THE APPROXIMATE VARIANCE OF CORRELATION MEASURES OF LINKAGE DISEQUILIBRIUM

by

Mary Roop

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science

at

Dalhousie University
Halifax, Nova Scotia
April 2016

To my parents David and Josephine for all the kind, patient love and support they have forever provided me. To my sister Liia and the horse loves of our youth; To Lance.

Table of Contents

List of Tables

List of Figures

Abstract

A delta method approximation is utilized to produce the variance formula of the correlation measure of Linkage Disequilibrium for four types of genetic data. These data types include gametic and genotypic counts, with different assumptions used to simplify the analysis of the genotypic counts. In each case the variance formula is derived and plotted for several choices of the allele frequencies and for all feasible values of ρ. Simulations are carried out to compare the variance of simulated Linkage Disequilibrium correlation values to the theoretical variance formula. Results indicate that the variance formulae are good approximations for each type of genetic data. Fisher's transformation improved the approximation in some cases.

List of Abbreviations and Symbols Used

D'_{AB}	Standardized linkage disequilibrium measure
D_{AB}	Linkage disequilibrium measure
Δ'_{AB}	Standardized linkage disequilibrium measure for composite data
Δ_{AB}	Linkage disequilibrium measure for composite data
r_C	Correlation measure, composite data
r_g	Correlation measure, gametic data
r_{pk}	Correlation measure, phase known genotypic data
r_{rm}	Correlation measure, random mating genotypic data
HWD	Hardy-Weinberg disequilibrium
HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
MLE	Maximum likelihood estimate

Chapter 1

Introduction

This thesis is concerned with the development of variance formulae for measures of association in genetics. To assist with understanding the genetic terminology, some definitions are given in Table 1.1.

1.1 Linkage Disequilibrium

Linkage disequilibrium (LD), also referred to as gametic disequilibrium, is the statistical association between the alleles at two genetic loci. LD occurs when the presence of a particular allele at one locus affects the probability of an allele at a second locus. This thesis is concerned with finding variance formulae for correlation measures of LD.

LD can exist because of a number of factors, as described in Weir (1996). LD may occur due to random factors in reproduction across time when relative frequencies of different genotypes will be affected by chance disappearance of other genotypes within the population *(genetic drift)*. A population may be founded by a small set of individuals in which two alleles frequently occur together. Their descendants will also exhibit this allelic association *(founder effects)*. Paticular genes may allow subjects to better adapt and thrive in unique and evolving environments. These genes allow for better survival rates and therefore are more prominent in future generations *(biological selection)*. If a population consists of two or more subpopulations with different allele frequencies, then the overall population can exhibit LD even if each subpopulation has no LD *(admixture)*. A *mutation* at one locus will occur with a particular allele at a nearby locus and this haplotype is passed on largely intact to future generations, which exhibit strong LD. Recombination erodes the initial haplotype combination and eventually reduces the LD.

Two approaches have been used to find the location of genes which predispose

Term	Definition
Linkage Disequilibrium	The statistical association between the alleles at two genetic loci.
gene	An inherited sequence of nucleotides that form chromosomes and determine an individual's genetic make up.
allele	A unique variation of a gene.
locus	The specific location or position of an allele on a chromosome.
haplotype	The composition of alleles on a single chromosome.
genotype	The composition of alleles resulting from sexual reproduction.
phase	The alleles that occur together on the two chromosomes in genotypic data.
heterozygote	The presence of two different alleles on a particular gene, which may or may not affect the variation of inherited traits.
gamete	A fully developed haploid cell capable of uniting with another haploid cell for the process of sexual reproduction.
in-replusion	Double heterozygote alleles occur in pairing as Ab/aB.
in-coupling	Double heterozygote alleles occur in pairing as AB/ab.
chromosome	The genetic make-up of a cell.
random mating	Mating under the assumption that the genotypic probabilities are a product of the haplotype probabilities.
Hardy-Weinberg equilibrium	The result of mating where there are no effects of digenic disequilibrium and allele frequencies will remain constant form one generation to the next.
Hardy-Weinberg disequilibrium	Mating under significant effects of digenic disequilibrium, which will affect allele frequencies from one generation to the next.

Table 1.1: Definitions of some genetic terminology.

subjects to disease. Linkage analysis uses pedigree data to determine the recombination rate, and therefore genetic distance, between a disease susceptibility locus, and a set of known marker loci. In this approach the actual alleles present at the loci are unimportant, only the recombination rate or distance between them. With LD mapping, the strength of association is determined between the disease status and the alleles at a set of loci. A large association is taken to imply proximity. Interest in LD has risen with the understanding that there is greater power for mapping common disease genes with association studies than for traditional linkage studies (Ardlie et al., 2002; Risch and Merikangas, 1996). Genome wide association studies typically look for LD at a very large number of single nucleotide polymorphisms, and have been widely used in recent years.

1.2 Measures of Linkage Disequilibrium

This thesis considers the case of two bi-allelic loci, locus $\mathcal{A}$ with its alleles A and a, with probabilities p_A and $p_a = 1 - p_A$, and the locus $\mathcal{B}$ with alleles B and b, with probabilities p_B and $p_b = 1 - p_B$.

Lewontin's 1964 measure of LD

$$D_{AB} = p_{AB} - p_A p_B \tag{1.1}$$

is the difference between the joint probability of alleles A and B (p_{AB}) and the product of the individual probabilities (p_A and p_B), and therefore is a measure of the alleles' divergence from statistical independence (Lewontin, 1964). Table 1.2 relates the gametic joint probabilities to the value of D_{AB} and the probabilities of the product of the alleles A and B.

	B	b	Total
A	$p_{AB} = p_A p_B + D_{AB}$	$p_{Ab} = p_A p_b - D_{AB}$	p_A
a	$p_{aB} = p_a p_B - D_{AB}$	$p_{ab} = p_a p_b + D_{AB}$	p_a
Total	p_B	p_b	1

Table 1.2: The gametic probabilities.

The range of D_{AB} depends on the allele probabilities. Since all of the probabilities in Table 1.2 must be positive, D_{AB} is bounded by

$[max(-p_A p_B, -p_a p_b), min(p_A p_b, p_a p_B)]$, or $D_{ABmin} = max(-p_A p_B, -p_a p_b)$ and $D_{ABmax} = min(p_A p_b, p_a p_B)$.

The greatest range D_{AB} occurs with $p_A = p_B = 0.5$ when $D_{ABmin} = -0.25$ and $D_{ABmax} = 0.25$. The maximum D_{AB} occurs with $p_{AB} = p_A = 0.5$ and the minimum occurs with $p_{AB} = 0$.

To facilitate comparison between two populations with differing allele frequencies, Lewontin (1964) suggests standardizing the value of D_{AB} to D'_{AB}, where

$$D'_{AB} = \begin{cases} D_{AB}/D_{ABmax}, & \textit{if } D_{AB} > 0 \\ D_{AB}/|D_{ABmin}|, & \textit{if } D_{AB} \leq 0. \end{cases}$$

The transformed D'_{AB} is a value between -1 and 1.

A second standardization of D_{AB} is the correlation measure. If we define binary random variables X and Y to be indicators of the major alleles A and B, then $E(XY) = p_{AB}$, $E(X) = p_A$ and $E(Y) = p_B$, so D_{AB} is the covariance between X and Y. Then, using the facts that $Var(X) = p_A(1 - p_A)$ and $Var(Y) = p_B(1 - p_B)$, the correlation is

$$\rho = \frac{D_{AB}}{\sqrt{p_A p_a p_B p_b}}.$$

The range of this measure is smaller than (-1,1) unless $p_A = p_B = 0.5$.

Weir (1996) provides a composite measure of LD to be used with genotypic data. The composite measure is the sum

$$\Delta_{AB} = D_{AB} + D_{A/B} \tag{1.2}$$

of the gametic disequilibrium, D_{AB}, and the non-gametic disequilibrium,

$$D_{A/B} = p_{A/B} - p_A p_B, \tag{1.3}$$

which measures departure from independence of the A allele on one chromosome to the B allele on the other.

The composite measure can be standardized as

$$\Delta' = \begin{cases} \Delta/\Delta_{max} \\ \Delta/|\Delta_{min}| \end{cases}$$

where the bounds Δ_{min} and Δ_{max} are provided in Hamilton and Cole (2004).

To obtain the correlation standardization for the composite measure, define the random variables X and Y to have values

$$X = \begin{cases} -1 & \text{if genotype } AA \\ 0 & \text{if genotype } Aa \\ 1 & \text{if genotype } aa \end{cases} \qquad Y = \begin{cases} -1 & \text{if genotype } BB \\ 0 & \text{if genotype } Bb \\ 1 & \text{if genotype } bb \end{cases}$$

Zaykin (2004) showed that

$$\Delta = \frac{Cov(X,Y)}{2}.$$

Using the facts that

$$Var(X) = 2[p_A p_a + D_A] \;\text{ and }\; Var(Y) = 2[p_B p_b + D_B].$$

It follows that

$$\rho_C = Cor(X,Y) = \frac{\Delta}{\sqrt{(p_A p_a + D_A)(p_B p_b + D_B)}}.$$

where the coefficients

$$D_A = P_A^A - p_A^2 \;\text{ and }\; D_B = P_B^B - p_B^2$$

are measures of Hardy-Weinburg (HWD) disequilibrium, with P_A^A the probability of allele A on both chromosomes and P_B^B the probability of allele B on both chromosomes.

For two bi-allelic loci, $\mathcal{A}$ and $\mathcal{B}$, there are ten different combinations of alleles from the two separate loci resulting in the following genotypes: ABAB, ABAb, AbAb, ABaB, ABab, AbaB, Abab, aBaB, aBab, and abab. For eight of these ten genotypes the *phase* is clear, that is which alleles occur together on the two chromosomes. Note that the double heterozygotes ABab and AbaB are comprised of the same component alleles but in different pairings, the *in-coupling phase* (AB/ab) and the *in-repulsion phase* (Ab/aB). *In-coupling* indicates that A and B occur together on the same chromosome, whereas *in-repulsion* implies A and B are on different chromosomes. It is usually not possible to determine the *phase* of the double heterozygotes, that is whether they are *in-coupling* or *in-repulsion*, unless genotypic data from relatives can provide *phase* determination.

The genotypic probabilities are shown in Table 1.3 with the middle cell containing both types of double heterozygotes. For example, P_{aB}^{Ab} indicates that

	BB	Bb	bb	Total
AA	P^{AB}_{AB}	$2P^{AB}_{Ab}$	P^{Ab}_{Ab}	P^{A}_{A}
Aa	$2P^{AB}_{aB}$	$2P^{AB}_{ab} + 2P^{Ab}_{aB}$	$2P^{Ab}_{ab}$	P^{A}_{a}
aa	P^{aB}_{aB}	$2P^{aB}_{ab}$	P^{ab}_{ab}	P^{a}_{a}
Total	P^{B}_{B}	P^{B}_{b}	P^{b}_{b}	1

Table 1.3: Genotypic probabilities.

alleles Ab occur together on one chromosome and aB occur on the other. By convention, probabilities for genotypes with different haplotypes are multiplied by two to account for the two possible parentages.

The AB haplotype probability is

$$p_{AB} = P^{AB}_{AB} + P^{AB}_{Ab} + P^{AB}_{aB} + P^{AB}_{ab} \tag{1.4}$$

which depends on the *in-coupling* double heterozygote probability P^{AB}_{ab}. This term cannot be estimated when only the total number of double heterzygotes is observed. Similarly the probability of allele A on one chromosome and B on the other is

$$p_{A/B} = P^{AB}_{AB} + P^{AB}_{Ab} + P^{AB}_{aB} + P^{aB}_{Ab} \tag{1.5}$$

which depends on the *in-repulsion* double heterozygote probability P^{Ab}_{aB}. The composite measure is based on the sum

$$\begin{aligned}\Delta &= p_{AB} + p_{A/B} - 2p_A p_B \\ &= 2P^{AB}_{AB} + P^{AB}_{Ab} + P^{AB}_{aB} + \frac{1}{2}(P^{AB}_{ab} + P^{Ab}_{aB}) - 2p_A p_B\end{aligned}$$

and the term in parentheses is the total probability of double heterozygotes, which can be estimated by the observed proportion. Note that $p_A = P^A_A + P^A_a$ and $p_B = P^B_B + P^B_b$ so the composite measure can be estimated from genotypic counts as described below.

Although a number of other LD measures exist, we focus on the standardized versions of Lewontin's (1964) D_{AB} measure.

1.3 Estimates of LD Measures

Estimates of LD measures are obtained in four scenarios. The first is when we have counts of a random sample of gametes. The remaining cases are for data in the form of genotypic counts when various assumptions are made.

1.3.1 Using Gametic Counts

Gametic count data is the most basic genetic data case for assessing potential allelic association between two loci. Gametic data may be available if individual chromosomes have been sampled from a population or if applicable haplotypes can be inferred from family members' genotypic information and random mating is assumed. Table 1.4 displays gametic counts for the two bi-allelic loci A and B.

	B	b	Total
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
Total	n_B	n_b	N

Table 1.4: The gametic counts for two bi-allelic loci $\mathcal{A}$ and $\mathcal{B}$.

The multinomial probabilities that apply to the gametic case are displayed in Table 1.2. The haplotype probabilities are estimated by their observed proportions, so $\hat{p}_{AB} = n_{AB}/N$. Allelic probabilities are the sum of two haplotype probabilities, so $\hat{p}_A = \hat{p}_{AB} + \hat{p}_{Ab}$, and $\hat{p}_B = \hat{p}_{AB} + \hat{p}_{aB}$. The maximum likelihood estimate of D_{AB} with gametic data is therefore

$$\hat{D}_{AB,g} = \hat{p}_{AB} - \hat{p}_A\hat{p}_B.$$

Similarly, estimates of the standardized measures are

$$\hat{D}'_{AB,g} = \begin{cases} \frac{\hat{D}_{AB,g}}{\hat{D}_{ABmax}}, & \textit{if } \hat{D}_{AB,g} > 0 \\ \frac{\hat{D}_{AB,g}}{|\hat{D}_{ABmin}|}, & \textit{if } \hat{D}_{AB,g} \leq 0 \end{cases}$$

where $\hat{D}_{ABmax}$ and $\hat{D}_{ABmin}$ are obtained by substituting observed proportions for probabilities, and the correlation ρ is estimated by

$$r_g = \frac{\hat{D}_{AB}}{\sqrt{\hat{p}_A\hat{p}_a\hat{p}_B\hat{p}_b}}.$$

1.3.2 Using Genotypic Counts

Information on gametes is not readily available in most instances. Instead, data are often provided in the form of genotypic counts. The data usually appear as genotypic counts as shown in Table 1.5.

Two assumptions can be made to simplify the analysis of these counts. These are the assumptions of random mating and of knowledge of phase. Without either assumption D_{AB} cannot be estimated and the composite measure Δ_{AB} must be used.

Random Mating Assumption

Under this assumption the genotypic probabilities are a product of the haplotype probabilities. For example, $P^{AB}_{ab} = p_{AB}p_{ab}$.

The data consists of nine genotypic counts, with the indistinguishable double heterozygotes counted together, as in Table 1.5. These counts are assumed to be multinomially distributed, with index n equal to the number of subjects and probabilities as shown in Table 1.3.

	BB	Bb	bb	Total
AA	n^{AB}_{AB}	n^{AB}_{Ab}	n^{Ab}_{Ab}	n^{A}_{A}
Aa	n^{AB}_{aB}	n_{AaBb}	n^{Ab}_{ab}	n^{A}_{a}
aa	n^{aB}_{aB}	n^{aB}_{ab}	n^{ab}_{ab}	n^{a}_{a}
Total	n^{B}_{B}	n^{B}_{b}	n^{b}_{b}	n

Table 1.5: Genotypic counts, with phase unknown.

Using the facts that $p_A = p_{AB} + p_{Ab}$ and $p_B = p_{AB} + p_{aB}$ it is possible to write all the genotypic probabilities in terms of p_A, p_B, and p_{AB} and obtain the log likelihood as

$$\begin{aligned} l(p_A, p_B, p_{AB}) &= 2n_{AB}^{AB} log(p_{AB}) + n_{Ab}^{AB}[log(p_{AB}) + log(p_{Ab})] + 2n_{Ab}^{Ab} log(p_{Ab}) \\ &\quad + n_{aB}^{AB}[log(p_{AB}) + log(p_{aB})] + n_{AaBb} log(p_{AB}p_{ab} + p_{Ab}p_{aB}) \\ &\quad + n_{ab}^{Ab}[log(p_{Ab}) + log(p_{ab})] + 2n_{aB}^{aB} log(p_{aB}) \\ &\quad + n_{ab}^{aB}[log(p_{aB}) + log(p_{ab})] + 2n_{ab}^{ab} log(p_{ab}) \\ &= [2n_{AB}^{AB} + n_{Ab}^{AB} + n_{aB}^{AB}] log(p_{AB}) + [n_{Ab}^{AB} + 2n_{Ab}^{Ab} + n_{ab}^{Ab}] log(p_{Ab}) \\ &\quad + [n_{aB}^{AB} + 2n_{aB}^{aB} + n_{ab}^{aB}] log(p_{aB}) + [n_{ab}^{Ab} + n_{ab}^{aB} + 2n_{ab}^{ab}] log(p_{ab}) \\ &\quad + n_{AaBb} log(p_{AB}p_{ab} + p_{Ab}p_{aB}) \\ &= [2n_{AB}^{AB} + n_{Ab}^{AB} + n_{aB}^{AB}] log(p_{AB}) + [n_{Ab}^{AB} + 2n_{Ab}^{Ab} + n_{ab}^{Ab}] log(p_A - p_{AB}) \\ &\quad + [n_{aB}^{AB} + 2n_{aB}^{aB} + n_{ab}^{aB}] log(p_B - p_{AB}) \\ &\quad + [n_{ab}^{Ab} + n_{ab}^{aB} + 2n_{ab}^{ab}] log(1 - p_A - p_B - p_{ab}) \\ &\quad + n_{AaBb} log(p_{AB}(1 - 2p_A - 2p_B) + p_{AB}^2 + p_A p_B). \end{aligned} \tag{1.6}$$

Once the maximum likelihood estimates (MLEs) for p_A, p_B and p_{AB} are obtained numerically, the LD measures are estimated as before,

$$\hat{D}_{AB,rm} = \hat{p}_{AB} - \hat{p}_A\hat{p}_B, \tag{1.7}$$

$$\hat{D}'_{AB,rm} = \begin{cases} \frac{\hat{D}_{AB,rm}}{\hat{D}_{ABmax}}, & \textit{if } \hat{D}_{AB,rm} > 0 \\ \frac{\hat{D}_{AB,rm}}{|\hat{D}_{ABmin}|}, & \textit{if } \hat{D}_{AB,rm} \leq 0 \end{cases}$$

and

$$r_{rm} = \frac{\hat{D}_{AB,rm}}{\sqrt{\hat{p}_A\hat{p}_a\hat{p}_B\hat{p}_b}}. \tag{1.8}$$

Phase Known Assumption

For phase known genotypic data the phase of the double heterzygotes is known to be either AB/ab, *in-coupling*, or Ab/aB, *in-repulsion*.

There are ten genotypes with counts as shown in Table 1.6. The counts are multinomially distributed with the genotypic probabilities listed in Table 1.7.

<table>
<tr><th></th><th>BB</th><th colspan="2">Bb</th><th>bb</th><th>Total</th></tr>
<tr><td>AA</td><td>n_{AB}^{AB}</td><td colspan="2">n_{Ab}^{AB}</td><td>n_{Ab}^{Ab}</td><td>n_{A}^{A}</td></tr>
<tr><td>Aa</td><td>n_{aB}^{AB}</td><td>n_{ab}^{AB}</td><td>n_{aB}^{Ab}</td><td>n_{ab}^{Ab}</td><td>n_{a}^{A}</td></tr>
<tr><td>aa</td><td>n_{aB}^{aB}</td><td colspan="2">n_{ab}^{aB}</td><td>n_{ab}^{ab}</td><td>n_{a}^{a}</td></tr>
<tr><td>Total</td><td>n_{B}^{B}</td><td colspan="2">n_{b}^{B}</td><td>n_{b}^{b}</td><td>n</td></tr>
</table>

Table 1.6: Genotypic counts when the phase of the double heterozygotes is known.

<table>
<tr><th></th><th>BB</th><th colspan="2">Bb</th><th>bb</th><th>Total</th></tr>
<tr><td>AA</td><td>P_{AB}^{AB}</td><td colspan="2">$2P_{Ab}^{AB}$</td><td>P_{Ab}^{Ab}</td><td>P_{A}^{A}</td></tr>
<tr><td>Aa</td><td>$2P_{aB}^{AB}$</td><td>$2P_{ab}^{AB}$</td><td>$2P_{aB}^{Ab}$</td><td>$2P_{ab}^{Ab}$</td><td>P_{a}^{A}</td></tr>
<tr><td>aa</td><td>P_{aB}^{aB}</td><td colspan="2">$2P_{ab}^{aB}$</td><td>P_{ab}^{ab}</td><td>P_{a}^{a}</td></tr>
<tr><td>Total</td><td>P_{B}^{B}</td><td colspan="2">P_{b}^{B}</td><td>P_{b}^{b}</td><td>1</td></tr>
</table>

Table 1.7: Genotypic probabilities when the phase of the double heterozygotes is known.

The genotypic probabilities are estimated using observed proportions, and the haplotype probability p_{AB} is estimated in the sum

$$\hat{p}_{AB} = \hat{P}_{AB}^{AB} + \hat{P}_{Ab}^{AB} + \hat{P}_{aB}^{AB} + \hat{P}_{ab}^{AB}$$

.

Similar expressions follow for $\hat{p}_{Ab}$, $\hat{p}_{aB}$, and $\hat{p}_{ab}$

$$\hat{p}_{Ab} = \hat{P}_{Ab}^{Ab} + \hat{P}_{Ab}^{AB} + \hat{P}_{aB}^{Ab} + \hat{P}_{ab}^{Ab}$$

$$\hat{p}_{aB} = \hat{P}_{aB}^{aB} + \hat{P}_{aB}^{AB} + \hat{P}_{aB}^{Ab} + \hat{P}_{ab}^{aB}$$

$$\hat{p}_{ab} = \hat{P}_{ab}^{ab} + \hat{P}_{ab}^{AB} + \hat{P}_{ab}^{Ab} + \hat{P}_{ab}^{aB}$$

allowing for the calculation of estimates of p_A, and p_B using $\hat{p}_A = \hat{p}_{AB} + \hat{p}_{Ab}$ and $\hat{p}_B = \hat{p}_{AB} + \hat{p}_{aB}$. The LD estimate is

$$\hat{D}_{AB,pk} = \hat{p}_{AB} - \hat{p}_A\hat{p}_B,$$

and the standardized measures are

$$\hat{D}'_{AB,pk} = \begin{cases} \frac{\hat{D}_{AB,pk}}{\hat{D}_{ABmax}}, & \textit{if } \hat{D}_{AB,pk} > 0 \\ \frac{\hat{D}_{AB,pk}}{|\hat{D}_{ABmin}|}, & \textit{if } \hat{D}_{AB,pk} \leq 0 \end{cases}$$

and

$$r_{pk} = \frac{\hat{D}_{AB,pk}}{\sqrt{\hat{p}_A\hat{p}_a\hat{p}_B\hat{p}_b}}. \tag{1.9}$$

Estimating the Composite Measure of LD

When neither assumptions from the previous two cases can be made we use the composite measure (1.2). These genotypic counts in Table 1.5 are multinomially distributed with index n equal to the number of subjects and the probabilities listed in Table 1.5. The genotype probabilities are estimated by the observed proportions.

The composite measure of LD is estimated by

$$\begin{aligned} \hat{\Delta} &= \hat{p}_{AB} + \hat{p}_{A/B} - 2\hat{p}_A\hat{p}_B \\ &= 2\hat{P}^{AB}_{AB} + \hat{P}^{AB}_{Ab} + \hat{P}^{AB}_{aB} + \frac{1}{2}(\widehat{P^{AB}_{ab} + P^{Ab}_{aB}}) - 2\hat{p}_A\hat{p}_B \end{aligned}$$

The term in parenthesis is (one half) the total double heterozygote probability, which is estimated using the observed proportion of double heterozygotes.

The HWD coefficients are estimated as

$$\hat{D}_A = \hat{P}^A_A + \hat{p}^2_A \text{ and } \hat{D}_B = \hat{P}^B_B + \hat{p}^2_B$$

so the standardized measures are estimated as

$$\hat{\Delta}' = \begin{cases} \hat{\Delta}/\hat{\Delta}_{max} \\ \hat{\Delta}/|\hat{\Delta}_{min}|, \end{cases}$$

and

$$\hat{r}_C = \frac{\hat{\Delta}}{\sqrt{(\hat{p}_A\hat{p}_a + \hat{D}_A)(\hat{p}_B\hat{p}_b + \hat{D}_B)}}. \tag{1.10}$$

1.4 Approximate Variance for LD Measures

Approximate variance formulas for the $\hat{D}_{AB}$ measures of LD can be found in the literature. All formulae for the four cases of genetic data are collected here before continuing with the approximate variance formulae for the correlation LD estimates in the following chapters. The following notation is used for simplification: $\pi_A = p_A p_a$, $\pi_B = p_B p_b$, $\tau_A = 1 - 2p_A$ and $\tau_B = 1 - 2p_B$.

Brown (1975) and Weir (1996) both give the variance formula for the $\hat{D}_{AB,g}$ estimate of LD for gametic data

$$V(\hat{D}_{AB,g}) = (\pi_A \pi_B + D_{AB}\tau_A\tau_B - D_{AB}^2)/N. \tag{1.11}$$

Brown (1975) provides the variance of $\hat{D}_{AB,rm}$ for the case of genotypic data with the assumption of random mating

$$V(D_{AB,rm}) = \frac{8n^3(det(I))^{-1} + D_{AB}^2[\pi_A(p_b - p_B)^2 + \pi_B(p_a - p_A)^2] - 2D_{AB}^3(p_b - p_B)(p_a - p_A)}{2n(\pi_A\pi_B - D_{AB}^2)} \tag{1.12}$$

where I is the information matrix for p_{AB}, p_A and p_B, and

$$det(I) = \frac{8n^3[p_{AB}p_{ab}(p_{AB} + p_{ab}) + p_{Ab}p_{aB}(p_{Ab} + p_{aB})]}{(p_{AB}p_{ab} + p_{Ab}p_{aB})p_{AB}p_{Ab}p_{aB}p_{ab}}.$$

Weir and Cockerham (1989) give the approximate variance of $\hat{D}_{AB,pk}$ for genotypic data with the phase known assumption

$$V(\hat{D}_{AB,pk}) = [\pi_A\pi_B + \tau_A\tau_B D_{AB} + D_A D_B - D_{AB}^2 + D_{AB}^2 + D_{AB}^{AB}]/2n. \tag{1.13}$$

Weir and Cockerham (1989) provide the approximate variance formula for the estimate of the composite measure

$$V(\hat{\Delta}) = [(\pi_A + D_A)(\pi_B + D_B) + \tau_A\tau_B\Delta_{AB}/2 + \tau_A D_{ABB} + \tau_B D_{AAB} + \Delta_{AABB}]/n.$$

Zapata (1997) derived the approximate variance for the standardized measure of LD, D'_{AB},

$$V(\hat{D}'_{AB}) = \begin{cases} 0 & if \quad D_{AB} = -1 \\ X & if \quad -1 < D_{AB} \leq 0 \\ Y & if \quad 0 < D_{AB} < 1 \\ 0 & if \quad D_{AB} = 1 \end{cases}$$

$$X = \left[\frac{1}{n(|D_{ABmin}|)^2}\right]\Big[(1-|D'_{AB}|)\{nVar(D_{AB}) - |D'_{AB}|\;|D_{ABmin}|(p_Ap_B + p_ap_b - 2|D_{AB}|)\} \\ +|D'_{AB}|x_i(1-x_i)\Big]$$

where $x_i = p_Ap_B$, or p_ap_b when $|D_{ABmin}|$ is p_Ap_B or p_ap_b, and

$$Y = \left[\frac{1}{n(D_{ABmax})^2}\right]\Big[(1-|D'_{AB}|)\{nVar(D_{AB}) - |D'_{AB}|D_{ABmax}(p_Ap_b + p_ap_B - 2|D_{AB}|)\} \\ +|D'_{AB}|y_i(1-y_i)\Big]$$

where $y_i = p_Ap_b$, or p_ap_B when D_{ABmax} is p_Ap_b or p_ap_B.

Hamilton et al. (2006) give an approximate variance for the standardized composite measure displayed in Tables 1.8, 1.9, and 1.10. Their variance formula was given for three cases: I ($P^A_A \leq P^B_B$, $P^a_a \geq P^b_b$, and $P^a_a + P^B_B \leq 1$), II ($P^A_A \leq P^B_B$, $P^a_a \geq P^b_b$, and $P^a_a + P^B_B > 1$), and III ($P^A_A \leq P^B_B$ and $P^a_a \leq P^a_a$). The three cases are based on the bounds for Δ' used in the standardization of the composite measure.

Many of these variance formulae depend on higher order disequilibrium coefficients D_{AAB}, D_{ABB}, D^{AB}_{AB} and Δ_{AABB} (Weir, 1996).

1.5 Outline of Thesis

The purpose of this thesis is to obtain the approximate variance of the various estimates of the correlation measure of LD for the different types of data assumptions. The gametic case will be covered in Chapter 2, while the genotypic data, in each of the cases: (1) assuming random mating, (2) assuming phase known, and (3) composite measure will be covered in Chapters 3, 4 and 5 respectively. Finally, Chapter 6 will consist of comparisons across the different cases and applications to real world data.

Term	Coefficient
Denominator	$n\Delta_{max}^2 = n[2p_A p_b - .5(1 - P_a^a - P_B^B)]^2$
Numerator	
1	$(\pi_A + D_A)(\pi_B + D_B)$
Δ'	$\tau_A \tau_B \Delta_{max}/2$
D_{AAB}	τ_B
D_{ABB}	τ_A
Δ_{AABB}	1
$(\Delta')^2$	X
$\Delta' D_{AAB}$	$2(p_A - 2p_b)$
$\Delta' D_{ABB}$	$2(p_B - \tau_A)$
$(\Delta')^3$	$(p_B - 2p_A)(p_A + \tau_B)\Delta_{max}$
$(\Delta')^2 D_{AAB}$	$p_B - 2p_A$
$(\Delta')^2 D_{ABB}$	$p_A + \tau_B$
$(\Delta')^2 \Delta_{AABB}$	$1/2$
$(\Delta')^4$	$\Delta_{max}/2$
X	$(5p_A^4 - 32p_A^3 p_B + 60p_A^2 p_B^2 - 32p_A p_B^3 + 5p_B^4 + 12p_A^3$ $-40p_A^2 p_B + 16p_A^2 D_B + 6p_A^2 D_A + 16p_A p_B^2 - 16p_A p_B D_B$ $-16p_A p_B D_A + 6p_B^2 D_B + 16p_B^2 D_A + 9p_A^2 + 4p_A D_A - p_B^2$ $-16p_B D_A + D_B^2 + 4D_A D_B + D_A^2 - 2p_A - D_B + 5D_A)/4$

Table 1.8: Approximate variance formula for $\hat{\Delta}'$, case I.

Term	Coefficient
Denominator	$n\Delta_{max}^2 = n(2p_A p_b)^2$
Numerator	
1	$(\pi_A + D_A)(\pi_B + D_B)$
Δ'	$\tau_A \tau_B \Delta_{max}/2$
D_{AAB}	τ_B
D_{ABB}	τ_A
Δ_{AABB}	1
$(\Delta')^2$	X
$\Delta' D_{AAB}$	$-4p_b/\Delta_{max}$
$\Delta' D_{ABB}$	$4p_A/\Delta_{max}$
$(\Delta')^3$	$-2\Delta_{max}$
$(\Delta')^2 D_{AAB}$	0
$(\Delta')^2 D_{ABB}$	0
$(\Delta')^2 \Delta_{AABB}$	0
$(\Delta')^4$	0
X	$2[p_A^2(5p_b^2 - p_b p_B + D_B) - p_b^2(p_A - D_B)]$

Table 1.9: Approximate variance formula for $\hat{\Delta}'$, case II.

Term	Coefficient
Denominator	$n\Delta_{max}^2 = n[2p_a p_B - (p_B - P_A^A)]^2$
Numerator	
1	$(\pi_A + D_A)(\pi_B + D_B)$
Δ'	$\tau_A \tau_B \Delta_{max}/2$
D_{AAB}	τ_B
D_{ABB}	τ_A
Δ_{AABB}	1
$(\Delta')^2$	X
$\Delta' D_{AAB}$	$4(p_A + p_B - 1)$
$\Delta' D_{ABB}$	$-2\tau_A$
$(\Delta')^3$	$2(p_A - p_B)\tau_A \Delta_{max}$
$(\Delta')^2 D_{AAB}$	$2\tau_A$
$(\Delta')^2 D_{ABB}$	0
$(\Delta')^2 \Delta_{AABB}$	0
$(\Delta')^4$	0
X	$(10p_A^4 - 32p_A^3 p_B + 24p_A^2 p_B^2 - 4p_A^3 + 24p_A^2 p_B + 4p_A^2 D_B$ $+12p_A^2 D_A - 24p_A p_B^2 - 16p_A p_B D_A + 4p_B^2 D_A - 4p_A p_B - 4p_A D_B$ $-4p_A D_A + 7p_B^2 + 4p_B D_A + 2D_A^2 - p_B + D_B)/2$

Table 1.10: Approximate variance formula for $\hat{\Delta}'$, case III.

Chapter 2

Approximate Variance for the Correlation Measure of LD Using Gametic Data

2.1 Variance Approximation Using the Delta Method

The delta method is used to calculate an asymptotically approximate variance of the correlation measure of LD for gametic data

$$r_g = \frac{\hat{D}_{AB}}{\sqrt{\hat{p}_A \hat{p}_a \hat{p}_B \hat{p}_b}}.$$

The correlation coefficient estimate is a function of the random vector

$$\mathbf{x} = (x_1, \ldots, x_4)^T = (n_{AB}, n_{Ab}, n_{aB}, n_{ab})^T$$

consisting of the 4 counts from Table 1.4.

These counts follow a multinomial distribution with index N, the number of gametes, and the probabilities

$$\mathbf{p} = (p_1, \ldots, p_4)^T = (p_{AB}, p_{Ab}, p_{aB}, p_{ab})^T.$$

The variance covariance matrix $\mathbf{V}$ is a 4×4 matrix composed of diagonal elements

$$V_{ii} = Cov(X_i, X_i) = Var(X) = Np_i(1 - p_i)$$

and off diagonal elements,

$$V_{ij} = Cov(X_i, X_j) = -Np_i p_j.$$

The variance approximation is calculated as

$$Var(r) \approx \mathbf{g}^T \mathbf{V} \mathbf{g}, \tag{2.1}$$

where

$$\mathbf{g} = (g_1, \ldots, g_4)^T$$

and

$$g_i = \left.\frac{\partial r_g}{\partial x_i}\right|_{N\mathbf{p}}$$

is the derivative of the correlation estimate with respect to the ith gametic count evaluated at the mean.

Maple software was utilized for all symbolic variance approximation calculations. The gradient $\mathbf{g}$ was constructed with the *diff* command calculating the partial derivatives of r_g with respect to each gametic count x_i. These derivatives were then evaluated at the expected values of the counts by replacing each occurrence of x_i by $Np_i, i = 1, \ldots, 4$, with the gametic probabilities expressed in terms of the allele frequencies and disequilibrium coefficient D_{AB} as shown in Table 1.2.

By pre-multiplying the variance covariance matrix $\mathbf{V}$ by the transpose of the gradient vector and multiplying that product by the gradient again as in (2.1), *Maple* produced an expression for the variance formula which filled nearly an entire page (Appendix A). Further work was carried out to simplify the formula to the form displayed in Table 2.1.

Terms were collected by powers of D_{AB}. These coefficients were then simplified using the notation $\pi_A = p_A p_a$, $\pi_B = p_B p_b$, $\tau_A = 1 - 2p_A$ and $\tau_B = 1 - 2p_B$.

Term	Coefficient
Denominator	
1	$4\pi_A^2\pi_B^2 N$
Numerator	
1	$4\pi_A^2\pi_B^2$
D_{AB}	$4\pi_A\pi_B\tau_A\tau_B$
D_{AB}^2	$-3(\pi_A + \pi_B) + 20\pi_A\pi_B$
D_{AB}^3	$2\tau_A\tau_B$

Table 2.1: Variance formula for r_g.

When $D_{AB} = 0$, the variance is $Var(r_g) = 1/N$ where N is the number of gametes in the sample.

2.2 Behaviour of the Asymptotic Variance

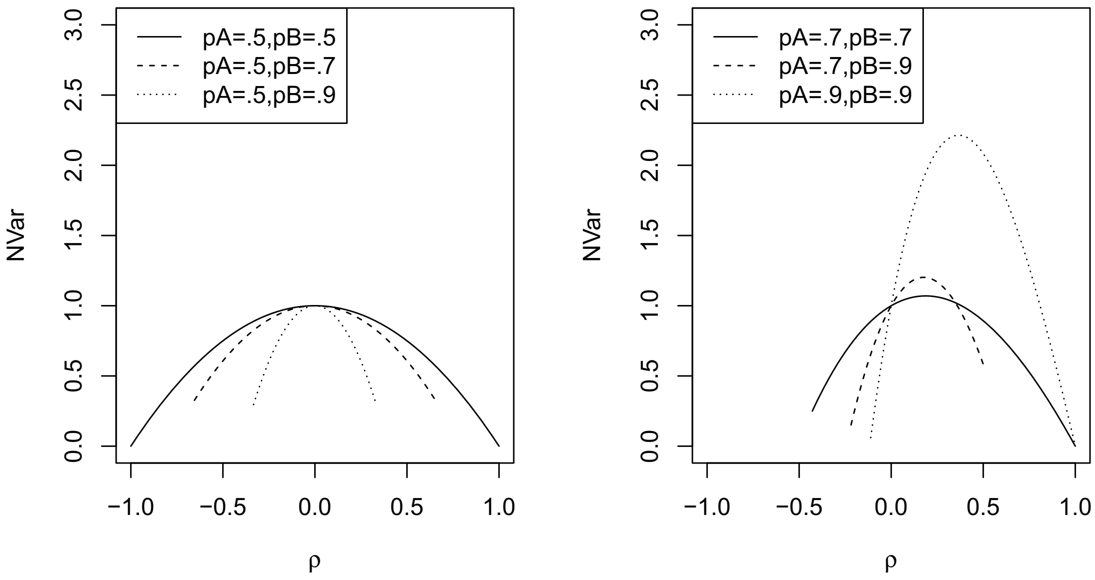

Figure 2.1: Asymptotic variance r_g as a function of ρ for several choices of p_A and p_B.

Figure 2.1 contains two panels displaying the approximate scaled variance as a function of p_A, p_B and ρ. The variance is multiplied by N to remove dependence on sample size.

In the more moderate allelic frequency case (left panel), there is a roughly parabolic shape with the maximum scaled variance of 1 reached at $\rho = 0$. The range of viable values of ρ is more restricted with increased p_B. For $p_A = p_B = 0.5$, the variance is zero at $|\rho| = 1$. In this case, $p_{Ab} = p_{aB} = 0$ so n_{Ab} and n_{aB} are always zero, giving $r_g = \pm 1$ without variation. This is a unique case and not true for $p_B \neq 0.5$ where only one of p_{Ab} or p_{aB} is zero at the extreme choices for ρ, so r_g is not always ± 1 and has variation.

For the more extreme allele frequencies (right panel), the plots are again roughly parabolic. As the allele frequencies become more extreme the maximum variance becomes larger and it occurs farther from $\rho = 0$.

2.3 Comparision of Asymptotic and Simulated Variance

In order to validate the variance formula, simulations were carried out using samples generated according to the multinomial gametic probabilities (Table 1.2). The correlation r_g was calculated for each simulated sample, and the variance of these estimates was obtained. The sampling distribution of the simulated correlations was examined and standard errors were calculated. Values for p_A and p_B included combinations of 0.5, 0.7 and 0.9, while values of ρ ranged over positive and negative values of 0.9, 0.5, 0.3, 0.1. Some of the combinations of ρ, p_A and p_B do not produce viable gametic probabilities. Asterisks indicate such combinations in the table. Simulations were run $m = 20,000$ times for gametic sample sizes of $N = 100$ and $N = 1000$. The value of $N = 100$ is quite small for genetic studies and might be too small for asymptotic formulae to be accurate. On the other hand $N = 1000$ is quite large and asymptotic results should be valid.

The sample variances of r_g were compared with the corresponding theoretical variances in ratio format for each ρ and allelic frequency combination. The results are displayed in Table 2.2. Overall, the ratios are close to one indicating that the variance formula is correct.

Standard errors were calculated for the ratios as follows. The ratio is calculated as

$$ratio = \frac{1}{Avar}\left(\frac{1}{m-1}\right)\sum(r_{g,i} - \bar{r}_g)^2 \approx \frac{\sum d_i^2}{mAvar}.$$

where $d_i = r_{gi} - \bar{r}_g$ and Avar is the asymptotic variance. Its variance is

$$Var(ratio) = \frac{Var(d_i^2)}{mAvar^2}$$

which is estimated by

$$\hat{V}ar(ratio) = \frac{s_{d^2}^2}{mAvar^2}$$

		ρ							
p_A	p_B	-0.9	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.9
$n = 100$									
.5	.5	0.99	1.02	1.01	1.00	1.01	1.01	1.02	1.00
	.7	*	1.01	1.00	1.01	0.99	1.01	1.02	*
	.9	*	*	1.01	1.00	1.02	1.04	*	*
.7	.5	*	1.01	1.01	1.01	1.02	1.01	1.01	*
	.7	*	*	1.02	1.01	1.02	1.01	1.01	0.99
	.9	*	*	*	1.04	1.01	1.03	1.07	*
.9	.5	*	*	1.04	1.00	1.01	1.05	*	*
	.7	*	*	*	1.00	1.01	1.03	1.08	*
	.9	*	*	*	1.01	1.02	1.06	1.10	1.18
$n = 1000$									
.5	.5	1.01	1.00	1.02	0.99	0.98	1.02	0.99	1.02
	.7	*	0.98	1.01	0.99	1.00	0.99	0.99	*
	.9	*	*	1.02	0.99	1.01	1.01	*	*
.7	.5	*	1.01	1.00	0.99	1.01	0.99	0.99	*
	.7	*	*	1.00	1.00	1.01	1.00	1.00	1.00
	.9	*	*	*	0.99	1.01	1.00	1.00	*
.9	.5	*	*	1.01	1.00	1.01	1.00	*	*
	.7	*	*	*	0.98	1.02	1.00	1.01	*
	.9	*	*	*	0.97	1.00	1.00	1.02	0.99

Table 2.2: Ratio of the simulated variance to the asymptotic variance of r_g for sample sizes $N = 100$ and $N = 1000$.

where $s^2_{d^2}$ is the sample variance of the d^2_i. The standard error of the ratio is

$$SE(ratio) \quad = \quad \frac{s_{d^2}}{\sqrt{m}Avar}.$$

For $N = 100$, the ratios range between 0.99-1.18 and are closer to one for the increased sample size (Table 2.2). Simulation standard errors calculated for each ratio all round to the value of 0.01 except for the case of (ρ, p_A, p_B)=(-0.1, 0.9, 0.9) and (ρ, p_A, p_B)=(0.9, 0.9, 0.9) where the standard error is 0.02. Several of the simulated ratios are more than two standard errors from unity, predominantly at the more extreme allele frequencies. Only 3 of the 47 variance ratios fall below unity indicating that the asymptotic variance tends to underestimate the true variance.

For $N = 1000$, all standard errors rounded to 0.01 and none of the ratios are more than two standard errors away from unity. The number of ratios falling below unity rises to 15 out of 47.

Upon examining the sampling distributions of the simulated correlations, it was noted that all distributions are approximately normal except when (ρ, p_A, p_B)=((-0.1,0.9),0.9,0.9) (Figure 2.2). For $N = 100$, the distributions are highly skewed for these cases, skewed right for $\rho = -0.1$ and skewed left for $\rho = 0.9$. This skewness is an evident sign that asymptotic conditions have not been reached with a sample size of $N = 100$. The skewness is greatly reduced with $N = 1000$, and the ratio of simulated to asymptotic variances is much closer to 1.0.

2.3.1 Fisher's Transformation of the Correlation

Because some of the sampling distributions in the simulations were very skewed, Fisher's transformation of the correlation coefficient

$$\delta = \frac{1}{2} \; log \left(\frac{1+\rho}{1-\rho} \right)$$

was considered.

For bivariate normal data, this transformation of the sample correlation coefficient is approximately normally distributed. For estimation, a factor of $1/N$ was added to both the numerator and denominator in order to prevent problems in evaluating the transformation when $|r_g| = 1$, giving

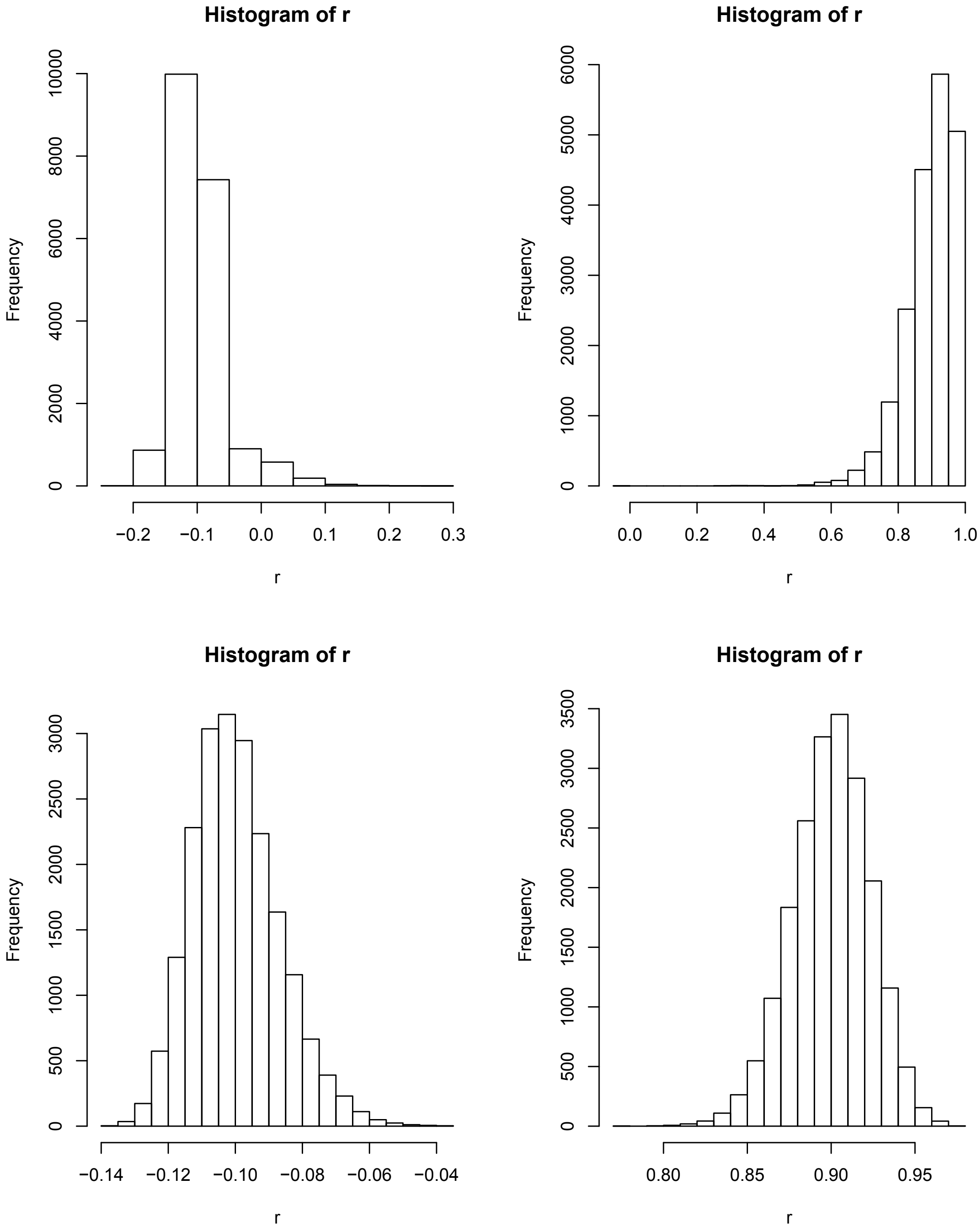

Figure 2.2: Histograms of r_g for allelic frequencies $p_A = p_B = 0.9$ with $\rho = -0.1$ (left) and $\rho = 0.9$ (right) for sample sizes of $N = 100$ (top) and $N = 1000$ (bottom).

$$\hat{\delta}_g = \frac{1}{2}\ log\left(\frac{1 + r_g + 1/N}{1 - r_g + 1/N}\right).$$

The asymptotic variance approximation for $\hat{\delta}_g$ using the delta method is

$$Var(\hat{\delta}) = \frac{1}{(1-(\rho)^2)^2}\ Var(r_g),$$

ignoring terms of $1/N^2$ and smaller.

The ratios of simulated to asymptotic variances of $\hat{\delta}_g$ are shown in Table 2.3. To simplify the comparison of the standardized and unstandardized values, the columns are labelled by their correlation values.

For $N = 100$, standard errors of the simulated ratios all round to 0.01 for every case other than for (ρ, p_A, p_B) = (-0.9, 0.5, 0.5), (ρ, p_A, p_B) = (0.9, 0.7, 0.7) and (ρ, p_A, p_B) = ((-0.1, 0.5, 0.9), 0.9, 0.9) where errors rounded to 0.02.

For $N = 100$, the Fisher transformation makes the ratio closer to 1 in 18 cases, makes it farther from 1 in 10 cases and keeps it the same in 19 cases. Some cases, for example (ρ, p_A, p_B) = ((-0.1, 0.5, 0.9), 0.9, 0.9), (0.9, 0.7, 0.7) and (-0.9, 0.5, 0.5) are made much worse by the transformation. Histograms of $\hat{\delta}_g$ (Figure 2.3 and 2.4) reveal that the value $r_g = 1$ is transformed to $\hat{\delta}_g = .5\ log(\frac{2+1/100}{1/100}) = .5\ log(201) = 2.65$. This value is detached from the remaining distribution of the ρ values giving a large sample variance and a large ratio. For small values of ρ the transformation is nearly linear and has little effect.

For $N = 1000$, the Fisher transformation improved the ratio in 3 cases, made it worse in 7, and kept it the same in 39.

		ρ							
p_A	p_B	-0.9	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.9
$n = 100$									
.5	.5	1.10	1.03	1.01	0.99	1.01	1.01	1.02	1.11
	.7	*	1.00	1.00	1.01	0.99	1.01	1.02	*
	.9	*	*	1.01	1.00	1.02	1.02	*	*
.7	.5	*	1.00	1.01	1.00	1.02	1.01	1.00	*
	.7	*	*	1.00	1.01	1.01	1.02	1.03	1.21
	.9	*	*	*	1.03	1.01	1.02	1.06	*
.9	.5	*	*	1.02	1.00	1.01	1.02	*	*
	.7	*	*	*	0.99	1.02	1.02	1.07	*
	.9	*	*	*	0.99	1.05	1.11	1.18	1.83
$n = 1000$									
.5	.5	1.02	1.00	1.02	0.99	0.98	1.02	1.00	1.03
	.7	*	0.98	1.01	1.00	1.00	0.99	0.99	*
	.9	*	*	1.02	0.99	1.01	1.01	*	*
.7	.5	*	1.01	1.00	0.99	1.01	0.99	0.98	*
	.7	*	*	1.00	1.00	1.01	1.00	1.00	1.02
	.9	*	*	*	0.99	1.01	1.00	1.00	*
.9	.5	*	*	1.00	1.00	1.01	1.00	*	*
	.7	*	*	*	0.98	1.02	1.00	1.00	*
	.9	*	*	*	0.97	1.00	1.01	1.02	1.04

Table 2.3: Ratio of the simulated variance to the asymptotic variance of Fisher's transformation of r_g.

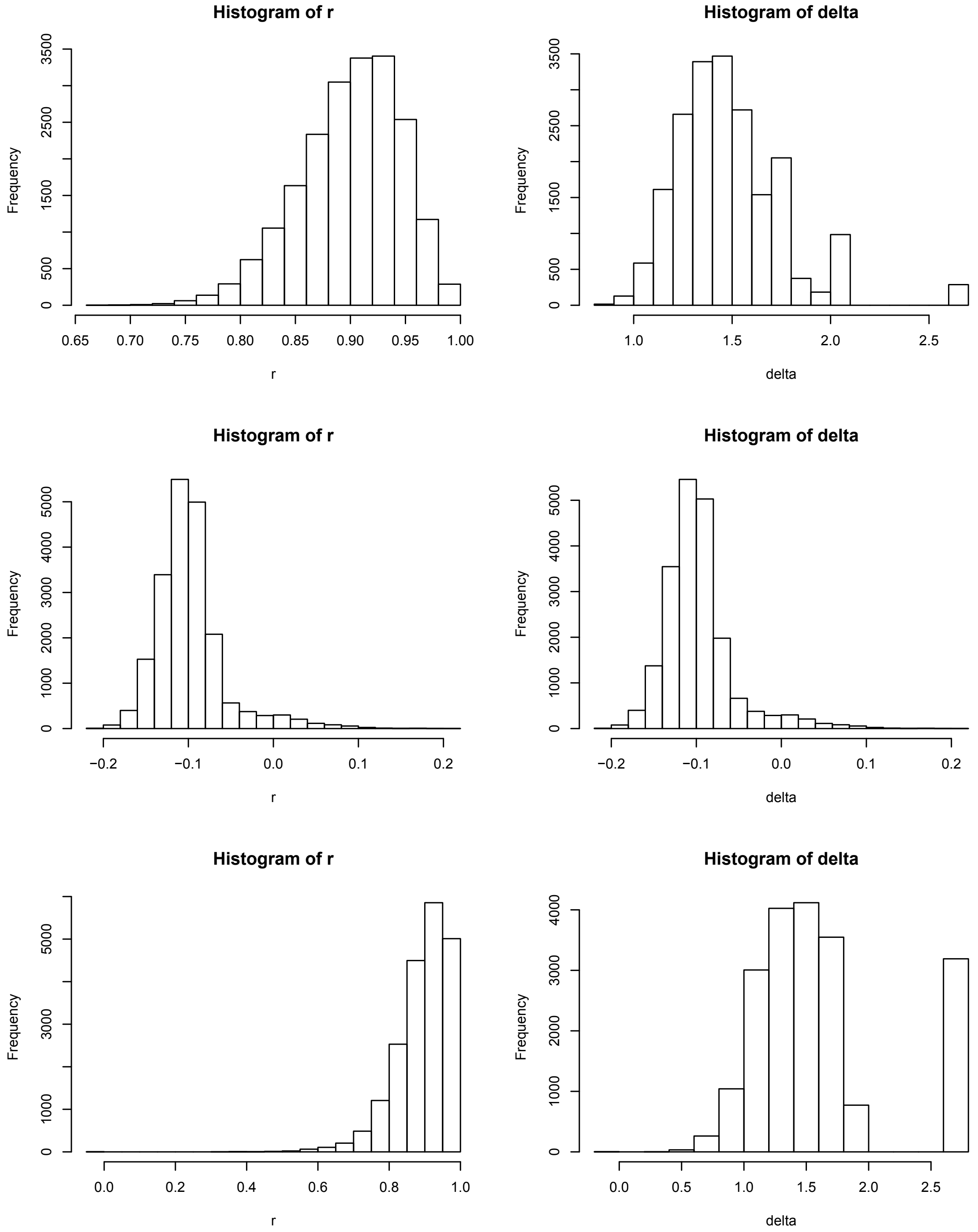

Figure 2.3: Histograms of simulated untransformed r_g values (left) and Fisher transformed $\hat{\delta}_g$ values (right) for (δ_g, p_A, p_B) = (0.9, 0.7, 0.7) (top), (δ_g, p_A, p_B) = (-0.1, 0.9, 0.9) (middle), (δ_g, p_A, p_B) = (0.9, 0.9, 0.9) (bottom) and sample size of $N = 100$.

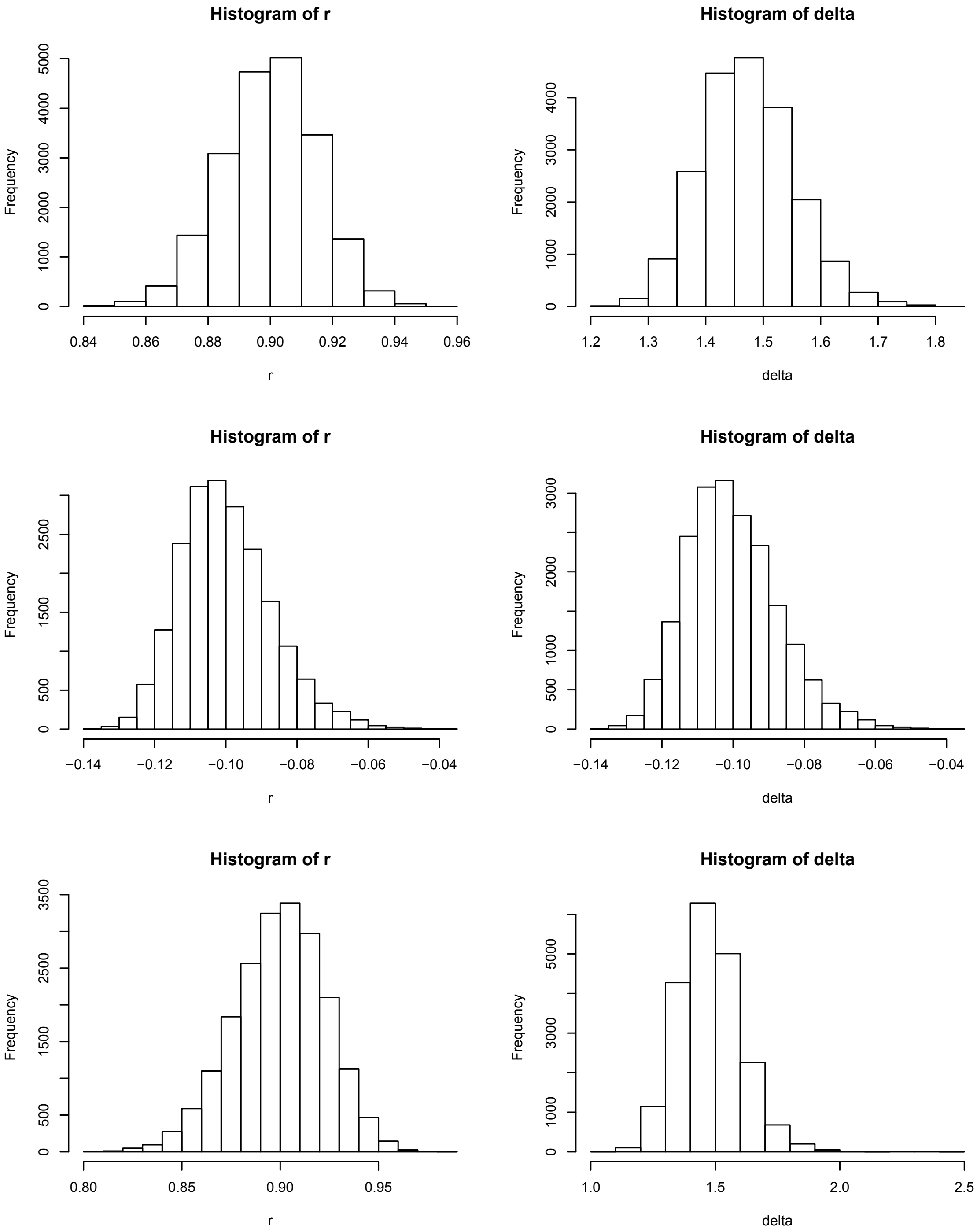

Figure 2.4: Histograms of simulated r_g values (left) and Fisher transformed $\hat{\delta}_g$ values (right) for (δ_g, p_A, p_B) = (0.9, 0.7, 0.7) (top), (δ_g, p_A, p_B) = (-0.1, 0.9, 0.9) (middle), (δ_g, p_A, p_B) = (0.9, 0.9, 0.9) (bottom) and sample size of $N = 1000$.

Chapter 3

Approximate Variance for the Correlation Measures of LD Using Genotypic Data Assuming Random Mating

3.1 Variance Approximation Using the Delta Method

The delta method and symbolic computation were used to calculate an asymptotically approximate variance of r_{rm} using *Maple* software. Recall (1.8) that r_{rm} is obtained using maximum likelihood estimates of

$$\boldsymbol{\theta} = (p_A, p_B, p_{AB})^T. \tag{3.1}$$

Likelihood theory tells us that the MLE $\hat{\boldsymbol{\theta}}$ is approximately unbiased with variance covariance matrix $\mathbf{V}$ given by the inverse of the expected information matrix.

$$I(\boldsymbol{\theta}) = E\left(-\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right) \tag{3.2}$$

where l is the log likelihood (1.6).

Using the delta method the correlation r_{rm} has an approximate variance

$$Var(r_{rm}) \approx \mathbf{q}^T \boldsymbol{I}^{-1}\mathbf{q}, \tag{3.3}$$

where $\mathbf{q}$ is the vector of derivatives of r_{rm} with respect to the elements of $\hat{\boldsymbol{\theta}}$ evaluated at their means. *Maple* commands and output for calculating the approximate variance are found in Appendix A. Derivatives are calculated using the *diff* operator. Because $E(\hat{p}_A)$, $E(\hat{p}_B)$, and $E(\hat{p}_{AB})$ are not known, occurrences of $\hat{p}_A$, $\hat{p}_B$ and $\hat{p}_{AB}$ in the derivatives were evaluated at p_A, p_B and $p_{AB} = p_A p_B + D_{AB}$. This ignores the bias in the MLEs but does not effect the accuracy of the variance approximation.

The *Maple* output for the variance formula filled numerous lines. Further work was carried out to simplify the expression to the form displayed in Table 3.1. The variance is written as a polynomial in D_{AB} with coefficients which are functions of

Term	Coefficient
Denominator	
1	$8\pi_A^3\pi_B^3 n$
D_{AB}	$8\pi_A^2\pi_B^2\tau_A\tau_B n$
D_{AB}^2	$24\pi_A^2\pi_B^2 n$
Numerator	
1	$8\pi_A^3\pi_B^3$
D_{AB}	$12\pi_A^2\pi_B^2\tau_A\tau_B$
D_{AB}^2	$\pi_A\pi_B(120\pi_A\pi_B - 23(\pi_A+\pi_B)+4)$
D_{AB}^3	$\tau_A\tau_B(38\pi_A\pi_B - 3(\pi_A+\pi_B))$
D_{AB}^4	$96\pi_A\pi_B - 17(\pi_A+\pi_B)+2$
D_{AB}^5	$6\tau_A\tau_B$

Table 3.1: Variance formula for r_{rm}.

p_A and p_B. For simplification, the following notation was used: $\pi_A = p_A p_a$, $\pi_B = p_B p_b$, $\tau_A = 1 - 2p_A$, and $\tau_B = 1 - 2p_B$. Note that when $D_{AB} = 0$, $Var(r_{rm}) = 1/n$, where n is the number of subjects.

3.2 Behaviour of the Asymptotic Variance

Figure 3.1 contains two panels displaying the approximate variance as a function of p_A, p_B and ρ. The variance is multiplied by n to remove dependence on sample size. Note that n subjects give $N = 2n$ gametes, so the variance of r_{rm} is approximately twice that of r_g, and is 2/N when $D_{AB} = 0$.

The plots of the variance versus ρ are roughly parabolic. When $p_A = 0.5$ (left panel) the maximum scaled variance of 1 occurs at $\rho = 0$ regardless of the value of p_B. As p_B increases from 0.5, the range of the feasible values for ρ decreases and the variance decreases more rapidly. As the allele frequencies become more extreme (right panel), the maximum variance increases and it occurs further from $\rho = 0$. These variance patterns are very similar to those for the gametic measure r_g.

3.3 Comparison of Approximate and Simulated Variance

To validate the variance formula, simulations were carried out using genotypic counts generated according to the multinomial genotypic probabilities in Table 1.3. Values for p_A and p_B were chosen as combinations of 0.5, 0.7 and 0.9, while values

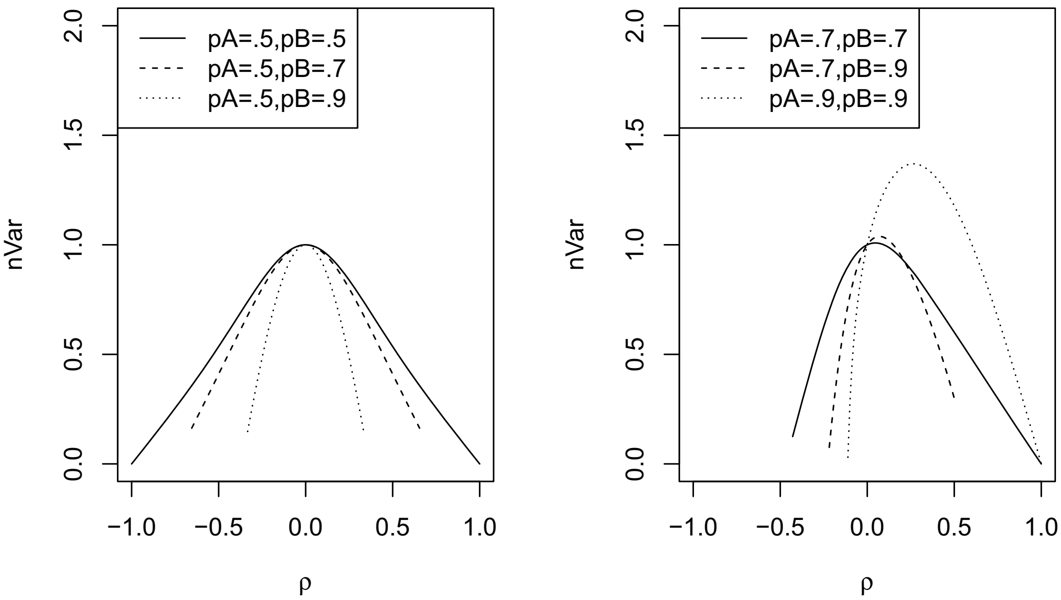

Figure 3.1: Asymptotic variance of r_{rm} as a function of ρ for several choices of p_A and p_B.

of ρ ranged over positive and negative values of 0.9, 0.5, 0.3, 0.1. Simulations were run $m = 20,000$ times for genotypic sample sizes of $n = 100$ and $n = 1000$. Correlations were calculated for each sample and their variance was calculated. Some of the combinations of ρ, p_A and p_B do not produce viable genotypic probabilities. These combinations are indicated in Table 3.2 by asterisks. The sample variances of r_{rm} were compared with the corresponding theoretical variances in ratio format for each ρ and allelic frequency combination.

		ρ							
p_A	p_B	-0.9	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.9
$n = 100$									
.5	.5	1.01	1.01	1.04	1.03	1.03	1.01	1.01	1.00
	.7	*	1.01	1.03	1.01	1.03	1.06	1.02	*
	.9	*	*	1.07	1.06	1.08	1.05	*	*
.7	.5	*	1.01	1.02	1.02	1.03	1.04	0.99	*
	.7	*	*	1.05	1.02	1.04	1.05	1.02	1.02
	.9	*	*	*	1.10	1.08	1.00	1.06	*
.9	.5	*	*	1.04	1.07	1.07	1.04	*	*
	.7	*	*	*	1.12	1.07	1.06	1.03	*
	.9	*	*	*	1.25	1.01	1.02	1.01	1.07
$n = 1000$									
.5	.5	0.99	0.99	1.01	1.00	0.99	1.00	1.00	1.00
	.7	*	1.00	1.01	1.01	0.99	1.02	1.00	*
	.9	*	*	0.99	1.02	1.00	0.98	*	*
.7	.5	*	0.99	0.98	1.00	1.00	1.00	1.00	*
	.7	*	*	1.00	0.99	0.99	1.00	1.01	1.01
	.9	*	*	*	1.01	1.01	1.01	0.99	*
.9	.5	*	*	1.00	1.00	1.02	1.01	*	*
	.7	*	*	*	1.01	1.02	1.00	0.99	*
	.9	*	*	*	1.05	1.00	1.02	0.99	1.02

Table 3.2: Ratio of simulated variance to the theoretical variance of r_{rm}.

Overall, the ratios are close to 1, especially for $n = 1000$, indicating the variance formula is correct. All ratios were greater than one for $n = 100$, indicating that the asymptotic formula tends to underestimate the variance and that asymptotic conditions have not been met. This underestimation of the variance is not an issue with $n = 1000$, where just over half (61%) of the ratios are above unity.

For $n = 100$, some of the ratios are quite large. Simulation standard errors of the ratios all round to 0.01 except for one. For (ρ, p_A, p_B) = (-0.1, 0.9, 0.9) the standard error is 0.03. More than three quarters of the ratios deviate from unity by more than two standard errors. The worst case is for $(\rho,\, p_A,\, p_B)$ = (-0.1, 0.9, 0.9), where the ratio is 1.25.

For $n = 1000$, most of the ratios are quite close to one. The standard errors of the ratios all round to 0.01 and only one case has a ratio more than two standard errors from one. As with $n = 100$, the worst case is when $(\rho,\, p_A,\, p_B)$=(-0.1, 0.9, 0.9). Histograms of the sampling distributions for this case are shown in Figure 3.2.

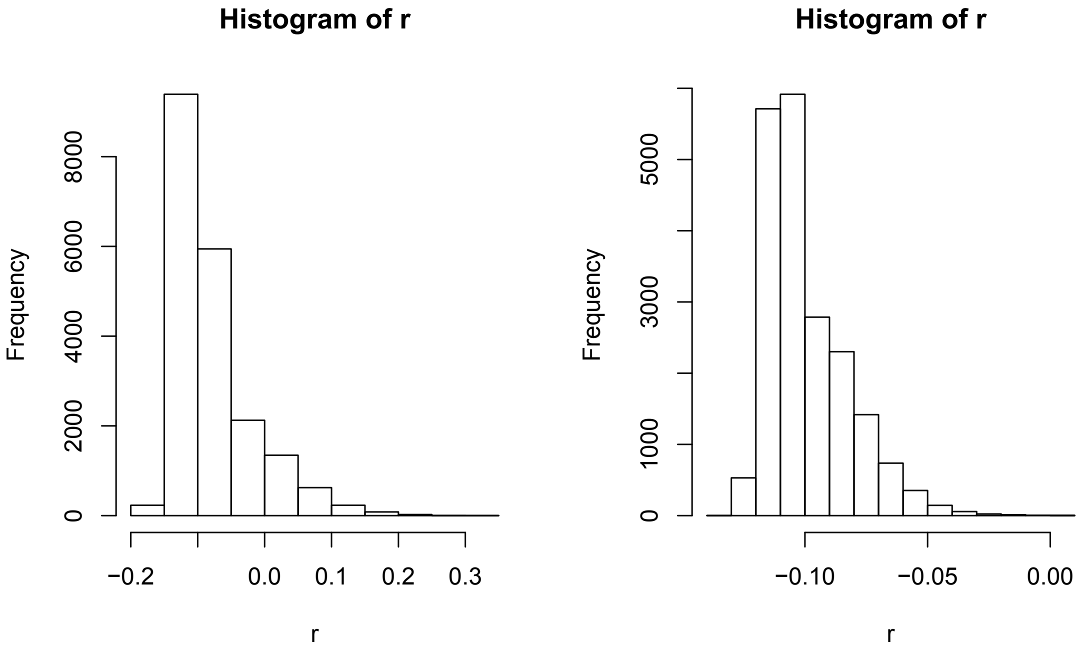

Figure 3.2: Histograms of simulated r_{rm} values for $(\rho,\, p_A,\, p_B)$ = (-0.1, 0.9, 0.9), $n = 100$ (left) and $n = 1000$ (right).

These histograms both show skewness which may indicate that asymptotic conditions have not been met.

3.3.1 Fisher Transformation of the Correlation

As with the gametic measure, we investigated the use of Fisher's transformation

$$\hat{\delta}_{rm} = \frac{1}{2}\ log \left(\frac{1 + r_{rm} + 1/n}{1 - r_{rm} + 1/n} \right)$$

which has variance approximation

$$Var(\hat{\delta}_{rm}) = \frac{1}{(1-\rho^2)^2}\ Var(r_{rm}).$$

The ratios of simulated to asymptotic variances of $\hat{\delta}_{rm}$ are shown in Table 3.3. These results are based on the same random samples as in Table 3.2.

Comparing Table 3.2 and Table 3.3 reveals that the transformation improves the ratios or keeps them the same in 40/47 of the cases when $n = 100$, and in 37/47 of the cases when $n = 1000$. Most cases which did not show improvement correspond to extreme choices of ρ, p_A, and p_B. When (ρ, p_A, p_B)=(0.9, 0.9, 0.9) some of the $r_{rm} = 1.0$ leading to the same problem described for r_g.

		ρ							
p_A	p_B	-0.9	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.9
$n = 100$									
.5	.5	0.95	0.99	1.03	1.03	1.04	1.00	0.99	0.94
	.7	*	0.99	1.01	1.03	1.03	1.02	0.99	*
	.9	*	*	1.03	1.07	1.07	1.00	*	*
.7	.5	*	0.99	1.01	1.00	1.04	1.01	0.99	*
	.7	*	*	1.01	1.04	1.05	1.04	1.02	0.96
	.9	*	*	*	1.10	1.09	1.03	1.00	*
.9	.5	*	*	1.01	1.07	1.06	1.01	*	*
	.7	*	*	*	1.09	1.09	1.01	1.01	*
	.9	*	*	*	1.23	1.01	1.04	1.07	1.30
$n = 1000$									
.5	.5	0.98	1.00	0.99	1.00	1.00	1.00	1.00	0.98
	.7	*	1.01	1.00	1.01	1.00	1.01	1.01	*
	.9	*	*	1.00	1.00	1.00	1.00	*	*
.7	.5	*	1.00	1.01	1.01	1.00	1.01	1.01	*
	.7	*	*	0.99	1.02	1.00	1.00	1.00	0.99
	.9	*	*	*	1.01	0.99	1.01	0.99	*
.9	.5	*	*	0.99	1.03	1.00	1.00	*	*
	.7	*	*	*	0.99	1.03	1.00	1.01	*
	.9	*	*	*	1.01	1.01	1.01	1.02	1.01

Table 3.3: Ratio of simulated variance to the asymptotic variance of Fisher's transformation of r_{rm}.

Chapter 4

Approximate Variance for the Correlation Measure of LD when the Genotypic Phase is Known

4.1 Variance Approximation Using the Delta Method

An approximate variance formula is derived for r_{pk}, the correlation measure of LD (1.9) for the phase known genotypic case.

This measure depends on the genotypic counts

$$\mathbf{x} = (n^{AB}_{AB}, n^{AB}_{Ab}, n^{Ab}_{Ab}, n^{AB}_{aB}, n^{AB}_{ab}, n^{Ab}_{aB}, n^{Ab}_{ab}, n^{aB}_{aB}, n^{aB}_{ab}, n^{ab}_{ab})^T,$$

which follow a multinomial distribution with index n and the probabilities

$$\begin{aligned}\mathbf{p} &= (p_1, \ldots, p_{10})^T \\ &= (P^{AB}_{AB}, 2P^{AB}_{Ab}, P^{Ab}_{Ab}, 2P^{AB}_{aB}, 2P^{AB}_{ab}, 2P^{Ab}_{aB}, 2P^{Ab}_{ab}, P^{aB}_{aB}, 2P^{aB}_{ab}, P^{ab}_{ab})^T\end{aligned}$$

and variance covariance matrix $\mathbf{V}$, which is a 10×10 matrix composed of diagonal elements

$$\begin{aligned}V_{ii} &= Cov(X_i, X_i) \\ &= Var(X) \\ &= np_i(1 - p_i)\end{aligned}$$

and off diagonal elements

$$V_{ij} = Cov(X_i, X_j) = -np_ip_j.$$

The variance approximation is calculated as

$$Var(r_{pk}) = \mathbf{g}^T\mathbf{V}\mathbf{g}$$

where $\mathbf{g} = (g_1, \ldots, g_{10})^T$ and

$$g_i = \left.\frac{\partial r_{pk}}{\partial x_i}\right|_{n\mathbf{p}_i}$$

Probability	Formula										
P_{AB}^{AB}	$P_A^A P_B^B$	$+2p_Ap_BD_{AB}$	$+D_{AB}^2$	$+2p_Ap_BD_{A/B}$	$+D_{A/B}^2$	$+2p_AD_{ABB}$	$+2p_BD_{AAB}$	$+p_B^2D_A$	$+p_A^2D_B$	$+D_AD_B$	$+D_{AABB}$
P_{Ab}^{AB}	$P_A^A P_b^B$	$+2p_A\tau_BD_{AB}$	$-2D_{AB}^2$	$+2p_A\tau_BD_{A/B}$	$-2D_{A/B}^2$	$-4p_AD_{ABB}$	$+2\tau_BD_{AAB}$	$+2\pi_BD_A$	$-2p_A^2D_B$	$-2D_AD_B$	$-2D_{AABB}$
P_{Ab}^{Ab}	$P_A^A P_b^b$	$-2p_Ap_bD_{AB}$	$+D_{AB}^2$	$-2p_Ap_bD_{A/B}$	$+D_{A/B}^2$	$+2p_AD_{ABB}$	$-2p_bD_{AAB}$	$+p_b^2D_A$	$+p_A^2D_B$	$+D_AD_B$	$+D_{AABB}$
P_{aB}^{AB}	$P_a^A P_B^B$	$+2\tau_Ap_BD_{AB}$	$-2D_{AB}^2$	$+2\tau_Ap_BD_{A/B}$	$-2D_{A/B}^2$	$+2\tau_AD_{ABB}$	$-4p_BD_{AAB}$	$-2p_B^2D_A$	$+2\pi_AD_B$	$-2D_AD_B$	$-2D_{AABB}$
P_{ab}^{AB}	$P_a^A P_b^B$	$+2\ XD_{AB}$	$+2D_{AB}^2$	$+2\ YD_{A/B}$	$+2D_{A/B}^2$	$-2\tau_AD_{ABB}$	$-2\tau_BD_{AAB}$	$-2\pi_BD_A$	$-2\pi_AD_B$	$+2D_AD_B$	$+2D_{AABB}$
P_{aB}^{Ab}	$P_a^A P_b^B$	$+2\ YD_{AB}$	$+2D_{AB}^2$	$+2\ XD_{A/B}$	$+2D_{A/B}^2$	$-2\tau_AD_{ABB}$	$-2\tau_BD_{AAB}$	$-2\pi_BD_A$	$-2\pi_AD_B$	$+2D_AD_B$	$+2D_{AABB}$
P_{ab}^{Ab}	$P_a^A P_b^b$	$-2\tau_Ap_bD_{AB}$	$-2D_{AB}^2$	$-2\tau_Ap_bD_{A/B}$	$-2D_{A/B}^2$	$+\tau_AD_{ABB}$	$+4p_bD_{AAB}$	$-2p_b^2D_A$	$+2\pi_AD_B$	$-2D_AD_B$	$-2D_{AABB}$
P_{aB}^{aB}	$P_a^a P_B^B$	$-2p_ap_BD_{AB}$	$+D_{AB}^2$	$-2p_ap_BD_{A/B}$	$+D_{A/B}^2$	$-2p_aD_{ABB}$	$+2p_BD_{AAB}$	$+p_B^2D_A$	$+p_a^2D_B$	$+D_AD_B$	$+D_{AABB}$
P_{ab}^{aB}	$P_a^a P_b^B$	$-2p_a\tau_BD_{AB}$	$-2D_{AB}^2$	$-2p_a\tau_BD_{A/B}$	$-2D_{A/B}^2$	$+4p_aD_{ABB}$	$+2\tau_BD_{AAB}$	$+2\pi_BD_A$	$-2p_a^2D_B$	$-2D_AD_B$	$-2D_{AABB}$
P_{ab}^{ab}	$P_a^a P_b^b$	$+2p_ap_bD_{AB}$	$+D_{AB}^2$	$+2p_ap_bD_{A/B}$	$+D_{A/B}^2$	$-2p_aD_{ABB}$	$-2p_bD_{AAB}$	$+p_b^2D_A$	$+p_a^2D_B$	$+D_AD_B$	$+D_{AABB}$

Table 4.1: Genotypic probabilities expressed in terms of allelic probabilities and disequilibrium coefficients when the genotypic phase is known (Notation: $X = p_Ap_B + p_ap_b$ and $Y = X - 1$).

is the derivative of the correlation with respect to a genotypic count evaluated at the mean.

Maple software was used for all symbolic variance approximation calculations. The gradient **g** was constructed from the partial derivatives with respect to each genotypic cell count (x_i) using the *diff* operator. The expected values were then evaluated by replacing each x_i by its mean $np_i, i = 1, \ldots, 10$. The genotypic probabilities were expressed in terms of the allele frequencies and a complete set of disequilibrium coefficients as shown in Table 4.1 (Weir and Cockerham, 1989).

As described in Weir (1996), these disequilibria include the single locus departures from Hardy Weinberg equilibrium (HWE)

$$D_A = P_A^A - p_A^2$$
$$D_B = P_B^B - p_B^2,$$

two locus gametic and non-gametic disequilibrium D_{AB} and $D_{A/B}$ described above (1.1 and 1.3), and three higher order disequilibria D_{AAB}, D_{ABB} and D_{AABB}.

The two trigenic disequilibrium measures are

$$D_{AAB} = p_{AAB} - p_A D_{AB} - p_A D_{A/B} - p_B D_A - p_A^2 p_B$$

and

$$D_{ABB} = p_{ABB} - p_B D_{AB} - p_B D_{A/B} - p_A D_B - p_A p_B^2.$$

where

$$p_{AAB} = P_{AB}^{AB} + \frac{1}{2} P_{Ab}^{AB} \tag{4.1}$$

and

$$p_{ABB} = P_{AB}^{AB} + \frac{1}{2} P_{aB}^{AB} \tag{4.2}$$

compare probabilities of three alleles at the two loci to the products of allele frequencies, these measures having removed any digenic disequilibria.

The quadrigenic disequilibrium measure

$$\begin{aligned} D_{AB}^{AB} = P_{AB}^{AB} &- 2p_A D_{ABB} - 2p_B D_{AAB} - 2p_A p_B D_{AB} - 2p_A p_B D_{A/B} \\ &- p_A^2 D_B - p_B^2 D_A - D_{AB}^2 - D_{A/B}^2 - D_A D_B - p_A^2 p_B^2 \end{aligned}$$

accounts for the remaining disequilibrium after all other forms of disequilibrium have been removed.

Maple produced output for the variance formula which filled numerous pages (Appendix A). Further work was carried out to simplify the formula to the form displayed in Table 4.2.

Term	Coefficient
Denominator	$\pi_A^3\pi_B^3 n$
Numerator	
1	$\pi_A^3\pi_B^3/2 + \pi_A^2\pi_B^2 D_A D_B/2$
D_{AB}	$\pi_A^2\pi_B^2\tau_A\tau_B/2$
D_{AB}^2	$\pi_A\pi_B(-3(\pi_A+\pi_B)+20\pi_A\pi_B)/8$ $+(\tau_A^2\pi_B^2 D_A + \tau_B^2\pi_A^2 D_B)/8$
D_{AB}^3	$\pi_A\pi_B\tau_A\tau_B/4$
$D_{AB}D_{AAB}$	$-\pi_B^2\pi_A\tau_A/2$
$D_{AB}D_{ABB}$	$-\pi_A^2\pi_B\tau_B/2$
$D_{A/B}D_{AB}^2$	$\pi_A\pi_B\tau_A\tau_B/4$
$D_{A/B}^2$	$\pi_A^2\pi_B^2/2$
D_{AABB}	$\pi_A^2\pi_B^2/2$

Table 4.2: Variance formula for r_{pk}.

This is a polynomial in the five higher order disequilibrium measures with coefficients which are functions of p_A, p_B, D_A and D_B. When all disequilbrium measures are zero, the variance is $1/2n = 1/N$.

4.2 Behaviour of the Asymptotic Variance

Figure 4.1 displays the approximate variance of r_{pk} as a function of p_A, p_B, and ρ in a state of HWE with all higher order disequilibria set to zero ($D_A = D_B = D_{A/B} = D_{AAB} = D_{ABB} = D_{AABB} = 0$). The approximate variance is multiplied by n to remove the dependence on sample size. These plots are similar to those of the previous chapters except that the maximum variance value is one half instead of one when p_A or p_B equals 0.5 and $\rho = 0$. When $p_A = 0.5$ (left panel), the variance reaches its maximum at $\rho = 0$ and decreases as ρ increases in magnitude. As p_B becomes more extreme, the range of possible ρ decreases and the variance decreases from its maximum at a faster rate than for the less extreme values of p_B. When p_A increases from 0.5 (right panel) the maximum variance increases, and it

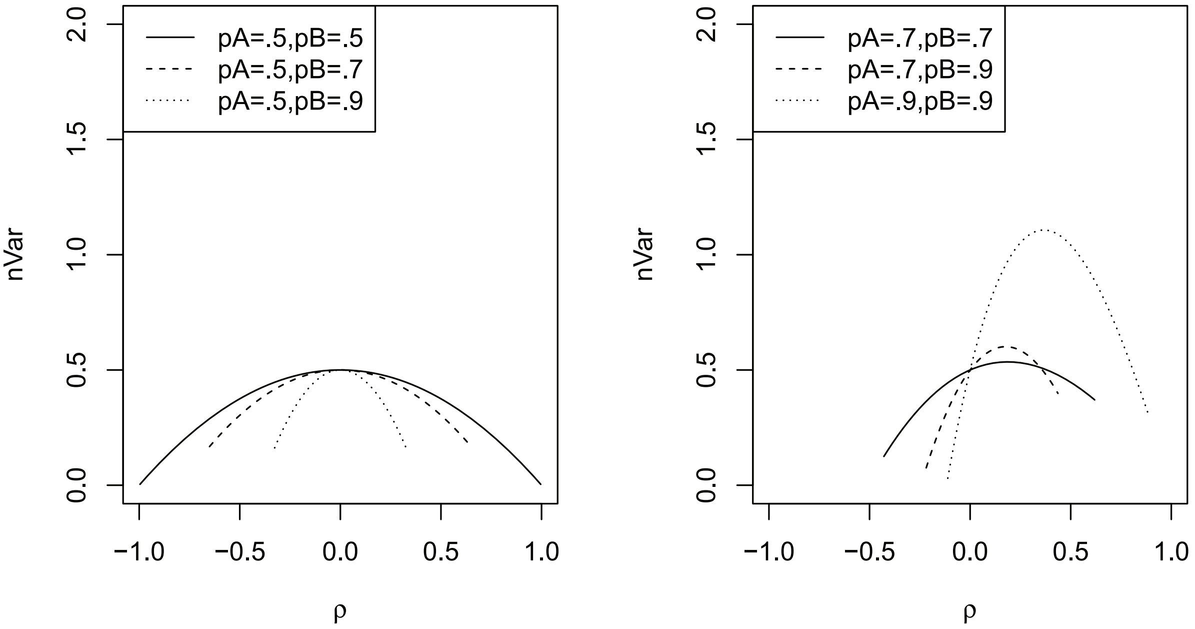

Figure 4.1: Asymptotic variance of r_{pk} as a function of ρ for several choices of p_A and p_B assuming random mating ($D_A = D_B = D_{A/B} = D_{AAB} = D_{ABB} = D_{AABB} = 0$).

occurs at larger values of ρ.

Figure 4.2 displays the variance of r_{pk} as a function of p_A, p_B, and ρ when there is HWD. The approximate variance is multiplied by n to remove the dependence on sample size. In these plots, the HWD coefficients are chosen to be a fraction of the maximum, where $D_{Amax} = p_A p_a$ and $D_{Bmax} = p_B p_b$. All higher order disequilibrium coefficients are taken to be zero.

When $p_A = p_B = .5$ (top panels), the variance is in $\rho = 0$. When $p_A = .7$, $p_B = .9$ (bottom panels) the variance function is not symmetric about its maximum and the maximum occurs to the right of $p = 0$. When $D_B = 0$ (left panels), the maximum variance does not change with D_A. When $D_B > 0$ (right panels) the maximum variance increases with D_A. The range of feasible values for ρ decreases as both D_A and D_B increases.

4.3 Comparison of Theoretical and Simulated Variance

The variance approximation was validated using simulation. Values for p_A and p_B were taken to be combinations of 0.5, 0.7 and 0.9, while ρ ranged over positive and negative values of 0.1, 0.3, 0.5, 0.9. There were $m = 20,000$ random samples of size $n = 100$ and $n = 1000$ generated for both the case of HWE and a HWD. The variance of the simulated correlation values was calculated and compared to the approximate variance in ratio format.

The HWD cases used ($D_A = .5D_{DAmax}$, $D_B = .5D_{DBmax}$) and ($D_A = 0.25D_{Amax}$, $D_B = 0.25D_{Bmax}$). All higher order disequilibrium values were set to zero ($D_{A/B} = D_{AAB} = D_{ABB} = D_{AABB} = 0$). Some of the combinations of ρ, p_A and p_B do not produce viable genotypic probabilities. These combinations are indicated in Table 4.3 and Table 4.4 by asterisks.

4.3.1 HWE Results

Table 4.3 displays the ratio of the simulated variance to the asymptotic variance when all other disequilibria are zero. The ratios are mostly very close to one indicating that the approximate variance formula is accurate.

For $n = 100$, six of the 47 ratios are more than two simulation standard errors (which were all .01) from one. All but one of these cases occurs with p_A or p_B equal

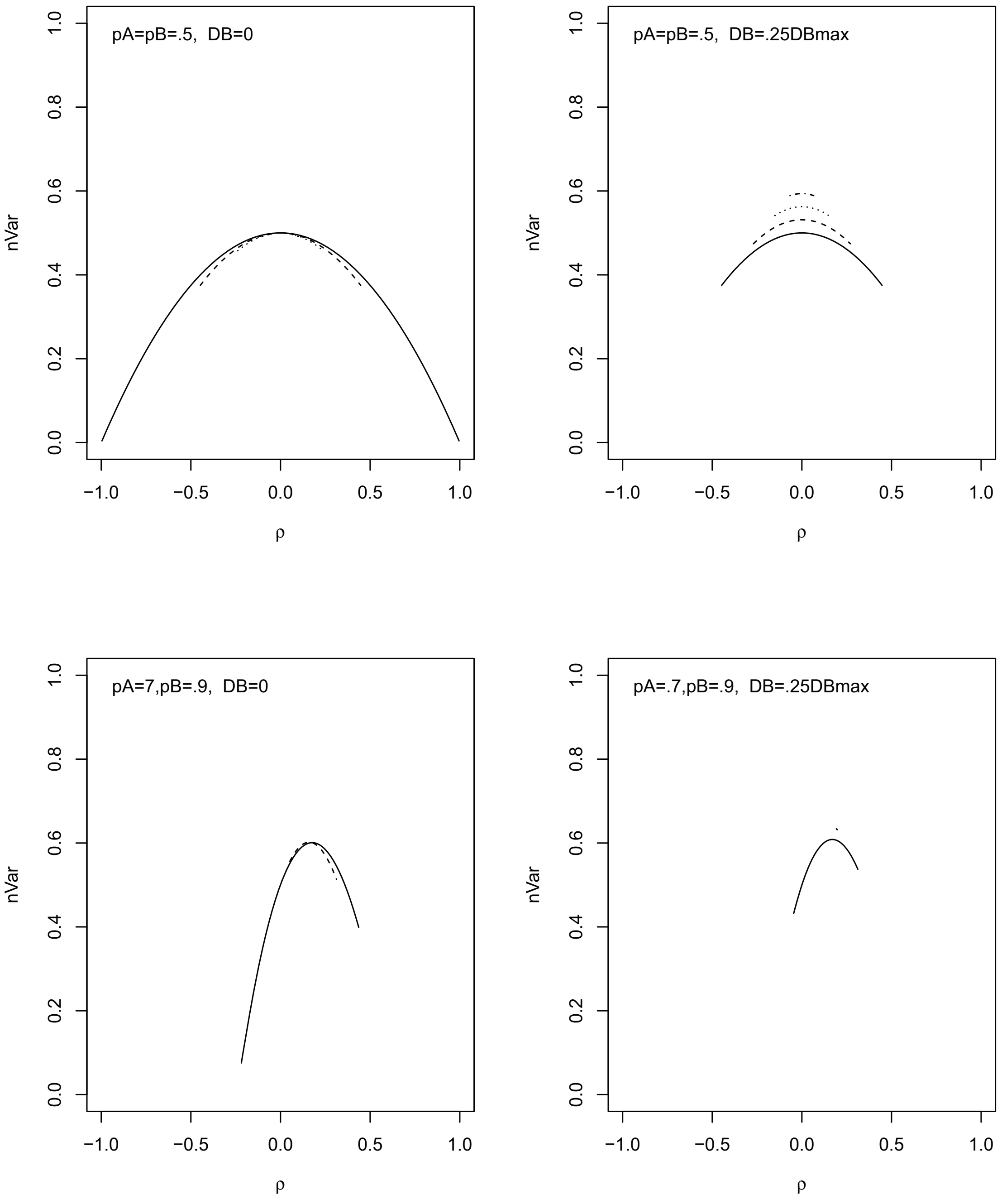

Figure 4.2: Asymptotic variance as a function of ρ_{pk}, p_A, p_B and of the HW disequilibrium coefficients D_A and D_B. D_A=0 (solid), D_A=.25D_{Amax} (dashed), D_A=.5D_{Amax} (dotted), D_A=.75D_{Amax} (dotted-dashed).

		ρ							
p_A	p_B	-0.9	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.9
$n = 100$									
.5	.5	1.02	1.01	1.00	1.01	0.98	1.00	0.99	0.99
	.7	*	1.01	0.98	0.99	0.99	1.00	1.00	*
	.9	*	*	1.00	0.99	1.01	1.03	*	*
.7	.5	*	1.00	1.00	1.01	1.02	1.00	1.03	*
	.7	*	*	1.01	1.02	1.00	1.00	1.00	1.01
	.9	*	*	*	1.01	1.00	1.02	1.01	*
.9	.5	*	*	1.01	0.99	0.99	1.01	*	*
	.7	*	*	*	1.01	0.99	1.01	1.04	*
	.9	*	*	*	1.01	1.00	1.03	1.04	1.06
$n = 1000$									
.5	.5	1.02	1.00	1.00	1.00	0.99	1.01	0.99	1.00
	.7	*	1.02	1.01	1.00	0.98	1.01	0.99	*
	.9	*	*	0.99	1.00	1.01	1.01	*	*
.7	.5	*	1.00	1.00	0.98	1.00	1.00	1.00	*
	.7	*	*	1.01	1.01	1.00	1.01	1.00	1.00
	.9	*	*	*	1.02	1.00	0.99	1.00	*
.9	.5	*	*	1.02	1.01	0.99	1.00	*	*
	.7	*	*	*	1.01	1.01	1.01	1.01	*
	.9	*	*	*	0.99	1.00	0.99	1.02	1.00

Table 4.3: Ratio of simulated variance to the theoretical variance of r_{pk} in the HWE case.

to 0.9. There are many more ratios greater than one (23) than are less than one (10), indicating that the variance formula tends to underestimate the true variance. An examination of the sampling distribution of the sample correlation showed it to be skewed when p_A, p_B or ρ are extreme, indicating that $n = 100$ may not be large enough for asymptotic conditions to apply.

For $n = 1000$, the ratios are much closer to one and all within two simulation standard errors. The balance of values below one (10) and above one (18) is better, and on examination of the sampling distributions of r_{pk} showed them to be all symmetric.

4.3.2 HWD Results

		ρ					
p_A	p_B	-0.1	0.1	-0.3	-0.1	0.1	0.3
$n = 100$		.5DAmax	.5DBmax		.25DAmax	.25DBmax	
.5	.5	1.01	1.01	1.03	1.02	1.01	0.99
	.7	0.99	1.00	*	1.02	1.00	*
	.9	*	*	*	0.99	0.99	*
.7	.5	1.00	1.00	*	1.00	1.01	*
	.7	*	1.00	*	1.00	0.99	1.01
	.9	*	1.00	*	1.00	1.01	*
.9	.5	*	*	*	0.99	1.02	*
	.7	*	1.00	*	0.99	1.01	*
	.9	*	1.02	*	*	1.01	1.03
$n = 1000$							
.5	.5	1.00	1.01	0.99	0.99	1.00	1.01
	.7	0.98	1.00	*	0.99	0.99	*
	.9	*	*	*	1.00	1.00	*
.7	.5	1.00	1.01	*	0.98	1.01	*
	.7	*	1.00	*	1.02	1.00	1.01
	.9	*	1.00	*	0.99	1.01	*
.9	.5	*	*	*	1.00	0.99	*
	.7	*	1.00	*	1.00	1.00	*
	.9	*	1.00	*	*	1.00	1.00

Table 4.4: Ratio of simulated variance to the theoretical variance of r_{pk} in the HWD case.

Table 4.4 shows the simulated ratios when there is HWD (D_A, D_B)=($0.5D_{Amax}$, $0.5D_{Bmax}$), and (D_A, D_B)=($0.25D_{Amax}$, $0.25D_{Bmax}$). With HWD the possible values for ρ are limited. The results (Table 4.4) are similar to when there is HWE. Most simulated ratios are close to one. The largest deviations from one occur when p_A or p_B is large and ρ is close to its feasible boundary. The ratios are closer to one when $n = 1000$ than when $n = 100$.

4.3.3 Simulations with Fisher Transformation

The simulations were also run using Fisher's transformation, to see whether the variance approximation is more accurate on this scale. The results, shown in Tables 4.5 and 4.6, indicate that the transformation has little effect when $n = 1000$ and tends to produce slightly worse ratios when $n = 100$.

		ρ							
p_A	p_B	-0.9	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.9
$n = 100$									
.5	.5	0.96	1.00	0.99	1.00	0.97	0.99	0.98	0.95
	.7	*	0.99	0.97	0.98	0.98	0.99	0.98	*
	.9	*	*	0.98	0.98	1.00	1.01	*	*
.7	.5	*	0.98	0.99	1.00	1.00	0.99	1.01	*
	.7	*	*	0.98	1.01	0.99	0.99	0.99	0.97
	.9	*	*	*	1.00	0.99	1.01	0.99	*
.9	.5	*	*	0.98	0.98	0.98	0.99	*	*
	.7	*	*	*	1.00	0.99	1.00	1.01	*
	.9	*	*	*	0.98	1.00	1.04	1.05	1.31
$n = 1000$									
.5	.5	1.01	1.00	1.00	1.00	0.99	1.01	0.99	0.99
	.7	*	1.02	1.01	1.00	0.98	1.00	0.99	*
	.9	*	*	0.99	1.00	1.01	1.01	*	*
.7	.5	*	1.00	1.00	0.98	1.00	1.00	1.00	*
	.7	*	*	1.02	1.01	1.00	1.01	1.01	1.00
	.9	*	*	*	1.02	1.00	0.99	1.00	*
.9	.5	*	*	1.02	1.01	0.99	1.00	*	*
	.7	*	*	*	1.01	1.01	1.01	1.01	*
	.9	*	*	*	0.99	1.00	0.99	1.02	1.02

Table 4.5: Ratio of the simulated Fisher transformed variance to the theoretical variance of r_{pk} in the case of HWE.

		ρ					
p_A	p_B	-0.1	0.1	-0.3	-0.1	0.1	0.3
$n = 100$		.5DAmax	.5DBmax		.25DAmax	.25DBmax	
.5	.5	1.00	1.01	1.02	1.01	1.00	0.98
	.7	0.99	0.99	*	1.01	0.99	*
	.9	*	*	*	0.98	0.98	*
.7	.5	1.00	0.99	*	0.99	1.00	*
	.7	*	1.00	*	0.99	1.00	1.00
	.9	*	1.00	*	0.99	1.00	*
.9	.5	*	*	*	0.98	1.01	*
	.7	*	0.99	*	0.98	1.00	*
	.9	*	1.03	*	*	1.02	1.06
$n = 1000$							
.5	.5	1.00	1.01	0.99	0.99	0.99	1.01
	.7	0.98	0.99	*	0.98	0.99	*
	.9	*	*	*	1.00	1.00	*
.7	.5	1.00	1.00	*	0.98	1.01	*
	.7	*	1.00	*	1.02	1.00	1.01
	.9	*	1.00	*	0.99	1.01	*
.9	.5	*	*	*	0.99	0.98	*
	.7	*	1.00	*	1.00	0.99	*
	.9	*	1.00	*	*	0.99	1.00

Table 4.6: Ratio of simulated Fisher transform variance to the theoretical variance of r_{pk} in the case of HWD.

Chapter 5

Approximate Variance for the Composite Correlation Measure of LD for Genotypic Data

5.1 Variance Approximation Using the Delta Method

An approximate variance formula is derived in this chapter for the composite correlation estimate of LD, r_C (1.10).

The sample correlation coefficient is a function of the nine genotypic counts which include the unphased double heterozygotes (Table 1.5). The genotypic counts

$$\mathbf{x} = (n^{AB}_{AB}, n^{AB}_{Ab}, n^{Ab}_{Ab}, nP^{AB}_{aB}, n^{AaBb}, n^{Ab}_{ab}, n^{aB}_{aB}, n^{aB}_{ab}, n^{ab}_{ab})^T$$

follow a multinomial distribution with probabilities from Table 1.3

$$\mathbf{p} = (p_1, \ldots, p_9)^T = n(P^{AB}_{AB}, 2P^{AB}_{Ab}, P^{Ab}_{Ab}, 2P^{AB}_{aB}, 2(P^{AB}_{ab} + P^{Ab}_{aB}), 2P^{Ab}_{ab}, P^{aB}_{aB}, 2P^{aB}_{ab}, P^{ab}_{ab})^T$$

and variance covariance matrix $\mathbf{V}$, which is a 9×9 matrix composed of diagonal elements

$$V_{ii} = Cov(X_i, X_i) = Var(X) = np_i(1 - p_i)$$

and off diagonal elements,

$$V_{ij} = Cov(X_i, X_j) = -np_ip_j.$$

The variance approximation is calculated as

$$Var(r_C) = \mathbf{g}^T\mathbf{V}\mathbf{g}$$

where $\mathbf{g} = (g_1, \ldots, g_9)^T$ and

$$g_i = \left.\frac{\partial r_C}{\partial x_i}\right|_{n\mathbf{p}}$$

is the derivative of the correlation with respect to the genotypic counts evaluated at the mean.

The gradient **g** was constructed of the partial derivatives with respect to each genotypic cell count (x_i) using the *diff* operator in *Maple*. The derivatives were then evaluated at the mean counts by replacing each count x_i by $np_i, i = 1, \dots, 9$. The genotypic probabilities were expressed in terms of the allele frequencies and disequilibrium coefficients, as shown in Table 5.1 (Weir and Cockerham, 1989), where the notation $\pi_A = p_A p_a$, $\pi_B = p_B p_b$, $\tau_A = 1 - 2p_A$, and $\tau_B = 1 - 2p_B$ is used to simplify the expressions.

Probability	Formula					
P^{AB}_{AB}	$P^A_A P^B_B$	$+2p_Ap_B\Delta_{AB}$	$+\Delta^2_{AB}$	$+2p_AD_{ABB}$	$+2p_BD_{AAB}$	$+\Delta_{AABB}$
P^{AB}_{Ab}	$P^A_A P^B_b$	$+p_A\tau_B\Delta_{AB}$	$-\Delta^2_{AB}$	$-2p_AD_{ABB}$	$+\tau_BD_{AAB}$	$-\Delta_{AABB}$
P^{Ab}_{Ab}	$P^A_A P^b_b$	$-2p_Ap_b\Delta_{AB}$	$+\Delta^2_{AB}$	$+2p_AD_{ABB}$	$-2p_bD_{AAB}$	$+\Delta_{AABB}$
P^{AB}_{aB}	$P^A_a P^B_B$	$+p_B\tau_A\Delta_{AB}$	$-\Delta^2_{AB}$	$+\tau_AD_{ABB}$	$-2p_BD_{AAB}$	$-\Delta_{AABB}$
P_{AaBb}	$P^A_a P^B_b$	$+\tau_B\tau_A\Delta_{AB}$	$+2\Delta^2_{AB}$	$-2\tau_AD_{ABB}$	$-2\tau_BD_{AAB}$	$+2\Delta_{AABB}$
P^{Ab}_{ab}	$P^A_a P^b_b$	$-\tau_Ap_b\Delta_{AB}$	$-\Delta^2_{AB}$	$+\tau_AD_{ABB}$	$+2p_bD_{AAB}$	$-\Delta_{AABB}$
P^{aB}_{aB}	$P^a_a P^B_B$	$-2p_ap_B\Delta_{AB}$	$+\Delta^2_{AB}$	$-2p_aD_{ABB}$	$+2p_BD_{AAB}$	$+\Delta_{AABB}$
P^{aB}_{ab}	$P^a_a P^B_b$	$-p_a\tau_B\Delta_{AB}$	$-\Delta^2_{AB}$	$+2p_aD_{ABB}$	$+\tau_BD_{AAB}$	$-\Delta_{AABB}$
P^{ab}_{ab}	$P^a_a P^b_b$	$+2p_ap_b\Delta_{AB}$	$+\Delta^2_{AB}$	$-2p_aD_{ABB}$	$-2p_bD_{AAB}$	$+\Delta_{AABB}$

Table 5.1: Genotypic probabilities expressed in terms of allelic probabilities and disequilibrium coefficients when the phase is unknown.

Weir (1996) derived these probabilities using the quadrigenic disequilibrium measure for the phase unknown case

$$\begin{aligned}\Delta_{AABB} &= D^{AB}_{AB} - 2D_{AB}D_{A/B} \\ &= P^{AB}_{AB} - 2p_AD_{ABB} - 2p_BD_{AAB} - 2p_Ap_B\Delta_{AB} - \Delta^2_{AB} \\ &\quad - p^2_AD_B - p^2_BD_A - D_AD_B - p^2_Ap^2_B,\end{aligned} \tag{5.1}$$

which accounts for the remaining disequilibrium after all other forms are removed.

Maple produced output for the variance formula which filled numerous pages (Appendix A). Further work was carried out to simplify the formula to the form displayed in Table 5.2. This formula is a polynomial in Δ, D_{AAB}, D_{ABB}, and Δ_{AABB} with coefficients which are functions of p_A, p_B, D_A and D_B.

Term	Coefficient
Denominator	$n(\pi_A + D_A)^3(\pi_B + D_B)^3$
Numerator	
1	$(\pi_A + D_A)^3(\pi_B + D_B)^3$
Δ	$.5(\pi_A + D_A)^2(\pi_B + D_B)^2\tau_A\tau_B$
Δ^2	$.375(-\pi_B\tau_B^2D_A^2 - \pi_A\tau_A^2D_B^2 + 2(\pi_A + \pi_B - 8\pi_A\pi_B)D_AD_B$ $+\pi_B(\pi_B - 2\pi_A)D_A + \pi_A(\pi_A - 2\pi_B)D_B - \pi_A\pi_B((\pi_A + \pi_B) - 4\pi_A\pi_B)$ $+(4D_AD_B + D_A + D_B)D_AD_B)$
Δ^3	$.25(\pi_A + D_A)(\pi_B + D_B)\tau_A\tau_B$
Δ^4	$.5(\pi_A + D_A)(\pi_B + D_B)$
D_{AAB}	$(\pi_A + D_A)^2(\pi_B + D_B)^2\tau_B$
D_{ABB}	$(\pi_A + D_A)^2(\pi_B + D_B)^2\tau_A$
ΔD_{AAB}	$-3(\pi_A + D_A)(\pi_B + D_B)^2\tau_A$
ΔD_{ABB}	$-3(\pi_A + D_A)^2(\pi_B + D_B)\tau_B$
$\Delta^2 D_{AAB}$	$.5(\pi_A + D_A)(\pi_B + D_B)\tau_B$
$\Delta^2 D_{ABB}$	$.5(\pi_A + D_A)(\pi_B + D_B)\tau_A$
Δ_{AABB}	$(\pi_A + D_A)^2(\pi_B + D_B)^2$
$\Delta^2\Delta_{AABB}$	$.5(\pi_A + D_A)(\pi_B + D_B)$

Table 5.2: Variance formula for r_C

When all the disequilibrium coefficients are zero, the variance is $1/n$.

5.2 Behaviour of the Approximate Variance

The behaviour of the variance formula was studied graphically with $D_{AAB} = D_{ABB} = \Delta_{AABB} = 0$ when there is HWE ($D_A = D_B = 0$, Figure 5.1) and when there is HWD (D_A, $D_B \neq 0$, Figure 5.2). These plots are similar to those of previous chapters. When there is HWE and $p_A = 0.5$ (left panel, Figure 5.1), the variance reaches its maximum at $\rho = 0$ and decreases as ρ increases in magnitude. As p_B becomes more extreme the range of possible values for ρ decreases and the variance decreases more rapidly from its maximum as ρ increases in magnitude. As p_A increases from 0.5 (right panel, 5.1) the maximum variance increases in magnitude, and it occurs at larger values of ρ.

Figure 5.2 displays the scaled variance of r_C as a function of p_A, p_B, and ρ_C when there is HWD. In the two left panels $D_B = 0$, and in the two right panels $D_B = 0.25D_{Bmax}$, where $D_{Bmax} = p_Bp_b$. Within each panel D_A is 0, 0.25, 0.5 and 0.75 of its maximum, $D_{Amax} = p_Ap_a$. The top two panels have allele frequencies

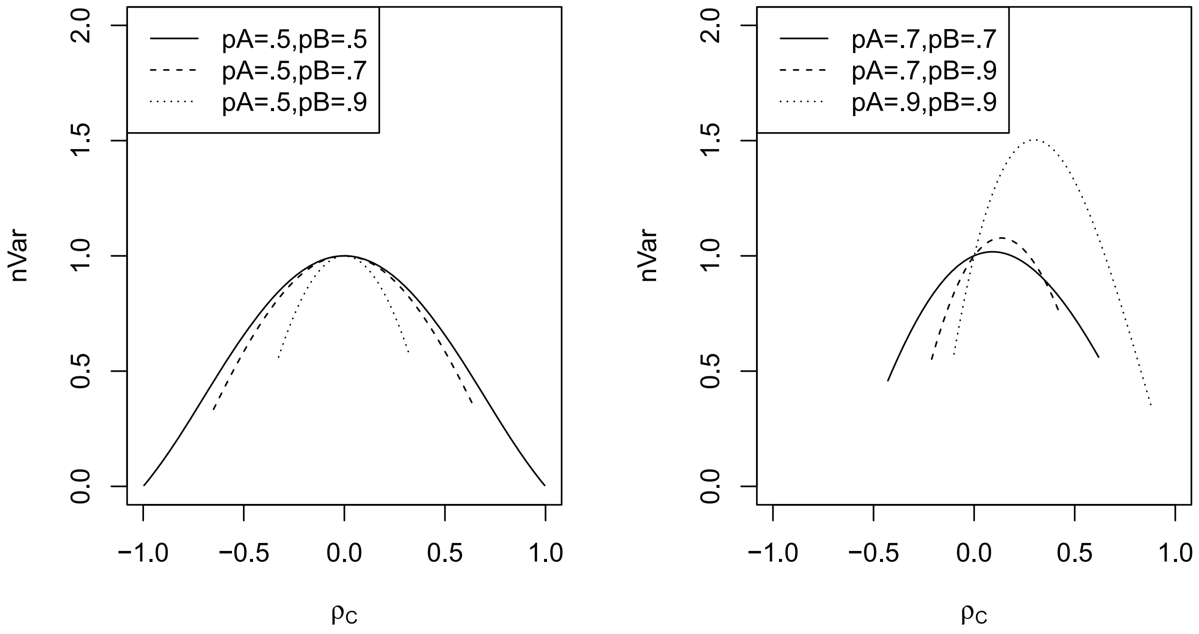

Figure 5.1: Asymptotic variance as a function of ρ_C and several choices of p_A and p_B assuming $(D_A = D_B = D_{AAB} = D_{ABB} = \Delta_{AABB} = 0)$.

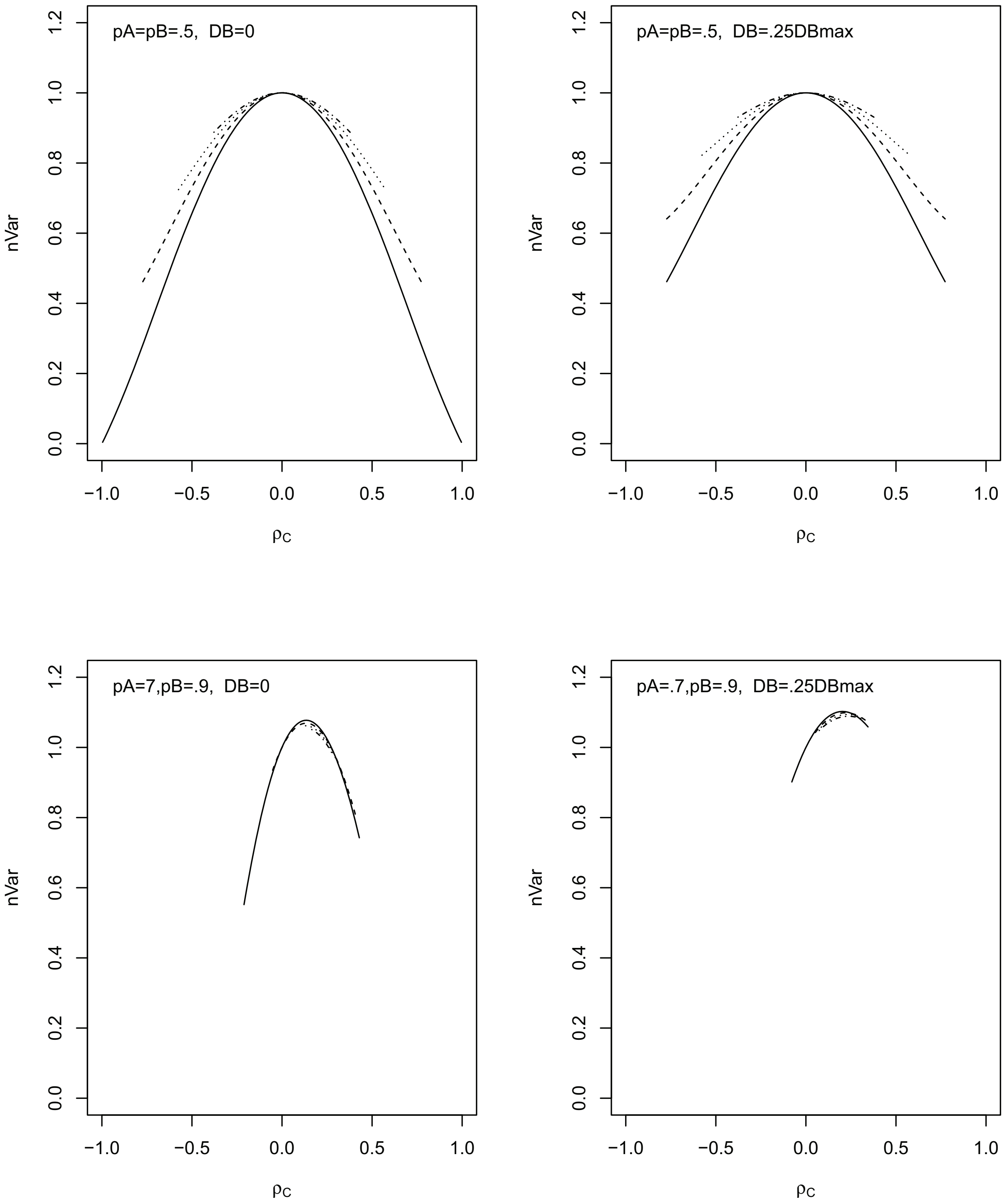

Figure 5.2: Asymptotic variance as a function of ρ_C and several choices of p_A and p_B the HW coefficients of D_A and D_B. Lines define the functional values for D_A=0 (solid), D_A=.25$D_{A,max}$ (dashed), D_A=.5D_{Amax} (dotted), D_A=.75D_{Amax} (dotted-dashed).

$p_A = p_B = 0.5$, while the bottom panels have $p_A = 0.7$, $p_B = 0.9$. In the top two panels where $p_A = p_B = 0.5$, the variance decreases more slowly from its maximum as a function of ρ_C as D_A increases. There is less dependence on D_A in the lower two panels and the range of feasible values for ρ_C is smaller.

5.3 Comparision of Theoretical and Simulated Variance

As in the previous chapters, the variance approximation is validated using simulation. Values for p_A and p_B were chosen to be combinations of 0.5, 0.7 and 0.9, while values of ρ_C ranged over positive and negative values of 0.1, 0.3, 0.5, 0.9. The simulation used $m = 20,000$ random samples of $n = 100$ and $n = 1000$ for both the HWE and HWD cases. The variance of the simulated correlation values was calculated and compared to the approximate variance using a ratio. Simulation standard errors were also calculated and the sampling distributions were examined. The HWE case was produced with all disequilibrium values D_A, D_B, D_{AAB}, D_{ABB} and Δ_{AABB} set to zero. The HWD case used $D_A = 0.5D_{Amax}$ and $D_B = 0.5D_{Bmax}$, with higher order disequilibriums set to zero ($D_{AAB} = D_{ABB} = \Delta_{AABB} = 0$). Some of the combinations of ρ_C, p_A and p_B do not produce viable genotypic probabilities. These combinations are indicated in Table 5.3 and 5.4 by asterisks.

HWE Results

Table 5.3 displays the ratios of the simulated variance to the asymptotic variance when all disequilibria are zero in the HWE case. All ratios are close to one so the variance formula appears to be valid. Simulation standard errors were all approximately 0.01 for each ratio. For $n = 100$, roughly a third of the ratios are greater than 1.02, and only five are less than one indicating the approximate variance tends to underestimate the true variance. The ratios above 1.02 usually correspond to large values of p_A, p_B, or ρ_C. For $n = 1000$, most of the ratios are closer to one and all cases have ratios inside the range 0.98 and 1.02.

For $n = 100$ the sampling distributions of the simulated correlation values are skewed in several cases corresponding to the more extreme allele frequencies and values of ρ. All sampling distributions are symmetric when $n = 1000$.

		ρ_C							
p_A	p_B	-0.9	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.9
$n = 100$									
.5	.5	1.04	1.01	1.03	1.01	1.03	1.00	1.01	1.02
	.7	*	1.02	1.02	1.00	0.98	1.02	1.03	*
	.9	*	*	1.03	1.02	1.00	1.02	*	*
.7	.5	*	1.03	1.02	1.01	1.02	1.02	1.00	*
	.7	*	*	1.03	0.99	1.02	1.04	1.01	1.04
	.9	*	*	*	1.02	1.01	1.02	1.02	*
.9	.5	*	*	1.03	1.00	0.99	1.03	*	*
	.7	*	*	*	1.03	1.00	1.02	1.04	*
	.9	*	*	*	1.04	1.01	0.99	0.99	1.07
$n = 1000$									
.5	.5	0.99	0.99	1.00	0.99	0.99	0.99	0.99	1.02
	.7	*	0.98	0.99	1.00	1.00	1.01	0.99	*
	.9	*	*	1.01	1.00	1.01	1.02	*	*
.7	.5	*	1.00	0.99	1.00	1.01	1.01	0.99	*
	.7	*	*	0.98	0.98	1.01	1.00	1.00	1.01
	.9	*	*	*	1.00	1.00	1.00	1.00	*
.9	.5	*	*	0.98	1.00	1.00	1.00	*	*
	.7	*	*	*	1.00	1.01	1.01	1.01	*
	.9	*	*	*	0.99	0.98	1.01	1.00	1.01

Table 5.3: Ratio of simulated variance to the theoretical variance of r_C when $D_A = D_B = D_{AAB} = D_{ABB} = \Delta_{AABB} = 0$.

HWD Results

		ρ_C					
p_A	p_B	-0.5	-0.3	-0.1	0.1	0.3	0.5
$n = 100$							
.5	.5	1.03	1.02	1.01	1.02	1.01	1.01
	.7	*	1.02	1.02	0.99	1.01	*
	.9	*	*	0.99	1.00	*	*
.7	.5	*	1.02	1.01	1.00	1.02	*
	.7	*	0.99	1.01	1.01	1.01	*
	.9	*	*	1.01	1.02	*	*
.9	.5	*	*	0.99	1.01	*	*
	.7	*	*	1.00	1.01	*	*
	.9	*	*	*	1.03	1.03	*
$n = 1000$							
.5	.5	1.00	1.00	1.02	1.00	0.99	1.00
	.7	*	1.01	0.99	1.00	1.00	*
	.9	*	*	1.00	1.02	*	*
.7	.5	*	1.00	0.99	1.01	0.99	*
	.7	*	1.01	1.01	1.00	0.99	*
	.9	*	*	0.99	1.02	*	*
.9	.5	*	*	0.99	1.01	*	*
	.7	*	*	1.00	0.99	*	*
	.9	*	*	*	0.99	1.01	*

Table 5.4: Ratio of simulated variance to the theoretical variance of r_C when there is HWD.

Table 5.4 shows that the simulated variance ratios are mostly near one when there is HWD. The larger number of asterisks reflects the fact that the boundaries of ρ_C are more restricted when D_A and D_B are not zero. For $n = 100$ all but three of the 29 ratios are within 2 simulation SEs from one ([0.98,1.02]), while fewer (4) ratios lay below one and than above one (21). For the case of $n = 1000$ all ratios are within 2 SEs from unity with a more even distribution above and below one.

For $n = 100$ the sampling distributions of the correlations are skewed in some cases, corresponding to values of ρ_C near its boundary (-0.5 and 0.5) or extreme allelic frequencies ($p_A = p_B = 0.7$ and $p_A = p_B = 0.9$). When $n = 1000$ all sampling distributions are symmetric indicating that asymptotic conditions are better met with the larger sample size.

5.3.1 Simulations with the Fisher Transformation

The Fisher transformation was again utilized in order to see whether the variances approximation was improved on this scale.

Simulated variance ratios for the Fisher transformed correlation are shown in Table 5.5 for the case of HWE and Table 5.6 for the case of HWD.

For the case of HWE when $n = 100$, the transformed ratios are the same as or better than the untransformed ratios two thirds the time. However, ratios were much poorer near the boundaries of ρ. When $n = 1000$ transformed variance ratios matched the untransformed ratio in 43 out of the 47 ratios and improved it in 3 of the ratios.

		ρ_C							
p_A	p_B	-0.9	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.9
$n = 100$									
.5	.5	0.93	0.99	1.02	1.02	1.03	1.00	0.99	0.92
	.7	*	0.99	1.00	1.00	1.00	1.00	0.97	*
	.9	*	*	1.02	0.98	1.01	1.02	*	*
.7	.5	*	1.01	1.00	1.01	1.03	1.00	1.02	*
	.7	*	*	1.02	1.02	1.02	1.02	1.01	0.94
	.9	*	*	*	1.03	1.01	1.01	1.01	*
.9	.5	*	*	1.01	1.00	0.98	1.02	*	*
	.7	*	*	*	1.01	1.01	0.99	1.00	*
	.9	*	*	*	1.02	1.01	1.01	1.03	1.27
$n = 1000$									
.5	.5	0.98	0.99	1.00	0.99	0.99	0.99	0.99	1.00
	.7	*	0.98	0.99	1.00	1.00	1.01	0.99	*
	.9	*	*	1.01	1.00	1.01	1.01	*	*
.7	.5	*	1.00	0.99	1.00	1.01	1.01	0.99	*
	.7	*	*	0.98	0.98	1.01	1.00	1.00	1.00
	.9	*	*	*	1.00	1.00	1.00	1.00	*
.9	.5	*	*	0.98	1.00	1.00	1.00	*	*
	.7	*	*	*	1.00	1.01	1.01	1.01	*
	.9	*	*	*	0.99	0.98	1.01	1.00	1.01

Table 5.5: Ratio of the simulated variance to the theoretical variance of the Fisher transform correlation when $D_A = D_B = D_{AAB} = D_{ABB} = \Delta_{AABB} = 0$.

For the case of HWD (Table 5.6) when $n = 100$ the transformation improved the ratio in 6 cases, made it worse in 6 cases and kept in the same in 16 cases. When $n = 1000$ the transformation made the ratio closer to one in 6 cases, further

		ρ_C					
p_A	p_B	-0.5	-0.3	-0.1	0.1	0.3	0.5
$n = 100$							
.5	.5	1.03	1.02	1.00	1.03	1.01	1.03
	.7	*	0.98	1,02	0.99	1.00	*
	.9	*	*	1.01	1.00	*	*
.7	.5	*	1.00	1.01	1.01	1.02	*
	.7	*	1.01	1.01	1.02	1.00	*
	.9	*	*	0.99	1.02	*	*
.9	.5	*	*	1.00	1.00	*	*
	.7	*	*	0.99	1.03	*	*
	.9	*	*	*	1.04	1.14	*
$n = 1000$							
.5	.5	1.00	1.00	1.01	1.02	1.00	1.00
	.7	*	1.01	1.00	1.03	0.99	*
	.9	*	*	0.99	0.99	*	*
.7	.5	*	0.99	1.01	1.00	0.98	*
	.7	*	1.02	1.00	1.00	0.98	*
	.9	*	*	1.01	1.02	*	*
.9	.5	*	*	0.99	0.98	*	*
	.7	*	*	1.04	0.98	*	*
	.9	*	*	*	1.01	1.02	*

Table 5.6: Ratio of simulated Fisher transform variance to the theoretical variance of r_C in the case of HWD.

from one in 12 cases and kept it the same in 10 cases.

Chapter 6

Discussion

6.1 Comparison of the Variance Formulae

The variance formulae are quite complicated but some comparisons can be made. Table 6.1 displays all four approximate variance formulae under the assumption of random mating. For the phase known and composite cases this requires equating to zero the coefficients D_A, D_B, D_{AAB}, D_{ABB}, $D_{A/B}$ and D_{AABB} or Δ_{AABB} in their respective variance formulae (Table 4.2 and Table 5.2). When comparing to the gametic variance note that n genotypes correspond to $2N$ haplotypes.

With this conversion the gametic variance and the phase known variances are equal. This reflects the fact that the same amount of genetic information is obtained from genotypic data as gametic data when there is random mating when the phase is assumed known.

The variance formula for the random mating case is the most complicated, with extra powers of D_{AB} in the numerator and denominator. The variance of the composite measure has several coefficients similar to those of the gametic and phase known cases.

When $D_{AB} = 0$ the gametic variance is $1/N = 1/2n$, the phase known variance is $1/2n$, the random mating variance is $1/n$ and the composite variance is $1/n$. The phase known assumption is a stronger assumption than the assumption of random mating leading to a smaller variance. The random mating and composite variances at $D_{AB} = 0$ are twice as big.

An examination of the variance formulae for the unstandardized LD measures (1.11), (1.12) and (1.13) is complicated by the fact that the formula (1.12) has not been simplified completely for the random mating case. This simplification was carried out in Maple and the result is shown in Table 6.2. When $D_{AB} = 0$, $V(\hat{D}_{ABrm}) = \pi_A \pi_B / n$ which equals $V(\hat{\Delta}_{AB})$ from the composite case. However the gametic and phase known variances are $V(\hat{D}_{ABg}) = V(\hat{D}_{ABpk}) = \pi_A \pi_B / 2n$, so the

	Formulae							
Cases	Denominator		Numerator					
	1		1	D_{AB}	D_{AB}^2	D_{AB}^3	D_{AB}^4	D_{AB}^5
Gametic	$\pi_A^2\pi_B^2 N$		$\pi_A^2\pi_B^2$	$\pi_A\pi_B\tau_A\tau_B$	$(-3(\pi_A+\pi_B)+20\pi_A\pi_B)/4$	$\tau_A\tau_B/2$		
Genotypic								
Phase Known	$\pi_A^2\pi_B^2 2n$		$\pi_A^2\pi_B^2$	$\pi_A\pi_B\tau_A\tau_B$	$(-3(\pi_A+\pi_B)+20\pi_A\pi_B)/4$	$\tau_A\tau_B/2$		
Random	1	$8\pi_A^3\pi_B^3 n$	$8\pi_A^3\pi_B^3$	$12\pi_A^2\pi_B^2\tau_A\tau_B$	$\pi_A\pi_B(120\pi_A\pi_B$	$\tau_A\tau_B(38\pi_A\pi_B$	$96\pi_A\pi_B+2$	$6\tau_A\tau_B$
Mating	D_{AB}	$8\pi_A^2\pi_B^2\tau_A\tau_B n$			$-23(\pi_A+\pi_B)+4)$	$-3(\pi_A+\pi_B))$	$-17(\pi_A+\pi_B)$	
	D_{AB}^2	$24\pi_A^2\pi_B^2 n$						
Composite	$\pi_A^2\pi_B^2 n$		$\pi_A^2\pi_B^2$	$\pi_A\pi_B\tau_A\tau_B/2$	$(-3(\pi_A+\pi_B)+12\pi_A\pi_B)/8$	$\tau_A\tau_B/4$	1/2	

Table 6.1: Comparison of each of the four approximate variance formulas under random mating.

doubling of variance when phase is unknown occurs for the unstandardized measures as it does for the correlation standardization.

Term	Coefficient
Denominator	
1	$2\pi_A\pi_B n$
D_{AB}	$2\tau_A\tau_B n$
D_{AB}^2	$6n$
Numerator	
1	$2\pi_A^2\pi_B^2$
D_{AB}	$3\tau_A\tau_B\pi_A\pi_B$
D_{AB}^2	$-p_A^2(24\pi_B+5)+p_A(24\pi_B-5)-5\pi_B+1$
D_{AB}^3	$3\tau_A\tau_B$
D_{AB}^4	-2

Table 6.2: Simplified variance formula for $\hat{D}_{ABrm}$.

6.2 Applications to Real Data

The approximate variance formulae are applied to real genotypic data in this section. Table 6.3 gives genotypic counts at the 990 and 986 loci of the calcium sensing receptor (CASR) gene for primary hyperparathyroidism (PHPT) patients and for a control group (Hamilton and Cole, 2008). Note that the variant alleles S and G do not occur together except possibly with the unphased double heterozygotes $ASRG$. The wild type allele frequencies are both fairly extreme in both groups. The correlation measures are shown, and the phase known approach was used twice, assuming the double heterozygotes are *in-repulsion* and *in-coupling.* The near equivalence of the *in-repulsion* measure $r_{pk,R}$ to the random mating measure r_{rm} suggests that the double heterozygotes are *in-repulsion.* The composite measure is bigger in magnitude, which may reflect the fact that it includes both inter-gametic and intra-gametic disequilbria. The variance for the phase known measure is the smallest when *in-repulsion* is assumed, but not too different from the random mating variance. The composite measure has the largest variance. The results are similar in the cases and controls.

Table 6.4 gives genotypic counts and summary statistics for a Xavante Indian population sample and a sample of Irish republicans at the MN and S blood group

Genotypes	Cases	Controls
AARR	116	247
AARG	18	35
AAGG	0	1
ASRR	80	124
ASRG	1	6
ASGG	0	0
SSRR	22	20
SSRG	0	0
SSGG	0	0
Total	237	433
Statistics	Cases	Controls
p_A	0.7362	0.8037
p_R	0.9600	0.9503
D_A	0.0233	.0077
D_R	-.0016	-.0002
D_{AR}	-.0105	-.0085
$D_{A/R}$	-.0097	-.0028
Δ_{AR}	-.0190	-.0126
$r_{pk,C}$	-.0979	-.0329
$r_{pk,R}$	-.1222	-.1130
r_{rm}	-.1223	-.1130
r_C	-.2125	-.1426
Variances		
$Vr_{pk,C}$	.00068345	.00090469
$Vr_{pk,R}$	.00021119	.00009087
Vr_{rm}	.00022332	.00009178
Vr_C	.00113113	.00096164

Table 6.3: Observed counts, disequilibrium coefficients, and approximate variances of r for PHPT cases and controls at the AS and RG loci (Hamilton and Cole, 2008).

loci (Hamilton and Cole, 2008). Note that the variant alleles N and s do occur together in this data set so we have an example of composite data; we cannot infer the phase of the double heterozygotes $MNSs$. The wild type allele frequencies are more moderate in this example. The variances of the correlation measures are smaller for the Irish Republicans, in part due to the larger sample size. For the Xavante Indians the random mating and composite variance are nearly twice as large as the phase known variances. For the Irish Republicans the composite measure has the largest variance and the *in-coupling* phase known measure has the smallest.

Genotypes	Xavante	Irish
MMSS	91	121
MMSs	147	248
MMss	85	164
MNSS	32	53
MNSs	78	422
MNss	75	375
NNSS	5	9
NNSs	17	65
NNss	7	241
Total	537	1698
Statistics	Xavante	Irish
p_M	.4637	.3242
p_S	.7737	.5642
D_M	.0234	.0027
D_S	.0028	-.0044
Δ_{MS}	.0273	.0782
$r_{pk,C}$	.2395	.4362
$r_{pk,R}$	-.1085	-.0992
r_{rm}	.1122	.3568
r_C	.1242	.3380
Variances		
$Vr_{pk,C}$	.00069747	.00017709
$Vr_{pk,R}$	.00083947	.00034275
Vr_{rm}	.00173746	.00035731
Vr_C	.00168318	.00047672

Table 6.4: Observed counts, disequilibrium coefficients, and approximate variances of r for the Xavante Indian and Irish Republican populations at the MN and S blood group loci.

6.3 Statistical Advantages of Correlation Measure of LD

Zapata (2011); Teare et al. (2002) deem D'_{AB} and ρ to be most commonly used measures of LD. While the standardized measure D'_{AB} has the advantage that it can take values from -1 to 1, in this section we will look at some statistical advantages of the correlation measure.

6.3.1 Smooth Variance Formulae when D' is not Smooth

The approximate variance formulae for ρ in this paper are all smooth functions of D_{AB} while those for D'_{AB} and Δ'_{AB} are not smooth functions (Zapata, 1997; Hamilton et al., 2006). A discontinuity in the variance occurs at D_{AB} or $\Delta_{AB} = 0$ because the standardizing constant is different for negative and positive values. The standardized value is the same in all cases for the correlation measures. The difference in the standardized constant also leads to unusual sampling distributions for D'_{AB} and Δ'_{AB} in some cases.

6.3.2 Testing for LD and Association

To test the null hypothesis that there is no LD, an obvious test statistic is $\hat{D}_{AB}$ divided by its standard error. For gametic data we get

$$X^2 = \frac{\hat{D}^2_{AB}}{\pi_A \pi_B / N} = Nr^2$$

after squaring and evaluating the variance (1.11) at $D_{AB} = 0$. This statistic is equivalent to the goodness of fit test for independence in a 2 × 2 table, and is approximately X^2_1 in large samples. The correlation measure, or its square, occurs naturally as a scaled version of the test statistic. With genotypic data, the test statistics for LD can all be written in terms of the correlation standardization.

In related work, Pritchard and Przeworski (2001) investigated the loss of information when testing for association between a disease phenotype and a marker rather than the disease susceptibility locus (DSL). They showed that the sample size required to achieve the same power is $1/r^2$ larger using the marker locus, where r is the correlation measure of LD between the marker and disease susceptibility locus, i.e.

$$n_{Marker} = n_{DSL}/r^2.$$

6.4 Summary

A delta method approximation to the variance formula of the correlation measure of LD was produced for gametic data and then for three different cases of genotypic data. In each case the variance formula was derived and plotted for several choices of the allele frequencies and for all feasible values of ρ. Simulations were then carried out to compare the variance of simulated LD correlation values to the theoretical variance formula. Results indicate that the variance formulae are good approximations for each type of genetic data. As expected, results were poorer for the small sample size, $n = 100$, than for the large sample size $n = 1000$. Fisher's transformation improved the approximation in some cases, made it worse in others and had little effect most of the time. In practise, use of Fisher's transformation may not be a sufficient improvement to warrant its use.

Appendix A

Maple Code

Gametic Case

```
pAh := (x1 + x2)/n; pah := 1 − pAh;
pBh := (x1 + x3)/n; pbh := 1 − pBh;
pABh := x1/n;
Dh := −pAh ∗ pBh + pABh;
vAh := pAh ∗ pah; vBh := pBh ∗ pbh;
rh := Dh/sqrt(vAh ∗ vBh);
a1 := simplify(diff(rh, x1)); a2 := simplify(diff(rh, x2));
a3 := simplify(diff(rh, x3)); a4 := simplify(diff(rh, x4));
avec :=< (a1, a2, a3, a4) >;
ps :=< (p1, p2, p3, p4) >;
with(LinearAlgebra);
vmat := n ∗ (Matrix(1..4, 1..4, ps, shape =
diagonal) − OuterProductMatrix(ps, ps, compact = false));
p1 := pA ∗ pB + D; p2 := pA ∗ pb − D;
p3 := pB ∗ pa − D; p4 := pa ∗ pb + D;
x1 := n ∗ p1; x2 := n ∗ p2; x3 := n ∗ p3; x4 := n ∗ p4;
pa := 1 − pA; pb := 1 − pB;
piA := pA ∗ pa; piB := pB ∗ pb;
tauA := 1 − 2 ∗ pA; tauB := 1 − 2 ∗ pB;
aVar := simplify(Transpose(avec).vmat.avec);
aVar := 4∗pA^4∗pB^4+16∗D∗pA^3∗pB^3−8∗pA^4∗pB^3−8∗pA^3∗pB^4+20∗D^2∗pA^2∗
pB^2−24∗D∗pA^3∗pB^2−24∗D∗pA^2∗pB^3+4∗pA^4∗pB^2+16∗pA^3∗pB^3+4∗pA^2∗pB^4+
8∗D^3∗pA∗pB−20∗D^2∗pA^2∗pB−20∗D^2∗pA∗pB^2+8∗D∗pA^3∗pB+36∗D∗pA^2∗
pB^2+8∗D∗pA∗pB^3−8∗pA^3∗pB^2−8∗pA^2∗pB^3−4∗D^3∗pA−4∗D^3∗pB+3∗D^2∗
pA^2+20∗D^2∗pA∗pB+3∗D^2∗pB^2−12∗D∗pA^2∗pB−12∗D∗pA∗pB^2+4∗pA^2∗pB^2+
```

$2*D^3-3*D^2*pA-3*D^2*pB+4*D*pA*pB/4*n*pA^2*pB^2*(-1+pB)^2*(-1+pA)^2$

den := denom(aVar);

num := numer(aVar);

num := collect(num, D); $num3 := coeff(num, D, 3);$

$simplify(num3/(2*tauA*tauB));$ 1

$num2 := coeff(num, D, 2);$

$simplify(num2/(-3*(piA+piB)+20*piA*piB));$

1

$num1 := coeff(num, D, 1);$

$simplify(num1/(4*piA*piB*tauA*tauB));$

1

$num0 := coeff(num, D, 0);$

$simplify(num0/(4*piA^2*piB^2));$

1

$simplify(den/(4*piA^2*piB^2*n));$

1

$simplify(aVar-(D^3*num3+D^2*num2+D*num1+num0)/den);$

0

Genotypic Data, MLE

$pa := 1-pA; pb := 1-pB;$

$pAb := pA-pAB; pab := pa-paB; paB := pB-pAB;$

$x11 := pAB*pAB; x12 := 2*pAB*pAb; x13 := pAb*pAb;$

$x21 := 2*pAB*paB; x22 := 2*pAB*pab+2*pAb*paB; x23 := 2*pAb*pab;$

$x31 := paB*paB; x32 := 2*paB*pab; x33 := pab*pab;$

simplify(x11+x12+x13+x21+x22+x23+x31+x32+x33);

1 $logLike := n11*log(x11)+n12*log(x12)+n13*log(x13)+n21*log(x21)+$ $n22*log(x22)+n23*log(x23)+n31*log(x31)+n32*log(x32)+n33*log(x33);$

d11 := simplify(-(diff(diff(logLike, pA), pA)));

d12 := simplify(-(diff(diff(logLike, pA), pB)));

d13 := simplify(-(diff(diff(logLike, pA), pAB)));

d21 := simplify(-(diff(diff(logLike, pB), pA)));

d22 := simplify(-(diff(diff(logLike, pB), pB)));

d23 := simplify(-(diff(diff(logLike, pB), pAB)));

d31 := simplify(-(diff(diff(logLike, pAB), pA)));

d32 := simplify(-(diff(diff(logLike, pAB), pB)));

d33 := simplify(-(diff(diff(logLike, pAB), pAB)));

with(LinearAlgebra);

Imat := Matrix(3, 3, [d11, d12, d13, d21, d22, d23, d31, d32, d33]);

$n11 := n \ast x11;\ n12 := n \ast x12;\ n13 := n \ast x13;$

$n21 := n \ast x21;\ n22 := n \ast x22;\ n23 := n \ast x23;$

$n31 := n \ast x31;\ n32 := n \ast x32;\ n33 := n \ast x33;$

simplify(n11+n12+n13+n21+n22+n23+n31+n32+n33);

invEImat := MatrixInverse(Imat);

$ro := (-pA \ast pB + pAB)/sqrt(pA \ast pB \ast pa \ast pb);$

dro1 := simplify(diff(ro, pA));

dro2 := simplify(diff(ro, pB));

dro3 := simplify(diff(ro, pAB));

pAB := pA*pB+D;

$roVec :=< (dro1, dro2, dro3) >;$

$roVar := simplify(Transpose(roVec).invEImat.roVec);$

$rovar := roVar[1, 1];$

$rovar := (1/8) \ast (8 \ast pA^6 \ast pB^6 + 48 \ast D \ast pA^5 \ast pB^5 - 24 \ast pA^6 \ast pB^5 - 24 \ast pA^5 \ast pB^6 + 120 \ast$
$D^2 \ast pA^4 \ast pB^4 - 120 \ast D \ast pA^5 \ast pB^4 - 120 \ast D \ast pA^4 \ast pB^5 + 24 \ast pA^6 \ast pB^4 + 72 \ast pA^5 \ast pB^5 +$
$24 \ast pA^4 \ast pB^6 + 152 \ast D^3 \ast pA^3 \ast pB^3 - 240 \ast D^2 \ast pA^4 \ast pB^3 - 240 \ast D^2 \ast pA^3 \ast pB^4 + 96 \ast D \ast$
$pA^5 \ast pB^3 + 300 \ast D \ast pA^4 \ast pB^4 + 96 \ast D \ast pA^3 \ast pB^5 - 8 \ast pA^6 \ast pB^3 - 72 \ast pA^5 \ast pB^4 - 72 \ast$
$pA^4 \ast pB^5 - 8 \ast pA^3 \ast pB^6 + 96 \ast D^4 \ast pA^2 \ast pB^2 - 228 \ast D^3 \ast pA^3 \ast pB^2 - 228 \ast D^3 \ast pA^2 \ast pB^3 +$
$143 \ast D^2 \ast pA^4 \ast pB^2 + 480 \ast D^2 \ast pA^3 \ast pB^3 + 143 \ast D^2 \ast pA^2 \ast pB^4 - 24 \ast D \ast pA^5 \ast pB^2 - 240 \ast$
$D \ast pA^4 \ast pB^3 - 240 \ast D \ast pA^3 \ast pB^4 - 24 \ast D \ast pA^2 \ast pB^5 + 24 \ast pA^5 \ast pB^3 + 72 \ast pA^4 \ast pB^4 + 24 \ast$
$pA^3 \ast pB^5 + 24 \ast D^5 \ast pA \ast pB - 96 \ast D^4 \ast pA^2 \ast pB - 96 \ast D^4 \ast pA \ast pB^2 + 88 \ast D^3 \ast pA^3 \ast pB +$
$342 \ast D^3 \ast pA^2 \ast pB^2 + 88 \ast D^3 \ast pA \ast pB^3 - 23 \ast D^2 \ast pA^4 \ast pB - 286 \ast D^2 \ast pA^3 \ast pB^2 - 286 \ast$
$D^2 \ast pA^2 \ast pB^3 - 23 \ast D^2 \ast pA \ast pB^4 + 60 \ast D \ast pA^4 \ast pB^2 + 192 \ast D \ast pA^3 \ast pB^3 + 60 \ast D \ast pA^2 \ast$
$pB^4 - 24 \ast pA^4 \ast pB^3 - 24 \ast pA^3 \ast pB^4 - 12 \ast D^5 \ast pA - 12 \ast D^5 \ast pB + 17 \ast D^4 \ast pA^2 + 96 \ast D^4 \ast$
$pA \ast pB + 17 \ast D^4 \ast pB^2 - 6 \ast D^3 \ast pA^3 - 132 \ast D^3 \ast pA^2 \ast pB - 132 \ast D^3 \ast pA \ast pB^2 - 6 \ast D^3 \ast$

```
pB^3+46*D^2*pA^3*pB+170*D^2*pA^2*pB^2+46*D^2*pA*pB^3-48*D*pA^3*pB^2-48*
D*pA^2*pB^3+8*pA^3*pB^3+6*D^5-17*D^4*pA-17*D^4*pB+9*D^3*pA^2+50*D^3*
pA*pB+9*D^3*pB^2-27*D^2*pA^2*pB-27*D^2*pA*pB^2+12*D*pA^2*pB^2+2*D^4-
3*D^3*pA-3*D^3*pB+4*D^2*pA*pB)/(pA^2*pB^2*(pA-1)^2*(pB-1)^2*n*(pA^2*
pB^2+4*D*pA*pB-pA^2*pB-pA*pB^2+3*D^2-2*D*pA-2*D*pB+pA*pB+D))
numrovar := simplify(numer(rovar));
denrovar := simplify(denom(rovar));
collect(numrovar, D);
tauA := 1 - 2 * pA; tauB := 1 - 2 * pB;
pa := 1 - pA; pb := 1 - pB;
piA := pa * pA; piB := pb * pB;
num5 := coeff(numrovar, D, 5);
simplify(num5/(6 * tauA * tauB));
1
num4 := coeff(numrovar, D, 4);
simplify(num4/(96 * piA * piB - 17 * (piA + piB) + 2));
1
num3 := coeff(numrovar, D, 3);
simplify(num3/(tauA * tauB * (38 * piA * piB - 3 * (piA + piB))));
1
num2 := coeff(numrovar, D, 2);
simplify(num2/(piA * piB * (120 * piA * piB - 23 * (piA + piB) + 4)));
1
num1 := coeff(numrovar, D, 1);
simplify(num1/(12 * piA^2 * piB^2 * tauB * tauA));
1
num0 := factor(coeff(numrovar, D, 0));
simplify(num0/(8 * piA^3 * piB^3));
1
den0 := factor(coeff(denrovar, D, 0));
simplify(den0/(8 * n * piA^3 * piB^3));
1
```

$den1 := factor(coeff(denrovar, D, 1));$

$simplify(den1/(8*n*piA^2*piB^2*tauA*tauB));$

1

$den2 := factor(coeff(denrovar, D, 2));$

$simplify(den2/(24*n*piA^2*piB^2*``));$

1

$finalcheck := simplify(rovar-(D^5*num5+D^4*num4+D^3*num3+D^2*num2+D*num1+num0)/(D^2*den2+D*den1+den0));$

0

Phase Informed Genotypic Data

$pAh := (x1+x4+x8+.5*(x2+x9+x5+x6))/n :$

$pBh := (x1+x2+x3+.5*(x4+x6+x5+x7))/n :$

pah:=1-pAh: pbh:=1-pBh:

$Dh := (x1+.5*(x4+x2+x5))/n-pAh*pBh :$

$roh := (Dh)/sqrt(pah*pbh*pAh*pBh) :$

g11:=diff(roh,x1):

sg11:= simplify(g11):

g12:=diff(roh,x2): sg12:=simplify(g12):

g13:=diff(roh,x3): sg13:=simplify(g13):

g14:=diff(roh,x4): sg14:=simplify(g14):

g15:=diff(roh,x5): sg15:=simplify(g15):

g16:=diff(roh,x6): sg16:=simplify(g16):

g17:=diff(roh,x7): sg17:=simplify(g17):

g18:=diff(roh,x8): sg18:=simplify(g18):

g19:=diff(roh,x9): sg19:=simplify(g19):

g110:=diff(roh,x10): sg110:=simplify(g110):

tauA:=1-2*pA: tauB:=1-2*pB:

piA := pA*(1-pA): piB:=pB*(1-pB):

pa:=1-pA: pb:=1-pB:

$x1 := n*(pA^2*pB^2+2*pA*DABB+pB*2*DAAB+pA*pB*2*DAB+pA*pB*2*DA_B+DAB^2+DA_B^2+pA^2*DB+pB^2*DA+DA*DB+DAABB) :$

$x2 :=$
$n*(2*piA*pB^2+2*tauA*DABB-pB*4*DAAB+tauA*pB*2*DAB+tauA*pB*2*DA_B-2*DAB^2-2*DA_B^2+2*piA*DB-2*pB^2*DA-2*DA*DB-2*DAABB) :$

$x3 := n*(pa^2*pB^2-2*pa*DABB+pB*2*DAAB-pa*pB*2*DAB-2*pa*pB*DA_B+DAB^2+DA_B^2+pa^2*DB+pB^2*DA+DA*DB+DAABB) :$

$x4 :=$
$n*(2*pA^2*piB-4*pA*DABB+2*tauB*DAAB+2*pA*tauB*DAB+pA*tauB*2*DA_B-2*DAB^2-2*DA_B^2-2*pA^2*DB+2*piB*DA-2*DA*DB-2*DAABB) :$

$x5 := n*(2*piA*piB-2*tauA*DABB-tauB*2*DAAB+2*DAB*(pA*pB+pa*pb)+2*DA_B*(pA*pB+pa*pb-1)+2*DAB^2+2*DA_B^2-2*piA*DB-2*piB*DA+2*DA*DB+2*DAABB) :$

$x6 := n*(2*piA*piB-2*tauA*DABB-2*tauB*DAAB+2*DAB*(pA*pB+pa*pb-1)+2*DA_B*(pA*pB+pa*pb)+2*DAB^2+2*DA_B^2-2*piA*DB-2*piB*DA+2*DA*DB+2*DAABB);$

$x7 := n*(2*pa^2*piB+4*pa*DABB+2*tauB*DAAB-pa*tauB*2*DAB-pa*tauB*2*DA_B-2*DAB^2-2*DA_B^2-2*pa^2*DB+2*piB*DA-2*DA*DB-2*DAABB) :$

$x8 := n*(pA^2*pb^2+2*pA*DABB-pb*2*DAAB-pA*pb*2*DAB-2*pA*pb*DA_B+DAB^2+DA_B^2+pA^2*DB+pb^2*DA+DA*DB+DAABB) :$

$x9 := n*(2*piA*pb^2+2*tauA*DABB+4*pb*DAAB-tauA*pb*2*DAB-tauA*pb*2*DA_B-2*DAB^2-2*DA_B^2+2*piA*DB-2*pb^2*DA-2*DA*DB-2*DAABB) :$

$x10 := n*(pa^2*pb^2-2*pa*DABB-pb*2*DAAB+2*pa*pb*DAB+2*pa*pb*DA_B+DAB^2+DA_B^2+pa^2*DB+pb^2*DA+DA*DB+DAABB) :$

temp:=simplify(x1+x2+x3+x4+x5+x6+x7+x8+x9+x10):

collect(temp, n);

collect(NN, DAB);

sg11m:=simplify(sg11):

sg12m:=simplify(sg12):

sg13m:=simplify(sg13):

sg14m:=simplify(sg14):

sg15m:=simplify(sg15):

sg16m:=simplify(sg16):

sg17m:=simplify(sg17):

sg18m:=simplify(sg18):

sg19m:=simplify(sg19):

sg110m := simplify(sg110);

$gvec :=<$
$sg11m, sg12m, sg13m, sg14m, sg15m, sg16m, sg17m, sg18m, sg19m, sg110m >$:

$ps :=< p1, p2, p3, p4, p5, p6, p7, p8, p9, p10 >$:

with(LinearAlgebra):

$vmat := n * (Matrix(1..10, 1..10, ps, shape =$
$diagonal) - OuterProductMatrix(ps, ps, compact = false)) :$

p1:=x1/n: p2:=x2/n: p3:=x3/n:
p4:=x4/n: p5:=x5/n: p6 := x6/n; p7:=x7/n:
p8:=x8/n: p9:=x9/n: p10:=x10/n:

var1:=simplify(Transpose(gvec).vmat.gvec):

$var1 := -DAAB*DAB*pA^3*pB^2 - DAB*pA^5*pB^2 - DAB*DABB*pA^2*pB^3 +$
$DAAB*DAB*pA*pB^3 + DAB*DABB*pA^3*pB + DAB^2*DB*pA^3*pB + DA*$
$DAB^2*pA*pB^3 - DA*DB*pA^3*pB^2 - DA*DB*pA^2*pB^3 - DAB^2*DB*pA^3*pB^2 -$
$DA*DAB^2*pA^2*pB^3 - DA*DB*pA^4*pB^3 - DA*DB*pA^3*pB^4 - DAAB*DAB*$
$pA^3*pB^4 + DAB^2*DA_B*pA^3*pB^3 - DAB*DABB*pA^4*pB^3 - DAB*pA^2*pB^5 -$
$DAABB*pA^3*pB^2 - DAABB*pA^2*pB^3 - DAABB*pA^4*pB^3 - DAABB*pA^3*$
$pB^4 - DA_B^2*pA^3*pB^2 - DA_B^2*pA^2*pB^3 + DAB^3*pA^3*pB^3 - DA_B^2*pA^4*pB^3 - DA_B^2*$
$pA^3*pB^4 + .5*DA*DB*pA^4*pB^2 + .5*DA*DAB^2*pA^2*pB^2 + .5*DAB^2*DA_B*pA^3*$
$pB - .5*DAB*DABB*pA^4*pB - .5*DA*DAB^2*pA*pB^2 - .75*DAB^2*DA_B*pA^2*$
$pB - .75*DAB^2*DA_B*pA*pB^2 - .5*DAB^2*DB*pA^2*pB + .5*DA*DB*pA^2*pB^4 -$
$.5*DAAB*DAB*pA*pB^4 + .5*DAB^2*DA_B*pA*pB^3 + .5*DAB^2*DB*pA^2*pB^2 +$
$1.5*DAAB*DAB*pA^2*pB^2 + 1.5*DAB*DABB*pA^2*pB^2 + 2.25*DAB^2*DA_B*$
$pA^2*pB^2 - 3.*DAB*DABB*pA^3*pB^2 - 3.*DAAB*DAB*pA^2*pB^3 - 1.500000000*$
$DAB^2*DA_B*pA^3*pB^2 - 1.500000000*DAB^2*DA_B*pA^2*pB^3 - .5*DAB^2*DB*pA^4*$
$pB + 1.5*DAB*DABB*pA^4*pB^2 + 2.*DAB*DABB*pA^3*pB^3 - .5*DA*DAB^2*pA*$
$pB^4 + 2.*DA*DB*pA^3*pB^3 + 2.*DAAB*DAB*pA^3*pB^3 + 1.5*DAAB*DAB*pA^2*$
$pB^4 + .5*DAB^2*DB*pA^4*pB^2 + .5*DA*DAB^2*pA^2*pB^4 + .5*DA*DB*pA^4*pB^4 +$
$.5*DA*DB*pA^2*pB^2 - .5*DAAB*DAB*pA*pB^2 + .25*DAB^2*DA_B*pA*pB - .5*$
$DAB*DABB*pA^2*pB + .5*DAB^3*pA^3*pB + .5*DAABB*pA^4*pB^2 - .375*DAB^2*$

$pA^4*pB+4.*DAB*pA^5*pB^3-.25*DAB^2*DB*pA^3+.5*DA_B^2*pA^4*pB^2-.25*DA*$
$DAB^2*pB^3+.125*DAB^2*DB*pA^4-.375*DAB^2*pA*pB^4+2.875*DAB^2*pA^4*$
$pB^2+4.*DAB*pA^3*pB^5+2.875*DAB^2*pA^2*pB^4+.5*DA_B^2*pA^2*pB^4+.125*DA*$
$DAB^2*pB^4+.5*DAABB*pA^2*pB^4+.5*DAB^3*pA*pB^3-.75*DAB^3*pA^2*pB-$
$.75*DAB^3*pA*pB^2+.75*DAB^2*pA^3*pB+.75*DAB^2*pA*pB^3-1.5*DAB^3*pA^3*$
$pB^2-1.5*DAB^3*pA^2*pB^3+10.*DAB^2*pA^3*pB^3-10.*DAB*pA^4*pB^3-10.*DAB*$
$pA^3*pB^4+2.*DA_B^2*pA^3*pB^3+2.*DAABB*pA^3*pB^3+2.250000000*DAB^3*pA^2*$
$pB^2-5.750000000*DAB^2*pA^3*pB^2-5.750000000*DAB^2*pA^2*pB^3+2.500000000*$
$DAB*pA^4*pB^2+2.500000000*DAB*pA^2*pB^4+2.500000000*DAB^2*pA^4*pB^4-5.*$
$DAB*pA^5*pB^4-5.*DAB*pA^4*pB^5+.5*DA_B^2*pA^4*pB^4+.5*DAABB*pA^4*pB^4-$
$5.*DAB^2*pA^4*pB^3-5.*DAB^2*pA^3*pB^4+12.5*DAB*pA^4*pB^4+2.*DAB*pA^5*$
$pB^5-2.*DAB*pA^3*pB^2-.375*DAB^2*pA^2*pB+.5*DAB*pA^2*pB^2+8.*DAB*$
$pA^3*pB^3+3.25*DAB^2*pA^2*pB^2+.5*DA_B^2*pA^2*pB^2+.125*DA*DAB^2*pB^2+.5*$
$DAABB*pA^2*pB^2+.25*DAB^3*pA*pB+.125*DAB^2*DB*pA^2-2.*DAB*pA^2*$
$pB^3-.375*DAB^2*pA*pB^2-.5*pA^6*pB^3+1.5*pA^6*pB^4+.5*pA^6*pB^6-1.5*pA^6*$
$pB^5-1.5*pA^5*pB^6+4.5*pA^5*pB^5-4.5*pA^5*pB^4-4.5*pA^4*pB^5+1.500000000*pA^5*$
$pB^3+1.500000000*pA^3*pB^5+1.500000000*pA^4*pB^6-.5*pA^3*pB^6-1.5*pA^3*pB^4-$
$1.5*pA^4*pB^3+.5*pA^3*pB^3+4.5*pA^4*pB^4/((-1.+pA)^3*(-1.+pB)^3*pA^3*pB^3*n)$

varden := simplify(denom(var1));

$simplify(varden/((pA*(1-pA))^3*(pB*(1-pB))^3*n));$

1.

varnum:=numer(var1):

simplify(varnum):

$collectDAABB := collect(varnum, DAABB):$

$DAABBc1 := simplify(coeff(collectDAABB, DAABB, 1));$

$simplify(2*DAABBc1/((pA*(1-pA))^2*(pB*(1-pB))^2));$

1.

$DAABBc0 := simplify(coeff(collectDAABB, DAABB, 0));$

$collectDABfromcollectDAABB := collect(DAABBc0, DAB);$

$DABc3 := simplify(coeff(collectDABfromcollectDAABB, DAB, 3));$

$simplify(4*DABc3/(pA*(1-pA)*pB*(1-pB)*(1-2*pA)*(1-2*pB)));$

1.000000000

$DABc2 := simplify(coeff(collectDABfromcollectDAABB, DAB, 2));$

$DABc2DA_B := simplify(coeff(DABc2, DA_B, 1));$

$simplify(4*DABc2DA_B/(pA*(1-pA)*pB*(1-pB)*(1-2*pA)*(1-2*pB)));$

1.000000000

$DABc2DA := coeff(DABc2, DA, 1);$

$simplify(8*DABc2DA/(((1-2*pA)*(1-2*pA))*(pB*(1-pB))^2));$

1.000000000

$DABc2DB := coeff(DABc2, DB, 1);$

$simplify(8*DABc2DB/(((1-2*pB)*(1-2*pB))*(1-pA)^2*pA^2));$

1.

$DABc2DA_B := simplify(coeff(DABc2, DA_B, 1));$

$simplify(4*DABc2DA_B/(pA*(1-pA)*pB*(1-pB)*(1-2*pA)*(1-2*pB)));$

1.000000000

$DABc2OtherStuff :=$

$-DA*DABc2DA-DABc2DB*DB-DA_B*DABc2DA_B+DABc2;$

$simplify(8*DABc2OtherStuff/(pA*(1-pA)*pB*(1-pB)*(-3*pB*(1-pB)-pA*(1-pA)*(3-20*pB*(1-pB)))));$

1.000000000

$simplify(-DA*DABc2DA-DABc2DB*DB-DA_B*DABc2DA_B+DABc2-DABc2OtherStuff);$

0

$DABc1 := simplify(coeff(collectDABfromcollectDAABB, DAB, 1));$

$DABc1DAAB := simplify(coeff(DABc1, DAAB, 1));$

$simplify((-2*DABc1DAAB)*(1/((1-2*pA)*pA*(1-pA)*(pB*(1-pB))^2)));$

1.000000000

$DABc1DABB := simplify(coeff(DABc1, DABB, 1));$

$simplify((-2*DABc1DABB)*(1/((1-2*pB)*(pA*(1-pA))^2*pB*(1-pB))));$

1.000000000

$DABc1otherStuff := -DAAB*DABc1DAAB-DABB*DABc1DABB+DABc1;$

simplify(DABc1otherStuff);

$simplify(2*DABc1otherStuff/((pA*(1-pA)^2*pB*(1-pB)^2*(1-2*pA)*(1-2*pB)*pA)*pB));$

1.000000000

$collectDA_Bfromcollect DAABB := collect(DAABBc0, DA_B);$

$DA_Bc2 := simplify(coeff(collectDA_Bfromcollect DAABB, DA_B, 2));$

$simplify(2 * DA_Bc2/((pA * (1 - pA))^2 * (pB * (1 - pB))^2));$

1.000000000

$DA_Bc1DAB2 := simplify(coeff(collectDA_Bfromcollect DAABB, DA_B, 1));$

$simplify(4 * DA_Bc1DAB2/(pA * (1 - pA) * pB * (1 - pB) * (1 - 2 * pA) * (1 - 2 * pB) * DAB^2));$

1.000000000

$otherStuff := simplify(-DA * DAB^2 * DABc2DA - DAB^3 * DABc3 - DAB^2 * DABc2DB * DB - DAAB * DAB * DABc1DAAB - DAB^2 * DABc2OtherStuff - DAB * DABB * DABc1DABB - DA_B^2 * DA_Bc2 - DAB * DABc1otherStuff - DA_B * DA_Bc1DAB2 + DAABBc0);$

$DADBc1 :=$

$simplify(coeff(coeff(otherStuff, DA, 1), DB, 1)); simplify(DADBc1/(.5 * piA^2 * piB^2));$

1.000000000

$one :=$

$simplify(-DA * DADBc1 * DB + otherStuff); simplify(one/(.5 * piA^3 * piB^3));$

1.000000000

$check1 := simplify(-DA * DAB^2 * DABc2DA - DAB^3 * DABc3 - DAB^2 * DABc2DB * DB - DA * DADBc1 * DB - DAAB * DAB * DABc1DAAB - DAB^2 * DABc2OtherStuff - DAB * DABB * DABc1DABB - DA_B^2 * DA_Bc2 - DAB * DABc1otherStuff - DA_B * DA_Bc1DAB2 + DAABBc0 - one);$

0.

$check2 := simplify(-DA * DAB^2 * DABc2DA - DAB^3 * DABc3 - DAB^2 * DABc2DB * DB - DAB^2 * DA_B * DABc2DA_B - DA * DADBc1 * DB - DAAB * DAB * DABc1DAAB - DAB^2 * DABc2OtherStuff - DAB * DABB * DABc1DABB - DA_B^2 * DA_Bc2 - DAB * DABc1otherStuff + DAABBc0 - one);$

0.

$finalcheck :=$

$simplify(varnum - (1/2) * piA^2 * piB^2 * DAABB - DA_Bc2 * DA_B^2 - DABc2DA_B *$

$DA_B * DAB^2 - DABc1DABB * DABB * DAB - DABc1DAAB * DAAB * DAB - DABc1otherStuff * DAB - DABc2DA * DA * DAB^2 - DABc2DB * DB * DAB^2 - DABc2OtherStuff * DAB^2 - DABc3 * DAB^3 - DADBc1 * DA * DB - one)$;

0.

Genotypic Data; Composite Measure

$pAh := (x1 + x4 + x7 + .5 * (x2 + x5 + x8))/n :$

$pBh := (x1 + x2 + x3 + .5 * (x4 + x5 + x6))/n :$

$pah := 1 - pAh : pbh := 1 - pBh :$

$delh := (2 * x1 + x4 + x2 + (x5/2))/n - (2 * pAh * pBh) :$

$Paah := (x3 + x6 + x9)/n : PBBh := (x1 + x2 + x3)/n :$

$yh := (x1 + x2 + x3 + x7 + x8 + x9)/n - ((x7 + x8 + x9)/n - (x1 + x2 + x3)/n)^2$

$vxh := (x1 + x4 + x7 + x3 + x6 + x9)/n - ((x3 + x6 + x9)/n - (x1 + x4 + x7)/n)^2$

$den := sqrt(vxh * vyh)$

roh:= (2*delh)/den:

g11:=diff(roh,x1): sg11:= simplify(g11):

g12:=diff(roh,x2): sg12:=simplify(g12):

g13:=diff(roh,x3): sg13:=simplify(g13):

g14:=diff(roh,x4): sg14:=simplify(g14):

g15:=diff(roh,x5): sg15:=simplify(g15):

g16:=diff(roh,x6): sg16:=simplify(g16):

g17:=diff(roh,x7): sg17:=simplify(g17):

g18:=diff(roh,x8): sg18:=simplify(g18):

g19:=diff(roh,x9): sg19:=simplify(g19):

tauA:=1-2*pA: tauB:=1-2*pB:

piA := pA*(1-pA): piB:=pB*(1-pB):

pa:=1-pA: pb:=1-pB:

$x1 := n * (pA^2 * pB^2 + 2 * pA * DABB + pB * 2 * DAAB + pA * pB * 2 * del + del^2 + pA^2 * DB + pB^2 * DA + DA * DB + delAABB) :$

$x2 := n * (2 * piA * pB^2 + 2 * tauA * DABB - pB * 4 * DAAB + tauA * pB * 2 * del - 2 * del^2 + 2 * piA * DB - 2 * pB^2 * DA - 2 * DA * DB - 2 * delAABB) :$

$x3 := n * (pa^2 * pB^2 - 2 * pa * DABB + pB * 2 * DAAB - pa * pB * 2 * del + del^2 +$

$pa^2 * DB + pB^2 * DA + DA * DB + delAABB):$

$x4 := n * (2 * piB * pA^2 - 4 * pA * DABB + tauB * 2 * DAAB + pA * tauB * 2 *$

$del - 2 * del^2 - 2 * pA^2 * DB + 2 * piB * DA - 2 * DA * DB - 2 * delAABB):$

$x5 := n * (4 * piA * piB - 4 * tauA * DABB - tauB * 4 * DAAB + tauA * tauB * 2 *$

$del + 4 * del^2 - 4 * piA * DB - 4 * piB * DA + 4 * DA * DB + 4 * delAABB):$

$x6 := n * (2 * pa^2 * piB + 4 * pa * DABB + 2 * tauB * DAAB - pa * tauB * 2 * del -$

$2 * del^2 - 2 * pa^2 * DB + 2 * piB * DA - 2 * DA * DB - 2 * delAABB):$

$x7 := n * (pA^2 * pb^2 + 2 * pA * DABB - pb * 2 * DAAB - pA * pb * 2 * del + del^2 +$

$pA^2 * DB + pb^2 * DA + DA * DB + delAABB):$

$x8 := n * (2 * piA * pb^2 + 2 * tauA * DABB + 4 * pb * DAAB - pb * tauA * 2 * del -$

$2 * del^2 + 2 * piA * DB - 2 * pb^2 * DA - 2 * DA * DB - 2 * delAABB):$

$x9 := n * (pa^2 * pb^2 - 2 * pa * DABB - pb * 2 * DAAB + pa * pb * 2 * del + del^2 +$

$pa^2 * DB + pb^2 * DA + DA * DB + delAABB):$

simplify(x1+x2+x3+x4+x5+x6+x7+x8+x9):

sg11m:=simplify(sg11):

sg12m:=simplify(sg12):

sg13m:=simplify(sg13):

sg14m:=simplify(sg14):

sg15m:=simplify(sg15):

sg16m:=simplify(sg16):

sg17m:=simplify(sg17):

sg18m:=simplify(sg18):

sg19m:=simplify(sg19):

gvec:=¡sg11m,sg12m,sg13m,sg14m,sg15m,sg16m,sg17m,sg18m,sg19m¿:

ps:=¡p1,p2,p3,p4,p5,p6,p7,p8,p9¿;:

with(LinearAlgebra):

vmat:=n*(Matrix(1..9,1..9,ps,shape=diagonal)-

OuterProductMatrix(ps,ps,compact=false)):

p1:=x1/n: p2:=x2/n: p3:=x3/n:

p4:=x4/n: p5:=x5/n: p6:=x6/n:

p7:=x7/n: p8:=x8/n: p9:=x9/n:

$var1 := simplify(Transpose(gvec).vmat.gvec);$

$var1 := -4.*DAAB*pA^2*pB^3-9.*DB*pA^4*pB^2-3.*DB*pA^6*pB^2-2.*DAAB*$
$DB^2*pA^3+.5*del*pA^2*pB^2+8.*del*pA^3*pB^3+3.*DA*pA^4*pB^3-.375*DA^2*del^2*$
$pB-2.*del*pA^2*pB^3-3.*DB^2*pA^3*pB^2+.375*DA*del^2*pB^2+.375*DA^2*DB*$
$del^2-.375*del^2*pA^2*pB+2.25*del^2*pA^2*pB^2+3.*DA*pA^2*pB^3+3.*DB*pA^3*$
$pB^2+.25*del^3*pA*pB+4.*del*pA^5*pB^3-9.*DA*pA^4*pB^4-2.*del*pA^3*pB^2-3.*$
$DA^2*pA^2*pB^3-4.*DAAB*pA^4*pB^3-9.*DA*pA^2*pB^4-4.*DABB*pA^3*pB^2+$
$.3750*DB*del^2*pA^2-.375*del^2*pA*pB^2-3.*DA^3*DB^2*pB^2+9.*DA^2*pA^2*pB^4-$
$2.*DABB*pA^5*pB^2+1.875*del^2*pA^4*pB^2+.375*DB*del^2*pA^4+.5*del^3*pA^3*pB-$
$.375*del^2*pA^4*pB+3.*DA^3*DB*pB^2-4.*DA^2*DAAB*pB^3+1.875*DA^2*del^2*$
$pB^2-3.*DA^2*DB^3*pA^2-2.*DA^2*DABB*pB^3+18.*DA*pA^3*pB^4+9.*DB*pA^5*$
$pB^2-6.*DA*pA^3*pB^3-9.*DA^2*pA^2*pB^5+2.500000000*DB^2*del*pA^4-6.*DB*$
$pA^3*pB^3+.5*del^3*pA*pB^3-.375*del^2*pA*pB^4+3.*DA*DB^3*pA^2-.75*DA*del^3*$
$pB^2-4.*DABB*DB^2*pA^3+1.875*DB^2*del^2*pA^2-.75*DB*del^3*pA^2+.5*DA^2*$
$DB^2*del+.375*DA*DB^2*del^2+.25*DA*DB*del^3-.375*DB^2*del^2*pA-3.*DA*$
$pA^2*pB^6-9.*DB^2*pA^5*pB^2-9.*DB*pA^4*pB^4+4.*del*pA^3*pB^5+9.*DA*pA^2*$
$pB^5-2.*DAAB*pA^2*pB^5-4.*DABB*pA^3*pB^4+9.*DB^2*pA^4*pB^2+18.*DB*$
$pA^4*pB^3+3.*DB*pA^3*pB^4+1.875000000*del^2*pA^2*pB^4+2.500000000*DA^2*del*$
$pB^4+.375*DA*del^2*pB^4-3.*DA*pA^4*pB^6-3.*DB*pA^6*pB^4+2.*del*pA^5*pB^5+$
$3.*DA^2*pA^2*pB^6+9.*DA*pA^4*pB^5+6.*DA*pA^3*pB^6-2.*DAAB*pA^4*pB^5-$
$2.*DABB*pA^5*pB^4+3.*DB^2*pA^6*pB^2+6.*DB*pA^6*pB^3+9.*DB*pA^5*pB^4+$
$1.500000000*del^2*pA^4*pB^4-5.*del*pA^5*pB^4-5.*del*pA^4*pB^5-3.*DA^2*pA*pB^6-$
$18.*DA*pA^3*pB^5+5.*DAAB*pA^4*pB^4+4.*DAAB*pA^3*pB^5+4.*DABB*pA^5*$
$pB^3+5.*DABB*pA^4*pB^4-3.*DB^2*pA^6*pB-18.*DB*pA^5*pB^3-3.*del^2*pA^4*$
$pB^3-3.*del^2*pA^3*pB^4+12.50000000*del*pA^4*pB^4+3.*DA^3*DB*pB^4-2.*DA^2*$
$DAAB*pB^5+1.500000000*DA^2*del^2*pB^4+9.*DA^2*pA*pB^5+3.*DA*DB^3*pA^4-$
$10.*DAAB*pA^3*pB^4-2.*DABB*DB^2*pA^5-10.*DABB*pA^4*pB^3+1.500000000*$
$DB^2*del^2*pA^4+9.*DB^2*pA^5*pB+.5*del^4*pA^2*pB^2-1.5*del^3*pA^3*pB^2-1.5*del^3*$
$pA^2*pB^3+6.*del^2*pA^3*pB^3-10.*del*pA^4*pB^3-10.*del*pA^3*pB^4-2.*delAABB*$
$pA^4*pB^3+5.*DABB*DB^2*pA^4-2.*delAABB*pA^3*pB^4-2.*DABB*pA^2*pB^3-$
$6.*DA^3*DB*pB^3+.5*DA*DB*del^4+5.*DABB*pA^4*pB^2+8.*DABB*pA^3*pB^3+$
$.5*DB*del^3*pA^3-3.*DB^2*del^2*pA^3+3.*DA^2*DB^3*pA-.5*del^4*pA^2*pB+.5*DA*$
$del^4*pB-3.75*del^2*pA^3*pB^2+5.*DA^2*DAAB*pB^4-.5*DB*del^4*pA^2-2.*DB^2*$

$delAABB*pA^3-2.*DA^2*del*pB^3-3.*DA^2*del^2*pB^3-.5*del^4*pA*pB^2-9.*DA^2*$
$pA*pB^4+2.5*del*pA^4*pB^2-6.*DA*DB^3*pA^3+2.25*del^3*pA^2*pB^2+1.500000000*$
$DA^2*DB^2*del^2-2.*DAAB*pA^3*pB^2-.5*DA*del^4*pB^2+8.*DAAB*pA^3*pB^3-$
$.75*DA*del^2*pB^3+4.*delAABB*pA^3*pB^3-9.*DB^2*pA^4*pB+3.*DB^2*pA^3*pB+$
$3.*DA^3*DB^2*pB-2.*DB^2*del*pA^3-3.75*del^2*pA^2*pB^3+.5*DA*del^3*pB^3+5.*$
$DAAB*pA^2*pB^4+2.5*del*pA^2*pB^4-2.*DA^2*delAABB*pB^3+3.*DA^2*pA*pB^3+$
$.5*DB*del^4*pA-.75*DB*del^2*pA^3+.5*del^4*pA*pB-.75*del^3*pA^2*pB-.75*$
$del^3*pA*pB^2+.75*del^2*pA^3*pB+.75*del^2*pA*pB^3-2.*delAABB*pA^3*pB^2-2.*$
$delAABB*pA^2*pB^3+.5*DA^2*del*pB^2+.25*DA*del^3*pB+.5*DB^2*del*pA^2+.25*$
$DB*del^3*pA+pA^3*pB^3+pA^6*pB^6-DA^3*pB^6-DB^3*pA^6+DA^3*DB^3-pA^3*pB^6+$
$DB^3*pA^3-pA^6*pB^3+DA^3*pB^3+9.*pA^4*pB^4-3.*pA^3*pB^4-3.*pA^4*pB^3-3.*pA^6*$
$pB^5-3.*pA^5*pB^6+9.*pA^5*pB^5-9.*pA^5*pB^4-9.*pA^4*pB^5+3.*DA^3*pB^5+3.*$
$DB^3*pA^5+3.*pA^5*pB^3+3.*pA^3*pB^5+3.*pA^4*pB^6-3.*DB^3*pA^4+3.*pA^6*pB^4-$
$3.*DA^3*pB^4+6.*DA*DAAB*DB^2*del*pA+2.*DA*DB*del*pA*pB+4.*DA*$
$DB*delAABB*pA^2*pB^2-8.*DA*DAAB*DB*pA^2*pB^3-8.*DA*DABB*DB*$
$pA^3*pB^2+8.*DA*DB*del*pA^3*pB^3-4.*DA^2*DB*del*pA*pB^3+6.*DA*DAAB*$
$del*pA*pB^4+12.*DA*DABB*del*pA^2*pB^3-4.*DA*DB^2*del*pA^3*pB-6.*DA*$
$DB*del^2*pA^2*pB^2-12.*DA*DB*del*pA^3*pB^2-12.*DA*DB*del*pA^2*pB^3+12.*$
$DAAB*DB*del*pA^3*pB^2+6.*DABB*DB*del*pA^4*pB+4.*DA^2*DABB*DB*$
$pA*pB^2+2.*DA^2*DB^2*del*pA*pB+6.*DA^2*DB*del*pA*pB^2+4.*DA*DAAB*$
$DB^2*pA^2*pB+12.*DA*DAAB*DB*pA^2*pB^2+8.*DA*DAAB*DB*pA*pB^3-$
$12.*DA*DAAB*del*pA*pB^3+8.*DA*DABB*DB*pA^3*pB+12.*DA*DABB*$
$DB*pA^2*pB^2-18.*DA*DABB*del*pA^2*pB^2-12.*DA*DABB*del*pA*pB^3+6.*$
$DA*DB^2*del*pA^2*pB+6.*DA*DB*del^2*pA^2*pB+6.*DA*DB*del^2*pA*pB^2+$
$18.*DA*DB*del*pA^2*pB^2-12.*DAAB*DB*del*pA^3*pB-18.*DAAB*DB*$
$del*pA^2*pB^2-12.*DABB*DB*del*pA^3*pB-4.*DA^2*DABB*DB*pA*pB-4.*$
$DA*DAAB*DB^2*pA*pB+6.*DA*DAAB*DB*del*pB^2-12.*DA*DAAB*DB*$
$pA*pB^2+6.*DA*DABB*DB*del*pA^2-12.*DA*DABB*DB*pA^2*pB-6.*DA*$
$DB*del^2*pA*pB-6.*DA*DB*del*pA^2*pB-6.*DA*DB*del*pA*pB^2-4.*DA*$
$DB*delAABB*pA^2*pB-4.*DA*DB*delAABB*pA*pB^2+4.*DA*DAAB*DB*$
$pA*pB+4.*DA*DABB*DB*pA*pB-6.*DA*DABB*del*pA*pB+4.*DA*DB*$
$delAABB*pA*pB-6.*DAAB*DB*del*pA*pB+4.*DA*DB*del*pA*pB^3-4.*$

$DA*DABB*DB*pA*pB^2+18.*DA*DABB*del*pA*pB^2-2.*DA*DB^2*del*pA*$
$pB+18.*DAAB*DB*del*pA^2*pB+6.*DAAB*DB*del*pA*pB^2+6.*DABB*$
$DB*del*pA^2*pB-6.*DA*DAAB*DB*del*pB-6.*DA*DABB*DB*del*pA+6.*$
$DA^2*DABB*DB*del*pB+4.*DA*DB*del*pA^3*pB-2.*DA^2*DB*del*pA*pB-$
$4.*DA*DAAB*DB*pA^2*pB+6.*DA*DAAB*del*pA*pB^2+6.*DA*DABB*del*$
$pA^2*pB-DB^2*del*pA^2*pB+DA*DB^2*del*pA+DA^2*DABB*pB^2-DAAB*del^2*$
$pA^2*pB^3-DABB*del^2*pA^3*pB^2-DB*del^3*pA^3*pB+DA^2*DB*del*pB-DA*$
$del^3*pA*pB^3+DABB*DB*del^2*pA^3+DABB*del^2*pA^3*pB-DA^2*DB^2*del*pA-$
$DA^2*DB^2*del*pB-DA^2*del*pA*pB^2+DA*DAAB*del^2*pB^3+DAAB*del^2*pA*$
$pB^3+DAAB*DB^2*pA^2+delAABB*pA^2*pB^2+DABB*pA^2*pB^2+DAAB*pA^2*$
$pB^2+DAAB*DB^2*pA^4+DA^2*DAAB*DB^2+DA^2*DAAB*pB^2+DA^2*delAABB*$
$pB^2-del*pA^5*pB^2+DAAB*pA^4*pB^2+delAABB*pA^4*pB^2+DA^2*DABB*DB^2+$
$DA^2*DABB*pB^4+DABB*pA^2*pB^4+delAABB*pA^2*pB^4+DABB*DB^2*pA^2+$
$DB^2*delAABB*pA^2-del*pA^2*pB^5+DB*del*pA^2*pB+DA*del*pA*pB^2+del^3*$
$pA^3*pB^3+delAABB*pA^4*pB^4-DA^2*del*pB^5-DB^2*del*pA^5+DA^2*delAABB*$
$pB^4+DB^2*delAABB*pA^4+DA^2*DB^2*delAABB-2.*DA*DAAB*pA^2*pB^2-$
$3.*DA*DAAB*del*pB^2+2.*DA*DAAB*pA*pB^2+2.*DA*DB^2*delAABB*pA+$
$.5*DA*DB*del^2*delAABB+.75*DA*DB*del^2*pA+.75*DA*DB*del^2*pB+.5*$
$DA*del^2*delAABB*pB-.75*DA*del^2*pA*pB+2.*DA*delAABB*pA*pB^2-3.*$
$DABB*DB*del*pA^2+2.*DABB*DB*pA^2*pB+.5*DB*del^2*delAABB*pA-.75*$
$DB*del^2*pA*pB+2.*DB*delAABB*pA^2*pB+.5*del^2*delAABB*pA*pB-.5*$
$DB*del^3*pA*pB+.75*DB*del^2*pA*pB^2-2.*DB*delAABB*pA^2*pB^2+2.*DA*$
$DABB*DB^2*pA+.5*DABB*DB*del^2*pA+.5*DABB*del^2*pA*pB-2.*DA*$
$del*pA*pB^5-8.*DB*del*pA^3*pB^3-9.*DA^2*DB^2*pA*pB^2+2.*DA*DABB*pA*$
$pB^4-9.*DA*DB^2*pA^2*pB^2-18.*DA*DB*pA^2*pB^3+4.*DAAB*DB*pA^2*pB^3-$
$3.*DAAB*del*pA*pB^4+8.*DABB*DB*pA^3*pB^2-6.*DABB*del*pA^2*pB^3+4.*$
$DB^2*del*pA^3*pB-.75*DB*del^2*pA^2*pB^2+2.*DB*del*pA^2*pB^3-3.*DA^2*DB*$
$del*pB^2+6.*DA*DAAB*del*pB^3-3.*DA*DB^2*del*pA^2+9.*DA*DB^2*pA^2*pB-$
$.75*DA*DB*del^2*pB^2-2.*DAAB*DB^2*pA^2*pB+6.*DABB*DB*del*pA^3-2.*$
$DABB*DB*pA^2*pB^2-.5*DABB*del^2*pA*pB^2-4.*DB*del*pA^5*pB^3-9.*DA^2*$
$DB*pA^2*pB^4+2.*DA^2*del*pA*pB^5+4.*DA*DAAB*pA^2*pB^5+4.*DA*DABB*$
$pA^3*pB^4-9.*DA*DB^2*pA^4*pB^2-18.*DA*DB*pA^4*pB^3-18.*DA*DB*pA^3*$

$pB^4+10.*DA*del*pA^3*pB^4+6.*DA*del*pA^2*pB^5+4.*DAAB*DB*pA^4*pB^3-$
$6.*DAAB*del*pA^3*pB^4+4.*DABB*DB*pA^5*pB^2-6.*DABB*del*pA^4*pB^3+$
$2.*DB^2*del*pA^5*pB-8.*DAAB*DB*pA^3*pB^3+12.*DAAB*del*pA^3*pB^3+9.*$
$DAAB*del*pA^2*pB^4-4.*DABB*DB*pA^5*pB-10.*DABB*DB*pA^4*pB^2+9.*$
$DABB*del*pA^4*pB^2+12.*DABB*del*pA^3*pB^3-5.*DB^2*del*pA^4*pB-15.*DB*$
$del*pA^4*pB^2-2.*DB*delAABB*pA^4*pB^2+4.*DA^2*DAAB*DB*pB^3+4.*DA^2*$
$DABB*pA*pB^3-4.*DA*DABB*pA*pB^3-2.*DA*DB^2*delAABB*pA^2-.5*DA*$
$DB*del^3*pA-.5*DA*DB*del^3*pB-.5*DA*del^2*delAABB*pB^2+.75*DA*del^2*$
$pA*pB^2-3.*DA*del*pA^2*pB^2-4.*DA*delAABB*pA*pB^3-4.*DAAB*DB*pA^3*$
$pB-6.*DAAB*DB*pA^2*pB^2-1.500000000*DAAB*del^2*pA*pB^2+9.*DAAB*$
$del*pA^2*pB^2+6.*DAAB*del*pA*pB^3-1.500000000*DABB*DB*del^2*pA^2-8.*$
$DABB*DB*pA^3*pB-1.500000000*DABB*del^2*pA^2*pB+6.*DABB*del*pA^3*$
$pB+9.*DABB*del*pA^2*pB^2-.5*DB*del^2*delAABB*pA^2+.75*DB*del^2*pA^2*$
$pB-3.*DB*del*pA^2*pB^2-4.*DB*delAABB*pA^3*pB-.5*del^2*delAABB*pA^2*$
$pB-.5*del^2*delAABB*pA*pB^2+2.*DA^2*DB*delAABB*pB+6.*DB*del*pA^5*$
$pB^2+10.*DB*del*pA^4*pB^3-2.*DA^2*DABB*pA*pB^4+9.*DA^2*DB^2*pA^2*pB^2+$
$9.*DA^2*DB*pA*pB^4-5.*DA^2*del*pA*pB^4-10.*DA*DAAB*pA^2*pB^4-4.*DA*$
$DAAB*pA*pB^5-8.*DA*DABB*pA^3*pB^3-6.*DA*DABB*pA^2*pB^4+9.*DA*$
$DB^2*pA^4*pB+36.*DA*DB*pA^3*pB^3-15.*DA*del*pA^2*pB^4-2.*DA*delAABB*$
$pA^2*pB^4-2.*DAAB*DB^2*pA^4*pB-6.*DAAB*DB*pA^4*pB^2+.5*DA*DABB*$
$del^2*pB-18.*DA*DB*pA^3*pB^2+9.*DA^2*DB*pA*pB^2+18.*DA*DB^2*pA^3*pB^2+$
$9.*DA*DB*pA^2*pB^4-3.*DABB*DB*del*pA^4+10.*DABB*DB*pA^4*pB+1.5*$
$DABB*del^2*pA^2*pB^2-18.*DABB*del*pA^3*pB^2+1.5*DB*del^3*pA^2*pB+5.*$
$DB*del*pA^4*pB+12.*DB*del*pA^3*pB^2+2.*DB*delAABB*pA^4*pB+4.*DB*$
$delAABB*pA^3*pB^2+.5*del^2*delAABB*pA^2*pB^2-2.*DA^2*DAAB*DB^2*pB-$
$6.*DA^2*DAAB*DB*pB^2-2.*DA^2*DABB*DB^2*pA+9.*DA^2*DB^2*pA*pB-2.*$
$DA^2*DB*delAABB*pB^2-1.5*DA*DAAB*del^2*pB^2-8.*DA*DAAB*pA*pB^3-$
$6.*DA*DABB*DB^2*pA^2-6.*DA*DABB*pA^2*pB^2+18.*DA^2*DB*pA^2*pB^3-$
$6.*DA^2*DABB*del*pB^3-9.*DA^2*DB^2*pA^2*pB-2.*DA^2*DABB*DB*pB^2+$
$9.*DA^2*DABB*del*pB^2-.5*DA*DABB*del^2*pB^2-3.*DA^2*DABB*DB*del+$
$2.*DA^2*DABB*DB*pB+.5*DA*DABB*DB*del^2-2.*DA^2*DABB*pA*pB^2-$
$.75*DA*DB*del^2*pA^2-.5*DA*del^3*pA*pB+.75*DA*del^2*pA^2*pB-2.*DA*$

$delAABB*pA^2*pB^2-.5*DAAB*del^2*pA^2*pB+2.*DA^2*DAAB*DB*pB+.5*DA*$
$DAAB*del^2*pB+.5*DAAB*del^2*pA*pB+2.*DA^2*DB*del*pB^3-18.*DA^2*DB*$
$pA*pB^3-3.*DA*DAAB*del*pB^4+10.*DA*DAAB*pA*pB^4+4.*DA*DABB*$
$DB^2*pA^3+12.*DA*DABB*pA^2*pB^3+2.*DA*DB^2*del*pA^3-18.*DA*DB^2*pA^3*$
$pB+1.5*DA*del^3*pA*pB^2+12.*DA*del*pA^2*pB^3+5.*DA*del*pA*pB^4+4.*DA*$
$delAABB*pA^2*pB^3+2.*DA*delAABB*pA*pB^4+4.*DAAB*DB^2*pA^3*pB+12.*$
$DAAB*DB*pA^3*pB^2+1.5*DAAB*del^2*pA^2*pB^2-18.*DAAB*del*pA^2*pB^3+9.*$
$DA*DB*pA^4*pB^4-4.*DA*del*pA^3*pB^5+9.*DA*DB*pA^4*pB^2-8.*DA*del*pA^3*$
$pB^3-2.*DB*del*pA^5*pB-9.*DA^2*DB*pA^2*pB^2+4.*DA^2*del*pA*pB^3+8.*DA*$
$DAAB*pA^2*pB^3+4.*DA*DABB*pA^3*pB^2-.75*DA*del^2*pA^2*pB^2+2.*DA*del*$
$pA^3*pB^2+2.*DAAB*DB*pA^4*pB-6.*DAAB*del*pA^3*pB^2-3.*DABB*del*pA^4*$
$pB-3.*DA^2*DABB*del*pB-3.*DAAB*DB^2*del*pA+9.*DAAB*DB^2*del*pA^2-$
$.5*DAAB*DB*del^2*pA^2-3.*DA*DAAB*DB^2*del+2.*DA*DAAB*DB^2*pA+$
$.50*DA*DAAB*DB*del^2+.5*DAAB*DB*del^2*pA-6.*DAAB*DB^2*del*pA^3-2.*$
$DA*DAAB*DB^2*pA^2-4.*DA*del*pA*pB^3+2.*DA*DABB*pA*pB^2+2.*DAAB*$
$DB*pA^2*pB-3.*DAAB*del*pA*pB^2-3.*DABB*del*pA^2*pB-4.*DB*del*pA^3*$
$pB+9.*DA*DB*pA^2*pB^2+DA*DABB*del^2*pA*pB^2+DA*DB*del^3*pA*pB+$
$DAAB*DB*del^2*pA^2*pB-DA*DAAB*DB*del^2*pB-DA*DABB*DB*del^2*$
$pA-DA*DABB*del^2*pA*pB-DAAB*DB*del^2*pA*pB-12.*DA*DAAB*DB*$
$del*pA*pB^2-12.*DA*DABB*DB*del*pA^2*pB+12.*DA*DAAB*DB*del*pA*$
$pB+12.*DA*DABB*DB*del*pA*pB/[(-pA^2+DA+pA)^3*(-pB^2+DB+pB)^3*n]$
$varden := simplify(denom(var1));$
$simplify(varden/((piA+DA)^3*(piB+DB)^3*n));$ 1 $varnum := numer(var1):$
$simplify(varnum): collectDel := collect(varnum, del):$
$del4 := simplify(coeff(collectDel, del, 4));$
$simplify(del4/((1/2)*(piA+DA)*(piB+DB)));$ 1.
$del3 := simplify(coeff(collectDel, del, 3));$
$simplify(del3/((1/4)*(piA+DA)*(piB+DB)*tauA*tauB));$ 1.

$del2 := simplify(coeff(collectDel, del, 2));$

$del2delAABB := coeff(del2, delAABB);$
$simplify(del2delAABB/((1/2)*(piA+DA)*(piB+DB)));$ 1.
$del2DAAB := simplify(coeff(del2, DAAB));$

$simplify(del2DAAB/((1/2)*(piA+DA)*(piB+DB)*tauB));$ 1.

$del2DABB := simplify(coeff(del2, DABB));$

$simplify(del2DABB/((1/2)*(piA+DA)*(piB+DB)*tauA));$ 1.

$del20 := simplify(eval(del2, [DABB=0, DAAB=0, delAABB=0]));$

$simplify(del20/(3/8*(-piB*tauB^2*DA^2-piA*tauA^2*DB^2+(2*(-8*piA*$
$piB+piA+piB))*DA*DB+piB*(piB-2*piA)*DA+piA*(piA-2*piB)*$
$DB-piA*piB*(piA*tauB^2+piB)+DA*DB*(4*DA*DB+DA+DB))));$
1. $del1 := simplify(coeff(collectDel, del, 1));$

$del1DABB := simplify(coeff(del1, DABB));$

$simplify(del1DABB/(-3*(piA+DA)^2*(piB+DB)*tauB));$ 1.

$del1DAAB := coeff(del1, DAAB);$

$simplify(del1DAAB/(-3*(piA+DA)*(piB+DB)^2*tauA));$ 1.

$del10 := simplify(eval(del1, [DABB=0, DAAB=0, delAABB=0]));$

$simplify(del10/(.5*(piA+DA)^2*(piB+DB)^2*tauA*tauB));$ 1.

$del0 := simplify(coeff(collectDel, del, 0));$

$del0DABB := coeff(del0, DABB);$

$simplify(del0DABB/((piA+DA)^2*(piB+DB)^2*tauA));$ 1.

$del0DAAB := coeff(del0, DAAB);$

$simplify(del0DAAB/((piA+DA)^2*(piB+DB)^2*tauB));$ 1.

$del0delAABB := coeff(del0, delAABB);$

$simplify(del0delAABB/((piA+DA)^2*(piB+DB)^2));$ 1.

$del00 := simplify(eval(del0, [DABB=0, DAAB=0, delAABB=0]));$

$simplify(del00/((piA+DA)^3*(piB+DB)^3));$ 1.

$simplify(var1-(del^4*del4+DAAB*del^2*del2DAAB+DABB*del^2*$
$del2DABB+del^3*del3+del^2*delAABB*del2delAABB+DAAB*del*$
$del1DAAB+DABB*del*del1DABB+del^2*del20+DAAB*del0DAAB+$
$DABB*del0DABB+del*del10+delAABB*del0delAABB+del00)/varden);$ 0.

Appendix B

R Functions for Approximate Variance

Gametic Case

$varR = function(PA = A, PB = B, D = 0) \;\; piA = PA * (1 - PA)$

$piB = PB * (1 - PB)$

$tauA = 1 - 2 * PA$

$tauB = 1 - 2 * PB$

$den = 4 * piA^2 * piB^2$

$a = 2 * tauA * tauB$

$c = 4 * piA * piB * tauA * tauB$

$d = 4 * piA^2 * piB^2$

$b = -3 * (piA + piB) + 20 * piA * piB$

$var = (a * D^3 + b * D^2 + c * D + d) * (1/den)$

$return(var)$

Genotypic Data, MLE

$mleVarR < -function(pA = .1, pB = .2, D = .02)$ # Define tau and pi

$tauA = 1 - 2 * pA$

$tauB = 1 - 2 * pB$

$piA = pA * (1 - pA)$

$piB = pB * (1 - pB)$

# Define the numerator

$num.D5 = 6 * tauA * tauB * D^5$

$num.D4 = (96 * piA * piB - 17 * (piA + piB) + 2) * D^4$

$num.D3 = (tauA * tauB * (38 * piA * piB - 3 * (piA + piB))) * D^3$

$num.D2 = (piA * piB * (120 * piA * piB - 23 * (piA + piB) + 4)) * D^2$

$num.D1 = (12 * piA^2 * piB^2 * tauB * tauA) * D$

$num.D0 = 8 * piA^3 * piB^3$

$\#$ Define the denominator

$den.D0 = 8 * piA^3 * piB^3$

$den.D1 = 8 * piA^2 * piB^2 * tauA * tauB * D$

$den.D2 = (24 * piA^2 * piB^2) * D^2$

$\#$ Total variance

$num = num.D0 + num.D1 + num.D2 + num.D3 + num.D4 + num.D5$

$den = den.D0 + den.D1 + den.D2$

$var = num/den$

$return(var)$

Phase Informed Genotypic Data

$phaseKnownVarR < -function(pA = .2, pB = .3, DA = pA/2, DB = pB/2,$

$DAB = 1, DA_B = 0, DABB = 0, DAAB = 0, DAABB = 0)$

$piA = pA * (1 - pA)$

$piB = pB * (1 - pB)$

$tauA = 1 - 2 * pA$

$tauB = 1 - 2 * pB$

$den = piA^3 * piB^3$

$one < -.5 * piA^3 * piB^3$

$num.DADB = DA * DB * .5 * piA^2 * piB^2$

$num.DAB3 = DAB^3 * .25 * piA * piB * tauA * tauB$

$num.DAB2 = DAB^2 * (-3 * (piA + piB) + 20 * piA * piB) * piA * piB * (1/8)$

$num.DAB = DAB * .5 * piA^2 * piB^2 * tauA * tauB * pA * pB$

$num.DAB2DA = DAB^2 * DA * tauA^2 * piB^2 * (1/8)$

$num.DAB2DB = DAB^2 * DB * tauB^2 * piA^2 * (1/8)$

$num.DABDAAB = DAB * DAAB * tauA * piA * piB^2 * (-.5)$

$num.DABDABB = DAB * DABB * tauB * piB * piA^2 * (-.5)$

$num.DAABB = DAABB * (1/2) * piA^2 * piB^2$

$num.DA/B2 = DA_B^2 * .5 * piA^2 * piB^2$

$num.DA/B.DAB2 = DA_B * DAB^2 * .25 * tauA * tauB * piA * piB$

$numlong = one + num.DADB + num.DAB3 + num.DAB2 + num.DAB$

$+num.DAB2DA + num.DAB2DB + num.DABDAAB + num.DABDABB$

$+num.DABDABB + num.DAABB + num.DA_B2 + num.DA_B$

# variance

$vr = numlong/den$

$return(vr)$

Composite Measure for Phase Unknown Genotypic Data

```
comVarR<-function(pA=.2,pB=.3,DA=pA/2,DB=pB/2,del=1,
DAAB=0,DABB=0,delAABB=0) {
piA=pA*(1-pA)
piB=pB*(1-pB)
tauA=1-2*pA
tauB=1-2*pB
```

$den = ((piA + DA)^3) * ((piB + DB)^3)$

$one.1 < -((piA + DA)^3 * (piB + DB)^3)$

$num.del4.1 = del^4 * (1/2) * ((piA + DA) * (piB + DB))$

$num.del3.1 = del^3 * (1/4) * ((piA + DA) * (piB + DB) * tauA * tauB)$

$num.del2.1 = del^2 * (3/8) * ((-1) * piB * tauB^2 * DA^2 + (-1) * piA * tauA^2 * DB^2$
$+2 * (piA + piB - 8 * piA * piB) * DA * DB + piB * (piB - 2 * piA) * DA$
$+piA * (piA - 2 * piB) * DB + (-1) * piA * piB * ((piA + piB) - 4 * piA * piB)$
$+DA * DB * (4 * DA * DB + DA + DB))$

$num.del.1 = del * ((piA + DA)^2 * (piB + DB)^2 * .5 * tauA * tauB)$

$num.DAAB.1 = DAAB * ((piA + DA)^2 * (piB + DB)^2 * tauB)$

$num.DABB.1 = DABB * ((piA + DA)^2 * (piB + DB)^2 * tauA)$

$num.delAABB.1 = delAABB * ((piA + DA)^2 * (piB + DB)^2)$

$num.delDAAB.1 = del * DAAB * (-3) * ((piA + DA) * (piB + DB)^2 * tauA)$

$num.delDABB.1 = del * DABB * (-3) * ((piA + DA)^2 * (piB + DB) * tauB)$

$num.del2DAAB.1 = del^2 * DAAB * (1/2) * ((piA + DA) * (piB + DB) * tauB)$

$num.del2DABB.1 = del^2 * DABB * (1/2) * ((piA + DA) * (piB + DB) * tauA)$

$num.del2delAABB.1 = del^2 * delAABB * (1/2) * ((piA + DA) * (piB + DB))$

$numlong.1 = one.1 + num.del.1 + num.del2.1 + num.del3.1 + num.del4.1$
$+num.DAAB.1 + num.DABB.1 + num.delAABB.1 + num.del2DAAB.1$
$+num.del2DABB.1 + num.del2delAABB.1 + num.delDAAB.1 + num.delDABB.1$

```
# variance
vr = numlong.1/den
return(vr)
}
```

Bibliography

K. G. Ardlie, L. Kruglyak, and M. Seielstad. Patterns of linkage disequilibirium in the human genome. *Genetics*, 3(299), 2002.

A. H. D. Brown. Sample sizes required to detect linkage disequilibrium between two of three loci. *Theoretical Population Biology*, 8(184), 1975.

D. C. Hamilton and D. E. C. Cole. Standardizing a composite measure of linkage disequilibrium. *Annals of Human Genetics*, 68(234):234–239, 2004.

D. C. Hamilton and D. E. C. Cole. Testing for equality of standardized composite measures of linkage disequilibrium. *Annals of Human Genetics*, 72(292):292–296, 2008.

D. C. Hamilton, Q. Liu, and D. E. C. Cole. Approximate variance of the standardized composite measure of linkage disequilibrium. *Annals of Human Genetics*, 70(535):535–540, 2006.

R. C. Lewontin. The interaction of selection and linkage I. general considerations; heterotic models. *Genetics*, 49(49):49–67, 1964.

J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics*, 69(1), 2001.

N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(1516), 1996.

M. D. Teare, A. M. Dunning, F. Durocher, G. Rennart, and D. F. Easton. Sampling distribution of summary linkage disequilibrium measures. *Annual Human Genetics*, 66(223), 2002.

B. S. Weir. *Genetic Data Analysis II.* Sinauer, 1996.

B. S. Weir and C. C. Cockerham. *Mathematical Evolution Theory*, chapter Complete Characterization of Disequilibrium at Two Loci. Princeton University Press, New Jersey, USA, 1989.

C. Zapata. Approximate variance of the standardized measure of gametic disequilibrium D'. *American Journal of Human Genetics*, 61(771), 1997.

C. Zapata. On the uses and applications of the most commonly used measures of linkage disequilibrium from the comparative analysis of their statistical properties. *Human Heredity*, 71(186), 2011.

D. V. Zaykin. Bounds and normalization of the composite linkage disequilibrium coefficient. *Genetic Epidemiology*, 27(3):253–257, 2004.