Weierstraß-Institut

für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

On an extended interpretation of linkage disequilibrium in genetic case-control association studies

Thorsten Dickhaus[1], Jens Stange[1], Haydar Demirhan[2]

submitted: October 30, 2014

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
E-Mail: Thorsten.Dickhaus@wias-berlin.de
Jens.Stange@wias-berlin.de

[2] Hacettepe University
Beytepe 06800
Ankara, Turkey
E-Mail: haydarde@hacettepe.edu.tr

No. 2029
Berlin 2014

2010 *Mathematics Subject Classification.* 62J15, 62P10, 62E20.

Key words and phrases. Asymptotic Gaussianity, chi-squared statistic, contingency table, correlation structure, Delta method, Fisher's exact test, odds ratio.

We thank Mette Langaas and Øyvind Bakke for some useful hints regarding Lemma A.1. This research was partly supported by the Deutsche Forschungsgemeinschaft via grant No. DI 1723/3-1 (Jens Stange).

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

We are concerned with statistical inference for $2 \times 2 \times K$ contingency tables in the context of genetic case-control association studies. Multivariate methods based on asymptotic Gaussianity of vectors of test statistics require information about the asymptotic correlation structure among these test statistics under the global null hypothesis. We show that for a wide variety of test statistics this asymptotic correlation structure is given by the linkage disequilibrium matrix of the K loci under investigation. Three popular choices of test statistics are discussed for illustration.

1 Introduction

Multivariate statistical methods based on asymptotic Gaussianity of test statistics are receiving more and more attention in the context of multiple test problems in genetics; see, e. g., Conneely and Boehnke (2007), Moskvina and Schmidt (2008), Dickhaus and Stange (2013), and Part II of Dickhaus (2014). The reason is that incorporating the (asymptotic) correlation structure of test statistics in the statistical analysis leads to an improvement of statistical power in comparison with a locus-by-locus analysis in combination with a (for instance, Bonferroni) correction for multiplicity. In genetics, correlations between the (expected) allele frequencies at genomic positions (loci) in the same chromosome are technically described by linkage disequilibrium (LD); see, for example, Chapter 10 of Ziegler and König (2006).

LD matrices for several target populations are publicly available from databases like those of The International HapMap Consortium (2005) or The 1000 Genomes Consortium (2010). Hence, they may be regarded as external structural information in the context of frequentist inference or as prior information in the context of Bayesian inference. In this work, we show that LD has a much broader interpretation when testing for association in $2 \times 2 \times K$ contingency tables occurring in genetic case-control studies, where the total number of loci under consideration is equal to K. Namely, under the null hypothesis of no associations, the LD coefficient of two loci i and j coincides with the asymptotic Pearson correlation coefficient of T_i and T_j (denoted by $\rho(T_i, T_j)$) for a wide variety of different test statistics T_i, T_j which are commonly used for testing association in the marginal contingency tables i and j, respectively.

The rest of the paper is structured as follows. In Section 2, we introduce basic notation and general assumptions. Section 3 contains our methodological results as well as three examples. We conclude with a discussion in Section 4. Some auxiliary results needed for Example 3.2 are deferred to the appendix.

Allele	A	a	$\sum$	B	b	$\sum$
Phenotype 1	$x_{11}^{(i)}$	$x_{12}^{(i)}$	$n_{1.}$	$x_{11}^{(j)}$	$x_{12}^{(j)}$	$n_{1.}$
Phenotype 0	$x_{21}^{(i)}$	$x_{22}^{(i)}$	$n_{2.}$	$x_{21}^{(j)}$	$x_{22}^{(j)}$	$n_{2.}$
Absolute count	${n_{.1}}^{(i)}$	${n_{.2}}^{(i)}$	N	${n_{.1}}^{(j)}$	${n_{.2}}^{(j)}$	N

Table 1: Schematic representation of data for an allelic association test problem at two genetic loci i and j.

2 Notation and preliminaries

Throughout the work, we let i and j denote two genomic positions in the same chromosome. We assume that the data collected for testing association between the allelic status of the respective locus and a given binary phenotype can be summarized in two contingency tables which are as in Table 1, where A (B) denotes the major allele at locus i (j) and a (b) the corresponding minor allele. The numbers $n_{1.}$ of cases (phenotype 1) and $n_{2.}$ of controls (phenotype 0) do not depend on the genomic position and are fixed by experimental design. Furthermore, we assume that all observational units have been sampled independently of each other from the same target population.

Notice that, conditional to all four marginal counts $n_{1.}$, $n_{2.}$, ${n_{.1}}^{(\gamma)}$, and ${n_{.2}}^{(\gamma)}$, the contingency table for locus γ can be reconstructed from $x_{11}^{(\gamma)}$ alone, $\gamma \in \{i, j\}$. Hence, conditionally to these marginal counts $X_{11}^{(\gamma)}$ is a (marginally) sufficient statistic for contingency table γ, where the capitalized notation indicates that the cell entry is regarded as a random variable. This is why essentially all (marginal) association tests which are commonly used in practice employ some transformation of $X_{11}^{(\gamma)}$ as a test statistic in contingency table γ. Letting f_γ denote such a (smooth) transformation, we will show in Section 3 that asymptotically ($N \to \infty$) the correlation coefficient of the test statistics $T_i = f_i(X_{11}^{(i)})$ and $T_j = f_j(X_{11}^{(j)})$ equals the LD coefficient $\mathrm{LD}(i, j)$ of loci i and j. This result has important consequences, because it enables one to carry out the multiple association test for all K loci simultaneously as a multivariate procedure which takes the asymptotic correlation structure among the locus-specific test statistics into account.

In all asymptotic considerations, we assume for convenience that

$$\lim_{N\to\infty} n_{1.}/N = \tau \in (0, 1).$$

3 Main results

Lemma 3.1. *Assume that the null hypothesis of no association between phenotype and allelic status is fulfilled at both loci i and j. Let $\gamma \in \{i, j\}$ and denote by p_γ the probability that a randomly chosen individual from the target population exhibits the major allele at locus γ, i. e., p_γ is the (expected) major allele frequency in the target population at locus γ. Finally, denote by p_{ij} the probability that a randomly chosen individual exhibits the major alleles at both loci i and j. Then, the following assertions hold true.*

(a) $X_{11}^{(\gamma)} \sim$ *Bin*$(n_{1.}, p_\gamma)$.

(b) *Cov*$(X_{11}^{(i)}, X_{11}^{(j)}) = n_{1.}\mathcal{D}_{AB}$, *where* $\mathcal{D}_{AB} = p_{ij} - p_i p_j$.

(c) $\rho(X_{11}^{(i)}, X_{11}^{(j)}) =$ *LD*$(i,j) = \frac{\mathcal{D}_{AB}}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}$.

(d) *Let*

$$\tilde{X}_{11}^{(\gamma)} = \sqrt{n_{1.}} \left(\frac{X_{11}^{(\gamma)}}{n_{1.}} - p_\gamma \right).$$

Then, the bivariate random vector $(\tilde{X}_{11}^{(i)}, \tilde{X}_{11}^{(j)})^\top$ *is asymptotically jointly normally distributed with zero expectation and covariance matrix*

$$\Sigma^* = \begin{pmatrix} p_i(1-p_i) & \mathcal{D}_{AB} \\ \mathcal{D}_{AB} & p_j(1-p_j) \end{pmatrix}. \tag{1}$$

Proof. Assertion (a) is obvious. For assertion (b), we employ the representation

$$X_{11}^{(\gamma)} = \sum_{k=1}^{n_{1.}} \mathbf{1}\{\text{Case } k \text{ exhibits the major allele at locus } \gamma\}.$$

This entails that $\mathbb{E}[X_{11}^{(i)} X_{11}^{(j)}] = n_{1.}p_{ij} + (n_{1.}^2 - n_{1.})p_i p_j$. Combining this with assertion (a) implies (b). Assertion (c) follows immediately from (a) and (b). Assertion (d) is an application of the binomial central limit theorem of de Moivre and Laplace in combination with the Cramér-Wold device. ■

Remark 3.1. *The statistic* $X_{11}^{(\gamma)}$ *is the test statistic employed by Fisher's exact test in contingency table* γ.

Theorem 3.1. *Let* $f = (f_i, f_j) : \mathbb{R}^2 \to \mathbb{R}^2$ *be a smooth transformation such that its Jacobian (matrix of partial derivaties)* ∇f*, evaluated at the point* $(\mathbb{E}[n_{1.}^{-1} X_{11}^{(i)}] = p_i, \mathbb{E}[n_{1.}^{-1} X_{11}^{(j)}] = p_j)$*, is a positive definite diagonal matrix. Then, the correlation coefficient among the two components of* $f(X_{11}^{(i)}/n_{1.}, X_{11}^{(j)}/n_{1.})$ *is asymptotically* $(N \to \infty)$ *equal to LD*(i,j).

Proof. We apply the bivariate Delta method in analogy to Section 3 of Wei and Higgins (2013). To this end, we let $\nabla f(u,v)$ denote the entry at position (u,v) of ∇f, evaluated at (p_i, p_j), where $1 \leq u, v \leq 2$, and let Σ stand for the asymptotic covariance matrix of the two components of $\sqrt{n_{1.}} f(X_{11}^{(i)}/n_{1.}, X_{11}^{(j)}/n_{1.})$, which asymptotically follow a bivariate normal distribution. Making use of the assumption regarding ∇f and of Σ^* from (1), the bivariate Delta method yields that

$$\begin{aligned} \Sigma &= \begin{pmatrix} \nabla f(1,1) & 0 \\ 0 & \nabla f(2,2) \end{pmatrix} \Sigma^* \begin{pmatrix} \nabla f(1,1) & 0 \\ 0 & \nabla f(2,2) \end{pmatrix} \\ &= \begin{pmatrix} \nabla f(1,1)\Sigma_{11}^* \nabla f(1,1) & \nabla f(1,1)\Sigma_{12}^* \nabla f(2,2) \\ \nabla f(1,1)\Sigma_{21}^* \nabla f(2,2) & \nabla f(2,2)\Sigma_{22}^* \nabla f(2,2) \end{pmatrix}. \end{aligned}$$

Hence, the correlation coefficient among the two components of $f(X_{11}^{(i)}/n_{1.}, X_{11}^{(j)}/n_{1.})$ is asymptotically $(N \to \infty)$ equal to

$$\frac{\nabla f(1,1)\mathcal{D}_{AB}\nabla f(2,2)}{\nabla f(1,1)\sqrt{\Sigma_{11}^*}\nabla f(2,2)\sqrt{\Sigma_{22}^*}} = \frac{\mathcal{D}_{AB}}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} = \mathsf{LD}(i,j).$$

■

Plainly phrased, the assertion of Theorem 3.1 means that the asymptotic correlation structure among the K marginal association test statistics is exactly given by the LD matrix among the K loci under consideration, provided that each test statistic T_γ is a smooth transformation of $X_{11}^{(\gamma)}$ only, without utilizing data from other loci.

Example 3.1 (Logarithmic odds ratios)**.** *One widely applied marginal test statistic is the (empirical) logarithmic odds ratio, say $\hat{\lambda}_\gamma$ for marginal contingency table γ. To avoid pathologies, assume that $x_{rc}^{(\gamma)} > 0$ for $1 \leq r, c \leq 2$ and $\gamma \in \{i, j\}$. In terms of $x_{11}^{(\gamma)}$, one can then write $\hat{\lambda}_\gamma = \log(x_{11}^{(\gamma)}) + \log(n_{.2}^{(\gamma)} - n_{1.} + x_{11}^{(\gamma)}) - \log(n_{1.} - x_{11}^{(\gamma)}) - \log(n_{.1}^{(\gamma)} - x_{11}^{(\gamma)}) = f_\gamma(x_{11}^{(\gamma)})$. Considering the so-defined function $f = (f_i, f_j)$, where we artificially include $x_{11}^{(j)}$ ($x_{11}^{(i)}$) as a second argument to f_i (f_j), it is straightforward to check that the assumptions of Theorem 3.1 are fulfilled. Hence, we have that $\rho(\hat{\lambda}_i, \hat{\lambda}_j)$ is asymptotically equal to $\mathsf{LD}(i,j)$. The application of the univariate Delta method to prove asymptotic Gaussianity of $\hat{\lambda}_\gamma$ is mentioned in Section 3.1.7 of Agresti (2002).*

We may remark here that the exact finite-sample correlation coefficient of $\hat{\lambda}_i$ and $\hat{\lambda}_j$ has been derived by Bagos (2012). As a sanity check for our asymptotic result, we derived Figure 1. The data for this figure have been taken from a genetic association study regarding a (dichotomized) behavioral measure of impulsiveness (yet unpublished data), consisting of $n_{1.} = 299$ cases (highly impulsive individuals) and $n_{2.} = 2430$ controls. The data points in Figure 1 correspond to $K = 10$ correlated genomic loci, leading to $\binom{10}{2} = 45$ pairwise LD coefficients. Although it is not guaranteed that the global null hypothesis of no genetic association with impulsiveness holds for the $K = 10$ loci displayed in Figure 1, the 10 estimated odds ratios were very close to 1, such that this assumption seemed justified. Qualitatively, the obvious agreement of abscissas and ordinates in Figure 1 has been confirmed by many other analogous graphs which we omit here.

Example 3.2 (Chi-squared statistics)**.** *Let $e_{rc}^{(\gamma)} = n_{r.}n_{.c}^{(\gamma)}/N$, $1 \leq r, c \leq 2$, and denote by*

$$Q^{(\gamma)} = \sum_{r=1}^{2}\sum_{c=1}^{2} \frac{(X_{rc}^{(\gamma)} - E_{rc}^{(\gamma)})^2}{E_{rc}^{(\gamma)}}$$

the chi-squared statistic for testing association in contingency table γ. It is well-known that $Q^{(\gamma)}$ is asymptotically chi-square distributed under the null with one degree of freedom.

Theorem 3.1 is not directly applicable in this case, because the representation (2) *given in Lemma A.1 below shows that the assumption of a positive definite Jacobian is violated here (diagonal elements of ∇f are equal to zero). However, Lemma 3.1 in combination with Lemma A.2 below yields that the correlation coefficient between $Q^{(i)}$ and $Q^{(j)}$ is asymptotically given by $\mathsf{LD}^2(i,j)$. We verify this result in Figure 2, in analogy to Figure 1.*

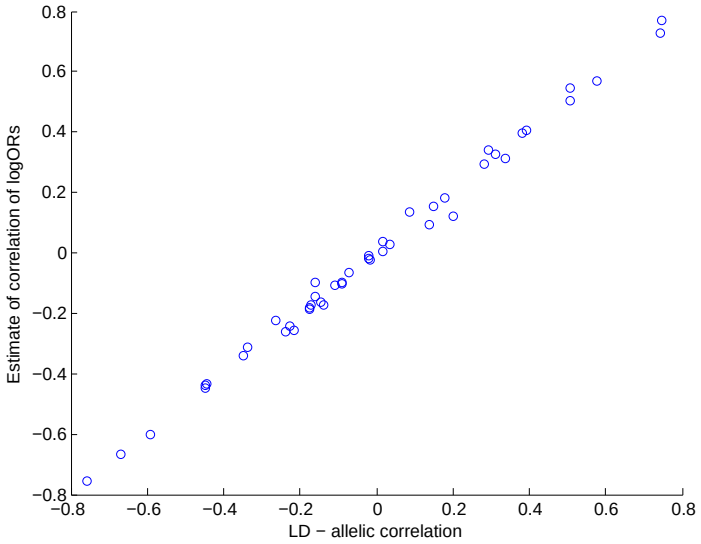

Figure 1: Empirical comparison of LD(i, j) and $\rho(\hat{\lambda}_i, \hat{\lambda}_j)$ on the basis of $K = 10$ correlated genomic loci, leading to 45 pairwise data points. The abcissas are given by pairwise LD coefficients, while the ordinates have been calculated by the exact finite-sample formulas by Bagos (2012). The sample consisted of $n_{1.} = 299$ cases and $n_{2.} = 2430$ controls.

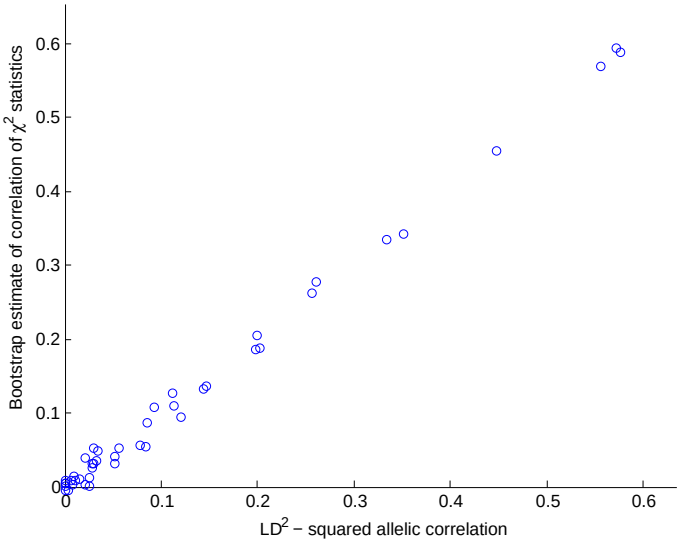

Figure 2: Empirical comparison of $\mathrm{LD}^2(i,j)$ and $\rho(Q^{(i)}, Q^{(j)})$ on the basis of $K = 10$ correlated genomic loci, leading to 45 pairwise data points. The abcissas are given by pairwise squared LD coefficients, while the ordinates have been calculated by nonparametric bootstrap. The sample consisted of $n_{1\cdot} = 299$ cases and $n_{2\cdot} = 2430$ controls.

4 Discussion

We have drawn a connection between the correlation structure among (marginal) test statistics for association in $2 \times 2 \times K$ contingency tables and the $(K \times K)$ LD matrix in genetic case-control studies. Asymptotically, these two quantities coincide under the global null hypothesis. This result can be exploited for multivariate statistical inference, because external LD information can be used to approximate the correlation matrix of the locus-specific test statistics, provided that the sample size N is large.

One concrete application of our Example 3.1, which we want to work out in a detailed manner in future research, consists in Bayesian inference for $2 \times 2 \times K$ contingency tables based on the K-dimensional vector of logarithmic odds ratios as considered by Demirhan and Hamurkaroglu (2008). Here, the LD matrix can be used as an informative prior for the correlation structure among the logarithmic odds ratios (which are regarded as random objects in the Bayesian setup).

As a possible extension of our work, the Delta method can also be employed to work out the asymptotic correlation structure of test statistics in $2 \times C \times K$ tables for $C > 2$. The special case of $C = 3$ is relevant in association studies if the locus-specific diploid allele pairs are considered instead of the mere alleles. Dickhaus and Stange (2013) referred to this setup as a multiple genotypic association test problem and derived the asymptotic correlation structure of the K marginal chi-squared statistics in that case; cf. the discussion around their Definition 4.2 and Lemma 4.2. In the case of a (2×3)-contingency table at each locus γ, the bivariate vector $(X_{11}^{(\gamma)}, X_{12}^{(\gamma)})^\top$ is sufficient conditional to all marginals, and a multinomial central limit theorem holds for $(X_{11}^{(\gamma)}, X_{12}^{(\gamma)})^\top$. Hence, considering two such loci i and j, the four-variate Delta method can be applied.

A Auxiliary results

Lemma A.1. *Let* $\mathbf{x} = \begin{pmatrix} x_{11} & \dots & x_{1C} \\ x_{21} & \dots & x_{2C} \end{pmatrix}$ *denote a* $(2 \times C)$*-contingency table with row marginals* $n_{1.}$*,* $n_{2.}$*, column marginals* $n_{.1}, \dots, n_{.C}$*, and total sample size* $N = n_{1.} + n_{2.}$*. Define* $e_{rc} = n_{r.} n_{.c} / N$ *for* $1 \leq r \leq 2$ *and* $1 \leq c \leq C$*, and let*

$$Q(\mathbf{x}) = \sum_{r=1}^{2} \sum_{c=1}^{C} \frac{(x_{rc} - e_{rc})^2}{e_{rc}}$$

denote the value of the chi-squared statistic for testing association based on $\mathbf{x}$*. Then the following two assertions hold true.*

(a)

$$Q(\mathbf{x}) = \frac{N}{n_{2.}} \sum_{c=1}^{C} \frac{(x_{1c} - e_{1c})^2}{e_{1c}}.$$

(b) In the special case of $C = 2$, it holds that

$$Q(\mathbf{x}) = \left[\sqrt{\frac{N}{n_{2\cdot}}} \frac{(x_{11} - e_{11})}{\sqrt{n_{1\cdot}\hat{p}(1-\hat{p})}} \right]^2, \tag{2}$$

where $\hat{p} = n_{\cdot 1}/N$ denotes the empirical major allele frequency.

Proof. For proving part (a), it suffices to show that for any given column $1 \leq c \leq C$, we have

$$\frac{(x_{1c} - e_{1c})^2}{e_{1c}} + \frac{(x_{2c} - e_{2c})^2}{e_{2c}} = \frac{N}{n_{2\cdot}} \frac{(x_{1c} - e_{1c})^2}{e_{1c}},$$

which can be verified by elementary calculations. Part (b) follows from part (a) and by noticing that $(x_{12} - e_{12})^2 = (x_{11} - e_{11})^2$. ■

Lemma A.2. *Let Z_1, Z_2 be two jointly normally distributed random variables with $\mathbb{E}[Z_1] = \mathbb{E}[Z_2] = 0$, Var$(Z_1) =$ Var$(Z_2) = 1$, and correlation coefficient $\rho(Z_1, Z_2) = \rho$. Then it holds that Cov$(Z_1^2, Z_2^2) = 2\rho^2$ and, consequently, $\rho(Z_1^2, Z_2^2) = \rho^2$.*

Proof. Let $(X_0, X_1, X_2)^\top \sim \mathcal{N}_3(0, I_3)$ and notice that

$$(Z_1, Z_2)^\top \stackrel{\mathsf{d}}{=} (\sqrt{1-\rho}X_1 + \sqrt{\rho}X_0, \sqrt{1-\rho}X_2 + \sqrt{\rho}X_0)^\top.$$

Elementary probabilistic calculus now yields that Cov$(Z_1^2, Z_2^2) = \rho^2$Var(X_0^2), which implies the assertion in view of Var$(X_0^2) = 2$. ■

References

Agresti, A., 2002. Categorical data analysis. 2nd ed. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Chichester: Wiley.

Bagos, P. G., Jun 2012. On the covariance of two correlated log-odds ratios. Stat. Med. 31 (14), 1418–1431.

Conneely, K. N., Boehnke, M., Dec 2007. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. Am. J. Hum. Genet. 81 (6), 1158–1168.

Demirhan, H., Hamurkaroglu, C., 2008. Bayesian estimation of log odds ratios from $R \times C$ and $2 \times 2 \times K$ contingency tables. Statist. Neerlandica 62 (4), 405–424.
URL http://dx.doi.org/10.1111/j.1467-9574.2008.00387.x

Dickhaus, T., 2014. Simultaneous Statistical Inference with Applications in the Life Sciences. Springer-Verlag Berlin Heidelberg.

Dickhaus, T., Stange, J., 2013. Multiple point hypothesis test problems and effective numbers of tests for control of the family-wise error rate. Calcutta Statistical Association Bulletin 65 (257-260), 123–144.

Moskvina, V., Schmidt, K. M., 2008. On multiple-testing correction in genome-wide association studies. Genetic Epidemiology 32, 567–573.

The 1000 Genomes Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

The International HapMap Consortium, Oct 2005. A haplotype map of the human genome. Nature 437 (7063), 1299–1320.

Wei, Y., Higgins, J. P., Mar 2013. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. Stat. Med. 32 (7), 1191–1205.

Ziegler, A., König, I. R., 2006. A Statistical Approach to Genetic Epidemiology. Wiley, Weinheim.