Robust statistical tests of genetic association for the case-control study design

Running title: Robust tests of genetic association

Kai Wang

Department of Biostatistics

College of Public Health

The University of Iowa

Iowa City, Iowa 52242, USA

Corresponding author:

Kai Wang, PhD

Department of Biostatistics, C227 GH

College of Public Health

University of Iowa

Iowa City, IA 52242

e-mail: kai-wang@uiowa.edu

phone: (319) 384-5175

fax: (319) 384-5018

The central theme in case-control genetic association studies is to efficiently identify genetic markers associated with case-control status. Powerful statistical methods are critical to accomplishing this goal. A popular statistical method is the model-free Pearson's chi-square test. To achieve increased power, model-based tests have been widely used despite their lack of robustness to model-misspecification. Much research has been carried out on increasing the robustness of model-based tests. A model-free analysis framework is proposed. It involves less degree of freedom than the Pearson's chi-square test. The likelihood ratio statistic, the score statistic, and the Wald statistic are introduced. In addition, these statistics are less affected by the confounding effect of population stratification on genetic association. The performance of these statistics are evaluated by computer simulation. Also introduced is a test for the existence of population stratification. This statistic is asymptotically uncorrelated with the proposed lieklkihood ratio statistic, the score statistic, and the Wald statistic. All these statistics are applied to a study of height in a European American population.

Keywords: Case-control design, genetic association, genetic model, population stratification, likelihood ratio statistic, score statistic, Wald statistic

INTRODUCTION

Case-control study design is very popular for detecting genetic factors associated with dichotomous traits. It has been widely used in genome-wide association studies (GWASs) and will continue to be so. In the past three years, there have been more than 300 GWASs conducted on many complex human disorders such as glaucoma and age-related macular degeneration (Hindorff et al. 2009). A fundamental issue in a genetic association study is to efficiently identify associated genetic markers (typically single neucleotide polymorphisms, or SNPs).

The efficiency of a genetic association study critically hinges on the statistical methods adopted. A popular method for case-control design is Pearson's chi-square test. This is a model-free method. It has the virtue of having adequate power for traits having a range of gene-trait relationship. However, it can be less powerful than methods derived from genetic models (e.g., dominance models, recessive models, or additive models). It is also sensitive to the effect of population stratification and results in excessive false positive rates.

Disadvantages of the Pearson's chi-square test have led to the popularity of model-based methods. The Cochran-Armitage test of trend, which assumes the allele effect on phenotype is additive, is a good example. The major disadvantage of model-based methods is that they are sensitive to model mis-specification. When the assumed trait model is different from the truth, which is almost always the case, their power can be low (Slager and Schaid 2001; Freidlin et al. 2002; Schaid et al. 2005).

Much research has been devoted to designing testing procedures that are more robust to model mis-specification than model-based methods. Freidlin et al. (2002) proposed a maximin efficiency robust test and a test (named MAX) based on the maximum of test statistics under several analysis models. They found that the MAX test is generally more powerful than the other one. In another study, Freidlin et al. (2002) assumed an *a priori* ordering of the mean genetic effects for the three genotypes that are induced from the allele to be tested for by assuming the marker allele associated with the disease allele is known.

Such an ordering may be difficult to make in reality. To remove this restriction, Zheng (2003) proposed a "max and min scores" approach. Wang and Sheffield (2005) proposed a method that allows the trait model to be in a certain model space instead of being fixed. Although more robust to model mis-specification, these methods typically have complicated null distributions, making it hard to evaluate the significance of test statistics.

We introduce a model-free approach for genetic association mapping using SNPs in case-conrol study design. This approach simply compares the frequencies of a putative allele in cases and in controls to see whether there exists a significant difference. This idea has been used in the allelic test. The Cochran-Armitage test for trend (Sasieni 1997) also turns out to be a test on this difference. However, they use additional assumptions. The allelic test assumes that the marker genotype are under Hardy-Weinberg equilibrium. The Cochran-Armitage test for trend assumes that the effect of the disease variant obeys an additive model. The use of this idea here is different as no additional assumptions are being made. The resulting testing procedure involves only one degree of freedom, less than that of the Pearson's chi-square test.

Our approach is introduced in a standard likelihood analysis framework. Starting from the likelihood function of the observed data, the likelihood ratio test, the Wald test, and the score test are introduced. We will discuss their connection to existing methods. Performance of these methods will be evaluated in comparison with existing methods using simulation studies. This new approach is applied to a study of height in a European American population.

THE MODEL

Consider a biallelic marker such as a single nucleotide polymorphism (SNP). Denote the two alleles by A and a, respectively. The frequencies of genotypes aa, aA, and AA are denoted by P_{10}, P_{11}, and P_{12}, respectively, in cases and by P_{20}, P_{21}, and P_{22}, respectively, in controls. Suppose that there are n_{10}, n_{11}, and n_{12} individuals of these genotypes, respectively,

in cases and n_{20}, n_{21}, and n_{22} individuals of these genotypes, respectively, in controls. Let $n_{1\cdot} = n_{10} + n_{11} + n_{12}$ be the total number of individuals in cases and $n_{2\cdot} = n_{20} + n_{21} + n_{22}$ in controls. The total number of individuals involved in the study is denoted by $n_{\cdot\cdot}$ $(= n_{1\cdot} + n_{2\cdot})$.

The vector of genotype counts (n_{10}, n_{11}, n_{12}) in cases follows a trinomial distribution with parameter $n_{1\cdot}$ and (P_{10}, P_{11}, P_{12}). The vector of genotype counts (n_{20}, n_{21}, n_{22}) in controls follows a trinomial distribution as well. Since $P_{11} = 1 - P_{10} - P_{12}$ and $P_{21} = 1 - P_{20} - P_{22}$, the likelihood function can be described by P_{10}, P_{12}, P_{20}, and P_{22}. Let $\mathbf{P} = (P_{10}, P_{12}, P_{20}, P_{22})$ be the vector of parameters. The natural parameter space for $\mathbf{P}$ is Θ_2 which is defined as

$$\Theta_2 = \{\mathbf{P} | 0 \leq P_{10}, P_{12}, P_{20}, P_{22} \leq 1, P_{10} + P_{12} \leq 1, P_{20} + P_{22} \leq 1\}.$$

The log-likelihood function is

$$l(\mathbf{P}) = l_1(P_{10}, P_{12}) + l_2(P_{20}, P_{22}),$$

where

$$l_i(P_{i0}, P_{i2}) = n_{i0} \log(P_{i0}) + n_{i1} \log(1 - P_{i0} - P_{i2}) + n_{i2} \log(P_{i2}), \quad i = 1, 2.$$

Regular association tests often test the null hypothesis $\mathbf{P} \in \Theta_0$, where

$$\Theta_0 = \{\mathbf{P} | P_{10} = P_{20}, P_{12} = P_{22}\} \cap \Theta_2,$$

against the alternative $\mathbf{P} \in \Theta_2$ (Pearson's chi-square statistic) or some other alternatives implied by an assumed disease model (for instance, additive, dominance, or recessive model). Particularly, the popular Cochran-Armitage test for trend is based on the model that the log of odds of disease is linear in the number of copies of A alleles. It is well known that, although not necessary to assume a disease model, Pearson's chi-square can be less powerful than model-based tests. On the other hand, model-based tests are less robust to model misspecification. Parameter space Θ_0 is stringent as it requires the frequency of each genotype in cases is the same as in controls.

Let Θ_1 denote a less stringent parameter space in which only the frequency of allele A is required to be the same in cases as in controls. In terms of genotype frequencies, Θ_1 is given by

$$\Theta_1 = \{\mathbf{P} | P_{10} - P_{12} = P_{20} - P_{22}\} \cap \Theta_2.$$

It is apparent that $\Theta_0 \subset \Theta_1 \subset \Theta_2$. Note that the frequency of A allele in cases is $P_{12}+P_{11}/2 = 1/2-(P_{10}-P_{12})/2$. Similar relationship holds in controls. So the parameter space Θ_1 indeed requires equality of A allele frequency in cases and in controls.

We thus consider statistical tests for the null hypothesis $\mathbf{P} \in \Theta_1$ against the alternative $\mathbf{P} \in \Theta_2$. We will introduce the likelihood ratio statistic, the score statistic, and the Wald statistic. According to standard asymptotic theory, all these tests are asymptotically equivalent to each other and asymptotically follow a chi-square distribution with 1 degree of freedom.

The likelihood ratio statistic is

$$\Lambda = 2 \left[\max_{\mathbf{P} \in \Theta_2} l(\mathbf{P}) - \max_{\mathbf{P} \in \Theta_1} l(\mathbf{P}) \right].$$

The maximum of $l(\mathbf{P})$ for $\mathbf{P} \in \Theta_2$ is easy to compute. It is reached at $\hat{P}_{10} = n_{10}/n_{1\cdot}, \hat{P}_{12} = n_{12}/n_{1\cdot}, \hat{P}_{20} = n_{20}/n_{2\cdot}$, and $\hat{P}_{22} = n_{22}/n_{2\cdot}$. However, there is no explicit solution to $\max_{\mathbf{P} \in \Theta_1} l(\mathbf{P})$. To solve this problem, we maximize a profile likelihood.

Let p_1 and p_2 be the frequency of allele A in cases and in controls, respectively, and let F_1 and F_2 be the coefficient of inbreeding for cases and controls, respectively. This F value can be interpreted as the probability that a pair of alleles in a population are identical by descent or the correlation coefficient bwtween the indicators of a pair of alleles when the mating is random(Weir and Hill 2002). The former interpretation necessarily requires F to be in the interval $[0, 1]$. We adopt the latter interpretation so F is allowed to be negative.

The genotype frequencies can be written

$$\begin{aligned} P_{i2} &= F_i p_i + (1-F_i)p_i^2 = p_i^2 + F_i p_i q_i, \quad i = 1, 2, \\ P_{i1} &= 2(1-F_i)p_i(1-p_i) = 2p_i q_i - 2F_i p_i q_i, \quad i = 1, 2, \\ P_{i0} &= F_i(1-p_i) + (1-F_i)(1-p_i)^2 = q_i^2 + F_i p_i q_i, \quad i = 1, 2. \end{aligned}$$

Under the null hypothesis $\mathbf{P} \in \Theta_1$, $p_1 = p_2 = p$. Let $q = 1 - p$. In order to maximize $l(\mathbf{P})$ for $\mathbf{P} \in \Theta_1$, consider the profile likelihood function $\tilde{l}(p)$ defined as

$$\tilde{l}(p) = \max_{F_1} l_1(q^2 + F_1 pq, p^2 + F_1 pq) + \max_{F_2} l_2(q^2 + F_2 pq, p^2 + F_2 pq).$$

An explicit solution to $\max_{F_1} l_1(q^2 + F_1 pq, p^2 + F_1 pq)$ and $\max_{F_2} l_2(q^2 + F_2 pq, p^2 + F_2 pq)$ is given in Appendix A. It is obvious that the maximum of $\tilde{l}(p)$ over $p \in [0, 1]$ is the same as $\max_{\mathbf{P} \in \Theta_1} l(\mathbf{P})$. Maximization of $\tilde{l}(p)$ can be achieved using regular numerical algorithms. Let $\dot{\mathbf{P}}$ denote the solution to $\max_{\mathbf{P} \in \Theta_1} l(\mathbf{P})$. The likelihood ratio test Λ equals

$$\Lambda = 2 \sum_{i=1,2} \sum_{j=0,1,2} n_{ij} \log(\hat{P}_{ij} / \dot{P}_{ij}),$$

where $\hat{P}_{11} = 1 - \hat{P}_{10} - \hat{P}_{12}$ and $\dot{P}_{21} = 1 - \dot{P}_{20} - \dot{P}_{22}$.

The score statistic, denoted by S, for testing $\mathbf{P} \in \Theta_1$ versus $\mathbf{P} \in \Theta_2$ is (Appendix B),

$$S = \sum_{i=1,2} \sum_{j=0,1,2} n_{ij} (\hat{P}_{ij} / \tilde{P}_{ij} - 1),$$

where $\tilde{P}_{ij}$s, $i = 1, 2, j = 0, 1, 2$ are consistent estimate of P_{ij} under the hypothesis $\mathbf{P} \in \Theta_1$. For instance, one can use the maximum likelihood estimate $\dot{\mathbf{P}}$ under $\mathbf{P} \in \Theta_1$. However, to avoid maximizing over the profile likelihood $\tilde{l}(p)$, we fix p at the allele frequency in the combined sample of cases and controls, which is $\hat{p} = (n_{12} + n_{22})/n_{..} + (n_{11} + n_{21})/2n_{..}$. Then we use the result in Appendix A to find out P_{10} and P_{12} that maximize $l_1(P_{10}, P_{12})$ and P_{20} and P_{22} that maximize $l_2(P_{20}, P_{22})$. Since $\hat{p}$ converges to the true value of p in probability under the null hypothesis, the genotype frequencies obtained this way also converges to their true values in probability, which means they are consistent.

The Wald statistic, denoted by W, is (Appendix B)

$$W = \frac{[(\hat{P}_{12} - \hat{P}_{10}) - (\hat{P}_{22} - \hat{P}_{20})]^2}{n_{1\cdot}^{-1}[\hat{P}_{12} + \hat{P}_{10} - (\hat{P}_{12} - \hat{P}_{10})^2] + n_{2\cdot}^{-1}[\hat{P}_{22} + \hat{P}_{20} - (\hat{P}_{22} - \hat{P}_{20})^2]}.$$

Let $\hat{p}_1 = \hat{P}_{12} + \hat{P}_{11}/2$ be the frequency of allele A in cases and $\hat{p}_2 = \hat{P}_{22} + \hat{P}_{21}/2$ in controls. In addition, define $\hat{q}_1 = 1 - \hat{p}_1$, $\hat{q}_2 = 1 - \hat{p}_2$, and $\hat{q} = 1 - \hat{p}$. Let $\hat{F}_1 = 1 - \hat{P}_{11}/2\hat{p}_1\hat{q}_1$, $\hat{F}_2 = 1 - \hat{P}_{21}/2\hat{p}_2\hat{q}_2$, and $\hat{F} = 1 - (\hat{P}_{11} + \hat{P}_{21})/2\hat{p}\hat{q}$. With these notations, it is easy to see that

$$W = \frac{2(\hat{p}_1 - \hat{p}_2)^2}{n_{1\cdot}^{-1}\hat{p}_1\hat{q}_1(1 + \hat{F}_1) + n_{2\cdot}^{-1}\hat{p}_2\hat{q}_2(1 + \hat{F}_2)}.$$

We note that, in these notations, the Cochran-Armitage trend test, denoted by G, can be expressed as

$$G = \frac{2(\hat{p}_1 - \hat{p}_2)^2}{(n_{1\cdot}^{-1} + n_{2\cdot}^{-1})\hat{p}\hat{q}(1 + \hat{F})},$$

which shares the same numerator with the Wald statistic W. When $n_{1\cdot} = n_{2\cdot}$ (i.e., the number of cases is the same as the number of controls), the Wald statistic W is greater than the statistic G. This is because the denominator of W becomes $\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2 - (\hat{P}_{11} + \hat{P}_{21})/4$ and the denominator of G becomes $2\hat{p}\hat{q} - (\hat{P}_{11} + \hat{P}_{21})/4 = (\hat{p}_1 + \hat{p}_2)(\hat{q}_1 + \hat{q}_2)/2 - (\hat{P}_{11} + \hat{P}_{21})/4$. The former is smaller than the latter as their difference (the former minus the latter) is

$$\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2 - (\hat{p}_1 + \hat{p}_2)(\hat{q}_1 + \hat{q}_2)/2 = -(\hat{p}_1 - \hat{p}_2)^2/2 < 0.$$

It is apparent that this difference converges to 0 in probability for $\mathbf{P} \in \Theta_1$ as sample sizes in cases and in controls both to to infinity. However, when $n_{1\cdot} \neq n_{2\cdot}$, no such relationship holds. For instance, when the genotype ratios $aa : aA : AA$ are 5:30:30 in cases and 3:2:3 in controls, statistic G (= 2.37) is larger than statistic W (= 1.48); when the genotype ratios are 17:95:80 in cases and 42:71:62 in controls, statistic G (= 8.43) is smaller than statistic W (= 8.48).

The likelihood ratio statistic for testing $\mathbf{P} \in \Theta_0$ against $\mathbf{P} \in \Theta_2$ is

$$2\left[\max_{\mathbf{P}\in\Theta_2} l(\mathbf{P}) - \max_{\mathbf{P}\in\Theta_0} l(\mathbf{P})\right].$$

It is natural to decompose it in the following way:

$$2\left[\max_{\mathbf{P}\in\Theta_2} l(\mathbf{P}) - \max_{\mathbf{P}\in\Theta_0} l(\mathbf{P})\right] = 2\left[\max_{\mathbf{P}\in\Theta_2} l(\mathbf{P}) - \max_{\mathbf{P}\in\Theta_1} l(\mathbf{P})\right] + 2\left[\max_{\mathbf{P}\in\Theta_1} l(\mathbf{P}) - \max_{\mathbf{P}\in\Theta_0} l(\mathbf{P})\right].$$

The first term of the right hand side is the likelihood ratio test Λ proposed in this paper. The second term is the likelihood ratio statistic for testing $\mathbf{P} \in \Theta_0$ against $\mathbf{P} \in \Theta_1$. Equivalently, it tests for the null that $F_1 = F_2$ against the alternative $F_1 \neq F_2$. That is, whether the departure from the Hardy-Weinberg equilibrium in cases, as measured by F_1, is the same as in controls as measured by F_2. When population stratification is a confounding factor to genetic association, the F coefficient in cases will be different from that in controls. Hence this likelihood ratio statistic can be used to test whether or not there exists population stratification effect that confounds with genetic association. It can be approximated by its score statistic S' which equals:

$$S' = \frac{n_{..}^2}{n_{1\cdot}n_{2\cdot}} \cdot \frac{(n_{12}/n_{\cdot 2} - 2n_{11}/n_{\cdot 1} + n_{10}/n_{\cdot 0})^2}{1/n_{\cdot 2} + 4/n_{\cdot 1} + 1/n_{\cdot 0}},$$

where $n_{\cdot 2} = n_{12} + n_{22}, n_{\cdot 1} = n_{11} + n_{21}$, and $n_{\cdot 0} = n_{10} + n_{20}$. The derivation of this statistic is similar to that of S. The detail is omitted.

Statistic S' is asymptotically independent of the Wald statistic W for $\mathbf{P} \in \Theta_0$. Here is a sketch of a proof. Let $\phi = n_{1\cdot}/n_{\cdot\cdot}$ be the proportion of cases out of the total number of study subjects. Define function $f(\hat{P}_{10}, \hat{P}_{11}, \hat{P}_{12}, \hat{P}_{20}, \hat{P}_{21}, \hat{P}_{22})$ as

$$f(\hat{P}_{10}, \hat{P}_{11}, \hat{P}_{12}, \hat{P}_{20}, \hat{P}_{21}, \hat{P}_{22}) = \frac{n_{\cdot\cdot}^{1/2}}{[\phi(1-\phi)]^{1/2}} \cdot \frac{A}{B^{1/2}},$$

where

$$\begin{aligned} A &= \frac{\phi\hat{P}_{12}}{\phi\hat{P}_{12} + (1-\phi)\hat{P}_{22}} - \frac{2\phi\hat{P}_{11}}{\phi\hat{P}_{11} + (1-\phi)\hat{P}_{21}} + \frac{\phi\hat{P}_{10}}{\phi\hat{P}_{10} + (1-\phi)\hat{P}_{20}} \\ B &= \frac{1}{\phi\hat{P}_{12} + (1-\phi)\hat{P}_{22}} + \frac{4}{\phi\hat{P}_{11} + (1-\phi)\hat{P}_{21}} + \frac{1}{\phi\hat{P}_{10} + (1-\phi)\hat{P}_{20}} \end{aligned}$$

It is easy to see that $[f(\hat{P}_{10}, \hat{P}_{11}, \hat{P}_{12}, \hat{P}_{20}, \hat{P}_{21}, \hat{P}_{22})]^2$ equals the score statistic S'. Under the hypothesis $\mathbf{P} \in \Theta_0$, let P_0, P_1, and P_2 be the frequencies of genotypes aa, aA, and AA, respectively, that are common to cases and controls. The function $f(\hat{P}_{10}, \hat{P}_{11}, \hat{P}_{12}, \hat{P}_{20}, \hat{P}_{21}, \hat{P}_{22})$

is 0 when $\mathbf{P} \in \Theta_0$. Using Taylor's expansion,

$$\begin{aligned} & f(\hat{P}_{10}, \hat{P}_{11}, \hat{P}_{12}, \hat{P}_{20}, \hat{P}_{21}, \hat{P}_{22}) \\ &= \left[\frac{1}{P_0} + \frac{4}{P_1} + \frac{1}{P_2}\right]^{-1/2} \left[\left(\frac{\hat{P}_{10}}{P_0} - \frac{2\hat{P}_{11}}{P_1} + \frac{\hat{P}_{12}}{P_2}\right) - \left(\frac{\hat{P}_{20}}{P_0} - \frac{2\hat{P}_{21}}{P_1} + \frac{\hat{P}_{22}}{P_2}\right)\right] + o_p(n_{..}^{-1/2}) \end{aligned}$$

When $\mathbf{P} \in \Theta_0$, The main part on the right hand side is not correlated with either $\hat{P}_{12} - \hat{P}_{10}$ or $\hat{P}_{22} - \hat{P}_{20}$, since

$$\begin{aligned} & Cov\left(\hat{P}_{12} - \hat{P}_{10}, \frac{\hat{P}_{10}}{P_0} - \frac{2\hat{P}_{11}}{P_1} + \frac{\hat{P}_{12}}{P_2}\right) \\ &= n_{1\cdot}[-P_2 + 2P_2 + (1 - P_2) - (1 - P_0) - 2P_0 + P_0] \\ &= 0, \end{aligned}$$

and similarly,

$$Cov\left(\hat{P}_{22} - \hat{P}_{20}, \frac{\hat{P}_{20}}{P_0} - \frac{2\hat{P}_{21}}{P_1} + \frac{\hat{P}_{22}}{P_2}\right) = 0.$$

Since the numerator of statistic W is the square of $(\hat{P}_{12} - \hat{P}_{10}) - (\hat{P}_{22} - \hat{P}_{20})$, W is asymptotically uncorrelated with the S'.

SIMULATION

We first investigate the distribution of the proposed test under the hypothesis $\mathbf{P} \in \Theta_1$. In the simulation study of type I error rate, of particular interest is to investigate the situation where the frequency of the putative allele is the same in cases as in controls but the genotype frequencies differ. So the allele frequency in cases is set to be the same as in controls while the F coefficients are allowed to be different. Sample sizes in cases and in controls are allowed to be different as well. The number of simulation replicates in all scenarios (including power study) is fixed at 10000. Simulation results are presented in Tables 1–4. The type I error rates of the proposed score statisti S, Wald statistic W, and the likelihood ratio statistic Λ are close to the nominal values, so does the type I error rate of Cochran-Armitage trend test statistic G. The Pearson's chi-square statistic X^2 tends to be inflated when $F_1 \neq F_2$ which causes the genotype frequencies in cases to be different from those in controls.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

The power study is conducted as follows. Let $P_i, i = 0, 1, 2$ be the frequency of genotype i. Let $K = \sum_{i=0}^{2} f_i P_i$ be the prevalence of the trait. The frequency of genotype i would be $f_i P_i / K$ in cases and $(1 - f_i)P_i/(1 - K)$ in controls. In the absence of association, $f_0 = f_1 = f_2 = K$. There is no difference in genotype frequencies between cases and controls. Let $\gamma_i = f_i/f_0, i = 1, 2$, be the relative risk of genotype i to genotype 0. In the simulation, we consider a dominance model ($\gamma_1 = \gamma_2$), a recessive model ($\gamma_1 = 1$), an additive model ($\gamma_1 = (1 + \gamma_2)/2$), and a multiplicative model ($\gamma_1 = \gamma_2^{1/2}$). Given population prevalence K and the relative risk γ_2, f_0 can be determined from $f_0 = K/(P_0 + \gamma_1 P_1 + \gamma_2 P_2)$, from which $f_1 = \gamma_1 f_0$ and $f_2 = \gamma_2 f_0$ can be computed for each model. The value of γ_2 is fixed at 2. It is assumed that Hardy-Weinberg equilibrium holds at the disease locus and the frequency of the disease allele is denoted by p.

At significance level 0.01, the power of statistics G, Λ, S, W, and the Pearson's chi-square statistic X^2 is simulated for allele frequency $p = 0.1, 0.3, 0.5$, prevalence $K = 0.01, 0.1, 0.3$ and sample size $(n_{1\cdot}, n_{2\cdot}) = (300, 100), (200, 200)$, and $(100, 300)$. Results are presented for the dominance model (Fig. **??**), the recessive model (Fig. **??**), the additive model (Fig. **??**), and the multiplicative model (Fig. **??**). The four statistics G, S, W, and Λ have similar power when the number of cases is the same as the number of controls (i.e., $n_{1\cdot} = n_{2\cdot} = 200$) in all four generating models while the statistic X^2 has less power for the additive model and the multiplicative model but has more power for the dominant model (when allele frequency p is 0.3 or 0.5) and the recessive model. There seems to be a tendency that the score statistic S is less powerful than the likelihood ratio statistic Λ and the Wald statistic W when there are

more cases than controls (i.e., $(n_{1\cdot}, n_{2\cdot}) = (300, 100)$) but is more powerful when there are less cases than controls (i.e., $(n_{1\cdot}, n_{2\cdot}) = (100, 300)$) for allele frequency $p = 0.1$ and 0.3. The performance of the Cochran-Armitage trend test G is very similar to that of the score statistic S except that for the dominant model, it is a bit more powerful when $(n_{1\cdot}, n_{2\cdot}) = (300, 100)$ and a bit less powerful when $(n_{1\cdot}, n_{2\cdot}) = (100, 300)$ while for the recessive model the situation is the reverse. The power of the Pearson's chi-square statistic X^2 is the lowest for the additive model and the multiplicative model when allele frequency is 0.3 or 0.5. For each model, the power of the likelihood ratio statistic Λ does not change as much as other test statistics as the ratio of the subject number between cases and controls changes while the Wald statistic W seems to change most.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

APPLICATION TO AN ADULT HEIGHT STUDY IN A EUROPEAN AMERICAN POPULATION

A study of adult height in a European American population involves 1057 "short" individuals and 1132 "tall" individuals(Campbell, Ogburn, Lunetta, Lyon, Freedman, Groop, Altshuler, Ardlie, and Hirschhorn 2005). The SNP marker LCT–13910 has shown significant association. However, if the data are divided according to whether the four grand parents are all US-born, predominantly born in Southeastern Europe, or predominantly born in Northwestern Europe, it significance is dramatically reduced in each sub-group, suggesting population stratification exists in this European American population. The genotype counts at marker marker LCT–13910 are shown in table **??**, so are the p-values of various tests. Results from the proposed likelihood ratio statistic Λ, score statistic S, and Wald statistic W

are very similar to the Cochran-Armitage test for trend G. In addition, the value of statistic S' is 1.74 which does not detect population stratification (p-value = 0.1871).

[Table 5 about here.]

DISCUSSION

Traditional tests of genetic association for case-control design typically test the null hypothesis that the genotype frequencies in cases are the same as in controls. Such a null hypothesis is too stringent making these tests are not robust to conditions such as population structure and cryptic relatedness. Allelic test may be the only test directly based on the difference of allele frequency between cases and controls but it relies on the assumption of Hardy-Weinberg equilibrium. The Cochran-Armitage test for trend turns out to be test based on allele frequency difference as well but it relies on an assumed additive genetic model. We have introduced a model-free likelihood analysis framework for comparing allele frequency between cases and controls. It requires neither Hardy-Weinberg equilibrium nor assumed disease models. Compared to the model-free Pearson's chi-square test, the likelihood ratio statistic, the score statistic, and the Wald statistic have the virtue that each of them has only 1 degree of freedom.

One advantage of comparing allele frequency, instead of genotype frequencies, between cases and controls is that the null hypothesis is less affected by differences in the composition between cases and controls. For instance, due to population stratification, there will be difference in the F coefficient between cases and controls. While F coefficient affects genotype frequencies, it does not affect the difference between the two homozygous genotype frequencies. For example, $F_{10} - F_{12} = q_1 - p_1$. This explains why in the simulation study of type I error rate, the Pearson's chi-square statistic is inflated when F_1 in cases is different than F_2 in controls while the others are not. However, this does not mean the proposed statistics are immune to population stratification as it will cause a difference in allele frequency between cases and controls as well.

The statistic W has been proposed previously as a statistic to be used when the Hardy-Weinberg equilibrium does not hold in cases and controls combined (Schaid and Jacobsen 1999). It was regarded as a test statistic needs further study(Knapp 2001). The current study shows that W is the Wald statistic for testing $\mathbf{P} \in \Theta_1$ versus $\mathbf{P} \in \Theta_2$ in this study.

It is interesting to note that although the null hypothesis for the Cochran-Armitage test for trend is that the genotype frequencies in cases are the same as in controls it is much less sensitive to violation of this null than the Pearson's chi-square test for the same null (Tables 1-4). This phenomenon may be explained by the fact that the Cochran-Armitage test for trend turns out to be a test based on the difference of disease allele frequencies between cases and controls.

Overall, the proposed likelihood ratio statistic, the score statistic, and the Wald statistic provide some attractive alternatives for genetic association studies. They require neither the assumption of Hardy-Weinberg equilibrium nor assuming disease models. They involves only one degree of freedom. These statistics should be useful for unraveling genetic factors underlying complex traits.

The proposed statistics have been implemented in R. The code is available from the author upon request.

REFERENCES

Amemiya, T. (1985). *Advanced Econometrics*. Harvard Press.

Campbell, C. D., E. L. Ogburn, K. L. Lunetta, H. N. Lyon, M. L. Freedman, L. C. Groop, D. Altshuler, K. G. Ardlie, and J. N. Hirschhorn (2005). Demonstrating stratification in a european american population. *Nature Genetics 37*, 868–872.

Freidlin, B., G. Zheng, Z. Li, and J. Gastwirth (2002). Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Hum Hered 53*, 146–152.

Hindorff, L. A., H. Junkins, J. P. Mehta, and T. A. Manolio (2009). A catalog of published

genome-wide association studies. *Available at: www.genome.gov/gwastudies*, Accessed Oct 28, 2009.

Knapp, M. (2001). Re: "Biased tests of association: Comparisons of allele frequencies when departing from Hardy-Weinberg proportions. *Am J Epidemiol 154*, 287.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics 53*, 1253–1261.

Schaid, D. J. and S. J. Jacobsen (1999). Biased tests of association: Comparisons of allele frequencies when departing from Hardy-Weinberg proportions. *Am J Epidemiol 149*, 706–711.

Schaid, D. J., S. K. McDonnell, S. J. Hebbring, J. M. Cunningham, and S. N. Thibodeau (2005). Nonparametric tests of association of multiple genes with human disease. *Am J Hum Genet 76*, 780–793.

Slager, S. and D. Schaid (2001). Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Hum Hered 52*, 149–153.

Wang, K. and V. Sheffield (2005). A constrained-likelihood approach to marker-trait association studies. *Am J Hum Genet 77*, 768–780.

Weir, B. S. and W. G. Hill (2002). Estimating F-statistics. *Annu. Rev. Genet. 36*, 721–750.

Zheng, G. (2003). Use of max and min scores for trend tests for association when the genetic model is unknown. *Statistics in Medicine 22*, 2657–2666.

APPENDIX A THE MAXIMUM LIKELIHOOD ESTIMATE OF THE F COEFFICIENT FROM A SAMPLE GIVEN ALLELE FREQUENCY P

Let P_0, P_1, and P_2 be the frequencies of genotypes AA, Aa, and aa, respectively, in the sample. Then $P_0 = q^2 + Fpq, P_1 = 2pq - 2Fpq$, and $P_2 = p^2 + Fpq$. Let $\gamma = Fpq$. The log-likelihood function is

$$l(\gamma, p) = n_0 \log(q^2 + \gamma) + n_1 \log(2pq - 2\gamma) + n_2 \log(p^2 + \gamma).$$

Since each $P_i, i = 0, 1, 2$ is a probability, γ naturally satisfies $\gamma \in [b_l, b_u]$, where

$$b_l = \max\{-p^2, -q^2, pq - 0.5\}, b_u = \min\{1 - p^2, 1 - q^2, pq\}.$$

Given the value of p, $l(\gamma, p)$ is concave in γ. So it has a unique global maximum which occurs either in the interior or on the boundaries of interval $[b_l, b_u]$.

First consider the case that none of n_0, n_1 and n_2 is 0. Given the value of p, the first-order equation with respect to γ is

$$\frac{n_0}{q^2 + \gamma} + \frac{n_2}{p^2 + \gamma} - \frac{n_1}{pq - \gamma} = 0,$$

which results in a univariate second-order equation in γ. Of the two roots, the one that converges to the true value of γ as the sample size goes to infinity is

$$\gamma = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

where

$$\begin{aligned} a &= n_0 + n_1 + n_2, \\ b &= (n_0 + n_1)p^2 + (n_2 + n_1)q^2 - (n_0 + n_2)pq, \end{aligned}$$

and

$$c = pq(n_1 pq - n_0 p^2 - n_2 q^2).$$

Once having the value of γ, the value of F is obtained by $F = \gamma/pq$.

The situation that one and only one of n_0, n_1, and n_2 is 0 is dealt with as follows. If $n_0 = 0$, the solution to the first-order condition is

$$\gamma = n_2 p/n - p^2.$$

If $n_2 = 0$, the solution to the first-order equation is

$$\gamma = n_0 q/n - q^2.$$

If $n_1 = 0$, $l(\gamma, p)$ is an increasing function in γ. So the MLE is $\gamma = b_u$.

If only $n_0 > 0$ or only $n_2 > 0$, $l(\gamma, p)$ is increasing in γ. So the MLE of γ is b_u. is determined by $q^2 + \gamma = 1$, which implies $\gamma = 1 - q^2$. If only $n_1 > 0$, $l(\gamma, p)$ is decreasing in γ. So the MLE of γ is b_l.

Since $P_{11} = 1 - P_{10} - P_{12}$ and $P_{21} = 1 - P_{20} - P_{22}$, the vector of first-order derivatives is

$$\begin{pmatrix} \partial l/\partial P_{10} \\ \partial l/\partial P_{12} \\ \partial l/\partial P_{20} \\ \partial l/\partial P_{22} \end{pmatrix} = \begin{pmatrix} n_{10}/P_{10} - n_{11}/P_{11} \\ n_{12}/P_{12} - n_{11}/P_{11} \\ n_{20}/P_{20} - n_{21}/P_{21} \\ n_{22}/P_{22} - n_{21}/P_{21} \end{pmatrix}$$

Define matrices

$$\mathbf{A}_i = \begin{pmatrix} n_{i0}/P_{i0}^2 + n_{i1}/P_{i1}^2 & n_{i1}/P_{i1}^2 \\ n_{i1}/P_{i1}^2 & n_{i2}/P_{i2}^2 + n_{i1}/P_{i1}^2 \end{pmatrix}, i = 1, 2$$

The expectation of $\mathbf{A}_i$ is

$$E(\mathbf{A}_i) = n_{i\cdot} \begin{pmatrix} 1/P_{i0} + 1/P_{i1} & 1/P_{i1} \\ 1/P_{i1} & 1/P_{i2} + 1/P_{i1} \end{pmatrix}$$

It is straightforward to compute that the Fisher Information matrix is the following block-diagonal matrix

$$\mathbf{I} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}$$

where $\mathbf{0}$ is a 2×2 matrix whose elements are all 0. The expectation of $\mathbf{I}$, $E(\mathbf{I})$, is a block-diagonal matrix with blocks $E(\mathbf{A}_1)$ and $E(\mathbf{A}_2)$. It is easy to compute that

$$\begin{aligned}
&\left[\frac{\partial l(\mathbf{P})}{\partial \mathbf{P}}\right]^t \cdot [E(\mathbf{I})]^{-1} \cdot \frac{\partial l(\mathbf{P})}{\partial \mathbf{P}} \\
&= n_{1\cdot}^{-1}\left[P_{10}P_{11}\left(\frac{n_{10}}{P_{10}} - \frac{n_{11}}{P_{11}}\right)^2 + P_{11}P_{12}\left(\frac{n_{11}}{P_{11}} - \frac{n_{12}}{P_{12}}\right)^2 + P_{10}P_{12}\left(\frac{n_{10}}{P_{10}} - \frac{n_{12}}{P_{12}}\right)^2\right] \\
&\quad + n_{2\cdot}^{-1}\left[P_{20}P_{21}\left(\frac{n_{20}}{P_{20}} - \frac{n_{21}}{P_{21}}\right)^2 + P_{21}P_{22}\left(\frac{n_{21}}{P_{21}} - \frac{n_{22}}{P_{22}}\right)^2 + P_{20}P_{22}\left(\frac{n_{20}}{P_{20}} - \frac{n_{22}}{P_{22}}\right)^2\right] \\
&= n_{1\cdot}^{-1}\left[\frac{n_{10}^2}{P_{10}} + \frac{n_{11}^2}{P_{11}} + \frac{n_{12}^2}{P_{12}} - n_{1\cdot}^2\right] + n_{2\cdot}^{-1}\left[\frac{n_{20}^2}{P_{20}} + \frac{n_{21}^2}{P_{21}} + \frac{n_{22}^2}{P_{22}} - n_{2\cdot}^2\right] \\
&= \frac{\hat{P}_{10}n_{10}}{P_{10}} + \frac{\hat{P}_{11}n_{11}}{P_{11}} + \frac{\hat{P}_{12}n_{12}}{P_{12}} + \frac{\hat{P}_{20}n_{20}}{P_{20}} + \frac{\hat{P}_{21}n_{21}}{P_{21}} + \frac{\hat{P}_{22}n_{22}}{P_{22}} - n_{\cdot\cdot} \\
&= \sum_{i=1,2}\sum_{j=0,1,2}(\hat{P}_{ij}/P_{ij} - 1)n_{ij}.
\end{aligned}$$

According to equation (4.5.5) of Amemiya(Amemiya 1985), the score statistic is obtained by substituting an estimate of $\mathbf{P}$ that is consistent under H_1.

Now we derive the Wald statistic. Let $\hat{\delta} = (\hat{P}_{12} - \hat{P}_{10}) - (\hat{P}_{22} - \hat{P}_{20}) = 2(\hat{p}_1 - \hat{p}_2)$ where $\hat{p}_1 = (2n_{12} + n_{11})/2n_{1\cdot}$ is the observed frequency of allele A in cases and $\hat{p}_2$ the observed frequency of allele A in controls. The variance of $\hat{\delta}$ is

$$\begin{aligned} Var(\hat{\delta}) &= Var(\hat{P}_{12} - \hat{P}_{10}) + Var(\hat{P}_{22} - \hat{P}_{20}) \\ &= n_{1\cdot}^{-1}[P_{12}(1 - P_{12}) + P_{10}(1 - P_{10}) + 2P_{12}P_{10}] \\ &\quad + n_{2\cdot}^{-1}[P_{22}(1 - P_{22}) + P_{20}(1 - P_{20}) + 2P_{22}P_{20}] \\ &= n_{1\cdot}^{-1}[P_{12} + P_{10} - (P_{12} - P_{10})^2] + n_{2\cdot}^{-1}[P_{22} + P_{20} - (P_{22} - P_{20})^2] \\ &= n_{1\cdot}^{-1} \cdot 2p_1q_1(1 + F_1) + n_{2\cdot}^{-1} \cdot 2p_2q_2(1 + F_2) \end{aligned}$$

So the Wald test statistic is

$$\begin{aligned} W &= \frac{\hat{\delta}^2}{\hat{Var}(\hat{\delta})} \\ &= \frac{2(\hat{p}_1 - \hat{p}_2)^2}{n_{1\cdot}^{-1}\hat{p}_1\hat{q}_1(1 + \hat{F}_1) + n_{2\cdot}^{-1}\hat{p}_2\hat{q}_2(1 + \hat{F}_2)} \end{aligned}$$

One can also apply the formal definition of Wald statistic (e.g., equation (4.5.4) of Amemiya(Amemiya 1985)). Let $\delta = (P_{12} - P_{10}) - (P_{22} - P_{20})$. The Wald statistic equals

$$-h(\mathbf{P})^t \left\{ \frac{\partial h}{\partial \mathbf{P}^t} \left[\frac{\partial^2 \log L}{\partial \mathbf{P} \partial \mathbf{P}^t} \right]^{-1} \frac{\partial h^t}{\partial \mathbf{P}} \right\}^{-1} h(\mathbf{P})$$

evaluated at $\hat{\mathbf{P}}$. Substituting $\partial^2 \log L/\partial \mathbf{P}\partial \mathbf{P}^t$ by $n_{\cdot\cdot}$plim $n_{\cdot\cdot}^{-1}\partial^2 \log L/\partial \mathbf{P}\partial \mathbf{P}^t = E(\mathbf{I})$ (Amemiya 1985), it becomes

$$\hat{\delta}\left[(1, -1, 1, -1)E(\mathbf{I}(\hat{P}))^{-1}(1, -1, 1, -1)^t\right]^{-1}\hat{\delta},$$

which turns out to be equal to W.

Figure 1: Power comparison for the dominant model. The bars in each group are for (in the order) Cochran-Armitage trend test G, likelihood ratio statistic Λ, score statistic S, Wald statistic W, and Pearson's chi-square statistic X^2.

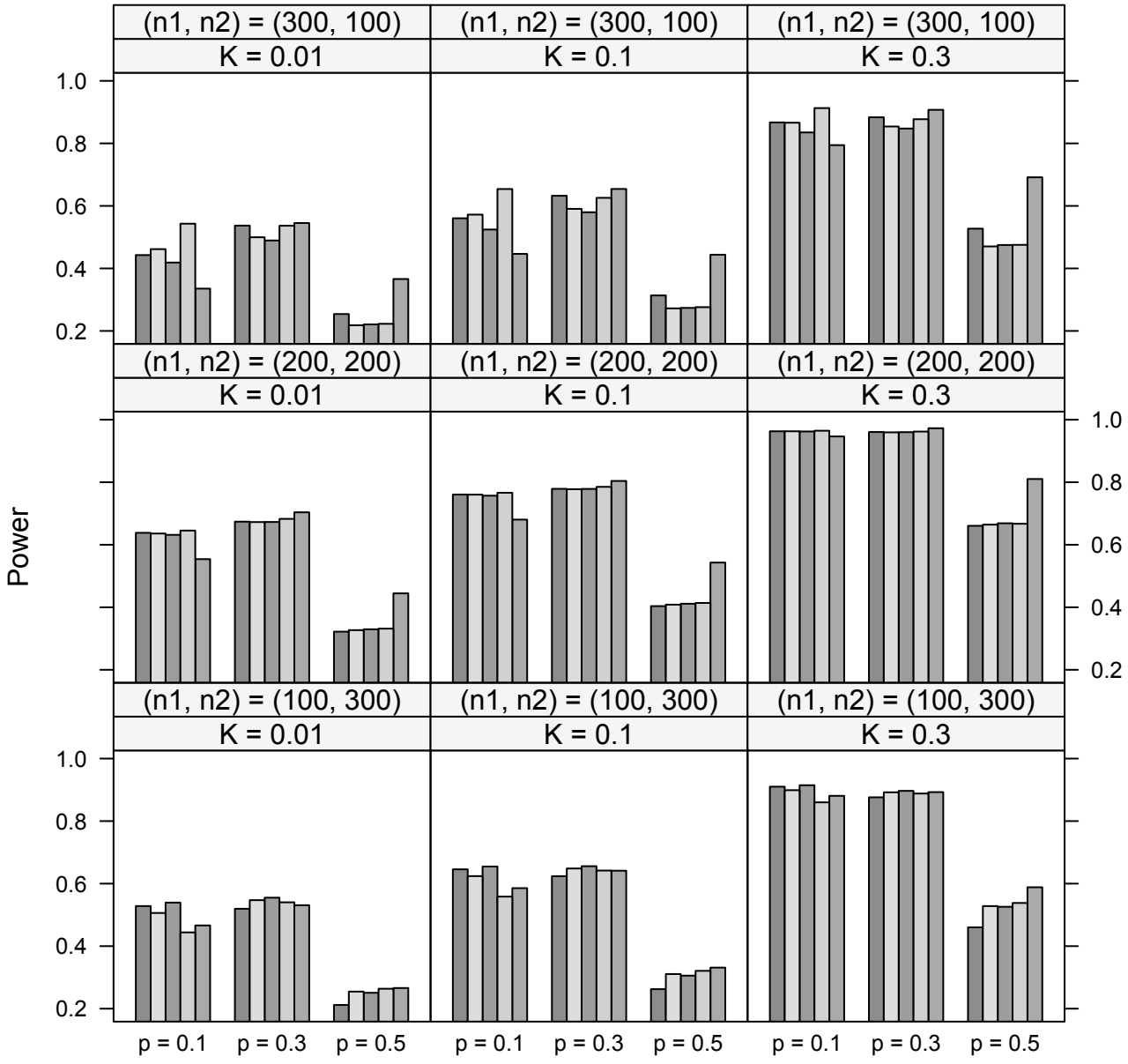

Figure 2: Power comparison for the recessive model. The bars in each group are for (in the order) Cochran-Armitage trend test G, likelihood ratio statistic Λ, score statistic S, Wald statistic W, and Pearson's chi-square statistic X^2.

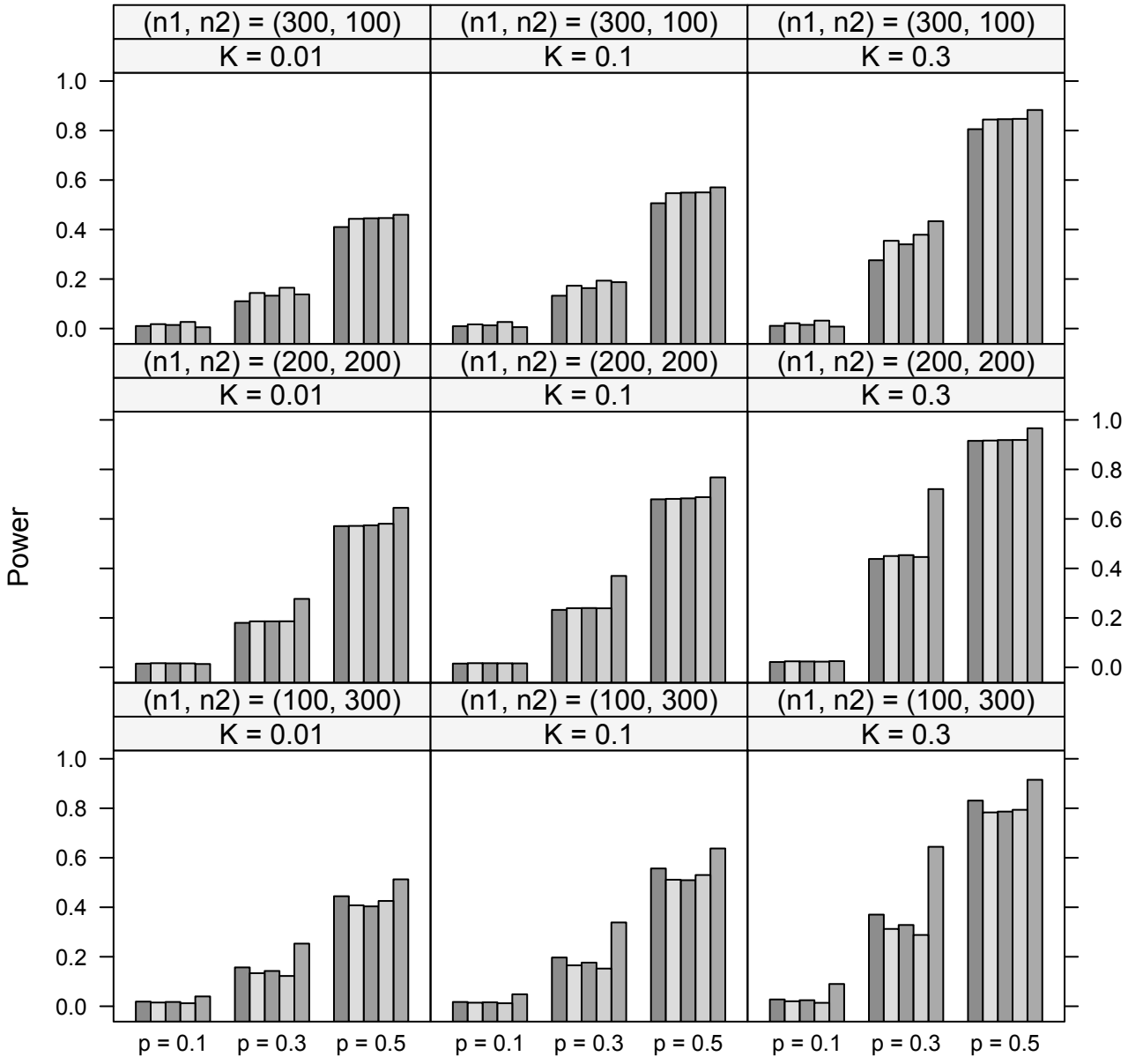

Figure 3: Power comparison for the additive model. The bars in each group are for (in the order) Cochran-Armitage trend test G, likelihood ratio statistic Λ, score statistic S, Wald statistic W, and Pearson's chi-square statistic X^2.

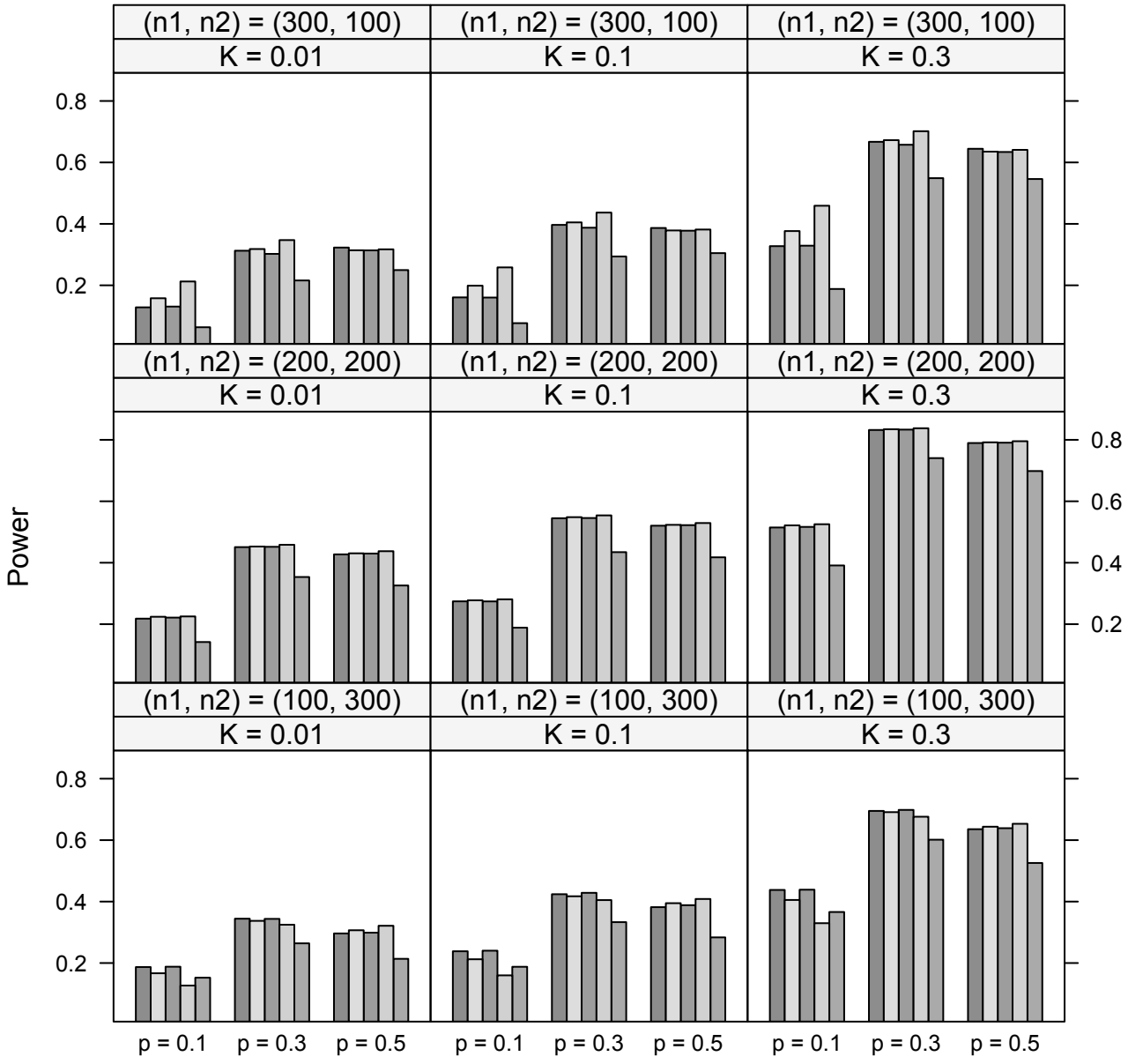

Figure 4: Power comparison for the multiplicative model. The bars in each group are for (in the order) Cochran-Armitage trend test G, likelihood ratio statistic Λ, score statistic S, Wald statistic W, and Pearson's chi-square statistic X^2.

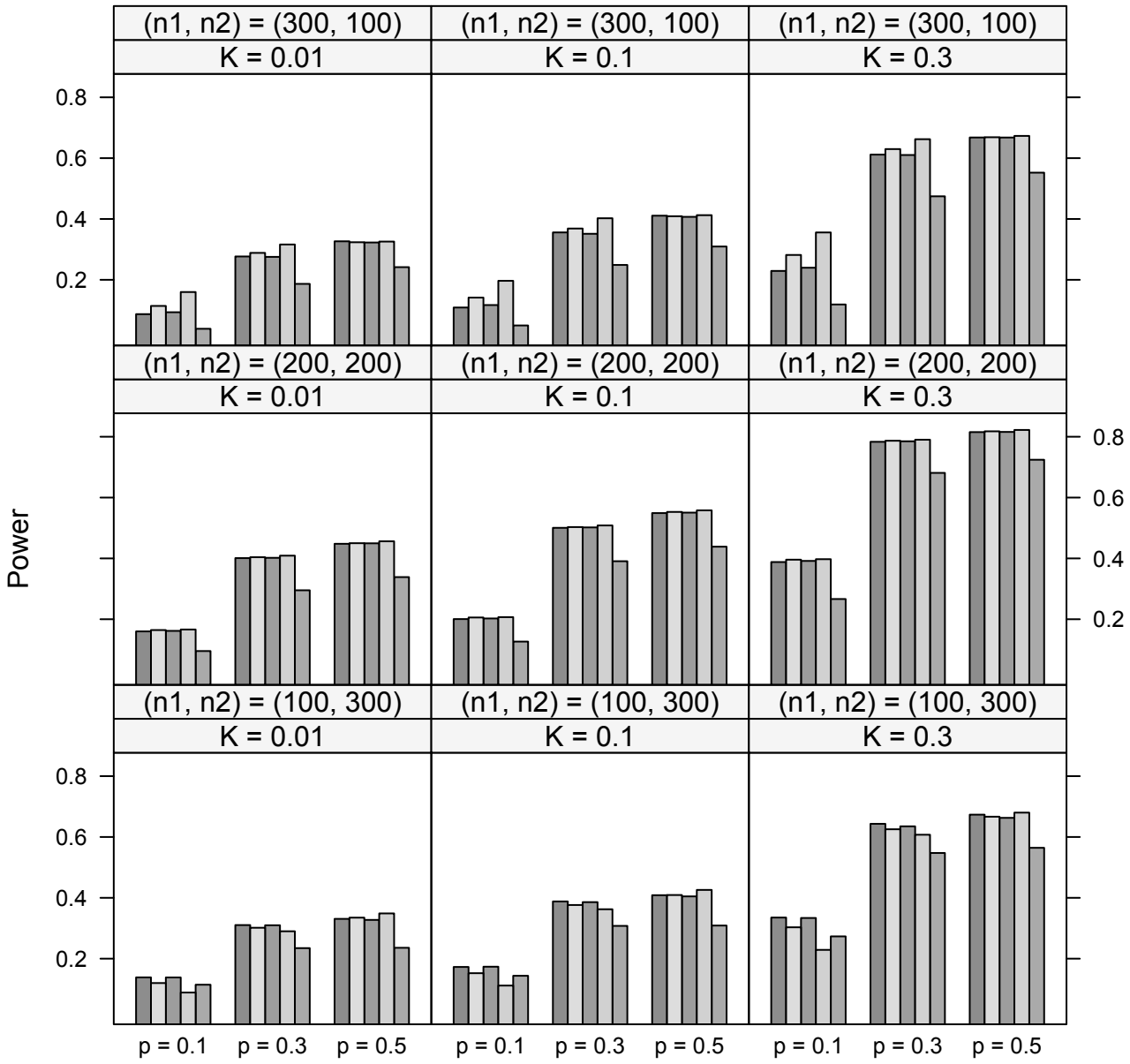

Table 1: Type I error rate. The vector of allele frequencies is $(p_1, p_2) = (0.1, 0.1)$ and significance level is $\alpha = 0.01$. X^2 is the Pearson's chi-square statistic.

		Statistic				
(F_1, F_2)	$(n_{1\cdot}, n_{2\cdot})$	X^2	G	S	W	Λ
(0.05, 0.05)	(100, 300)	0.0092	0.0104	0.0108	0.0131	0.0109
	(200, 200)	0.0056	0.0080	0.0086	0.0085	0.0087
	(300, 100)	0.0096	0.0088	0.0087	0.0142	0.0103
(0.1, 0.05)	(100, 300)	0.0179	0.0115	0.0109	0.0137	0.0116
	(200, 200)	0.0115	0.0116	0.0115	0.0123	0.0120
	(300, 100)	0.0093	0.0073	0.0080	0.0113	0.0095
(0.3, 0.05)	(100, 300)	0.1454	0.0129	0.0102	0.0161	0.0106
	(200, 200)	0.1484	0.0128	0.0127	0.0129	0.0131
	(300, 100)	0.0672	0.0076	0.0110	0.0118	0.0109
(0.1, −0.05)	(100, 300)	0.0999	0.0137	0.0112	0.0151	0.0123
	(200, 200)	0.0445	0.0096	0.0100	0.0103	0.0103
	(300, 100)	0.0185	0.0078	0.0096	0.0120	0.0102
(0.2, −0.05)	(100, 300)	0.2367	0.0133	0.0099	0.0135	0.0106
	(200, 200)	0.1700	0.0100	0.0109	0.0105	0.0107
	(300, 100)	0.0653	0.0071	0.0105	0.0134	0.0114

Table 2: Type I error rate. The vector of allele frequencies is $(p_1, p_2) = (0.1, 0.1)$ and significance level is $\alpha = 0.001$. X^2 is the Pearson's chi-square statistic.

		Statistic				
(F_1, F_2)	$(n_{1\cdot}, n_{2\cdot})$	X^2	G	S	W	Λ
(0.05, 0.05)	(100, 300)	0.0009	0.0014	0.0011	0.0030	0.0014
	(200, 200)	0.0009	0.0005	0.0004	0.0006	0.0005
	(300, 100)	0.0016	0.0012	0.0011	0.0020	0.0015
(0.1, 0.05)	(100, 300)	0.0023	0.0009	0.0008	0.0027	0.0012
	(200, 200)	0.0010	0.0016	0.0013	0.0019	0.0020
	(300, 100)	0.0006	0.0007	0.0007	0.0019	0.0010
(0.3, 0.05)	(100, 300)	0.0464	0.0010	0.0002	0.0032	0.0010
	(200, 200)	0.0306	0.0012	0.0012	0.0015	0.0015
	(300, 100)	0.0144	0.0003	0.0007	0.0026	0.0012
(0.1, −0.05)	(100, 300)	0.0243	0.0022	0.0015	0.0047	0.0023
	(200, 200)	0.0063	0.0009	0.0009	0.0011	0.0011
	(300, 100)	0.0020	0.0003	0.0005	0.0021	0.0005
(0.2, −0.05)	(100, 300)	0.0835	0.0013	0.0006	0.0029	0.0008
	(200, 200)	0.0289	0.0009	0.0007	0.0010	0.0008
	(300, 100)	0.0116	0.0010	0.0014	0.0021	0.0013

Table 3: Type I error rate. The vector of allele frequencies is $(p_1, p_2) = (0.3, 0.3)$ and significance level is $\alpha = 0.01$. X^2 is the Pearson's chi-square statistic.

		Statistic				
(F_1, F_2)	$(n_{1\cdot}, n_{2\cdot})$	X^2	G	S	W	Λ
(0.05, 0.05)	(100, 300)	0.0084	0.0090	0.0090	0.0111	0.0098
	(200, 200)	0.0091	0.0104	0.0105	0.0112	0.0108
	(300, 100)	0.0113	0.0106	0.0111	0.0128	0.0115
(0.1, 0.05)	(100, 300)	0.0141	0.0099	0.0086	0.0103	0.0096
	(200, 200)	0.0150	0.0098	0.0099	0.0109	0.0102
	(300, 100)	0.0138	0.0096	0.0106	0.0122	0.0109
(0.3, 0.05)	(100, 300)	0.2393	0.0133	0.0093	0.0116	0.0091
	(200, 200)	0.3272	0.0112	0.0119	0.0116	0.0115
	(300, 100)	0.2224	0.0073	0.0117	0.0127	0.0116
(0.1, −0.05)	(200, 200)	0.0959	0.0107	0.0111	0.0114	0.0111
	(200, 200)	0.0999	0.0119	0.0123	0.0127	0.0121
	(200, 200)	0.0937	0.0096	0.0101	0.0103	0.0100
(0.2, −0.05)	(100, 300)	0.2556	0.0153	0.0111	0.0132	0.0112
	(200, 200)	0.3506	0.0088	0.0092	0.0094	0.0091
	(300, 100)	0.2171	0.0064	0.0108	0.0124	0.0108

Table 4: Type I error rate. The vector of allele frequencies is $(p_1, p_2) = (0.3, 0.3)$ and significance level is $\alpha = 0.001$. X^2 is the Pearson's chi-square statistic.

		Statistic				
(F_1, F_2)	$(n_{1\cdot}, n_{2\cdot})$	X^2	G	S	W	Λ
(0.05, 0.05)	(100, 300)	0.0007	0.0003	0.0005	0.0006	0.0006
	(200, 200)	0.0005	0.0008	0.0008	0.0009	0.0008
	(300, 100)	0.0006	0.0010	0.0008	0.0010	0.0009
(0.1, 0.05)	(100, 300)	0.0016	0.0008	0.0007	0.0013	0.0009
	(200, 200)	0.0022	0.0010	0.0009	0.0013	0.0011
	(300, 100)	0.0030	0.0012	0.0014	0.0015	0.0012
(0.3, 0.05)	(100, 300)	0.0787	0.0020	0.0009	0.0019	0.0012
	(200, 200)	0.1224	0.0012	0.0015	0.0014	0.0014
	(300, 100)	0.0696	0.0002	0.0007	0.0011	0.0007
(0.1, −0.05)	(100, 300)	0.0206	0.0009	0.0009	0.0011	0.0009
	(200, 200)	0.0223	0.0008	0.0008	0.0009	0.0008
	(300, 100)	0.0211	0.0013	0.0014	0.0014	0.0014
(0.2, −0.05)	(100, 300)	0.0914	0.0029	0.0018	0.0025	0.0019
	(200, 200)	0.1348	0.0010	0.0011	0.0010	0.0011
	(300, 100)	0.0602	0.0005	0.0012	0.0021	0.0013

Table 5: Genotype counts at marker LCT–13910 (CC:CT:TT)

		All	Four US-born	Southeastern	Northwestern
Data	Tall	161:474:489	66:265:314	54:55:18	41:154:157
	Short	231:444:380	76:278:282	128:86:13	27:79:86
p-value	X^2	3.74E−06	0.263	0.00585	0.678
	G	1.42E−06	0.106	0.00188	0.717
	S	1.42E−06	0.106	0.00201	0.719
	W	1.31E−06	0.106	0.00248	0.720
	Λ	1.38E−06	0.106	0.00214	0.720