = 6p, \text{ and}$$

$$E(\#\ \text{Heterozygotes}) = \sum_{g=1}^{10} q_g numH_g$$

$$= 3p_{11}p_{12} + 2p_{11}p_{22} + 3p_{12}p_{22} + \frac{5}{2}p_{12}^2$$

$$= 6p(1-p) - 4fp(1-p).$$

The expected number of the allele *A* for a family using the familywise 2P1O genotype data is 6*p*. It is three times of 2*p*, the expected number of the allele *A* for an individual using the individual genotype data. Treating the 2P1O as if they are independent individuals and ignoring the within family correlation, provides an unbiased estimator of the allele frequency. However, the expected number of the heterozygotes for a 2P1O family is $6p(1-p) - 4fp(1-p) = 2(3-2f)p(1-p)$, which is different from the expected number of the heterozygotes, $2(1-f)p(1-p)$ among individuals without genetic correlations. Ignoring the within family genetic correlation will produce a biased estimator of the inbreeding coefficient.

Under a simple random sampling setting, suppose there are 2P1O families of size of *n*. Let *totalH* be the total number of the heterozygotes, *totalA* be the total number of the allele *A* from these *n* families (3*n* individuals). Let $\hat{p}$ be the estimator of the allele frequency. Based on the above calculation, the estimator of the inbreeding coefficient with respect to the random sampling of families ($\hat{f}_{\text{family}}$) as well as the estimator ignoring the within family correlations ($\hat{f}_{\text{individual}}$) are

$$\hat{f}_{\text{family}} = \frac{3}{2} - \frac{\frac{totalH}{n}}{2 \times 2\hat{p}(1-\hat{p})} = \frac{1}{2}\left[3 - \frac{totalH}{n \times 2\hat{p}(1-\hat{p})}\right],$$

$$\hat{f}_{\text{individual}} = 1 - \frac{totalH}{2 \times (3n) \times \hat{p}(1-\hat{p})} = \frac{1}{3}\left[3 - \frac{totalH}{n \times 2\hat{p}(1-\hat{p})}\right],$$

where $\hat{p} = \frac{totalA}{6\times n} = \frac{totalA}{2\times 3n} = \frac{totalA}{6n}$.

From above calculation, ignoring the within-family sampling design results in a 1/3 decrease of the measure of *f*. As we discussed in Chapter 1, the inbreeding coefficient is a correlation measure between two homologous genes in uniting gametes. Therefore, considering the within family correlation inflates the inbreeding coefficient (correlation) estimate. Although the final estimator could be corrected through this factor under our simplest scenario with only 2P1O families in the population, the justification for the standard error calculation is not trivial even under the simple random sampling.

With the introduction of different sample selection probability as well as within family genotype correlation, the discussion in this chapter fills the gap in estimating the inbreeding coefficient from complex survey which collects family level genotype data.

4.1 Direct Design-Based Estimator

Assume a finite population consists of *N* families, let G_i (i=1, 2, …, N) be the 2P1O family genotypes, $numA_i$ be the number of the allele *A* and $numH_i$ be the number of the heterozygotes for the i^{th} family in the population. If this finite population is assumed to be selected from a superpopulation with family genotype frequencies described in Table 17, the finite population mean of the number of the allele *A* is therefore $\overline{numA} = \frac{1}{N}\sum_{i=1}^{N} numA_i = 6p$. Similarly, the finite population mean of the number of the heterozygotes is $\overline{numH} = \frac{1}{N}\sum_{i=1}^{N} numH_i = 6p(1-p) - 4fp(1-p)$. The finite population parameter *f* can be viewed as a function of the finite population averages of the familywise allele *A* count ($\overline{numA}$) and heterozygotes count ($\overline{numH}$):

$$f = \frac{3}{2} - \frac{9 * \overline{numH}}{\overline{numA} * (3 - \overline{numA})}.$$

Under unequal probability sampling, each sampled families of size *n* has a design weight d_i, which is the inverse of the family unit sampling inclusion probability. From survey like the NHANES, the family level weights can be constructed by taking the average of the remaining weights (equal

to the final sample weight divided by the within-family weight) of 2 parents and 1 child in each of the sampled families (Korn & Graubard, 2003). The design-based direct survey estimator for the finite population inbreeding coefficient can be expressed as

$$\hat{f} = \frac{3}{2} - \frac{9 * \widehat{numH}}{\widehat{numA} * \left(3 - \widehat{numA}\right)},$$

where $\widehat{numH}$ and $\widehat{numA}$ are Horvitz–Thompson type estimators of population means. The standard error of the mean can be estimated from jackknife method as described earlier.

4.2 Parametric Bayesian Estimator with Survey Weight

Again, let $\tilde{d}_i = \frac{d_i}{\sum_{i \in s} d_i}$ be the normalized family sampling weight and is assumed to be independent of the observation of family genotype. We incorporate the complex sampling design through the effective sample size and effective family genotype counts in the Bayesian multinomial model:

$$Pr(\boldsymbol{n}^*|\boldsymbol{q}) = Pr\left((n_1{}^*, n_2{}^*, \dots, n_{10}{}^*)\middle|(q_1, q_2, \dots, q_{10})\right) = \frac{n^*!}{\prod_{g=1}^{10} n_g{}^*} \prod_{g=1}^{10} q_g{}^{n_g{}^*},$$

where $n_g{}^* = n^* \sum_{i=1}^{n} \tilde{d}_i \, I_{[G_i=g]}$. The design-based variance is calculated as $Var_{\text{design}}(\widehat{numA})$. Under simple random sampling without replacement, the variance is $Var_{\text{srswor}}(\widehat{numA})$. The design effect is Deff $= Var_{\text{design}}(\widehat{numA}) / Var_{\text{srswor}}(\widehat{numA})$ and therefore the effective sample size is $n^* = n/\text{Deff}$. Same as in Chapter 3, we use a non-informative proper uniform prior for the allele frequency *p*. Conditional on *p*, we apply a uniform prior on its parameter space for the inbreeding coefficient *f*. The OPENBUGS Bayesian model is listed below:

Level 1: $\boldsymbol{n}^* \sim Multinomial(n^*, \boldsymbol{q})$, $\boldsymbol{n}^* = (n_1{}^*, n_2{}^*, \dots, n_{10}{}^*)$, $\boldsymbol{q} = (q_1, q_2, \dots, q_{10})$,

Reparametrize $\boldsymbol{q}$ as a function of $\tilde{\boldsymbol{p}} = (p_{11}, p_{12}, p_{22})$ shown in Table 17;

$$\begin{cases} p_{11} = p^2 + p(1-p)f, \\ p_{12} = 2p(1-p)(1-f), \\ p_{22} = (1-p)^2 + p(1-p)f. \end{cases}$$

Level 2: $p \sim Uniform[0, 1]$,

$$f|p \sim Uniform\left[\frac{-\min(p,1-p)}{1-\min(p,1-p)}, 1\right].$$

```
model=function(){
    q11 <- (1 - f) * p * p + f * p
    q12 <- 2 * (1 - f) * p * (1 - p)
    q22 <- (1 - f) * (1 - p) * (1 - p) + f * (1 - p)
    q[1] <- q11 * q11
    q[2] <- q11 * q12
    q[3] <- q11 * q12
    q[4] <- 2 * q11 * q22
    q[5] <- q12 * q12/4
    q[6] <- q12 * q12/2
    q[7] <- q12 * q12/4
    q[8] <- q12 * q22
    q[9] <- q12 * q22
    q[10] <- q22 * q22
    y[ ] ~ dmulti(q[ ], n)
    p ~ dunif(0, 1)
    f <- w * (1 - f.min) + f.min
    f.min <- max(-p/(1 - p), -(1 - p)/p)
    w ~ dunif(0, 1)
    n <- sum(y[ ])
}
```

Alternatively by substituting the joint probability q_g (g=1,…,10) with (p_{11}, p_{12}, p_{22}) as shown in Table 17, the likelihood is then simplified as

$$Pr(\boldsymbol{n}^*|\boldsymbol{p}) = \frac{n^*!}{\prod_{g=1}^{10} {n_g}^*!} \prod_{g=1}^{10} {q_g}^{n_g^*}$$

$$\propto {p_{11}}^{2n_1^*+n_2^*+n_3^*+n_4^*} {p_{12}}^{n_2^*+n_3^*+2(n_5^*+n_6^*+n_7^*)+n_8^*+n_9^*} {p_{22}}^{2n_{10}^*+n_9^*+n_8^*+n_4^*}.$$

Therefore, the data can be modelled through the multinomial with parameter $\boldsymbol{p} = (p_{11}, p_{12}, p_{22})$ and the consolidated effective number of genotype count. The total count of $n_2^* + n_3^* + 2(n_5^* + n_6^* + n_7^*) + n_8^* + n_9^*$ is the number of parents with heterozygotes *Aa*; the total count of $2n_1^* + n_2^* + n_3^* + n_4^*$ is the number of parents with genotype *AA*; and the total count of $2n_{10}^* + n_9^* + n_8^* + n_4^*$ is the number of parents with genotype *aa*. In another words, the child genotype is not included in the above simplified multinomial likelihood. This is consistent with

the biological fact that the child genotype is produced under the Mendel's law conditioning on the parents' genotypes. Therefore, no additional parameters need be included in the hierarchical Bayesian model. Accordingly, we expect that analyzing the whole family genotype data using the method discussed in this chapter will produce a similar result as analyzing a subset of the data that only includes parents' genotypes.

4.3 Bayesian Pseudo-Empirical Likelihood Estimator

Let $l_{PEL} = n\sum_{i\in s}\widetilde{d}_i \log(\pi_i)$ be the pseudo empirical log-likelihood for the above design, where *S* is a randomly selected families. In order to maximize the likelihood l_{PEL} with the following constraints on the total probability and the unbiased estimating function:

$$\pi_i > 0, \sum_{i\in s}\pi_i = 1, \sum_{i\in s}\pi_i[numH_i - (6p(1-p) - 4fp(1-p))] = 0.$$

Lagrange method can therefore be used. Again, a survey design-based unbiased estimator $\hat{p}$ of the allele frequency is used to replace *p*. BPELE is derived by maximizing the profiled pseudo-empirical log-likelihood with modified constraints:

$$\pi_i > 0, \sum_{i\in s}\pi_i = 1, \sum_{i\in s}\pi_i[numH_i - (6\hat{p}(1-\hat{p}) - 4f\hat{p}(1-\hat{p}))] = 0.$$

Similarly as in Chapter 3, define $u_i = numH_i - (6\hat{p}(1-\hat{p}) - 4f\hat{p}(1-\hat{p}))$ and thus $\sum_{i\in s}\pi_i u_i = 0$. Lagrange multipliers and grid search method are used to derive the BPELE from the posterior mean.

4.4 Simulation Study

A simulation study for a single biallelic locus with alleles *A* and *a* is conducted to evaluate the performance of the above three estimators. Let a finite population of size *N*= 4,800 families of each two parents and one child (2P1O) consisting of two stratum with subpopulation size ratio of 5:1. For each family, the genotypes for the mother and the father are generated independently from the multinomial model with frequencies of $P(aa) = (1-p)^2 + p(1-p)f$, $P(Aa) = 2(1-f)p(1-p)$, and $P(AA) = p^2 + p(1-p)f$. Conditional on the parents' genotypes within a randomly generated family, the child's genotype is generated under the Mendel's law.

The parameters for the simulation are varied with the allele A frequency p=0.1, 0.3, or 0.5 and the inbreeding coefficient f=-0.1, 0, 0.05, 0.1, or 0.2. For the realized data at each simulation, estimators of the inbreeding coefficient are calculated by the methods described above. The bias, variance, mean square error (MSE) of the estimators are calculated similarly as in Chapter 2, except that the true population parameter in computing the bias and the MSE is based on the finite population of size N=4,800 families. When the number of sampled families is 20, there is small chance that all sampled families have the same family genotypes, which is replaced by extra simulation runs.

4.4.1 Stratified Simple Random Sampling

At each of 1000 iterations, a same sample size of twenty (20), forty (40) or eighty (80) families are randomly selected from each stratum. Therefore, the weights, computed as the inverse of selection probabilities, are different across the two strata.

Table 18, Table 19, Table 20, and Figure 9 report the simulation results to compare the design-based estimator and two Bayesian estimators for the bias, MSE, coverage probability (CP), out of range probability (ORP), and average interval length (AL). The design-based estimator generally has a negative bias. Both Bayesian estimators tend to have negative bias expect when the allele frequency is small (p=0.1). As in Chapter 3, when the allele frequency is small (p=0.1), the Bayesian methods have larger bias especially when the inbreeding coefficient is small (f=-0.1, 0). In general for all simulation scenario, the parametric Bayesian estimator has the smallest MSE when the sample size is small. When sample size is larger, all methods have relative similar MSE. For coverage, the BPELE maintains the best coverage but with larger average interval length in all three simulations. When the allele frequency is p=0.5, the BPELE does not maintain the nominal coverage probability. As expected, the design-based estimator is likely to lay outside of the parameter space, especially when the allele frequency is small (p=0.1). Increasing the sample size slightly reduces the out of range possibility. The situation gets even worse when the actual inbreeding coefficient is small (f=-0.1, 0).

Table 18. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Stratified Simple Random Sampling of Families，n=20 Families per Strata)

p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0071	0.0107	83	0.2778	63.6	0.1025	0.0139	74.4	0.3792	0.2578	0.0723	40.8	0.6836
	0	-0.0102	0.0275	65.8	0.4761	31.7	0.0716	0.0184	91.7	0.4391	0.1883	0.0435	99.3	0.7384
	0.05	-0.0252	0.0351	78.3	0.5975	17.3	0.037	0.0211	93.9	0.462	0.1283	0.0258	99.6	0.7654
	0.1	-0.0222	0.0338	76.4	0.5878	18.2	0.0495	0.0214	93.3	0.4713	0.1377	0.0278	100	0.7663
	0.2	-0.0284	0.0467	84.8	0.7281	6.5	0.0028	0.024	92	0.5081	0.0665	0.0183	99.4	0.8129
0.3	-0.1	-0.0109	0.0246	88.4	0.5158	2.8	0.0275	0.0152	89.5	0.3921	0.0281	0.0166	97.8	0.5749
	0	-0.0162	0.0237	91.3	0.5378	0.6	0.0109	0.0155	89.5	0.4063	0.0004	0.0177	97.3	0.6147
	0.05	-0.0163	0.0293	88	0.5518	0.3	-0.0095	0.0177	89	0.418	-0.008	0.0235	95.9	0.6305
	0.1	-0.0126	0.0271	89.2	0.5496	0	0.0051	0.0171	88.8	0.4162	-0.0076	0.0224	95.6	0.6313
	0.2	-0.0108	0.0296	88.9	0.5813	0	-0.0069	0.0182	88.7	0.4178	-0.0146	0.0261	95.2	0.6479
0.5	-0.1	-0.0199	0.0276	89.4	0.5506	0	0.0035	0.0178	87.1	0.4096	-0.008	0.0289	89.7	0.5459
	0	-0.0226	0.0268	90.5	0.5568	0	-0.0076	0.0174	89	0.4189	-0.0188	0.0261	90.6	0.5482
	0.05	-0.0147	0.0288	89.4	0.5504	0	-0.0069	0.0174	89.6	0.4175	-0.0193	0.0284	90	0.5468
	0.1	-0.0174	0.0276	89.4	0.5426	0	-0.0002	0.0159	90.9	0.4189	-0.0239	0.0277	90.6	0.5474
	0.2	-0.0096	0.0258	91	0.5415	0	-0.0023	0.017	88.8	0.4083	-0.0351	0.0318	91.2	0.5604

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 19. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Stratified Simple Random Sampling of Families，n=40 Families per Strata)

p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0018	0.0056	79.7	0.2175	56.5	0.0573	0.0047	80.4	0.2334	0.1887	0.0378	44.4	0.5137
	0	-0.0152	0.0122	81.2	0.3485	20.3	0.0369	0.009	93.5	0.3079	0.1214	0.0181	100	0.5775
	0.05	-0.0113	0.0174	86.4	0.4459	5	0.0142	0.0115	91.5	0.3435	0.0748	0.0114	99.9	0.6216
	0.1	-0.0107	0.0162	87.2	0.4377	5.3	0.026	0.012	91.1	0.3496	0.079	0.0114	99.9	0.6179
	0.2	-0.0187	0.0245	87.9	0.5153	0.7	-0.0122	0.0154	88.8	0.3886	0.0232	0.0113	99.4	0.6756
0.3	-0.1	0.0008	0.0116	89	0.3626	0	0.021	0.0073	90.2	0.2862	0.0113	0.0094	98.5	0.4532
	0	-0.0089	0.0132	89.7	0.3778	0	0.0087	0.0086	87.9	0.296	-0.007	0.012	96.7	0.4699
	0.05	-0.008	0.0133	89.9	0.388	0	-0.0067	0.0088	89.5	0.3064	-0.009	0.0125	96.5	0.4724
	0.1	-0.0057	0.0139	89.2	0.3857	0	0.0088	0.0087	89.2	0.3051	-0.0074	0.0133	94.9	0.4694
	0.2	-0.0133	0.0149	89.5	0.4065	0	-0.003	0.0095	88.7	0.3042	-0.017	0.0146	95.1	0.468
0.5	-0.1	-0.0103	0.013	91	0.3875	0	0.0083	0.0085	89.5	0.2973	-0.009	0.0127	90.1	0.3791
	0	-0.0054	0.0136	90	0.3928	0	-0.0025	0.0086	89.4	0.3045	-0.0055	0.0133	89.9	0.3854
	0.05	-0.0116	0.0137	90.1	0.3891	0	-0.01	0.0093	88.4	0.3023	-0.012	0.0135	89.7	0.3862
	0.1	-0.0093	0.0126	90.8	0.3799	0	0.0038	0.0085	89.9	0.303	-0.0099	0.0124	91.5	0.3776
	0.2	-0.0156	0.0132	89.2	0.38	0	-0.0068	0.0089	88.9	0.2946	-0.0171	0.013	89.5	0.372

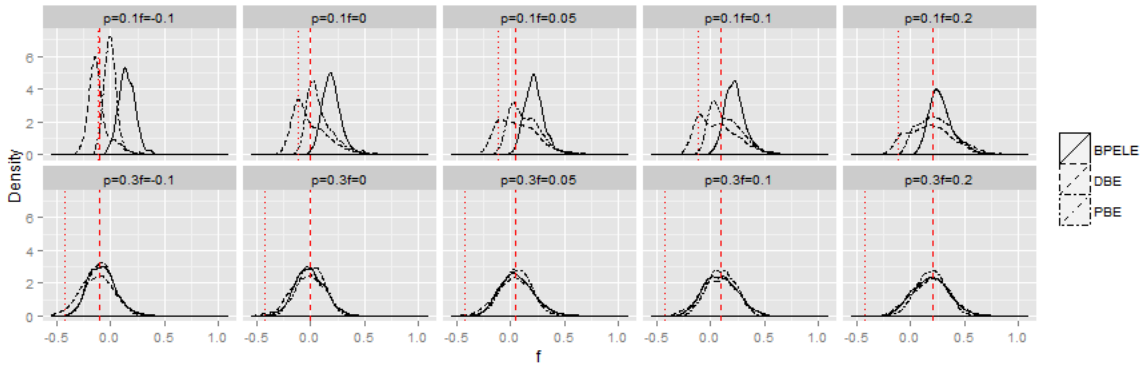
* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 20. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators (Stratified Simple Random Sampling of Families，n=80 Families per Strata)

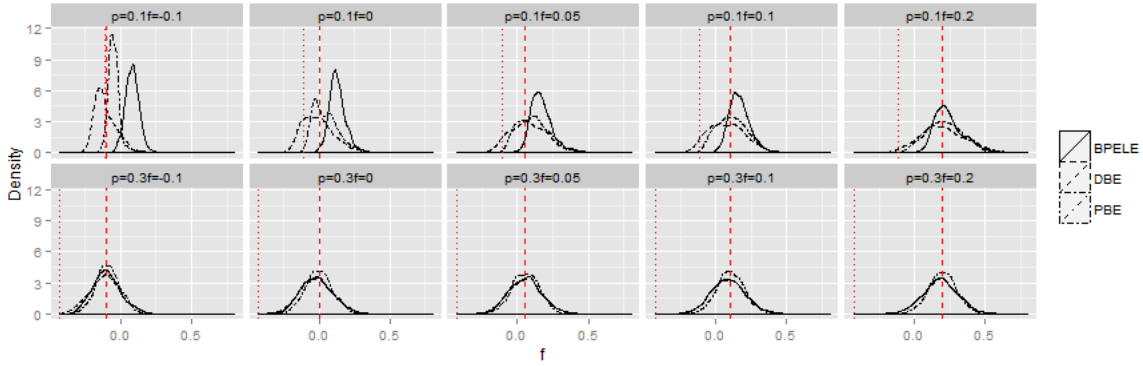
p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0001	0.0029	82.1	0.1599	51.2	0.0322	0.0017	81.7	0.1373	0.1379	0.02	44.1	0.3793
	0	-0.0046	0.0066	86.3	0.2571	7.8	0.0188	0.0044	91.9	0.2203	0.0738	0.0074	100	0.449
	0.05	-0.0027	0.0096	88.2	0.3229	0.4	0.0017	0.0067	87.2	0.2517	0.0369	0.0057	100	0.4988
	0.1	0.0008	0.0085	90	0.3171	0.7	0.0186	0.0062	91.1	0.2606	0.0413	0.0054	99.8	0.4967
	0.2	-0.0059	0.012	91.5	0.3705	0.1	-0.014	0.0085	89.1	0.294	0.0016	0.0078	99.4	0.5556
0.3	-0.1	-0.0037	0.0059	89.7	0.2568	0	0.0128	0.004	89.3	0.2041	-0.0021	0.0054	97.7	0.3418
	0	-0.0038	0.0068	87.8	0.2654	0	0.0089	0.0046	88.7	0.2129	-0.0044	0.0067	96.7	0.3395
	0.05	-0.006	0.0065	89.3	0.271	0	-0.0099	0.0043	89.7	0.2206	-0.0073	0.0065	96	0.3382
	0.1	-0.0015	0.0069	89.5	0.2715	0	0.0083	0.0047	89.1	0.2198	-0.0027	0.0068	95.5	0.3332
	0.2	-0.0084	0.0076	88.8	0.2885	0	-0.0038	0.005	87.1	0.2184	-0.0104	0.0075	93.8	0.3338
0.5	-0.1	-0.0038	0.0064	91.1	0.2729	0	0.0066	0.0046	89.1	0.2128	-0.0032	0.0063	90.6	0.2677
	0	-0.0058	0.0071	89.8	0.2773	0	-0.0047	0.0045	89.2	0.2179	-0.0059	0.0071	90	0.2736
	0.05	-0.0033	0.0068	89.7	0.2727	0	-0.0071	0.0044	89.3	0.2168	-0.0035	0.0068	90.1	0.272
	0.1	0.0004	0.0064	90.4	0.2685	0	0.0111	0.0045	89	0.2174	0.0001	0.0063	90.8	0.2682
	0.2	-0.0013	0.0063	90.5	0.2673	0	0.0013	0.0042	89.5	0.2109	-0.0022	0.0062	90.1	0.2638

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

A. *n*=20 families



B. *n*=40 families



C. *n*=80 families

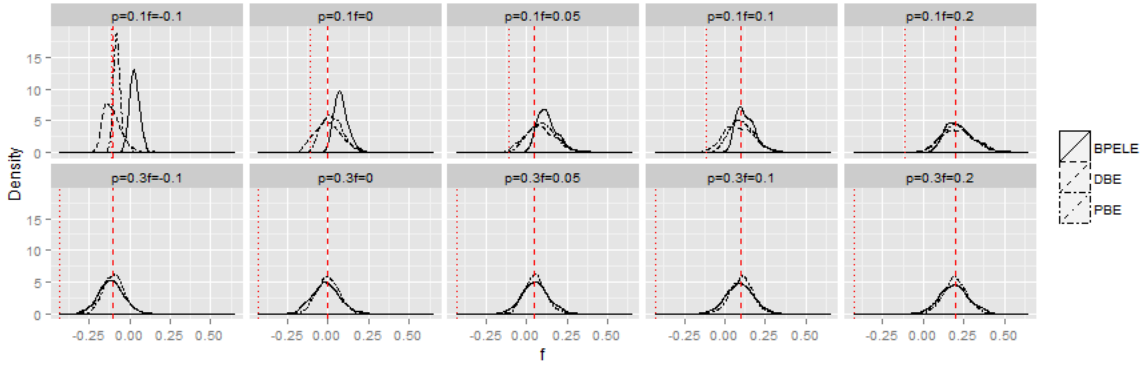


Figure 9. Density Plot of Simulated Estimators under Different Settings (Stratified Simple Random Sampling). BPELE=Bayesian Pseudo-Empirical Likelihood Estimator, DBE=Design-Based Estimator, PBE=Parametric Bayesian Estimator. Parameter p is the allele frequency and f is the inbreeding coefficient in the simulation. Panel A, B and C are for stratum family sample size 20, 40 and 80 families respectively. Red dashed vertical line is the true parameter f; red dotted vertical line is the lower bound of the parameter space for f conditional on allele frequency p.

4.4.2 Proportional to Population Size Sampling

For the second simulation study, we used proportional to size (PPS) sampling. For each family, the size variable X is generated by the number of the allele A count plus a log normal distributed random error (mean 0, standard deviation 0.2). A total of 20, 40 or 80 families are randomly selected from this population using the PPS sampling. Therefore, the sample selection probability is the sample size times the relative proportion of the size measure. The sampling weight for analysis is the inverse of the sample selection probability, $wt_i = 1/(n \times \frac{x_i}{\sum_{i=1}^{N} x_i})$, where i is the index of each family in the population and n is the sample size. The sampling weight from this PPS sampling design depends on the number of the allele A, therefore, is informative.

Table 21, Table 22, Table 23 and Figure 10 report the simulation results to compare design-based estimator and two Bayesian estimators for the bias, MSE, coverage probability (CP), out of range probability (ORP), and average interval length (AL). The design-based estimator generally has a negative bias, while both Bayesian methods tend to have negative bias when the allele frequency is medium or large (p=0.3, 0.5). When the allele frequency is small (p=0.1), the Bayesian methods have larger bias especially when the inbreeding coefficient is small (f=-0.1, 0). On average from all simulation scenario, the parametric Bayesian estimator has the smallest MSE. For coverage, the design-based estimator does not maintain the nominal level. The BPELE maintains the coverage closest to the nominal level, but with largest average interval length in all three simulations. The design-based estimator is likely to lay outside of the parameter space when the allele frequency is small (p=0.1, 0.3). The situation gets worse when the actual inbreeding coefficient is small (f=-0.1, 0). Increasing the sample size slightly reduces the out of range possibility.

Table 21. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators

(Proportional to Size Sampling, n=20 Families)

p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0179	0.018	84.3	0.2789	68.8	0.1945	0.0456	75.6	0.6294	0.3402	0.1238	37	0.8548
	0	-0.0514	0.0428	25.9	0.4569	49.7	0.1376	0.0339	96.5	0.665	0.2579	0.0763	98.5	0.8826
	0.05	-0.0677	0.0594	38.2	0.5993	37.1	0.0952	0.0268	97.7	0.689	0.196	0.0487	98.8	0.9029
	0.1	-0.0505	0.0631	39.1	0.6204	37.9	0.1069	0.0325	97	0.6871	0.2042	0.0529	99.2	0.9011
	0.2	-0.0919	0.0912	49.9	0.7968	24.7	0.0301	0.0257	97.9	0.7147	0.1122	0.0269	98.8	0.919
0.3	-0.1	-0.0394	0.0464	84	0.6649	9.6	0.0394	0.0241	94.2	0.5634	0.0659	0.0239	97.8	0.7266
	0	-0.0265	0.0522	86.2	0.7144	2.2	0.0276	0.0284	91.8	0.5893	0.0357	0.0262	97	0.7807
	0.05	-0.0428	0.06	86.1	0.7407	2.3	0.002	0.0294	90.3	0.6015	0.0057	0.0275	97.9	0.8006
	0.1	-0.0426	0.0616	85.5	0.7577	0.8	0.008	0.0289	93.1	0.6067	-0.0028	0.0315	96.9	0.807
	0.2	-0.0604	0.0709	83.7	0.7828	0.5	-0.0203	0.0325	91.7	0.617	-0.0438	0.0383	96.6	0.8425
0.5	-0.1	-0.0395	0.0518	87.5	0.725	0	0.0105	0.0268	93.4	0.586	0.056	0.0599	97	0.8912
	0	-0.035	0.0557	87.5	0.7402	0	-0.0128	0.0273	93.3	0.5971	-0.0355	0.049	94.2	0.8253
	0.05	-0.0309	0.0547	86.9	0.746	0	-0.0171	0.0284	92.8	0.5987	-0.0426	0.0494	94.8	0.8263
	0.1	-0.0418	0.0562	87.8	0.7305	0	-0.0045	0.0268	93.3	0.5975	-0.0565	0.0512	95	0.8452
	0.2	-0.039	0.055	87	0.7182	0	-0.0225	0.0276	93.2	0.5932	-0.113	0.063	94.2	0.8908

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

Table 22. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators

(Proportional to Size Sampling, n=40 Families)

p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0115	0.0116	76.3	0.2279	69.1	0.122	0.0185	76.9	0.4344	0.2583	0.0715	39.6	0.6818
	0	-0.0271	0.0287	46.3	0.406	39.6	0.0824	0.0175	95.8	0.495	0.1867	0.042	98.7	0.7321
	0.05	-0.038	0.0374	61.2	0.519	23.5	0.0472	0.0166	97.4	0.5294	0.1317	0.0258	99.4	0.7641
	0.1	-0.0309	0.0365	62.4	0.5292	25.1	0.0629	0.0203	96.4	0.5307	0.1386	0.0278	99.2	0.7642
	0.2	-0.0489	0.0573	72.4	0.6679	12.7	0.0072	0.021	97.8	0.5747	0.068	0.0196	98.2	0.8044
0.3	-0.1	-0.0151	0.0243	86.3	0.475	2.2	0.0256	0.0144	92.8	0.4214	0.0298	0.015	97.5	0.5697
	0	-0.0313	0.0307	84.5	0.5029	0.6	0.0062	0.0162	91.8	0.4428	-0.0053	0.0204	96.8	0.6044
	0.05	-0.0249	0.0342	83.1	0.5242	0.2	-0.0111	0.0179	91.6	0.4562	-0.0097	0.0248	94.9	0.6166
	0.1	-0.0433	0.0322	85.6	0.5233	0.2	-0.0056	0.0158	93.2	0.4603	-0.0313	0.0236	96	0.6212
	0.2	-0.0347	0.0379	84.8	0.5565	0	-0.0147	0.0177	93.4	0.4741	-0.0343	0.0317	94.2	0.6453
0.5	-0.1	-0.0246	0.0285	87.4	0.5131	0	0.006	0.015	92.5	0.4384	-0.0108	0.0311	90.8	0.5501
	0	-0.0219	0.0305	85.4	0.522	0	-0.0035	0.0164	92.3	0.4485	-0.0228	0.0294	89.1	0.5506
	0.05	-0.0202	0.0287	87.9	0.5273	0	-0.017	0.016	92.6	0.4485	-0.0251	0.0284	91.1	0.5501
	0.1	-0.0377	0.0298	86.4	0.5158	0	-0.0114	0.0156	93.6	0.4483	-0.0462	0.0305	89.5	0.5459
	0.2	-0.0343	0.0274	86.5	0.5124	0	-0.0183	0.0143	93.5	0.4443	-0.0562	0.0323	89.3	0.5626

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

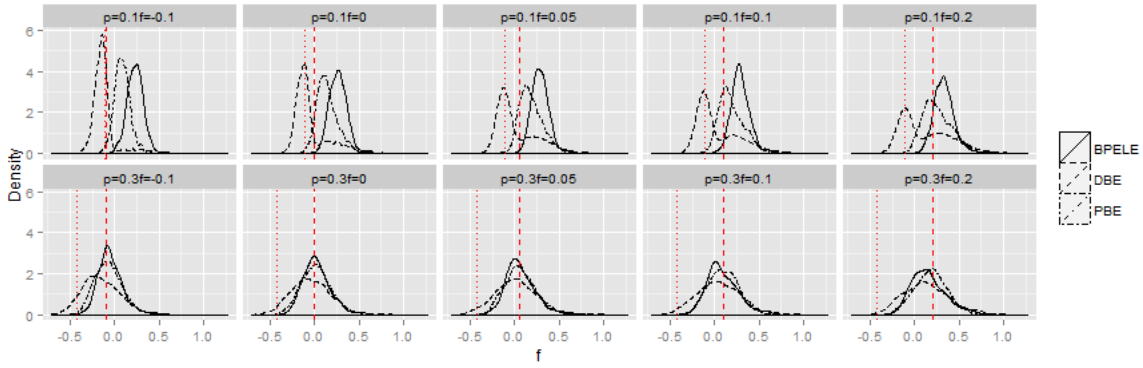
Table 23. Comparison of Design-Based, Parametric Bayesian and Bayesian Empirical Likelihood Estimators

(Proportional to Size Sampling, n=80 Families)

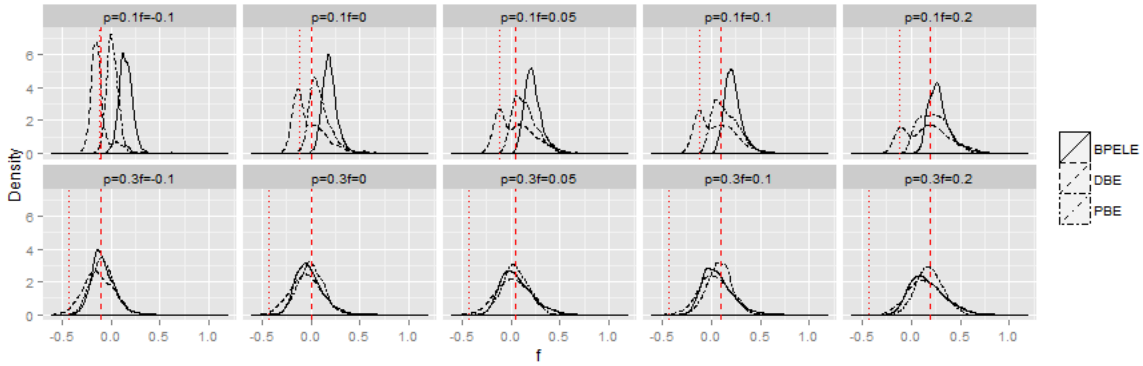
p	f	Design-Based Estimator					Parametric Bayesian Estimator				Bayesian Pseudo-Empirical Likelihood Estimator			
		Bias	MSE	CP	AL	ORP	Bias	MSE	CP	AL	Bias	MSE	CP	AL
0.1	-0.1	-0.0097	0.0055	66.5	0.178	65	0.067	0.0059	83.8	0.2694	0.1867	0.0366	42.6	0.5094
	0	-0.0085	0.0162	72.4	0.3361	22.4	0.0461	0.01	94.8	0.3522	0.126	0.0203	99.4	0.5781
	0.05	-0.0192	0.0218	80	0.4117	10.7	0.0184	0.0115	95.8	0.3911	0.0742	0.0125	98.7	0.6127
	0.1	-0.0154	0.0195	81.5	0.4087	10.1	0.0326	0.0121	95.2	0.3944	0.0802	0.012	99.6	0.6134
	0.2	-0.0191	0.0293	84.4	0.5172	3	0.0023	0.0151	92.9	0.4493	0.0282	0.0137	98.3	0.6725
0.3	-0.1	-0.0016	0.0142	85.1	0.3402	0.3	0.023	0.008	92.4	0.3116	0.0134	0.0108	96.4	0.4454
	0	-0.017	0.0148	85.4	0.3568	0	0.0033	0.0084	93	0.3278	-0.0128	0.0129	95.7	0.4639
	0.05	-0.0176	0.0168	84.1	0.3703	0	-0.0093	0.0096	92.4	0.3393	-0.0164	0.0152	94.7	0.4683
	0.1	-0.0159	0.0184	82.4	0.3732	0	0.0033	0.0102	92	0.342	-0.0157	0.0171	92.6	0.4624
	0.2	-0.0247	0.0215	80.2	0.3896	0	-0.0094	0.011	91.8	0.3514	-0.0273	0.0208	91	0.4683
0.5	-0.1	-0.0088	0.0129	88.4	0.3631	0	0.0123	0.0079	92.2	0.3214	-0.0078	0.0127	90.5	0.3789
	0	-0.007	0.0142	87.3	0.3718	0	-0.0028	0.0079	93.5	0.3279	-0.0074	0.014	89	0.3861
	0.05	-0.0126	0.014	87.6	0.3699	0	-0.0065	0.0082	93.5	0.3259	-0.0132	0.0138	90	0.3845
	0.1	-0.0162	0.015	86.6	0.3643	0	0.0016	0.008	93.4	0.3278	-0.017	0.0148	87.7	0.3785
	0.2	-0.0114	0.0151	84	0.3614	0	-0.0021	0.0079	92.7	0.3232	-0.0146	0.0153	85.4	0.3746

* MSE=mean square error, CP=coverage probability, ORP=out of range probability, AL=average length

A. *n*=20 families



B. *n*=40 families



C. *n*=80 families

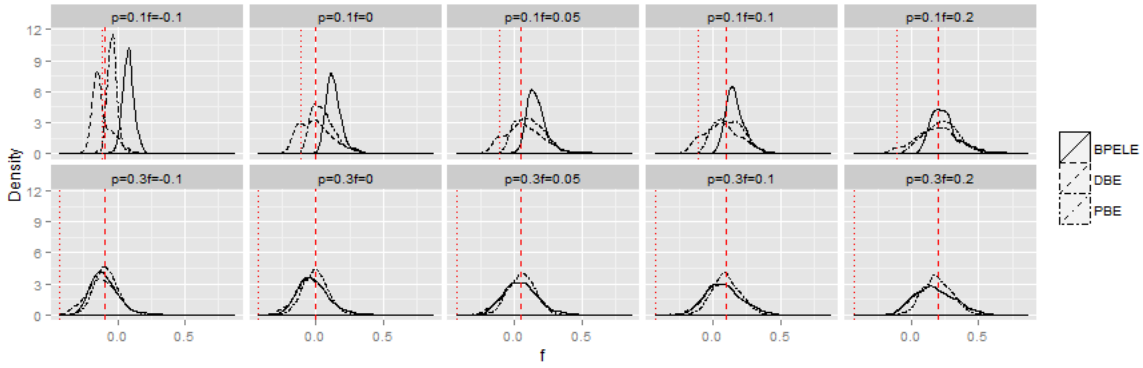


Figure 10. Density Plot of Simulated Estimators under Different Settings (Proportional to Population Size Sampling). BPELE=Bayesian Pseudo-Empirical Likelihood Estimator, DBE=Design-Based Estimator, PBE=Parametric Bayesian Estimator. Parameter p is the allele frequency and f is the inbreeding coefficient in the simulation. Panel A, B and C are for stratum family sample size 20, 40 and 80 families respectively. Red dashed vertical line is the true parameter f; red dotted vertical line is the lower bound of the parameter space for f conditional on allele frequency p.

Chapter 5: Estimation of the Inbreeding Coefficient Incorporating Between Subject Correlation

In Chapter 4, we incorporated within family level correlation in estimating the inbreeding coefficient. Inspired by the NHANES survey design, we now discuss the feasibility in incorporating the between family correlation to estimate the inbreeding coefficient. Because of the similarity in analysis between surveyed individual genotypes and surveyed family genotypes as noted earlier, it is reasonable to concentrate on the between subject correlation only. Most complex surveys usually have a clustered sampling design. Within cluster, it is expected to have between subject correlations. Similarly, between family correlations can happen when families are sampled from clusters. The hierarchical Bayesian model discussed in the following can be generalized to the family level genotype data.

5.1 Population Subdivision

Population subdivision means that the study population is not panmictic, e.g., individuals do not mate at random regardless of any environmental, genetic, or social preference. This departure from panmixia introduces amount of inbreeding at various levels of subdivision. Therefore, the genetic difference among individuals is a result of difference among members of the subpopulation, among members of different subpopulations in the same geographical region, and among members of subpopulations in the same geographical region (Holsinger K. , 1999). In another words, individuals who are from the same cluster, therefore, tend to be more genetically similar than those who are not. Failure to incorporate population cluster or stratification can severely affect statistical test of the genetic association (Pritchard, Stephens, & Donnelly, 2000). There are two types of variability to consider: variability due to sampling and that due to the population substructure. Even if two subpopulations are maintained under the same evolutionary conditions, they may still have different allele frequencies because of the stochastic nature (Weir & Hill, 2002). When the sample size is large and the population substructure is non-negligible, simulation studies show that the estimates from the subpopulations are closer to the true values than the pooled estimates. The opposite is also true for small sample size from more homogeneous subpopulations (Lockwood, Roeder, & Devlin, 2001).

The fixation index F is a measure of correlation between homologous genes relative to a pair of genes randomly selected from a population. In a simple two-level sampling hierarchy where individuals are within subpopulations and subpopulations are within total population, Wright classified the F statistics into three categories: F_{IT}, the correlation between genes within individual relative to the genes from population; F_{IS}, the correlation between genes within individual relative to the genes from subpopulation; F_{ST}, the correlation between genes within subpopulation relative to the genes from total population (Wright S. , 1965). By partitioning the genetic diversity and using the modeling approach, we can study whether any difference among populations are statistically important. Let H_I be the actual heterozygosity for individuals within subpopulation, H_S be expected heterozygosity within subpopulation assuming HWE within subpopulations, and H_T be expected heterozygosity in the total population assuming HWE for the total population. Based on the definition of F statistics, it maintains the following relationship:

$$F_{IT} = 1 - \frac{H_I}{H_T}, \qquad \bar{F}_{IS} = 1 - \frac{H_I}{H_S}, \qquad F_{ST} = 1 - \frac{H_S}{H_T},$$

$$1 - F_{IT} = (1 - \bar{F}_{IS})(1 - F_{ST}),$$

where $\bar{F}_{IS}$ is unweighted average of F_{IS} values across subpopulations with proportional to the subpopulation size. A weighted average calculation which includes relative deme effect sizes and subpopulation allele frequencies can be found in Wright (1969).

Genetic subdivision can be estimated based on the above observed and expected heterozygosity. The genetic diversity due to the allele frequency difference among subpopulation is, therefore, $G_{ST} = (H_T - H_S) / H_T$ (Nei, 1973). From its hierarchical population structure, under the framework of the analysis of variance (ANOVA), Cockerham decomposed the total variance of gene frequencies into variance components associated with various levels of subpopulation to estimate the correlation of genes (Cockerham, 1973). For a simple case of two alleles from a single locus, F_{ST} is equal to the variance among populations divided by the maximum possible variance given the observed mean allele frequencies, e.g., $\frac{var_S(p)}{\bar{P}(1-\bar{p})}$ (Weir & Cockerham, 1984). Thus, it is a measure of proportion of genetic diversity in the whole population that is due to the allele frequency difference among subpopulations. Small F_{ST} means that the allele frequencies within each population are similar. When the number of populations sampled is relatively large, both

estimators of F_{ST} – Nei's G_{ST} and Wei & Cockerham's θ, are typically very similar (Holsinger K. , 1999).

5.2 Bayesian Hierarchical Model

Based on the similarities between Bayesian modeling and hierarchical approach to partition the genetic diversity, Holsinger developed a Bayesian simple multi-locus bi-allelic beta-binomial model to estimate the fixation index (1999). The allele frequency p in each subpopulation has a beta distribution *Beta(a, b)* with $a = x(1-\theta)/\theta$ and $b = (1-x)(1-\theta)/\theta$, where θ corresponds to F_{ST} that can be calculated from the allele frequencies across all subpopulations and x corresponds to the empirical estimate of the mean allele frequency (Rorder, Escobar, Kadane, & Balazs, 1998). With appropriate priors for x and θ, the allele frequencies for each subpopulation and θ can be estimated from the posterior distribution. Then G_{ST} can be further estimated from a random effects model.

From Wright's classification, F_{IS} is the inbreeding coefficient f for the subpopulation and is often used as a measure of the genotype frequencies departure from HWE within subpopulations. Within each subpopulation, the genotype counts follow a multinomial distribution with subpopulation specific inbreeding coefficient f_i and allele frequency p_i similar as previously discussed. If the genotype frequencies in the total population differs from the HWE, expectations can be modeled based on the average allele frequency. For the bi-allelic case, they are given by:

$$\begin{cases} p_{11} = p^2 + p(1-p)F, \\ p_{12} = 2p(1-p)(1-F), \\ p_{22} = (1-p)^2 + p(1-p)F, \end{cases}$$

where p is the average allele frequency and F is the inbreeding coefficient in the total population F_{IT}. If there is no population subdivision, the above model is reduced to the one we used in previous chapters, therefore $F=F_{IS}=f$. For the subdivided population, let f be the average within-subpopulation departure from the HWE. Then, we have $(1-F) = (1-f)(1-\theta)$.

We will use the following notations:

s: index for the subpopulation (s = 1, 2, …, S);

k: index for the allele at a single locus (k = 1, 2, …, K);

m: index for the genotype (m = 1, 2, …, M), M=K(K+1)/2;

A_{ij}: index for the genotype with the alleles A_i and A_j (i =1, 2, …, K; j = i, i+1, …, K);

n_{sij}: index for the genotype count of A_{ij} for the subpopulation s;

$\boldsymbol{n_s}$: a vector of size M for the genotype counts $(n_{s11}, n_{s12}, \dots, n_{s1K}, n_{s22}, \dots, n_{sKK})$ for the subpopulation s;

$\widetilde{\boldsymbol{p}}_{\boldsymbol{s}}$: a vector of size M for the genotype frequencies $(p_{s11}, p_{s12}, \dots, p_{s1K}, p_{s22}, \dots, p_{sKK})$ for the subpopulation s;

$\boldsymbol{p_s}$**:** a vector of size K for the allele frequencies $(p_{s1}, p_{s2}, \dots, p_{sK})$ for the subpopulation s.

In the first level of hierarchy, the genotype counts of each subpopulation are modeled as $Multinomial(\boldsymbol{n_s}, \widetilde{\boldsymbol{p}}_{\boldsymbol{s}})$, e.g.,

$$P(\boldsymbol{n_s}|\widetilde{\boldsymbol{p}}_{\boldsymbol{s}}) = \frac{n_s!}{\prod_{i=1}^{K}\prod_{j=i}^{K} n_{sij}!}\prod_{i=1}^{K}\prod_{j=i}^{K} {p_{sij}}^{n_{sij}}.$$

The genotype frequencies can be re-parameterized using the inbreeding model as

$$p_{sii} = (1-f){p_{si}}^2 + f p_{si} = {p_{si}}^2 + p_{si}(1-p_{si})f, i = 1, \dots, k;$$
$$p_{sij} = 2(1-f)p_{si}p_{sj}, i = 1, \dots, k; j = i+1, \dots, k.$$

We assume a common f, which is the same as the average within-population inbreeding coefficient F_{IS}. A subpopulation specific inbreeding coefficient f_s can also be modelled when local inbreeding is of interest.

In the second level, the allele frequency $\boldsymbol{p_s}$ is modeled by $Dirichlet(\boldsymbol{a})$; $\boldsymbol{a} = (a_1, a_2, \dots, a_K)$ (*Beta* when K=2) (Lange, 1995):

$$P(\boldsymbol{p_s}|\boldsymbol{a}) = \frac{1}{B(\boldsymbol{a})}\prod_{k=1}^{K} {p_{sk}}^{a_k-1},$$

where $a = \sum_{k=1}^{K} a_k$ is the dispersion parameter with large value implying smaller dispersion among subpopulations. Therefore a_k/a is the expected value of the allele A_k frequency (Navarroa, Griffithsb, Steyversc, & Lee, 2006). Since the dispersion parameter a represents the amount of

variation that we expect to see in a finite sample, it can be approximated by the Wright's F_{ST} (or Wei & Cockerham's θ) as a=(1- θ)/ θ (Lockwood, Roeder, & Devlin, 2001). Therefore, our first re-parameterization of $\boldsymbol{a}$ is as follows:

$$a_k = \frac{1-\theta}{\theta} p_k, k = 1, \dots, K,$$

where $\boldsymbol{p} = (p_1, \dots, p_K)$ corresponds to the mean of the allele frequencies distribution across subpopulations. From the property of *Dirichlet* distribution, the mean and variance of the allele frequency estimator is given below (Lockwood, Roeder, & Devlin, 2001):

$$E(\hat{p}_{sk}) = \frac{a_k}{a} = \frac{\frac{1-\theta}{\theta} p_k}{\sum_{k=1}^{K} \frac{1-\theta}{\theta} p_k} = p_k;$$

$$Var(\hat{p}_{sk}) = \frac{a_k(a-a_k)}{(a+1)a^2} = \frac{\frac{1-\theta}{\theta} p_k (\frac{1-\theta}{\theta} - \frac{1-\theta}{\theta} p_k)}{(\frac{1-\theta}{\theta} + 1)(\frac{1-\theta}{\theta})^2} = p_k(1-p_k)\theta.$$

Let $\varphi(f), \varphi(\theta), \varphi(\boldsymbol{p})$ be the prior distribution for f, θ, and $\boldsymbol{p}$, respectively. The joint posterior distribution for $\boldsymbol{p}$, θ, and f is given by

$$\varphi(\boldsymbol{p}, f, \theta | \boldsymbol{n_s}) \propto \left\{ \prod_{s=1}^{S} \varphi(\boldsymbol{n_s} | \boldsymbol{p_s}) \, \varphi(\boldsymbol{p_s} | \boldsymbol{a}) \right\} \varphi(\boldsymbol{p}) \varphi(\theta) \varphi(f).$$

If $\boldsymbol{X}$ has a *Dirichlet* distribution, $\boldsymbol{X} = (X_1, \dots, X_K) \sim Dir(a)$, the vector $\boldsymbol{X}$ is neutral in the sense that X_K is independent of $X^{(-K)} = (\frac{X_1}{1-X_K}, \frac{X_2}{1-X_K}, \dots, \frac{X_{K-1}}{1-X_K})$. Any permutation of $\boldsymbol{X}$ is also neutral (Connor & Mosimann, 1969). Therefore, for k = 1, 2, …, K-1, it is possible to take log-transformation of the *Dirichlet* distribution parameters. Let $\gamma_k = log(\frac{a_k}{a_K})$ and $\beta = log(a)$. Under this re-parameterization, the prior distribution of $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{K-1})$, $\varphi(\boldsymbol{\gamma})$, can be assumed to have a multivariate normal distribution with mean and variance-covariance matrix derived from the observed data. The prior distribution of β, $\varphi(\beta)$, can also be assumed to have a normal distribution with parameters estimated form the data. The posterior distribution therefor is given by

$$\varphi(\boldsymbol{p}, f, \theta | \boldsymbol{n_s}) \propto \left\{ \prod_{s=1}^{S} \varphi(\boldsymbol{n_s} | \boldsymbol{p_s}) \, \varphi(\boldsymbol{p_s} | \boldsymbol{a}) \right\} \varphi(\boldsymbol{\gamma}) \varphi(\beta) \varphi(f).$$

One of the advantages of this *Multinomial-Dirichlet* hierarchical model is that the estimated minor allele frequency will not be forced to be zero if there is no observed allele count in a particular subpopulation. We will use this model parameterization in our future research.

The above model can be generalized to samples with unequal sample selection probabilities. In such case, we use the effective sample size $\boldsymbol{n_s}^*$as discussed in Chapters 3 and 4. When K=2, our first OPENBUGS hierarchical Bayesian model for the cluster sampling is as follows:

Level 1: $\boldsymbol{n_s}^* \sim Multinomial(n_s^*, \boldsymbol{p_s})$, $\boldsymbol{n_s}^* = (n_{s0}^*, n_{s1}^*, n_{s2}^*)$, $\boldsymbol{p_s} = (p_{s0}, p_{s1}, p_{s2})$, s=1, 2, …, S

$$n_s^* = n_{s0}^* + n_{s1}^* + n_{s2}^*, \begin{cases} p_{s0} = p_s^2 + p_s(1-p_s)f, \\ p_{s1} = 2p_s(1-p_s)(1-f), \\ p_{s2} = (1-p_s)^2 + p_s(1-p_s)f. \end{cases}$$

Level 2: $p_s \sim Beta[\frac{(1-\theta)p}{\theta}, \ \frac{(1-\theta)(1-p)}{\theta}]$.

Level 3: $p \sim$ *Uniform* [0, 1],

$f|p \sim$ *Uniform* [-min(p, 1-p)/(1-min(p,1-p)) ,1],

$\theta \sim$ *Uniform* [0, 1].

```
model=function() {
  for (i in 1:nsub){
    q[i,1]<-(1-f)*p[i]*p[i] + f*p[i]
    q[i,2]<-2*(1-f)*p[i]*(1-p[i])
    q[i,3]<-(1-f)*(1-p[i])*(1-p[i]) + f*(1-p[i])

    y[i,1:3] ~ dmulti(q[i,],n[i])

    n[i]<-sum(y[i,])
    p[i] ~ dbeta(alpha, beta)
  }

  alpha<-(1-theta)*pai/theta
  beta<-(1-theta)*(1-pai)/theta
  pai~dunif(0,1)
  f <- w*(1 - f.min) + f.min
  f.min <- max(-pai/(1-pai), -(1-pai)/pai)
```

```
  w ~ dunif(0, 1)
  theta~dunif(0,1)
}
```

Instead of using a common inbreeding coefficient at level one, cluster specific inbreeding models can be applied. The population average of the inbreeding coefficient, therefore, can be estimated from the average of the posterior cluster specific f_s, s=1, …, S (Cavalli-Sforza & Bodmer, 1971). The following is our second OPENBUGS hierarchical Bayesian model for the cluster sampling:

Level 1: $\boldsymbol{n_s}^* \sim Multinomial(n_s^*, \boldsymbol{p_s})$, $\boldsymbol{n_s} = (n_{s0}^*, n_{s1}^*, n_{s2}^*)$, $\boldsymbol{p_s} = (p_{s0}, p_{s1}, p_{s2})$, s=1, 2, …, S;

$$n_s^* = n_{s0}^* + n_{s1}^* + n_{s2}^*, \begin{cases} p_{s0} = (1-p_s)^2 + p_s(1-p_s)f_s, \\ p_{s1} = 2p_s(1-p_s)(1-f_s), \\ p_{s2} = p_s^2 + p_s(1-p_s)f_s. \end{cases}$$

Level 2: $p_s \sim Beta\left[\frac{(1-\theta)p}{\theta}, \ \frac{(1-\theta)(1-p)}{\theta}\right]$,

$$f_s|p_s \sim Uniform\left[-\frac{\min(p_s,1-p_s)}{1-\min(p_s,1-p_s)}, 1\right].$$

Level 3: $p \sim$ *Uniform* [0, 1],

$\theta \sim$ *Uniform* [0, 1].

```
model=function() {
  for (i in 1:nsub){
    q[i,1]<-(1-f[i])*p[i]*p[i] + f[i]*p[i]
    q[i,2]<-2*(1-f[i])*p[i]*(1-p[i])
    q[i,3]<-(1-f[i])*(1-p[i])*(1-p[i]) + f[i]*(1-p[i])

    y[i,1:3] ~ dmulti(q[i,],n[i])
    p[i] ~ dbeta(alpha,beta)
    f[i]<-w[i]*(1 - f.min[i]) + f.min[i]
    f.min[i]<- -min(p[i],1-p[i])/(1-min(p[i],1-p[i]))
    w[i]~dunif(0,1)
    n[i]<-sum(y[i,])
  }
  alpha<-(1-theta)*pai/theta
  beta<-(1-theta)*(1-pai)/theta
  pai~dunif(0,1)
```

```
  theta~dunif(0,1)
  pf<-mean(f[])
}
```

5.3 Simulation Study

We compare the parametric Bayesian estimator with the direct sample estimator using a Monte Carlo simulation study. First, a direct sample estimate $f^{(s)}$ is calculated based on simulated total sample genotypic data. This is compared with the Bayesian estimator with a common inbreeding coefficient model. Secondly, $\bar{F}_{IS}$, an average of F_{IS} values across subpopulations with proportional to the subpopulation size is derived with respect to the clustered data. This is compared with the Bayesian estimator of the population average of cluster specific inbreeding coefficients. The following parameters are used in the simulation:

(i) H=23, the number of sample strata;

(ii) L=2, the number of sample clusters per sample stratum;

(iii) m=60, the number of individuals sampled from each cluster;

(iv) $\rho = 0.017$, the intra-class correlation coefficient measuring the degree of correlation due to the clustering of the sample; This correspondent to an inflation of variance of $1 + \rho(m + 1) = 2$ (Brier, 1980) for proportion compared to a simple random sample with the same sample size. Therefore, the design effect at this setting is Deff=2.

(v) w, the non-informative sample weight, which is randomly assigned to one third each of the samples with values 1, 3 or 5.

(vi) G=5, the number of analyzing groups;

(vii) $\boldsymbol{p_G} = (p_1, p_2, p_3, p_4, p_5) = (0.2, 0.25, 0.3, 0.35, 0.4)$, the allele frequencies for G groups; We use this parameter setting to represent slightly different allele frequencies among subpopulations.

(viii) f=-0.1, 0, 0.05, 0.1 or 0.2.

These simulation settings are in line with the NHANES III study design (NCHS, 1994) and are modified from Li (2009) when the performance of testing HWE under complex sampling is studied. For the g-th group in each cluster, genotype frequencies are first generated from

independent *Dirichlet-Multinomial* distribution, then individual genotypes are simulated from multinomial with those generated genotype frequencies. The group size is generated by multinomial with cluster sample size and equal probabilities among groups.

For the direct sample estimator, the variance is estimated by the jackknife method. For the Bayesian approach, effective sample size and effective number of genotype counts are used in the multinomial model. Table 24 and Table 25 display the simulated (with 100 replications) bias, MSE, relative root mean square error (RRMSE), coverage probability and average interval length for the direct sample estimator and hierarchical Bayesian estimator. For the model with common f, the direct sample estimator is calculated based on the overall sample; while for the cluster specific f_s model, it is calculated as the mean of the subpopulation inbreeding coefficients weighted by the subpopulation size.

In all our simulations, only average direct sample estimator using cluster specific inbreeding coefficients can maintain the nominal coverage probability. The Bayesian estimator has less coverage than the design-based estimator. However, it has smaller simulated bias, MSE and average interval length. For example, when the true inbreeding coefficient f=0, the direct sample estimator has a relative large positive bias of 0.0228 with MSE of 0.001 when treating the total population without subdivision; while the Bayesian estimator has a small negative bias of -0.0012 with MSE of 0.0005. The RRMSE between the Bayesian and design-based estimator is 71%. The large positive bias observed for the direct sample estimator in Table 24 is likely due to the Wahlund effect caused by subpopulation structure. As we discussed in Chapter 1, the inbreeding coefficient is a measure of correlation between two homologous genes in uniting gametes. With the introduction of the intra-class correlation coefficient due to the clustering of the sample, it inflates the overall correlation between genes. By estimating the average subpopulation inbreeding coefficients, the bias of the direct sample estimator is reduced to -0.002. The bias for the Bayesian method is also reduced slightly to -0.0003.

To compare the design-based estimator with the Bayesian estimator under clustered survey sampling, a second simulation study is conducted. Instead of generating a random sample as above in each simulation iteration, we first create a finite population with 1000 PSUs with other simulation parameters remaining the same. The finite population inbreeding coefficient is

calculated as the average of the inbreeding coefficients across clusters. Next at each one of 1000 simulation iteration, we randomly select 50 PSUs from the finite population. All units within the selected PSUs are included in the final analysis sample. Based on the result from the first simulation study, we focus on the sample direct estimator $\bar{F}_{IS}$ and the Bayesian estimator from a cluster specific inbreeding coefficient model. The effective sample size and genotype counts are used in the Bayesian model. Table 26 indicates that both estimators have similar bias and MSE.

Table 24. Comparison of Direct Sample Estimator and Parametric Bayesian Estimators (Single Common f Model)

	Direct Sample Estimator				Parametric Bayesian Estimator				
f	*Bias*	*MSE*	*CP*	*AL*	*Bias*	*MSE*	*CP*	*AL*	*RRMSE*
-0.1	0.0222	0.001	88	0.0979	-0.0044	0.0005	77	0.0561	0.71
0	0.0228	0.001	90	0.1029	-0.0012	0.0005	85	0.0606	0.71
0.05	0.0244	0.0012	86	0.1041	0.0011	0.0005	82	0.062	0.65
0.1	0.0203	0.001	93	0.1061	-0.0015	0.0006	81	0.0628	0.77
0.2	0.0202	0.0009	90	0.1059	0.0000	0.0005	87	0.0635	0.75

*MSE=mean square error, CP=coverage probability, AL=average length, RRMSE=relative root mean square error (vs. Direct Sample Estimator)

Table 25. Comparison of Direct Sample Estimator and Parametric Bayesian Estimators (Cluster Specific f Model)

	Direct Sample Estimator				Parametric Bayesian Estimator				
f	*Bias*	*MSE*	*CP*	*AL*	*Bias*	*MSE*	*CP*	*AL*	*RRMSE*
-0.1	-0.0056	0.0006	95	0.0942	-0.0033	0.0005	74	0.0566	0.91
0	-0.002	0.0005	97	0.1028	-0.0003	0.0005	83	0.0606	1.0
0.05	0.0000	0.0005	96	0.1042	0.0015	0.0005	81	0.0618	1.0
0.1	-0.0019	0.0006	95	0.1063	-0.0008	0.0006	80	0.0628	1.0
0.2	-0.0004	0.0005	98	0.1066	0.0000	0.0005	87	0.0635	1.0

*MSE=mean square error, CP=coverage probability, AL=average length, RRMSE=relative root mean square error (vs. Direct Sample Estimator)

Table 26. Comparison of Design-Based Estimator and Parametric Bayesian Estimator (Cluster Specific f Model)

	Design-Based Estimator		Parametric Bayesian Estimator	
f	*Bias*	*MSE*	*Bias*	*MSE*
-0.1	0.0002	0.0003	0.0005	0.0003
0	-0.001	0.0004	-0.0008	0.0004
0.05	-0.0007	0.0004	-0.0005	0.0004
0.1	-0.0009	0.0004	-0.0008	0.0004
0.2	-0.0009	0.0004	-0.0008	0.0004

Chapter 6: Application to Health and Retirement Study

To illustrate our proposed Bayesian methods to estimate the inbreeding coefficient, we analyzed the HRS genetic data that consist of SNPs from two widely studied candidate genes. HRS employs a multi-stage area probability sampling design. The details of HRS study design can be found in Chapter 1.

The main HRS survey data is public and can be download once registered at its website https://ssl.isr.umich.edu/hrs/start.php. To obtain the access to HRS restricted genetic data, we first applied to *dbGaP* (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) genetic data repository. Once access to *dbGaP* was granted, we further applied to HRS restrict data access (http://hrsonline.isr.umich.edu/index.php?p=xxgen0) directly for the access to HRS-*dbGaP* cross-reference file as well as a well-selected subset of candidate genes and SNPs. The cross-reference file provides the ID crosswalk (HHID and PN) between the HRS genetic data files and the HRS public data files. After all datasets were downloaded from HRS data repository, we merged all data files by those unique identifiers. A total of 12,507 subjects were included in our merged data file. Among them, 6658 subjected were genotyped in 2006 and the remaining 5849 subjects were genotyped in 2008.

From released 2006 genetic data, sample weights are provided for the subjects whose biomarkers were selected for genotyping. Respondents with at least one valid biomarker result were assigned biomarker weights. The initial weights were calculated by dividing the HRS 2006 sample weights by the predicted probabilities of responses to each component. These interim weights were further trimmed at the 5th and 95th percentiles and were finally post-stratified back to the entire 2006 HRS sample by age, sex, and race/ethnicity. A similar process was conducted for the 2008 released genetic data. However, for this chapter we only focus on the 2006 genetic data for our application.

For the genetic data download, the value provided for each SNP is actually the dosage (number of coded alleles) for a person. Since the genotypes are further imputed using the external 1000 Genomes data (1000 Genomes, 2015), the dosage ranges from 0 to 2 and may not be exactly the integer values for each person. The data file also includes an INFO metric, which is the observed statistical information associated with the allele frequency estimate. It is considered as a measure

of SNP imputation quality (1000 Genomes Project Reference Panel, 2012). INFO score greater than 0.8 is considered as a conservative filter, while greater than 0.3 is a liberal one. As recommended in the Quality Control document, we used the following algorithm to define final genotype: genotype=0 if dosage value ≤0.5; genotype=1 if 0.5<dosage≤1.5; genotype=2 if dosage >1.5.

As documented in the Quality Control Report for Genotypic Data (2012), the inbreeding coefficient is given by 1 - (number of observed heterozygotes) / (number of expected heterozygotes), same as we discussed in Chapter 1 and Chapter 2 under the simple random sampling. The report shows an approximately symmetric distribution of the inbreeding coefficient estimates for all autosomal SNPs, with mean of 0.0023 for European and mean of 0.0019 for African (Figure 11). The results do not suggest an excess of positive inbreeding coefficient, or, equivalently, an excess of homozygotes. As a benchmark for the comparison, we computed the estimate of the inbreeding coefficient same as the one described in the Quality Report. Moreover, we computed the survey design-based estimator, parametric Bayesian estimator, and Bayesian pseudo-empirical likelihood estimator as discussed in Chapter 3.

Among many genes recorded in the HRS genetic data file, we selected two genes to demonstrate our application. First, we selected SNPs on APOC1 gene that encodes Apolipoprotein C1, a component of lipoproteins. This protein is responsible for the activation of esterified lecithin cholesterol with an important role in the exchange of esterified cholesterol between lipoproteins and in removal of cholesterol from tissues (Tata, et al., 1985). We also studied SNPs on BDNF gene that encodes brain-derived neurotrophic factor (BDNF), a protein mainly expressed in the central nervous system. Neurotrophic factors promote the survival of neurons by preventing associated signals that initiate programmed cell death. Study suggests that this gene is associated with depression (Binder & Scharfman, 2004; Zuccato, 2009).

We did not apply the recommended composite quality filter documented in the QC report. Such filter is typically used as initial screen tool for the downstream gene-disease association studies to exclude those potential questionable SNPs. The QC document reports that the percentage of all SNPs with minor allele frequency (MAF) less than 1% is 24.5% for autosomes and 17.2% for the X chromosome. To reduce computation time, we only studied a subset of the total reported SNPs

for both genes. First, we selected SNPs with MAF greater than 10%. For APOC1, there are a total of 105 SNPs in the HRS data repository. After we applied the MAF filter, 27 SNPs remained in the final analysis. For BDNF, there are a total of 357 SNPs. Because of the limitation of R memory size for our computer, we only selected first two hundred SNPs in the file. After the application of the MAF filter, there are 49 SNPs included in the comparison. In order to show results when MAF is just over the limit for SNP, our second study selected 38 SNPs from BDNF gene with MAF restricted from 1% to 5%.

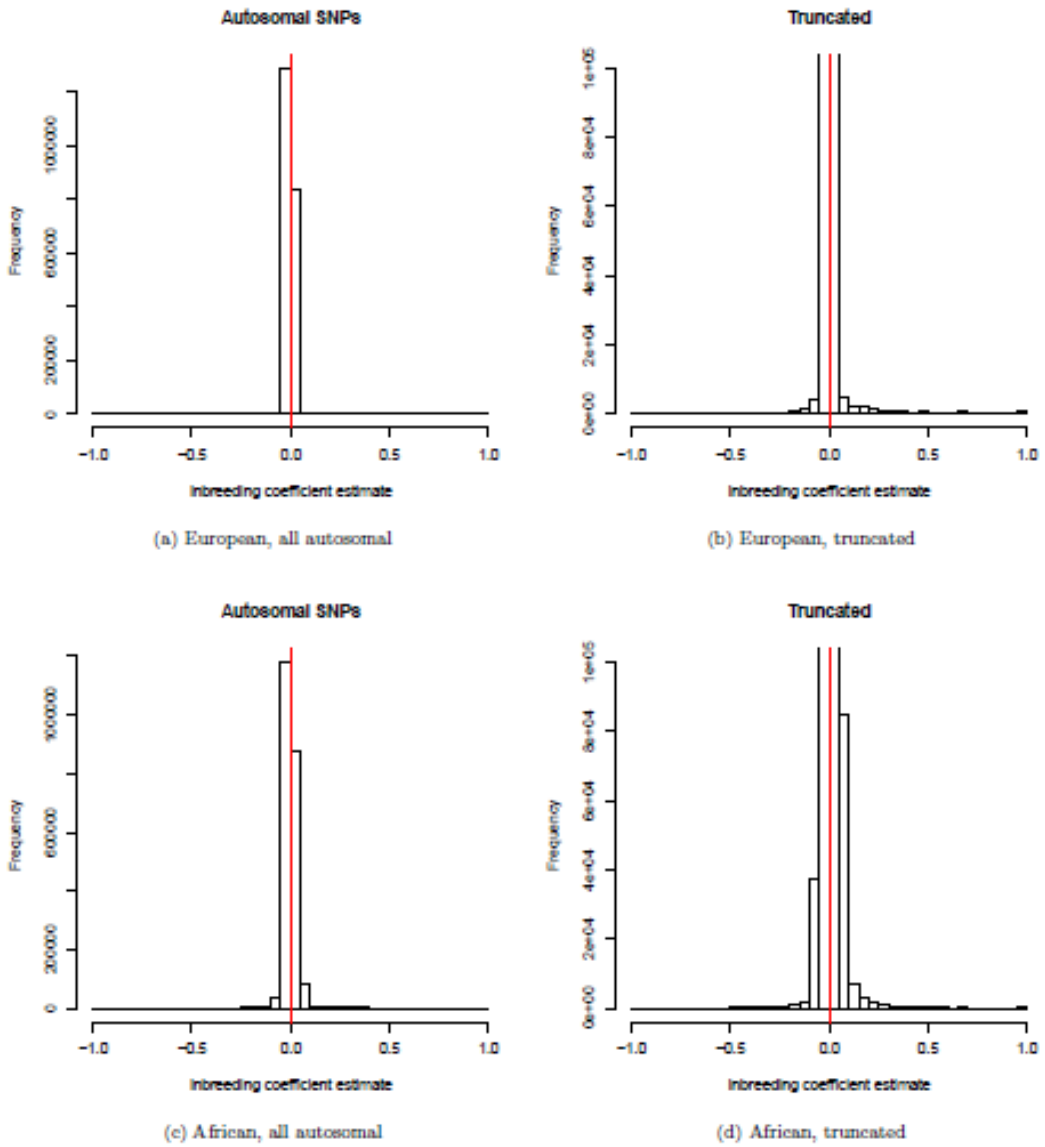


Figure 11. Distribution of Estimated Inbreeding Coefficients for all Autosomal SNPs ("Figure 26" in Quality Control Report for Genotypic Data, 2012)

Table 27, Table 28 and Figure 12 Panel A & B list four estimates of inbreeding coefficient for the selected SNPs with MAF>10%: naïve estimator for a simple random sampling, design-based survey sample estimator, parametric Bayesian estimator under a multinomial likelihood, and Bayesian pseudo-empirical likelihood estimator. When the inbreeding coefficient is small, ignoring the sampling design is more likely to generate a statistical significant estimator under a nominal 95% confidence level, e.g., the 95% confidence interval does not cover the zero. This observation is consistent with the QC report, which recommends to use a stricter filter threshold of p=0.0001 to exclude SNPs from further downstream analysis. Furthermore, for autosomal SNPs, the QQ plot of all p-values for the testing of HWE suggests deviation of observed from expected p-values is between 0.001 and 0.01. Such filter results in some potential good SNPs being screened out from further downstream analysis. Our analysis suggests that most of these estimates with small values are not considered statistically significant at 0.05 level any more when taking the sampling design into account. For these SNPs with all four estimates are consistently significant, we also notice that the size of the inbreeding coefficient is generally smaller when the sampling design feature is considered in analysis. Furthermore, the Bayesian methods generally have smaller credible intervals and, therefore, a more precise estimation of the inbreeding coefficient. The Bayesian pseudo-empirical likelihood estimator is the most conservative estimator among all four estimators we considered.

Table 29 and Figure 12 Panel C lists same four estimates for the selected SNPs with MAF greater than 1% but less than 5%. For all inbreeding coefficient estimate greater than 0.1, all four methods detect such large deviation from HWE. For smaller coefficient less than 0.1, Bayesian pseudo-empirical likelihood estimate suggests that all those are not worth mention at the 0.05 confidence level, while the results are sporadic for all other three methods.

Table 27. Estimated Inbreeding Coefficients for Selected SNPs from APOC1 Gene (MAF>10%).

SNP	***Unweighted Estimator***	***Design-Based Estimator***	***Parametric Bayesian Estimator***	***Bayesian Pseudo-Empirical Likelihood Estimator***
kgp658335	0.0136 (-0.0107, 0.0378)	-0.0004 (-0.0473, 0.0477)	-0.0004 (-0.0238, 0.0234)	-0.0004 (-0.0258, 0.0262)
kgp7807118	0.0074 (-0.017, 0.0319)	-0.0065 (-0.0576, 0.0446)	-0.0067 (-0.0296, 0.0175)	-0.0067 (-0.0437, 0.0293)
kgp8411531	0.0068 (-0.018, 0.0315)	-0.024 (-0.0627, 0.0146)	**-0.0227 (-0.044, -0.004)**	-0.0245 (-0.074, 0.024)
rs10414043	0.0103 (-0.0145, 0.0352)	0.0156 (-0.0276, 0.0587)	0.017 (-0.0081, 0.0435)	0.0152 (-0.0308, 0.0602)
rs111789331	0.0142 (-0.0109, 0.0393)	0.0107 (-0.0343, 0.0556)	0.0118 (-0.0124, 0.037)	0.0103 (-0.0361, 0.0559)
rs12721046	0.0168 (-0.0085, 0.042)	0.0132 (-0.0288, 0.0551)	0.014 (-0.0099, 0.042)	0.0128 (-0.033, 0.058)
rs12721051	0.0125 (-0.0123, 0.0373)	0.0056 (-0.0386, 0.0497)	0.0066 (-0.0189, 0.0326)	0.0053 (-0.0356, 0.0454)
rs12721056	0.006 (-0.0183, 0.0302)	-0.005 (-0.0505, 0.0405)	-0.004 (-0.0284, 0.0236)	-0.005 (-0.0315, 0.0215)
rs157594	-0.0102 (-0.0344, 0.014)	-0.0262 (-0.0759, 0.0235)	**-0.0259 (-0.05, -0.0015)**	**-0.0262 (-0.05, -0.002)**
rs157595	**0.0404 (0.0161, 0.0647)**	0.018 (-0.0294, 0.0655)	0.0178 (-0.0067, 0.0425)	0.018 (-0.0075, 0.0435)
rs3826688	**0.041 (0.0166, 0.0653)**	0.0223 (-0.0282, 0.0729)	0.0217 (-8e-04, 0.0449)	0.0223 (-0.0034, 0.0486)
rs390082	**0.0298 (0.0039, 0.0557)**	-0.0081 (-0.0495, 0.0332)	-0.0078 (-0.0296, 0.0146)	-0.0085 (-0.0552, 0.0368)
rs3925681	0.0126 (-0.0116, 0.0368)	0.0219 (-0.0153, 0.0592)	0.0222 (-0.0015, 0.0467)	0.0219 (-0.0034, 0.0466)
rs438811	**0.0415 (0.0167, 0.0663)**	0.0158 (-0.028, 0.0595)	0.0159 (-0.0079, 0.0399)	0.0156 (-0.0152, 0.0468)
rs439401	**0.0363 (0.0119, 0.0607)**	0.0191 (-0.031, 0.0693)	0.019 (-0.0053, 0.0442)	0.0191 (-0.0072, 0.0458)
rs4420638	0.0177 (-0.0071, 0.0426)	0.0125 (-0.0283, 0.0533)	0.0138 (-0.0111, 0.0378)	0.0123 (-0.0272, 0.0508)
rs445925	**0.0273 (0.0015, 0.053)**	-0.011 (-0.0543, 0.0323)	-0.011 (-0.0341, 0.0138)	-0.0114 (-0.0584, 0.0346)
rs483082	**0.0299 (0.0053, 0.0546)**	0.0078 (-0.0365, 0.0522)	0.0083 (-0.0168, 0.0332)	0.0077 (-0.0239, 0.0391)
rs484195	**0.0412 (0.0169, 0.0656)**	0.0229 (-0.0231, 0.0689)	0.0229 (-8e-04, 0.0489)	0.0229 (-0.0028, 0.0482)
rs5117	0.0134 (-0.0111, 0.0379)	-0.0112 (-0.0613, 0.0388)	-0.0102 (-0.0338, 0.0154)	-0.0114 (-0.0446, 0.0224)
rs56131196	0.0179 (-0.007, 0.0428)	0.0126 (-0.0282, 0.0534)	0.0132 (-0.011, 0.0386)	0.0123 (-0.0272, 0.0508)
rs584007	**0.0332 (0.0089, 0.0576)**	0.0158 (-0.0343, 0.0659)	0.0153 (-0.0091, 0.0391)	0.0158 (-0.0107, 0.0423)
rs59325138	0.0114 (-0.0129, 0.0356)	-3e-04 (-0.0447, 0.044)	-1e-04 (-0.025, 0.0229)	-4e-04 (-0.0261, 0.0249)
rs7256200	0.0109 (-0.014, 0.0357)	0.016 (-0.0272, 0.0593)	0.0171 (-0.0078, 0.0435)	0.0156 (-0.0306, 0.0614)
rs73052335	0.023 (-0.003, 0.049)	0.0152 (-0.0261, 0.0565)	0.0161 (-0.009, 0.0426)	0.0147 (-0.0347, 0.0633)
rs78959900	-0.001 (-0.0252, 0.0232)	-0.0079 (-0.0544, 0.0387)	-0.0073 (-0.0318, 0.0174)	-0.0079 (-0.0347, 0.0183)
*rs814573**	**0.0363 (0.011, 0.0615)**	0.0347 (-0.0093, 0.0786)	**0.0351 (0.0116, 0.0593)**	0.0345 (-0.0031, 0.0709)

* The INFO score is 0.786.

Table 28. Estimated Inbreeding Coefficients for Selected SNPs from BDNF Gene (MAF>10%).

SNP	***Unweighted Estimator***	***Design-Based Estimator***	***Parametric Bayesian Estimator***	***Bayesian Pseudo-Empirical Likelihood Estimator***
kgp10709149	0.0199 (-0.0057, 0.0455)	-0.0027 (-0.0495, 0.0442)	-0.0012 (-0.0238, 0.0229)	-0.0031 (-0.0527, 0.0453)
kgp2818969	0.0049 (-0.0195, 0.0292)	0.0045 (-0.0401, 0.0492)	0.0048 (-0.0189, 0.0316)	0.0043 (-0.0317, 0.0403)
kgp3196024	0.0049 (-0.0195, 0.0292)	0.0045 (-0.0401, 0.0492)	0.0048 (-0.0189, 0.0316)	0.0043 (-0.0317, 0.0403)
rs10767658	0.0069 (-0.0174, 0.0312)	0.0081 (-0.0344, 0.0506)	0.0088 (-0.0178, 0.034)	0.008 (-0.0205, 0.0365)
rs10767659	**0.1108 (0.0866, 0.1351)**	**0.0805 (0.0375, 0.1234)**	**0.0803 (0.0567, 0.1035)**	**0.0804 (0.0551, 0.1061)**
rs10767662	0.0214 (-0.0028, 0.0456)	0.017 (-0.0275, 0.0615)	0.0176 (-0.0061, 0.0414)	0.017 (-0.0075, 0.0415)
rs10835210	**0.0578 (0.0335, 0.0821)**	**0.0452 (0.0079, 0.0825)**	**0.0457 (0.0212, 0.0694)**	**0.0452 (0.0205, 0.0705)**
rs10835211	0.0221 (-0.0026, 0.0468)	0.013 (-0.0214, 0.0475)	0.0134 (-0.0098, 0.0379)	0.0129 (-0.0199, 0.0451)
rs10835213*	-**0.0381 (-0.0619,-0.0144)**	-0.0318 (-0.0717, 0.0081)	**-0.0319 (-0.0546, -0.0082)**	**-0.0319 (-0.0616, -0.0026)**
rs11030101	**0.0561 (0.0318, 0.0803)**	0.0424 (6e-04, 0.0842)	**0.0426 (0.0185, 0.0668)**	**0.0424 (0.0182, 0.0672)**
rs11030102	**0.0261 (0.0014, 0.0509)**	0.0168 (-0.0198, 0.0533)	0.017 (-0.0066, 0.0418)	0.0167 (-0.0162, 0.0488)
rs11030104	0.0084 (-0.016, 0.0329)	0.0109 (-0.0323, 0.0541)	0.0112 (-0.0121, 0.0358)	0.0107 (-0.0255, 0.0465)
rs11030107	**0.0263 (0.0016, 0.0511)**	0.017 (-0.0196, 0.0535)	0.0174 (-0.0065, 0.0427)	0.0168 (-0.0161, 0.0489)
rs11030108	0.0109 (-0.0134, 0.0352)	0.0082 (-0.0359, 0.0523)	0.008 (-0.0163, 0.0344)	0.0081 (-0.021, 0.037)
rs11030112	0.0137 (-0.0107, 0.038)	0.0091 (-0.0342, 0.0524)	0.0092 (-0.0146, 0.0342)	0.0091 (-0.0199, 0.0381)
rs114907865	**0.1072 (0.083, 0.1315)**	**0.0811 (0.0349, 0.1273)**	**0.0809 (0.0579, 0.1045)**	**0.081 (0.0559, 0.1069)**
rs11500197	0.0233 (-0.0014, 0.048)	0.0139 (-0.0216, 0.0494)	0.0138 (-0.0107, 0.0392)	0.0138 (-0.0183, 0.0457)
rs117567901	0.0052 (-0.0192, 0.0295)	0.0035 (-0.0364, 0.0433)	0.0035 (-0.0211, 0.0286)	0.0033 (-0.0322, 0.0388)
rs12575096	0.0126 (-0.012, 0.0372)	0.0136 (-0.0302, 0.0574)	0.0147 (-0.0106, 0.0384)	0.0134 (-0.0233, 0.0497)
rs12790234	0.0052 (-0.0192, 0.0295)	0.0035 (-0.0364, 0.0433)	0.0035 (-0.0211, 0.0286)	0.0033 (-0.0322, 0.0388)
rs138385919	**0.0263 (0.0016, 0.0511)**	0.017 (-0.0196, 0.0535)	0.0174 (-0.0065, 0.0427)	0.0168 (-0.0161, 0.0489)
rs1401635	0.0106 (-0.0137, 0.0349)	0.0066 (-0.0381, 0.0513)	0.0071 (-0.0174, 0.0333)	0.0065 (-0.0218, 0.0352)
rs143956188	0.0052 (-0.0192, 0.0295)	0.0035 (-0.0364, 0.0433)	0.0035 (-0.0211, 0.0286)	0.0033 (-0.0322, 0.0388)
rs1519480	**0.1086 (0.0843, 0.1329)**	**0.079 (0.0374, 0.1206)**	**0.0789 (0.0549, 0.1018)**	**0.0789 (0.0531, 0.1051)**
rs16917237	0.0106 (-0.0139, 0.0351)	0.0133 (-0.0307, 0.0573)	0.0134 (-0.0108, 0.0375)	0.0131 (-0.0228, 0.0492)
rs1808124	**0.0333 (0.0089, 0.0577)**	0.0283 (-0.0085, 0.0651)	**0.0285 (0.0031, 0.0548)**	**0.0283 (0.001, 0.055)**
rs189740576	-0.0033 (-0.0273, 0.0206)	-0.0212 (-0.0642, 0.0218)	-0.0194 (-0.0403, 0.0043)	-0.0217 (-0.0722, 0.0288)
rs1949513	**0.0652 (0.0409, 0.0896)**	**0.0512 (0.0125, 0.09)**	**0.0515 (0.0274, 0.0754)**	**0.0512 (0.0257, 0.0767)**
rs2049045	0.0084 (-0.0161, 0.033)	0.0119 (-0.0334, 0.0573)	0.012 (-0.0134, 0.0372)	0.0117 (-0.0283, 0.0507)
rs34379767	0.0052 (-0.0192, 0.0295)	0.0035 (-0.0364, 0.0433)	0.0035 (-0.0211, 0.0286)	0.0033 (-0.0322, 0.0388)
rs35038967	0.0124 (-0.0122, 0.037)	0.0133 (-0.0304, 0.057)	0.0141 (-0.0107, 0.0399)	0.0131 (-0.0235, 0.0485)
rs4378341	0.0248 (6e-04, 0.049)	0.0208 (-0.0242, 0.0657)	0.021 (-0.0031, 0.0452)	0.0208 (-0.0036, 0.0454)
rs4385847	0.0244 (2e-04, 0.0486)	0.0207 (-0.0238, 0.0653)	0.0207 (-0.0035, 0.045)	0.0207 (-0.0037, 0.0453)
rs4517468	**0.1007 (0.0764, 0.1249)**	**0.0689 (0.0255, 0.1122)**	**0.0688 (0.0462, 0.0933)**	**0.0688 (0.0433, 0.0943)**
rs4542361	0.0251 (9e-04, 0.0493)	0.022 (-0.0229, 0.0668)	0.0218 (-0.003, 0.0462)	0.022 (-0.0024, 0.0466)
rs4633417	0.0207 (-0.0035, 0.045)	0.0148 (-0.0321, 0.0617)	0.0144 (-0.0088, 0.0382)	0.0148 (-0.01, 0.039)
rs4923463	0.0045 (-0.0198, 0.0289)	0.0032 (-0.0411, 0.0475)	0.0042 (-0.0208, 0.0297)	0.003 (-0.0336, 0.0394)
rs4923464	0.0091 (-0.0154, 0.0336)	0.0109 (-0.0325, 0.0543)	0.0115 (-0.0134, 0.0363)	0.0107 (-0.0252, 0.0468)
rs4923466	0.0052 (-0.0192, 0.0295)	0.0035 (-0.0364, 0.0433)	0.0035 (-0.0211, 0.0286)	0.0033 (-0.0322, 0.0388)
rs61888762	0.0126 (-0.0118, 0.0369)	0.0087 (-0.0356, 0.0531)	0.0089 (-0.0144, 0.033)	0.0087 (-0.0204, 0.0376)
rs6265	0.0124 (-0.0123, 0.037)	0.012 (-0.0345, 0.0585)	0.0122 (-0.011, 0.0365)	0.0118 (-0.0264, 0.0496)
rs6484320	0.0037 (-0.0206, 0.028)	0.0062 (-0.0327, 0.0452)	0.0067 (-0.0184, 0.032)	0.006 (-0.0292, 0.0408)
rs6484321	0.0052 (-0.0191, 0.0296)	0.0065 (-0.0333, 0.0464)	0.007 (-0.0171, 0.0329)	0.0064 (-0.0287, 0.0413)
rs7103411	0.0037 (-0.0206, 0.028)	0.0061 (-0.0327, 0.0449)	0.0064 (-0.017, 0.0308)	0.0059 (-0.0293, 0.0407)
rs7103873	0.024 (-2e-04, 0.0482)	0.0208 (-0.0226, 0.0642)	0.0205 (-0.0035, 0.045)	0.0208 (-0.0039, 0.0451)
rs7104207	0.0237 (-5e-04, 0.048)	0.0207 (-0.0227, 0.0641)	0.0208 (-0.0038, 0.0452)	0.0207 (-0.004, 0.045)
rs7124442	**0.0347 (0.0103, 0.0591)**	0.032 (-0.0062, 0.0702)	**0.0319 (0.0069, 0.0572)**	**0.032 (0.0042, 0.0592)**
rs7482752	0.0251 (9e-04, 0.0493)	0.0199 (-0.0253, 0.0651)	0.0209 (-0.0031, 0.0458)	0.0199 (-0.0042, 0.0438)
rs7926362	0.0037 (-0.0206, 0.028)	0.0063 (-0.0327, 0.0453)	0.0073 (-0.0165, 0.0318)	0.0061 (-0.0292, 0.0408)

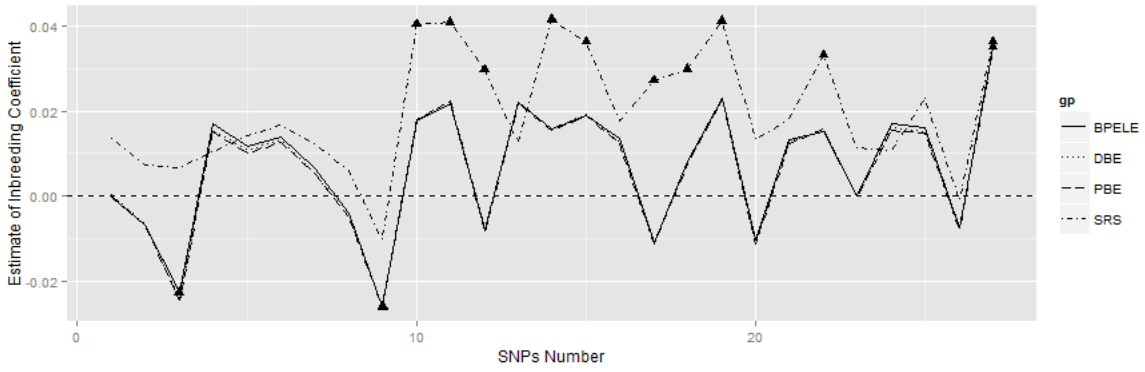
* The INFO score is 0.726.

Table 29. Estimated Inbreeding Coefficients for Selected SNPs from BDNF Gene (<1%MAF<5%).

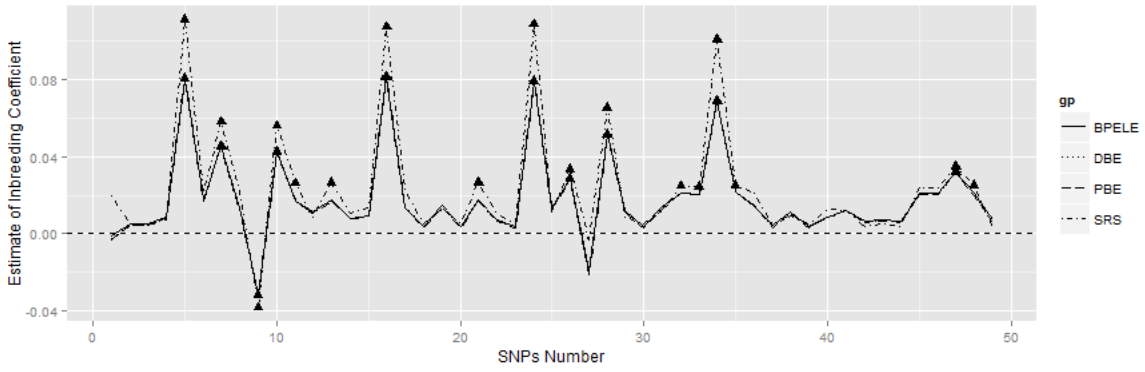
SNP	*Unweighted Estimator*	*Design-Based Estimator*	*Parametric Bayesian Estimator*	*Bayesian Pseudo-Empirical Likelihood Estimator*
kgp12568216	0.0081 (-0.0192, 0.0355)	-0.0063 (-0.0362, 0.0236)	6e-04 (-0.0186, 0.0292)	0.0214 (-0.0252, 0.0998)
kgp12594720	0.038 (-0.0063, 0.0822)	0.0649 (-0.0329, 0.1628)	0.0728 (0.0267, 0.1297)	0.0844 (-0.0049, 0.2111)
kgp12631228	0.0998 (0.0534, 0.1462)	**0.095 (0.0209, 0.1691)**	**0.0986 (0.0555, 0.1501)**	0.0946 (-0.0019, 0.1951)
kgp12991896	**0.0475 (0.004, 0.091)**	0.0434 (-0.0232, 0.1099)	0.0487 (0.0108, 0.0996)	0.0638 (-0.011, 0.173)
kgp1796534	-0.002 (-0.0244, 0.0205)	-0.0043 (-0.03, 0.0213)	0.0084 (-0.0112, 0.045)	0.0468 (-0.0107, 0.1493)
kgp370455	0.0092 (-0.0218, 0.0403)	-0.0045 (-0.0295, 0.0205)	0.0078 (-0.0113, 0.0457)	0.0472 (-0.0105, 0.1505)
kgp782640	**0.0485 (0.0118, 0.0853)**	0.0638 (-0.0108, 0.1384)	**0.0667 (0.0294, 0.1087)**	0.0639 (-0.0171, 0.1489)
rs112989286	0.0325 (-0.0109, 0.0758)	0.0545 (-0.0434, 0.1523)	**0.0697 (0.0246, 0.1356)**	0.0822 (-0.0045, 0.2125)
rs114813081	**0.0541 (0.0028, 0.1055)**	0.0761 (-0.0481, 0.2002)	**0.0807 (0.0306, 0.1475)**	0.0965 (-0.0027, 0.2353)
rs116590817	**0.0581 (0.0112, 0.105)**	0.0462 (-0.0219, 0.1142)	**0.05 (0.0142, 0.1012)**	0.0664 (-0.0104, 0.1776)
rs117314519	9e-04 (-0.0237, 0.0255)	-0.0113 (-0.0363, 0.0136)	-0.0065 (-0.0225, 0.0194)	0.0191 (-0.0253, 0.0957)
rs117392835*	1e-04 (-0.0242, 0.0245)	0.0239 (-0.0573, 0.1051)	**0.0332 (0.0014, 0.0779)**	0.0486 (-0.0148, 0.1472)
rs11819808	**0.2048 (0.1521, 0.2574)**	**0.1902 (0.1013, 0.2792)**	**0.1944 (0.1441, 0.2466)**	**0.187 (0.0819, 0.2859)**
rs11826087	**0.0733 (0.0193, 0.1274)**	0.0789 (-0.0356, 0.1933)	**0.0868 (0.039, 0.1509)**	0.0953 (-0.0029, 0.2291)
rs12284158	**0.2036 (0.1511, 0.2561)**	**0.1885 (0.0996, 0.2775)**	**0.1899 (0.1428, 0.242)**	**0.1854 (0.0807, 0.2837)**
rs12284724	**0.2036 (0.1511, 0.2561)**	**0.1885 (0.0996, 0.2775)**	**0.1899 (0.1428, 0.242)**	**0.1854 (0.0807, 0.2837)**
rs140500062	**0.0585 (0.0114, 0.1056)**	0.0465 (-0.0217, 0.1148)	**0.0519 (0.0159, 0.0984)**	0.0668 (-0.0104, 0.1786)
rs140893479*	**-0.0106 (-0.0124, -0.0089)**	**-0.011 (-0.0141, -0.0078)**	0.0029 (-0.0111, 0.0403)	0.0524 (-0.008, 0.163)
rs142615576	-0.0015 (-0.0254, 0.0224)	0.0081 (-0.0372, 0.0534)	0.0119 (-0.0112, 0.0396)	0.0132 (-0.0433, 0.0837)
rs145216856	**0.0585 (0.0114, 0.1056)**	0.0465 (-0.0217, 0.1148)	**0.0519 (0.0159, 0.0984)**	0.0668 (-0.0104, 0.1786)
rs145596917	0.0195 (-0.0139, 0.0529)	-0.0026 (-0.0283, 0.0231)	0.0012 (-0.0165, 0.035)	0.0319 (-0.0186, 0.1184)
rs146296986	**0.1825 (0.1258, 0.2392)**	**0.1886 (0.0711, 0.3062)**	**0.1919 (0.1373, 0.2541)**	**0.1855 (0.0808, 0.2838)**
rs183649896*	**-0.0117 (-0.0136, -0.0099)**	**-0.0111 (-0.0144, -0.0077)**	0.0017 (-0.0111, 0.037)	0.0521 (-0.0081, 0.1619)
rs190640500*	**-0.0121 (-0.014, -0.0102)**	**-0.0106 (-0.0132, -0.0079)**	0.0026 (-0.0104, 0.0375)	0.0539 (-0.0076, 0.1664)
rs34043390*	-0.0079 (-0.0259, 0.01)	**-0.0165 (-0.02, -0.013)**	-0.0076 (-0.017, 0.0165)	0.0359 (-0.0135, 0.1275)
rs4528281	0.0321 (-0.011, 0.0753)	0.0538 (-0.0432, 0.1508)	**0.0686 (0.0231, 0.128)**	0.0816 (-0.0046, 0.2114)
rs55699391	**0.1181 (0.0702, 0.1659)**	**0.1359 (0.0538, 0.2181)**	**0.1384 (0.0923, 0.1885)**	**0.1333 (0.0314, 0.2304)**
rs56274324	**0.1072 (0.0602, 0.1542)**	**0.1086 (0.0225, 0.1947)**	**0.112 (0.0691, 0.1637)**	**0.107 (0.0064, 0.2064)**
rs57083135	**0.0574 (0.0108, 0.1039)**	0.0456 (-0.0228, 0.114)	**0.0501 (0.0129, 0.1016)**	0.0659 (-0.0105, 0.1765)
rs58597651	**0.0574 (0.0108, 0.1039)**	0.0456 (-0.0228, 0.114)	**0.0501 (0.0129, 0.1016)**	0.0659 (-0.0105, 0.1765)
rs7124665	**0.0966 (0.0401, 0.1532)**	0.051 (-0.0121, 0.114)	**0.0563 (0.0176, 0.1043)**	0.0797 (-0.0057, 0.2083)
rs72878164	**0.1181 (0.0702, 0.1659)**	**0.1359 (0.0538, 0.2181)**	**0.1384 (0.0923, 0.1885)**	**0.1333 (0.0314, 0.2304)**
rs72878175	**0.0574 (0.0108, 0.1039)**	0.0456 (-0.0228, 0.114)	**0.0501 (0.0129, 0.1016)**	0.0659 (-0.0105, 0.1765)
rs72878179	**0.0574 (0.0108, 0.1039)**	0.0456 (-0.0228, 0.114)	**0.0501 (0.0129, 0.1016)**	0.0659 (-0.0105, 0.1765)
rs72878181	**0.0574 (0.0108, 0.1039)**	0.044 (-0.0221, 0.1101)	**0.0494 (0.0133, 0.0966)**	0.0644 (-0.0109, 0.1741)
rs72878196	**0.0958 (0.0504, 0.1411)**	**0.0908 (0.0186, 0.163)**	**0.0944 (0.0546, 0.14)**	0.0906 (-0.0047, 0.1893)
rs78531552	-0.0052 (-0.0263, 0.0158)	-0.0151 (-0.0327, 0.0025)	-0.0063 (-0.0195, 0.0201)	0.028 (-0.019, 0.112)
rs79292978	0.0178 (-0.0177, 0.0534)	0.0055 (-0.0337, 0.0448)	0.0158 (-0.008, 0.0576)	0.0483 (-0.0117, 0.1513)

* The INFO scores are less than 0.8.

A



B



C

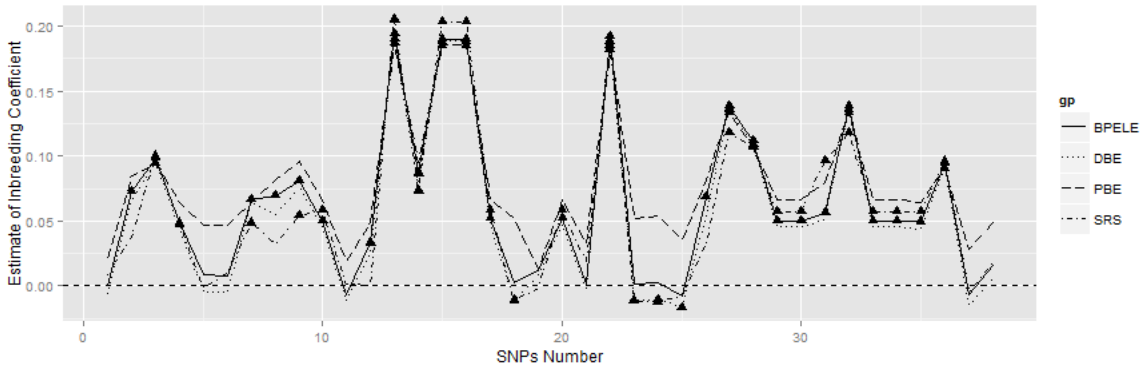


Figure 12. Comparison of Inbreeding Coefficient Estimators from HRS Genetic Data. BPELE=Bayesian Pseudo-Empirical Likelihood Estimator, DBE=Design-Based Estimator, PBE=Parametric Bayesian Estimator, SRS=Simple Random Sampling Estimator. Panel A, B and C are for APOC1 SNPs with allele frequency >10%, BDNF SNPs with allele frequency >10%, and BDNF SNPs with allele frequency between 1% and 5%, respectively. Solid triangles represent estimators with 95% CI do not cover zero (a significant deviation from HWE at 5% level).

Chapter 7: Summary and Future Research

In this dissertation, both parametric Bayesian method with multinomial likelihood and non-parametric Bayesian method with empirical likelihood are proposed for genetic data collected from a complex survey sampling design. The proposed methods are evaluated through simulations under various sampling designs, including stratified simple random sampling, proportional to size sampling and clustered sampling. Using real genetic data from HRS study, we have demonstrated the utility of the proposed Bayesian methods for SNPs screening in GWAS based on national level complex survey.

Design-based estimator is popular with survey practitioners since it automatically takes the survey design into account and provides a reliable estimate for large samples, which is often the case for large scale national survey. For non-standard statistics such as the inbreeding coefficient from genetic data, the standard error of the computed statistic is often complicated. However, it can be estimated by the Taylor linearization or widely used resampling techniques. For estimation of the inbreeding coefficient induced from general biological inbreeding model, design-based estimator may be outside of the parameter space limit. The use of such estimator needs be carefully reviewed when the minor allele frequency (MAF) is small, especially if the estimated inbreeding coefficient is also small.

Bayesian method provides an alternative approach for the estimation of the inbreeding coefficient. This dissertation fills the gap in the Bayesian application to GWAS where subjects may have different sample selection probabilities. Our proposed hierarchical Bayesian multinomial model coincides with the biological inbreeding model for the genotype distribution. Unlike the design-based estimator, our parametric Bayesian estimator is naturally restricted to the parameter space of the inbreeding coefficient induced by the general biological inbreeding model.

Without assuming a multinomial likelihood for the Bayesian model, our proposed non-parametric Bayesian estimator extended the widely used method proposed by Owen, Sitter, Rao & Wu to the nominal genetic data. Although the computation to maximize the empirical likelihood under the constraint of estimating equation is non-trivial, with a carefully selected algorithm and

approximate grid search method, it is reasonably efficient to produce a non-parametric Bayesian estimator.

Our simulation studies suggests that the design-based estimator and two Bayesian estimators have similar performance when the allele frequency is not small or the sample size is large enough. However, when allele frequency is small, the design-based method often produces estimator outside of the parameter space, especially when the sample size is also small. Such limitation can be easily addressed through Bayesian prior distribution for the parameter. Furthermore, when the data is very much skewed, the Bayesian method tends to produce a more reliable estimator based on its data augmentation nature. Our simulation studies also remind us that the computation of empirical likelihood on the boundary of the parameter space is not adequately stable. Therefore, a special attention needs be paid when the observed inbreeding coefficient is closer to the lower bound of the parameter space induced by the inbreeding model. On the other hand, Sforza & Bodmer (1971) and HRS study (Weir D. R., 2012) reported that inbreeding coefficient is not likely close to the lower boundary of parameter space for most human SNPs. Furthermore, practitioners are more concerned with a positive inbreeding coefficient for more heterozygotes which may indicate a deviation from HWE due to population structure. Such population structure, often exists in national level complex survey, needs be carefully taken into account when estimating the inbreeding coefficient in GWAS.

In terms of the choice of estimator to be used in practice, our recommendation should be based on the purpose of the inbreeding coefficient estimation. If the main purpose is to screen out the SNPs for genotyping error, it is reasonable to use the naïve estimator as stated in the QC report with a more conservative confidence level such as 0.0001. On the other hand, if the main purpose is to have a more robust estimator for the inbreeding coefficient, we recommend to incorporate the sampling design. Our demonstration using the HRS genetic data indicates that estimator ignoring the sampling design often overestimates the inbreeding coefficient, therefore, more likely to deviate from HWE. When minor allele frequency is small, we suggested to consider the Bayesian estimators to ensure it is within the parameter space.

As an initial assessment of Bayesian method in the estimation of inbreeding coefficient for genetic data from complex survey, we limit our discussion on the single locus SNPs. Perhaps the next

important step is to extend the method to include multiple loci SNPs. These loci share the same population history and therefore the degree of subpopulation heterogeneity across loci is expected to be similar (Lockwood, Roeder, & Devlin, 2001). Using an appropriate Bayesian model that borrows strength across loci, more accurate estimates of subpopulation allele frequency and the inbreeding coefficient can be obtained.

Introducing empirical likelihood in the Bayesian method may appear to be not practically useful in our preliminary study. However, without any model assumption, it is easier to include auxiliary variables into analysis. Although the focus of this dissertation is to estimate inbreeding coefficient, our immediate future research is gene-disease association study under complex survey setting. The robustness of model based Bayesian estimator is greatly dependent on its model specification. Serious misspecification of the model for gene-disease association will lead to misleading and often bad inferences. Bayesian method with the empirical likelihood does not require a parametric model. It is relatively straightforward to include candidate variables possibly related to the disease and borrow information from other sources such as Census data. For example, for most genetic inherited disease, the onset of disease is often related to aging. While it is possible to parametrically model the gene-disease association with the consideration of age and other environmental factors, the exact functional form may not be clear. On the other hand, finite population measure for age is easily accessible from Census. Using such auxiliary variable as another constraint to maximize the empirical likelihood, we can compare the genetic traits between subjects with the disease (case) and those without (control).

Appendices: R Programs

```
#--------------------------- Newton method to derive lamda from equation sum(ds*(u-mu)/(1+lamda*(u-mu)))=0
# adapted from Rao & Wu
Lag1<-function(u,ds,mu){
  L<--1/max(u-mu)
  R<--1/min(u-mu)
  dif<-1
  tol<-1e-8
  if(L>=0 | R<=0){
    dif=0
    M=0
  }
  while(dif>tol){
    M<-(L+R)/2
    glam<-sum((ds*(u-mu))/(1+M*(u-mu)))
    if(glam>0) L<-M
    if(glam<0) R<-M
    dif<-abs(glam)
  }
  return(M)
}


#--------function to calculate Bayesian posterior mean for inbreeding coefficient f and #confidence interval.
# adapted from Rao & Wu
#ds- normalized weight, ys- data
#phat- profiled allele frequency, use design-based estimator
Bay2=function(ds,ys,phat,to=0.00001,inc=0.001){
  pmin=min(phat,1-phat)
  R= -pmin/(1-pmin)
  mu=sum(ds*ys)
  lam=Lag1(ys,ds,mu)
  elm=exp(-nss*sum(ds*log(1+lam*(ys-mu))))
  el=NULL
  pf=NULL
  while(R<1-inc){
    R=R+inc
    mu=2*(1-R)*phat*(1-phat)
    lam=Lag1(ys,ds,mu)
    elr=exp(-nss*sum(ds*log(1+lam*(ys-mu))))
    if (elr/elm > to){
      el=c(el,elr)
      pf=c(pf,R)}
  }
  A=sum(el)*inc
  el=el/A
  ng=length(el)

  a=0
  k=0
  alpha=0.05
  while(a<=alpha/2){
    k=k+1
```

```
   a=sum(el[(ng-k):ng])*inc }
  Rhat=pf[ng-k]

  a=0
  k=1
  while(a<=alpha/2){
    k=k+1
    a=sum(el[1:k])*inc }
  Lhat=pf[k]

  bhat=sum(el*pf*inc)
  #plot(pf,el,type="l",ylab="Posteria Empirical Likelihood",xlab="f")
  return(cbind(bhat=bhat,bl=Lhat,br=Rhat,len=Rhat-Lhat))
}

#-----------------------------------------Jackknife variance estimator for f-----------------------------------------
jnse=function(df){
  ns=dim(df)[1]
  je=rep(0,ns)
  for (i in 1:ns){
    repdsgn = as.svrepdesign(dstrat, type="JKn", fay.rho=0.5)
    repwt=repdsgn$repweights[,i]
    repdf=data.frame(geno=df$geno,df$n1,wt=repwt$weights)
    jdstrat=svydesign(ids=~1,strata=NULL, weights=~wt,data=repdf)
    jp=svymean(~geno, jdstrat)[1]
    jn=svymean(~n1, jdstrat)[1]
    je[i]=1-jn/(2*(1-jp/2)*(jp/2))
  }
  return(jnse=sqrt((ns-1)*sum((je-mean(je))^2)/ns))
}

#---------------------------------R program to compare the MLE with MELE----------------------------------------
library(emplik)

totn=50 #sample size
p=0.1
f=0.05
p.alle=c(1-p,p)

nsim=250
set.seed=1234
k.alle=length(p.alle)
pij=c( (1-f)*p.alle[1]^2+f*p.alle[1],
       2*(1-f)*p.alle[1]*p.alle[2],
       (1-f)*p.alle[2]^2+f*p.alle[2] )
pij.cumsum=cumsum(pij)

fin=matrix(rep(0,nsim*8),nrow=nsim,ncol=8)
myfun=function(y){
  y1=y
  y2=y*(2-y)
  return(cbind(y1,y2))
}
for (k in 1：nsim){
  tmp=runif(totn)
```

```
  geno=sapply(1:(totn), function(k) min(which(pij.cumsum>=tmp[k])))
  geno=geno - 1
  print(c(k,table(geno)))
  if (length(table(geno))==3) fin[k,11:13]=table(geno)
  if (length(table(geno))==2) fin[k,11:13]=c(table(geno),0)
  p.bar=sum(geno)/(2*totn) #sample allele freq
  f.bar=1-(sum(geno*(2-geno))/totn)/(2*p.bar*(1-p.bar))
  f.sd=sqrt((1-f.bar)*(1-2*p.bar*(1-p.bar)*(1-f.bar) -
                      (2*p.bar-1)^2*(1-f.bar)^2)/(2*p.bar*(1-p.bar)*totn))
  fin[k,5]=f.bar
  fin[k,6]=f.sd

  nss=totn
  fin[k,1:4]=Bay2(ds=1/totn,ys=geno*(2-geno),phat=p.bar,to=0.00001,inc=0.001)

  x=geno*(2-geno)
  d=rep(1,totn)
  np=100
  nf=100
  mu1=mean(geno)
  mu2=mean(x)
  sp=sf=smu=logL=rep(0,(np*nf))
  for (i in 1:np){
    for (j in 1:nf){
      sp[(i-1)*nf+j]=i/(2*np)
      fmin=-sp[(i-1)*nf+1]/(1-sp[(i-1)*nf+1])
      sf[(i-1)*nf+j]=fmin+(j-1)*(1-fmin)/(nf-1)
      smu[(i-1)*nf+j]=2*(1-sf[(i-1)*nf+j])* sp[(i-1)*nf+j]*(1- sp[(i-1)*nf+j])
      #     temp=el.cen.EM(x, d, fun=function(t){t}, mu=smu[(i-1)*np+j])
      #     logL[(i-1)*np+j]=temp$loglik
      mu1=2*sp[(i-1)*nf+j]
      mu2=smu[(i-1)*nf+j]
      temp=el.cen.EM2(geno, d, fun=myfun, mu=c(mu1,mu2))
      logL[(i-1)*nf+j]=temp$loglik
    }
    print(i)
  }
  lh=data.frame(sp=sp,sf=sf,smu=smu,logL=logL)
  lmax=max(lh$logL)
  res=subset(lh,logL==lmax)
  fin[k,7]=res$sf
  fin[k,8]=res$sp
}

x=round(c(MLE_Bias=mean(fin[,5]-f,na.rm=TRUE),
        MLE_MSE=(mean(fin[,5])-f)^2+var(fin[,5]),
        MLE_Cover=100*sum(fin[,5]-1.96*fin[,6]<=f & f<=fin[,5]+1.96*fin[,6],na.rm=TRUE)/nsim,
        MLE_Len=mean(2*1.96*fin[,6],na.rm=TRUE),
        MLE_Out=100*(sum(fin[,5]>1,na.rm=TRUE)+sum(fin[,5] < -min(p,1-p)/(1-min(p,1-
p)),na.rm=TRUE))/nsim,

        EL_Bias=mean(fin[,7]-f,na.rm=TRUE),
        EL_MSE=(mean(fin[,7],na.rm=TRUE)-f)^2+var(fin[,7],na.rm=TRUE),
        EL_Out=100*(sum(fin[,7]>1,na.rm=TRUE)+sum(fin[,7] < -min(p,1-p)/(1-min(p,1-
p)),na.rm=TRUE))/nsim), 3)
```

```
#--------------------Code to analyze the HRS genetic data for APOC gene---------------------------------------
options(stringsAsFactors=FALSE)
source("C:/Research/R/allfunction.R")
library(survey)
library(arm)
library(rms)
library (BRugs)
library(coda)
library(dplyr)

#######Attach main survey, crosslink and genetic data#######
attach("R/Longevity.RData")
attach("R/crosslink.RData")
attach("R/track.RData")

survey=merge(htrack,crosslink,by=c("HHID","PN"))
survey=subset(survey,GENETICS06==1 | GENETICS08==1,
          select=c(STRATUM,SECU,GENETICS06,GENETICS08,KBIOWGTR,LBIOWGTR,
               HHID,PN,LOCAL_ID))
survey$LBIOWGTRn=as.numeric(as.character(survey$LBIOWGTR))
survey$KBIOWGTRn=as.numeric(as.character(survey$KBIOWGTR))
survey$LOCAL_ID=as.numeric(as.character(survey$LOCAL_ID))

#######################genetic data#########################
varlist=names(df.APOC1.SNP)
nvar=length(varlist)
eval(parse(text=paste("df.APOC1.SNP$",varlist[2])))

#######assign SNP to 0/1/2 according QC report guidance#######
df.APOC1.SNP=as.data.frame(cbind(df.APOC1.SNP[,1],sapply(df.APOC1.SNP[,-1], function(x)
  ifelse(x<=0.5,0,ifelse(x>0.5 & x<=1.5,1,2)))))
sapply(df.APOC1.SNP[,-1], function(x) table(x))
tt=stack(df.APOC1.SNP,select=-V1)

#######preselect SNP based on MAF#######
sumct=data.frame(unclass(table(tt$ind,tt$values)))
lsnp=sumct %>% mutate(pX0=100*X0/(X0+X1+X2),pX1=100*X1/(X0+X1+X2),pX2=100*X2/(X0+X1+X2),
  maf=100*(X1+2*X2)/(2*(X0+X1+X2)),
  snpname=row.names(sumct)) %>%
  filter(maf>10 & maf<90)
lsnp
lnsnp=lsnp$snpname

df.APOC1.SNP$LOCAL_ID=df.APOC1.SNP$V1

#######combine survey and genetic data#######
dim(survey)
dim(df.APOC1.SNP)
fdf=merge(survey,df.APOC1.SNP[,c("LOCAL_ID",lnsnp)],by="LOCAL_ID")
fdf06=subset(fdf,GENETICS06==1 & is.na(KBIOWGTRn)==FALSE)
summary(fdf06$KBIOWGTRn)


############Parametric Bayesian Model#######
```

```
model1=function() {
  q[1]<-(1-f)*p*p + f*p
  q[2]<-2*(1-f)*p*(1-p)
  q[3]<-(1-f)*(1-p)*(1-p) + f*(1-p)

  y[1:3] ~ dmulti(q[],n)

  #prior
  p ~ dunif(0, 1)
  f <- w*(1 - f.min) + f.min
  f.min <- max(-p/(1-p), -(1-p)/p)
  w ~ dunif(0, 1)

  n<-sum(y[])
}
writeModel(model1, "model1.txt")

nss=dim(fdf06)[1]
myAna=function(var,data=fdf06){
  geno=eval(substitute(var),data, parent.frame())
  n1=geno*(2-geno)
  wt=fdf06$KBIOWGTRn
  stype=fdf06$STRATUM
  psu=fdf06$SECU
  gtype=paste("N",geno,sep="")
  df=data.frame(geno,psu,n1,wt,stype,gtype)
  dstrat=svydesign(ids=~psu,strata=~stype, weights=~wt, data=df,nest=TRUE)
  p.svy=svymean(~geno, dstrat)
  n1.svy=svymean(~n1, dstrat)
  f_design=1-n1.svy/(2*(1-p.svy/2)*(p.svy/2))

 ns=length(table(df$stype))
 lst=unique(df$stype)
 je=rep(0,ns*2)
for (i in 1:ns){
  for (j in 1:2){
  #sdf=subset(df,stype!=lst[i] & psu!=j)
    sdf=df
  sdf$wt[sdf$stype==lst[i] & sdf$psu==j]=0
  sdf$wt[sdf$stype==lst[i] & sdf$psu!=j]=sdf$wt[sdf$stype==lst[i] & sdf$psu!=j]*2
  jdstrat=svydesign(ids=~psu,strata=~stype, weights=~wt,data=sdf,nest=TRUE)
  jp=svymean(~geno, jdstrat)[1]
  jn=svymean(~n1, jdstrat)[1]
  je[(i-1)*2+j]=1-jn/(2*(1-jp/2)*(jp/2))
  }
}
jnse=sqrt((2*ns-1)*sum((je-mean(je))^2)/(2*ns))

  nef=round(length(geno)/(svyvar(~geno, dstrat)/(sd(geno))^2),0)
  nys=data.frame(svytotal(~gtype,dstrat))
  ny=c(ifelse(is.na(nys["gtypeN0","total"]),0,nys["gtypeN0","total"]),
       ifelse(is.na(nys["gtypeN1","total"]),0,nys["gtypeN1","total"]),
       ifelse(is.na(nys["gtypeN2","total"]),0,nys["gtypeN2","total"]))
   ny=round(nef*ny/(sum(ny)),0)
```

```
  bdata=list(y=ny[3:1])
  inits1 = list(p=0.5,w=0.5)
  pbl=BRugsFit(data = bdata, inits = inits1,
               para = c("f"), modelFile = "model1.txt",
               numChains = 1,nBurnin = 1000, nIter = 10000, nThin = 10)$Stats

#srs
  p_srs=sum(geno)/(2*nss) #sample allele freq
  f_srs=1-mean(n1)/(2*p_srs*(1-p_srs))
  f_sd=sqrt((1-f_srs)*(1-2*p_srs*(1-p_srs)*(1-f_srs) -
                    (2*p_srs-1)^2*(1-f_srs)^2)/(2*p_srs*(1-p_srs)*nss))
  return(cbind(sn0=ny[3],sn1=ny[2],sn2=ny[1],f_srs=f_srs,srsl=f_srs-1.96*f_sd,srsr=f_srs+1.96*f_sd,
         f_design=f_design,dl=f_design-1.96*jnse, dr=f_design+1.96*jnse,
         f_eb=Bay2(ds=wt/sum(wt),ys=geno*(2-geno),phat=1-p.svy/2,to=0.00001,inc=0.001),
         f_pbl=pbl$mean,pbll=pbl$val2.5pc,pblr=pbl$val97.5pc))
}
nv=length(lnsnp)
suma=matrix(rep(0,16*nv),ncol=16,nrow=nv)
for (i in 1:nv){
  suma[i,]=myAna(eval(parse(text=paste(lnsnp[i]))),fdf06)
}
fsum=as.data.frame(suma)
names(fsum)=c("sn0","sn1","sn2","f_srs","srsl","srsr","f_design","dl","dr","bhat",
              "bl","br","len","f_pbl","pbll","pblr")
dfsum=data.frame(snp=lnsnp,
  srs=paste(round(fsum$f_srs,4), " (",round(fsum$srsl,4), ", ", round(fsum$srsr,4), ")",sep=""),
  des=paste(round(fsum$f_design,4), " (",round(fsum$dl,4), ", ", round(fsum$dr,4), ")",sep=""),
  bel=paste(round(fsum$bhat,4), " (",round(fsum$bl,4), ", ", round(fsum$br,4), ")",sep=""),
  bpb=paste(round(fsum$f_pbl,4), " (",round(fsum$pbll,4), ", ", round(fsum$pblr,4), ")",sep=""))

write.csv(dfsum,file="APOC1.csv")

#######Plot#######
library(ggplot2)
srs=fsum[,c("f_srs","srsl", "srsr")]
names(srs)=c("p","l","h")
srs$seq=1:dim(srs)[1]
srs <- srs[order(srs$p),]
srs$gp=rep("SRS",dim(srs)[1])

des=fsum[,c("f_design","dl", "dr")]
names(des)=c("p","l","h")
des$seq=1:dim(des)[1]
des <- des[order(des$p),]
des$gp=rep("DBE",dim(des)[1])

pbe=fsum[,c("bhat","bl","br")]
names(pbe)=c("p","l","h")
pbe$seq=1:dim(pbe)[1]
pbe <- pbe[order(pbe$p),]
pbe$gp=rep("PBE",dim(pbe)[1])

pbl=fsum[,c("f_pbl","pbll","pblr")]
names(pbl)=c("p","l","h")
pbl$seq=1:dim(pbl)[1]
```

```
pbl <- pbl[order(pbl$p),]
pbl$gp=rep("BPELE",dim(pbl)[1])
df=rbind(srs,des,pbe,pbl)

df_s=subset(df,df$l>0 | df$h<0)
ggplot(df, aes(x=seq, y=p, group=gp,linetype=gp))+
  geom_line()+
  scale_linetype_manual(values = c("solid","dotted","longdash","dotdash"))+
  geom_hline(aes(yintercept=0),linetype=2) +
  geom_point(data=df_s,aes(x=seq,y=p),shape=17,size=3)+
  xlab("SNPs Number")+ylab("Estimate of Inbreeding Coefficient")
```

Bibliography

1000 Genomes. (2015). Retrieved from http://www.1000genomes.org/data

1000 Genomes Project Reference Panel. (2012). *CIDR Health Retirement Study Imputation Report*. Retrieved from http://hrsonline.isr.umich.edu/sitedocs/genetics/HRS_1000G_IMPUTE2_REPORT_AUG2012.pdf

Ayres, K., & Balding, D. (1998). Measuring departure from Hardy-Weinberg: a Markov Chain Monte Carlo method for estimting the inbreeding coefficient. *Heredity*, 80: 769 - 777.

Balding, D., Bishop, M., & Cannings, C. (2007). *Handbook of Statistical Genetics.* West Sussex, England: John Wiley & Sons, Ltd.

Binder, D., & Scharfman, H. (2004). Brain-derived Neurotrophic Factor. *Growth Facotrs*, 22(3): 123-131.

Brier, S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, 67: 591–596.

Brooks, S., & Gelman, A. (1998). Alternative Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7:434-455.

Cavalli-Sforza, L., & Bodmer, W. (1971). *The Genetics of Human Populations.* USA: W.H. Freeman and Company.

Chakraborty, R., & Jin, L. (1992). Heterozygote deficiency, population substructure and their implications in DNA fimgureprinting. *Human. Genet.*, (88): 267-72.

Chang, I., & Mukerjee, R. (2008). Bayesian and frequentist confidence intervals arising from empirical-type likelihoods. *Biometrika* , 95(1): 139–147.

Chang, M., Lindegren, M., Butler, M., & et.al. (2009). Prevalence in the United States of selected candidate gene variants: Third National Health and Nutrition Examination Survey,1991–1994. *Am J Epidemiol*, 169(1):54–66.

Chaudhuri, S., & Ghosh, M. (2011). Empirical likelihood for small area estimation. *Biometrika*, 98: 473-480.

Chen, C., Wakefield, J., & Lumely, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spat Spatiotemporal Epidemiol*, 11: 33–43.

Chen, J., & Sitter, R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, (9): 385-406.

Chen, J., Sitter, R., & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, (89): 230-237.

Chen, S., & Lahiri, P. (2012). Inferences on small area proportions. *Journal of the Indian Society of Agriculture Statistics*, 66(1): 121-124.

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

Cohen, M. (1982). Estimating inbreeding in a natural population: a comparison of sampling properties. *Genetics*, 100: 339-358.

Connor, R., & Mosimann, J. (1969). Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association* , 64 (325): 194–206.

Consonni, G., Morenob, E., & Venturinic, E. (2011). Testing Hardy–Weinberg equilibrium: an objective Bayesian analysis. *Statistics in Medicine*, 30: 62-74.

Consonni, G.; Gutierrez-Pena, E.; Veronese, P. (2008). Compatiable priors for Bayesian model comparison with an application to the Hardy-Weinberg equilirium model. *Test*, 17: 585-605.

Crawford, D., & et.al. (2006). Genetic variation is associated with C-reactive protein levels in the Third National Health and Nutrition Examination Survey. *Circulation*, 114(23):2458-65.

Curie-Cohen, M. (1982). Estimates of inbreeding in a natural population: a comparison of sampling properties. *Genetics*, 100(2): 339–358.

Doeder, K., Escobar, M., Kadane, J., & Balazs, I. (1998). Measuring heterogeneity in forensic databases using hierachical Bayes models. *Biometrika*, (85): 269-287.

Ezzati, T., Massey, J., Waksberg, J., Chu, A., & Maurer, K. (1992). Sample design: Third National Health and Nutrition Examination Survey. *Vital and Health Statistics 2*, pp. (113):1-35.

Fang, K.-T., & Mokerjee, R. (2006). Empirical type likelihood allowing posterior credible sets with frequentist validity: high-order asymptotics. *Biometrika*, 93: 723-33.

Fang, K.-T., & Mukerjee, R. (2005). Expected length of confidence intervals based on empirical discrepancy statistics. *Biometrika*, 92: 499-503.

Fesinmeyer, M., & et.al. (2013). Genetic risk factors for BMI and obesity in an ethnically diverse population: results from the population architecture using genomics and epidemiology (PAGE) study. *Obesity*, 21(4):835-46.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, (22): 153-164.

Ghosh, M., & Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *BIometrika*, 91: 95-112.

Gorroochurn, P., & Hodge, S. (2006). Estimating Allele Frequencies and Inbreeding Coefficients in K-Allele Models. *Mathematical Population Studies*, 13:83–103.

Gunel, E., & Wearden, S. (1995). Bayesian estimation and testing of gene frequencies. *Theor Appl Genet*, 91:534-543.

Hartley, H., & Rao, R. (1968). A new estimation theory for sample surveys. *Biometrika*, (55):547-557.

Health and Retirement Study Survey Design. (2008, December). Retrieved from http://hrsonline.isr.umich.edu/sitedocs/surveydesign.pdf

Heeringa, S., Connor, J., & Darrah, D. (1986). *1980 SRC National Sample - Design and Development.* Ann Arbor: Institute for Social Research.

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

Holsinger, K., & Wallace, L. (2004). Bayesian approaches for the analysis of population genetic structure: an example from Platanthera leucophaea. *Molecular Ecology*, 13: 887 - 894.

Holsinger, K., & Weir, B. (2009). Geneetics in gographically structured populations: defining, estimating and interpreting F. *Nature Reviews Genetics*, (10): 639-650.

HRS 1992 (Wave 1) Documentation. (1992). Retrieved from http://hrsonline.isr.umich.edu/modules/meta/1992/core/codebook/03_core.htm

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15 (1): 1-96.

Koehler, E., Brown, E., & Haneuse, S. (2009). On the assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *Am Stat.*, 63(2): 155-162.

Korn, E., & Graubard, B. (2003). Estimating variance components by using survey data. *J R Stat Soc*, 65: 175-190.

Lange, K. (1995). Applications of the Dirichlet distribution to forensic match probabilities. *Genetica*, 96:107–17.

Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika*, 90: 319-26.

Li, C., & Horvitz, D. (1953). Some Methods of Estimating the Inbreeding Coefficient. *The American Journal of Human Genetics*, 5(2): 107-117.

Li, Y. (2013). A comparison of tests for Hardy-Weinberg Equilibrium in national genetic household surveys. *BMC Genetics*, 14: 14.

Li, Y., & Graubard, B. (2009). Testing Hardy-Weinberg Equilibrium and Homogeneity of Hardy-Weinberg Disequilibrium using complex survey data. *Biometrics*, 65(4):1096-104.

Li, Y., Li, Z., & Graubard, B. (2011). Testing for Hardy Weinberg equilibrium in national household surveys that collect family-based genetic data. *Hum Genet*, 75:732–41.

Li, Y.; Graubard, B.I. (2011). Testing for Hardy Weinberg Equilibrium in National Household Surveys that collect family-based genetic data. *Annals of Human Genetics*, 75(6):732-41.

Lockwood, J., Roeder, K., & Devlin, B. (2001). A Bayesian hierarchical model for allele frequencies. *Genetic Epidemiology*, 20:17–33.

Masel, J. (2011). Genetic drift. *Current Biology*, 21 (20): R837–R838.

Mengersen, K., Pudlo, P., & Robert, C. (2013). Bayesian computation via empirical likelihood. *PNAS*, 110(4): 1321-26.

Monahan, J., & Boos, D. (1992). Proper likelihood for Bayesian analysis. *Biometrika*, 79: 271-8.

Monohan, J. F. (2011). *Numeric Methods of Statistics, 2nd Edition.* Cambridge University Press.

Moonesinghe, R., Yesupriya, A., Chang, M., Dowling, N., Khoury, M., & Scott, A. (2010). A Hardy-Weinberg Equilibrium Test for analyzing population genetics surveys with complex sample designs. *American Journal Epidemiology*, 171: 932-941.

National Research Council. (1996). *The evolution of forensic DNA evidence.* Washington, DC: National Academy Press.

Navarroa, D., Griffithsb, T., Steyversc, M., & Lee, M. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, 50:101–122.

NCHS. (1994). Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. *Vital and Health Statistics*, 1 (32).

Nei, M. (1973). Analysis of gene diversity in subdivided population. *Proc. Natl. Acad. Sci.*, 70: 3321-3323.

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

NHANES. (2010, June 7). *National Health and Nutrition Examination Survey (NHANES) DNA Quality Control Protocol.* Retrieved from http://www.cdc.gov/nchs/data/nhanes/genetics/Quality_Control_Public.pdf

NHANES. (2015, March 27). *Types of Genetic Data Sets Available.* Retrieved from Anonymized Genetic Data: http://www.cdc.gov/nchs/data/nhanes/nhanes3/anonymized_genetic_data_list.pdf

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86(6):641-7.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for single functional. *Biometrika*, (75): 237-249.

Owen, A. (2001). *Empirical Likelihood.* Boca Raton, London, New York, Washington, D.C.: Chapman & Hall.

Pongpanich, M., Sullivan, P., & Tzeng, J. (2010). A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics*, 26 (14): 1731–1737.

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155: 945–959.

Qin, J., & Lawless, J. (1994). Empirical likelihood and general estimation equations. *The Annals of Statistics*, (22): 300-325.

Rao, J. (2011). Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal. *Statistical Science*, (26): 240–256.

Rao, J., & Wu, C. (2010). Bayesian pseudo-empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society*, (72): 533-544.

Rorder, K., Escobar, M., Kadane, J., & Balazs, I. (1998). Mesuring heterogeneity in foresic database using hierachical Bayes model. *Biometrika*, 85: 269-287.

Rust, K., & Rao, J. (1996). Variance estimatio for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5: 283-310.

She, D., Zhang, H., & Li, Z. (2009). Testing Hardy-Weinberg equilibrium using family data from complex surveys. *Hum Genet*, 73:449–455.

Shoemaker, J., Painter, I., & Weir, B. (1998). A Bayesian Characterization of Hardy-Weinberg Disequilibrium. *Genetics*, 149: 2079–2088.

Shoemaker, J., Painter, I., & Weir, B. (1998). A Bayesian Characterization of Hardy-Weinberg Disequilibrium. *Genetics*, 149: 2079–2088.

Si, Y., Pillai, N., & Gelman, A. (2015). Bayesian Nonparametric Weighted Sampling Inference. *Bayesian Analysis*, 1(1): 1-21.

Sitter, R., & Wu, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of American Statistical Association*, (97):535-543.

Skinner, C. J., Holt, D., & Smith, T. (1989). *Analysis of Complex Surveys.* New York: Wiley.

Steinberg, K., & et.al. (2001). Prevalence of C282Y and H63D mutations in the hemochromatosis (HFE) gene in the United States. *JAMA*, 285(17):2216-22.

Tata, F., Henry, I., AF, M., SC, W., D, W., KH, G., . . . SE, H. (1985). Isolation and characterisation of a cDNA clone for human apolipoprotein CI and assignment of the gene to chromosome 19. *Hum. Genet.*, 69 (4): 345.

University of Washington. (2012, March 5). *Quality Control Report for Genotypic Data.* Retrieved from http://hrsonline.isr.umich.edu/sitedocs/genetics/HRS_QC_REPORT_MAR2012.pdf

Wakefield, J. (2010). Bayesian Methods for Examining Hardy–Weinberg Equilibrium. *Biometrics*, 66(1):257-65.

Weir, B. (2010). Statistical genetic issues for genome-wide association studies. *Genome*, 53: 869-875.

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

Weir, B., & Hill, W. (2002). Estimating F-statsitics. *Annu Rev Genet*, 36:721-50.

Weir, B., Hill, W., & Cardon, L. (2004). Allelic association patterns for a dense SNP map. *Genet. Epidemiol.*, 27(4): 442-450.

Weir, D. R. (2012, March 5). *Quality Control Report for Genotypic Data.* Retrieved from http://hrsonline.isr.umich.edu/sitedocs/genetics/HRS_QC_REPORT_MAR2012.pdf

Wheeler, D. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 35(Database issue): D5–D12. Retrieved from National Institute of Health (NIH): www.ncbi.nlm.nih.gov/snp

Wolf, A., Caselli, R., Reiman, E., & Valla, J. (2013). APOE and neuroenergetics: an emerging paradigm in Alzheimer's disease. *Neurobiol Aging*, 34(4):1007-17.

Wright, S. (1921). Systems of Mating. V. General Considerations. *Genetics*, 6(2): 167–178.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to system of mating. *Evolution*, 19: 395–420.

Wright, S. (1969). *The Theory of Gene Frequencies.* Chicago, IL: The University of Chicago Press.

Wu, C., & Rao, J. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, (34): 359-375.

Yang, B., Qin, G., & Qin, J. (2011). Empirical likelihood-based inferences for a low income proportion. *The Canadian Journal of Statistics*, (39): 1-16.

Yang, Q., & et.al. (2010). Racial/ethnic differences in association of fasting glucose-associated genomic loci with fasting glucose, HOMA-B, and impaired fasting glucose in the U.S. adult population. *Diabetes Care*, 33(11):2370-7.

Zhang, L., & et.al. (2013). Association of functional polymorphism rs2231142 (Q141K) in the ABCG2 gene with serum uric acid and gout in 4 US populations: the PAGE Study. *Am J Epidemiol.*, 177(9):923-32.

Zhou, M. (2005). Empirical likelihood ratio with arbitrary censored/truncated data by EM algorithm. *Journal of Computational and Graphical Statistics*, 643-656.

Zuccato, C. (2009). Brain-derived neurotrophic factor in neurodegenerative diseases. *Nat Rev Neurol*, 5 (6): 311–22.