Original Paper

Human Heredity

Hum Hered 2012;73:14–17
DOI: 10.1159/000334085

Received: June 6, 2011
Accepted: September 24, 2011
Published online: December 30, 2011

Cochran-Armitage Test versus Logistic Regression in the Analysis of Genetic Association Studies

Stefan Wellek[a, b] Andreas Ziegler[c]

[a]Department of Biostatistics, Central Institute of Mental Health Mannheim, University of Heidelberg, Mannheim, [b]Department of Medical Biometry, Epidemiology and Informatics, University of Mainz, Mainz, and [c]Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany

Key Words
Logistic regression · Score statistic · Point and interval estimation · Wald statistic

Abstract
Objective: The Cochran-Armitage trend test based on the linear regression model has become a standard procedure for association testing in case-control studies. In contrast, the logistic regression model is generally used for estimating effect sizes. The aim of this paper is to propose an approach that allows for association testing and parameter estimation by means of the same statistic. ***Methods/Results:*** The trend test is recommendable as a test of no association between genotype and risk of disease. It is a two-sample test for differences between cases and controls with respect to the average number of risk alleles occurring in the genotype of an individual. We argue that this difference is not of primary interest in genetic association studies. It should be replaced with the disease odds ratio, which can be assessed under both cohort sampling and case-control sampling. ***Conclusion:*** The Cochran-Armitage trend test should be replaced by the Wald statistic from a logistic regression model for hypothesis testing and estimation in genetic association studies.

Introduction

Since the seminal paper by Sasieni [1], the trend test named after Cochran and Armitage (CA) has become a standard procedure for the confirmatory statistical analysis of genetic association studies following the case-control design. In the most common setting of a binary phenotype and a single nucleotide polymorphism (SNP), the trend test is a two-sample test for differences between cases and controls with respect to the average number of risk alleles occurring in the genotype of an individual. Here, we show that this difference is not of primary interest in genetic association studies. For more elaborate analyses involving estimation of a parameter of interest and subsequent computation of confidence limits, we argue that the difference should be replaced with the disease odds ratio, which can be assessed under both cohort and case-control sampling. We will point out that the natural basis of drawing statistical inferences about this latter parameter is logistic regression analysis by means of likelihood-based principles and techniques.

KARGER

E-Mail karger@karger.ch
www.karger.com/hhe

0001–5652/12/0731–0014$38.00/0

Prof. Dr. Andreas Ziegler
Institute of Medical Biometry and Statistics, University of Lübeck
University Hospital Schleswig-Holstein, Campus Lübeck
Maria-Goeppert-Strasse 1, DE–23562 Lübeck (Germany)
Tel. +49 451 500 2780, E-Mail ziegler@imbs.uni-luebeck.de

Table 1. Observed genotype frequencies and theoretical probabilities (in parentheses) for cases and controls at a diallelic marker with alleles a and A

Phenotype	Genotype			
	aa	aA	AA	Σ
Case	r_0 (p_{10})	r_1 (p_{11})	r_2 (p_{12})	r ($p_{1.}$)
Control	s_0 (p_{20})	s_1 (p_{21})	s_2 (p_{22})	s ($p_{2.}$)
Σ	n_0 ($p_{.0}$)	n_1 ($p_{.1}$)	n_2 ($p_{.2}$)	n (1.0)

Methods

Using the entries of the 2 × 3 contingency table of table 1, the CA trend test statistic is given by:

$$T_{CA} = \sqrt{\frac{n}{r(n-r)}}\,\frac{n(r_1+2r_2)-r(n_1+2n_2)}{\sqrt{n(n_1+4n_2)-(n_1+2n_2)^2}}. \quad (1)$$

Starting from equation 1, the CA statistic can be rewritten in a number of ways. Out of these equivalent representations, the following two are particularly illuminating for the present considerations. The algebraic identity of the respective expressions to the right-hand side of equation 1 is shown in the Appendix.

(I) Let X_k and Y_l denote the number of A alleles contained in the genotype of the k-th member of the group of cases ($k = 1, ..., r$) and the l-th control ($l = 1, ..., s$), respectively. Then, the CA statistic T_{CA} can be written as:

$$T_{CA} = \frac{\overline{X}-\overline{Y}}{\left[\widehat{Var}_0\left(\overline{X}-\overline{Y}\right)\right]^{1/2}}. \quad (2)$$

In equation 2, the estimator of the variance of the observed difference between the mean allele counts has to be computed treating the entries in both rows of table 1 as observations from two trinomial distributions whose parameters $\pi_{1j} = p_{1j}/p_{1.}$ and $\pi_{2j} = p_{2j}/p_{2.}$ satisfy the null hypothesis H_0: $\pi_{1j} = \pi_{2j}$ for all j. Accordingly, the CA test coincides with the usual (asymptotic) test for equality of the means of two distributions with common variance from which independent random samples have been taken, as applied to the special case where the range of all individual random variables representing the sample values is restricted to {0, 1, 2}.

(II) Let $\hat{\pi}_{1j}$ and $\hat{\pi}_{2j}$ denote the relative frequencies of genotypes obtained in the sample of cases and controls, respectively, so that $\hat{\pi}_{1j} = r_j/r$ and $\hat{\pi}_{2j} = s_j/s$. Equation 1 can then be rewritten as:

$$T_{CA} = 2\,\frac{\left(\hat{\pi}_{11}/2+\hat{\pi}_{12}\right)-\left(\hat{\pi}_{21}/2+\hat{\pi}_{22}\right)}{\sqrt{\frac{n}{rs}\left(\left(\hat{p}_{.1}+4\hat{p}_{.2}\right)-\left(\hat{p}_{.1}+2\hat{p}_{.2}\right)^2\right)}}. \quad (3)$$

In the denominator of equation 3, $\hat{p}_{.j}$ stands for the pooled estimate of the frequency of the j-th genotype given by $\hat{p}_{.j} = n_j/n$ for $j = 0, 1, 2$. Furthermore, $\hat{\pi}_{11}/2 + \hat{\pi}_{12}$ and $\hat{\pi}_{21}/2 + \hat{\pi}_{22}$ are well known [see, e.g. 2, § 2.4] to be the standard estimates of the frequency of allele A among cases and controls, respectively. Thus, equation 3 shows that the CA procedure tests for differences of the two populations sampled in a case-control study with respect to the frequency of the allele of interest.

(III) From the facts stated under (I) and (II), it does not become obvious why the CA test is an appropriate inferential procedure for assessing the association between the SNP under consideration and the risk of having or developing a disease. In fact, the goal of a genetic association study is to make inferences about the penetrance

$$\delta_j = P(D = 1 \mid J = j) \quad (4)$$

associated with the j-th genotype, i.e. the conditional probability that a randomly selected individual is diseased given the number j of risk alleles. The standard approach to modeling this function uses the log odds scale and is based on the assumption that the increase in disease risk is independent of the baseline value of j. This condition is satisfied if, and only if, there are constants α and β, say, such that:

$$\text{logit}(\delta_j) = \alpha + \beta \times j. \quad (5)$$

This logistic regression model arises under cohort rather than case-control sampling. However, likelihood inferences can also be made about all parameters except for the intercept in the case-control design and are the same under both sampling schemes [3, 4].

One of the standard likelihood-based tests of the null hypothesis H_0: $\beta = 0$ for any logistic regression model is the score test [5, § 6.3.2], and it is fairly easy to show [6] that in the case under consideration (equation 5), the score statistic is algebraically identical to the CA statistic in equation 2.

Results and Discussion

The key conclusions to be drawn from properties (I) to (III) are as follows: first, the CA test is recommendable as a test of the null hypothesis H_0 of no association between genotype and risk of disease. Actually, when used for that purpose, the CA statistic is asymptotically equivalent both to Wald's maximum likelihood statistic and the likelihood-ratio statistic [see, e.g. 7, § 6e.3]. Second, the CA statistic cannot be used for constructing confidence intervals for the odds ratio

$$\frac{\delta_{j+1}}{\left(1-\delta_{j+1}\right)} \Big/ \frac{\delta_j}{\left(1-\delta_j\right)}.$$

These facts are reflected in the common practice of reporting the results of genetic association studies even in our own work [8]: the presentation of p values obtained by means of the CA statistic is supplemented by giving estimates (point and interval) taken from standard logistic regression. This practice is, however, not sound be-

cause in the analysis of any study, the point estimate and the confidence interval should be based on the same statistic as the significance test. Actually, most studies on the use of confidence intervals or p value functions in addition to significance tests [see 9–16] were written with the understanding that both parts of the analysis of a given data set rely on the same statistic.

The aim of the final part of this article is to illustrate that this kind of consistency can easily be attained by replacing the CA statistic with Wald's maximum likelihood statistic for testing arbitrary hypotheses about the log odds ratio parameter β, specifying that $\beta = \beta_0$ for any fixed real number β_0, including $\beta_0 = 0$ as a special case. We note that Wald's testing procedure consumes slightly more computer processor time than the CA trend test because the maximum likelihood estimator has to be determined numerically by an iterative algorithm. However, despite the fact that not a few practitioners in genetic epidemiology feel to the contrary, it is by no means true that performing Wald's test in the analysis of a data set of the complexity encountered in a contemporary genome-wide association study is computationally infeasible. Using an implementation for logistic regression analysis restricted to the case of a single categorical covariate, computing all quantities involved in a complete analysis based on Wald's statistic for 1,000,000 SNPs takes about 70 min of execution time on a standard Intel PC of the present generation. We measured this value with a SAS IML function which we will be happy to provide to any interested reader. Translating the code in any other sufficiently rich programming language like *R* is a straightforward exercise so that from a computational perspective, there is no compelling reason to prefer CA-based over Wald-type logistic analysis.

Illustration

In order to illustrate the differences in the inferential scope of both approaches, we re-analyze data from an association study of APOE promoter polymorphisms with Alzheimer's disease (table 2) [17].

Analysis Based on the Score Statistic

Determining the mean difference of the number of T alleles counted in cases as compared with controls yields $\bar{X} - \bar{Y} = 0.2370$, the score statistic is computed to be $T_{CA} = 2.8920$, and the corresponding two-sided asymptotic p value is 0.0038. The confidence interval derived from the two-sided test refers to the parameter $E(\bar{X} - \bar{Y}) = 2(p_A^{(1)} - p_A^{(2)})$, with $p_A^{(1)}$ and $p_A^{(2)}$ denoting the frequency of allele A in the population of cases and controls, respectively, and has limits (0.0763, 0.3977).

Table 2. Observed genotype frequencies for the SNP –219 G>T (rs405509) in a case-control study of genetic risk factors for Alzheimer's disease [17]

Affection status	Number of T alleles			
	0	1	2	Σ
Case	31 (0.2844)	49 (0.4495)	29 (0.2661)	109 (1.0)
Control	67 (0.3564)	102 (0.5426)	19 (0.1011)	188 (1.0)
Σ	98	151	48	297

Analysis Based on the Wald-Type Statistic

The unconstrained ML estimator is computed to be $\hat{\beta} = 0.5175$, and the corresponding Wald statistic is 2.8595, yielding a two-sided asymptotic p value of 0.0041. This p value is slightly larger than its counterpart based on the score statistic. However, what really matters is the change in the interpretation of the confidence interval obtained using the observed value of the Wald test statistic T_W as a pivot. It holds for the common value e^{β} of the disease odds ratios associated with increasing the number of risk alleles by 1. At the nominal confidence level 95%, its limits turn out to be 1.1770 and 2.392 for the present data set.

Conclusion

All in all, we feel that there is a real need for revising the current practice of analyzing standard genetic association studies by means of two different statistics, of which the first one is used for computing a p value and the other one for doing interval estimations. A unified procedure based on Wald's maximum likelihood statistic seems to be the better choice – the more as computational feasibility of the latter approach is warranted even in the context of very large studies.

Appendix: Proof of the Equivalence of Equations 1, 2 and 3

Comparison of equations 1 and 2: by definition, X_k and Y_l count the number of A alleles observed in the k-th case and the l-th control, respectively. In view of the basic notation introduced in table 1, this implies:

$$\bar{X} - \bar{Y} = (r_1 + 2r_2)/r - (s_1 + 2s_2)/s. \quad \text{(A1)}$$

Furthermore, (r_0, r_1, r_2) and (s_0, s_1, s_2) are independent trinomially distributed vectors with totals r and s and both parameter vectors being equal to $(p_{.0}, p_{.1}, p_{.2})$ under H_0. From the basic properties of the multinomial family of distributions, it thus follows that we can write:

$$Var_0(\bar{X} - \bar{Y}) = (1/r + 1/s)(p_{.1}(1 - p_{.1}) + 4p_{.2}(1 - p_{.2}) - 4p_{.1}p_{.2}) \\ = (1/r + 1/s)(p_{.1} + 4p_{.2} - (p_{.1} + 2p_{.2})^2). \quad \text{(A2)}$$

Replacing $p_{.1}$ and $p_{.2}$ with the homologous sample proportions yields:

$$\widehat{Var}_0\left(\bar{X} - \bar{Y}\right) = \left(1/r + 1/s\right)\left(\left(n_1 + 4n_2\right)/n - \left(\left(n_1 + 2n_2\right)/n\right)^2\right). \quad \text{(A3)}$$

From equations A1 and A3, equality of equations 2 to 1 follows by elementary algebra, making use of the identities $r + s = n$, $r_j + s_j = n_j$ $(j = 1, 2)$.

Comparison of equations 1 and 3: that the denominator of equation 3 equals that of equation 1 (except for a multiplicative constant) is immediately obvious from the identities $\hat{p}_{.1} = n_1/n$, $\hat{p}_{.2} = n_2/n$, both of which hold true by definition. By the definition of π_{1j} and π_{2j}, we have:

$$(\hat{\pi}_{11}/2 + \hat{\pi}_{12}) - (\hat{\pi}_{21}/2 + \hat{\pi}_{22}) = (r_1/2 + r_2)/r - (s_1/2 + s_2)/s \\ = (1/rs)(s(r_1/2 + r_2) - r(s_1/2 + s_2)). \quad \text{(A4)}$$

In view of $r + s = n$, $r_j + s_j = n_j$ $(j = 1, 2)$, the last parenthesized expression in equation A4 is the same as $(n(r_1/2 + r_2) - r(n_1/2 + n_2))$. The rest follows through cancelling some common factors from the numerator and denominator of the fraction of equation 3 as rewritten by means of these identities.

Acknowledgement

This work was supported by the German Ministry of Education and Research (0315536F).

References

1 Sasieni PD: From genotypes to genes: doubling the sample sizes. Biometrics 1997;53:1253–1261.
2 Ziegler A, König IR: A Statistical Approach to Genetic Epidemiology: Concepts and Applications, ed 2. Weinheim, Wiley-VCH, 2010.
3 Anderson JA: Separate sample logistic discrimination. Biometrika 1972;59:19–35.
4 Prentice RL, Pyke R: Logistic disease incidence models and case-control studies. Biometrika 1979;66:403–411.
5 Bickel PJ, Doksum KA: Mathematical Statistics: Basic Ideas and Selected Topics, ed 2, vol 1. Upper Saddle River, Prentice Hall, 2001.
6 Williams PL: Trend test for counts and proportions; in Armitage P, Colton T (eds): Encyclopedia of Biostatistics. Chichester, John Wiley & Sons, 1998, vol 6, pp 4573–4584.
7 Rao CR: Linear Statistical Inference and Its Applications, ed 2. New York, John Wiley & Sons, 1973.
8 Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, König IR, Stevens S, Szymczak S, Trégouët DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Brænne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H, WTCCC and the Cardiogenics Consortium: Genomewide association analysis of coronary artery disease. N Engl J Med 2007;357:443–453.
9 Altman D, Machin D, Bryant T, Gardner S: Statistics with Confidence, ed 2. London, Wiley-Blackwell, 2000.
10 Birnbaum A: A unified theory of estimation, I. Ann Math Stat 1961;32:112–135.
11 Burton PR: Helping doctors to draw appropriate inferences from the analysis of medical studies. Stat Med 1994;13:1699–1713.
12 Greenland S: Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 1987;9:1–30.
13 Poole C: Beyond the confidence interval. Am J Public Health 1987;77:195–199.
14 Savitz DA, Tolo KA, Poole C: Statistical significance testing in the American Journal of Epidemiology, 1970–1990. Am J Epidemiol 1994;139:1047–1052.
15 Simon R: Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986;105:429–435.
16 Witte JS, Thomas DC, Langholz B: Re: 'Statistical significance testing in the American Journal of Epidemiology, 1970–1990.' Am J Epidemiol 1995;142:101.
17 Lambert JC, Araria-Goumidi L, Myllykangas L, Ellis C, Wang JC, Bullido MJ, Harris JM, Artiga MJ, Hernandez D, Kwon JM, Frigard B, Petersen RC, Cumming AM, Pasquier F, Sastre I, Tienari PJ, Frank A, Sulkava R, Morris JC, St. Clair D, Mann K, Wavrant-DeVrièze F, Ezquerra-Trabalon M, Amouyel P, Hardy J, Haltia M, Valdivieso F, Goate AM, Pérez-Tur J, Lendon CL, Chartier-Harlin MC: Contribution of APOE promoter polymorphisms to Alzheimer's disease risk. Neurology 2002;59:59–66.