Meta Gene 9 (2016) 225–229

Contents lists available at ScienceDirect

Meta Gene

journal homepage: www.elsevier.com/locate/mgene

An enhanced version of Cochran-Armitage trend test for genome-wide association studies

Mansi Ghodsi [a], Saeid Amiri [b], Hossein Hassani [c,*], Zara Ghodsi [a]

[a] Translational Genetics Group, Bournemouth University, UK
[b] University of Wisconsin-Green Bay, Department of Natural and Applied Sciences, Green Bay, WI, USA
[c] Institute for International Energy Studies, Tehran, 1967743 711, Iran

ARTICLE INFO

Article history:
Received 2 April 2016
Revised 30 June 2016
Accepted 1 July 2016
Available online 22 July 2016

Keywords:
Bootstrap method
Monte Carlo simulation
Chi-squared test
Contingency table
Genetic association
p-values

ABSTRACT

Genome-wide association studies the evaluation of association between candidate gene and disease status is widely carried out using Cochran-Armitage trend test. However, only a small number of research papers have evaluated the distribution of p-values for the Cochran-Armitage trend test. In this paper, an enhanced version of Cochran-Armitage trend test based on bootstrap approach is introduced. The achieved results confirm that the distribution of p-values of the proposed approach fits better to the uniform distribution, and it is thus concluded that the proposed method, which needs less assumptions in comparison with the conventional method, can be successfully used to test the genetic association.

1. Introduction

A central goal of genome wide association studies (GWAS) is to identify genetic risk factors for complex disorders. In order to find the disease genetic risk factors in a population, GWAS measures DNA sequence variations across human genome (Bush and Moore, 2012). Practitioners in medical sciences and bioinformatics use GWAS to investigate the relations in different disorders; GWAS of different cancers (Easton and Eeles, 2008), GWAS of pancreatic cancer (Amundadottir et al., 2009). The idea of genetic variations with alleles that are common in the population may explain much of the heritability of common diseases, see (Reich and Lander, 2001) and (Schork et al., 2009). Review of GWAS can be found in several texts and papers, see (Moore et al., 2010) among others.

In the simplest form of association mapping, a set of markers are genotyped in both sample of cases and sample of unrelated controls and then using different association tests, allele frequency differences or genotype frequency differences at each marker will be studied (Pritchard and Donnelly, 2001). The main idea behind GWAS studies relies on the fact that if a mutation has positive correlation with susceptibility of a disease, then that mutation is expected to be more frequent among affected individuals than those unaffected individuals (Pritchard and Donnelly, 2001). Hence, considering the existence of linkage disequilibrium (LD) between the marker locus and the susceptibility mutation, the marker close to the disease mutation may also present a frequency difference between case and control group of study (Pritchard and Donnelly, 2001).

Case-control traits can be analysed using either logistic regression or contingency table techniques (Bush and Moore, 2012). Contingency table methods examine the deviation from independence that is expected under the null hypothesis of observing no association between the disease under study and the measured allelic/genotyping frequency differences (Bush and Moore, 2012). Pearson chi-squared test and the related Fisher's exact test are the most widely used tests for independence of the rows and columns of the contingency table (Bush and Moore, 2012).

It should be noted that the association tests are performed separately for each individual marker and depending on the aim of study, the data for each marker with minor allele *a* and major allele *A* can be represented either as genotype count (e.g., *a*/*a*, *A*/*a* and *A*/*A*) or allele count (e.g., *a* and *A*) (Clarke et al., 2011). It is widely believed that the allelic association test with 1 degrees of freedom (df) is more reliable than the genotypic test with 2 df. However, it is imperative to note that this superior performance can only be considered for the case of having the penetrance of the heterozygote genotype between the penetrance of the two homozygote genotypes (Clarke et al., 2011). When the distribution of genotypes in the population deviates from Hardy-Weinberg proportions (HWE), of which additive, dominant and recessive models are all examples (Clarke et al., 2011), the frequency of genotypes rather than alleles should be compared by the Cochran-Armitage test for trend

* Correspoding author.
E-mail address: hassani.stat@gmail.com (H. Hassani).

http://dx.doi.org/10.1016/j.mgene.2016.07.001
2214-5400/

(Sasieni, 1997). For more information on different models see (Clarke et al., 2011).

Thus, the advantage of the Cochran-Armitage trend test in comparison to Pearson's Chi-Square test is that it possesses the superior conservation and is not dependent on the HWE assumption (Sasieni, 1997). Therefore a number of authors have recommended to use the Cochran-Armitage trend test as the genotype-based test for association (Sasieni, 1997; Corcoran et al., 2000; Li, 2008; Risch and Merikangas, 1996; Risch, 2000). It should also be noted that the allelic and trend statistic are equivalent when the combined sample is in HWE (Sasieni, 1997).

However, a major drawback of model based methods is that the statistical properties depend on the choice of weights. Thus, the model miss-specifications minimize the power of the test (Sasieni, 1997; Corcoran et al., 2000; Li, 2008; Risch and Merikangas, 1996; Risch, 2000). Furthermore, Escott-Price et al. (2013) showed that, although in most scenarios the Cochran-Armitage trend test is more powerful than the chi-squared test of genotype counts, the advantage is not substantial. Even, when the disease locus is extremely biased from the additive model, the chi-squared test of genotype counts can be more powerful than the Cochran-Armitage trend test due to the choice of scores for each genotype in the trend test (Escott-Price et al., 2013).

Although, there are considerable studies about the advantages and disadvantages of Cochran-Armitage trend test, to the best of our knowledge, there is a small number of researches which evaluated the distribution of p-values for this association test. In this paper the distribution of the p-values derived by the Cochran-Armitage trend test has been studied and it has been shown that unlike the considered presumption those p-values obtained by this test are not uniformly distributed. To overcome this issue, we introduce a new method, based on the bootstrap technique, for computing the p-value of the Cochran-Armitage trend test.

The bootstrap method has become a standard tool in statistical analysis and is an indispensable tool for testing statistical hypotheses. Using resampling, bootstrap approximates the sampling distribution of a statistic under the null (or the alternative) hypothesis. Bootstrap provides a practical complement to asymptotic parametric inference, hence have attracted many attentions in the applied. The efficiency of the nonparametric bootstrap method has also been shown by Amiri and von Rosen (2011) in which for example in the case of the Pearson chi-squared statistic with a Yates' correction and Fisher's exact test, remarkable improvement has been achieved. The Pearson chi-squared statistic with a Yates' correction and Fisher's exact test, are quite conservative and fail to reject the null hypothesis and can not be recommended to test independence with small sample sizes.

The remainder of this paper is organized as follows. The concept of Cochran-Armitage trend test is explained in Section 2. Section 3 studies the alternative approach to draw the inference including the bootstrap version of Cochran-Armitage trend test. Section 4 investigates the proposed method using the Monte Carlo simulation, which show they are the accurate tests in terms of the significant level and statistical power. Section 4 also demonstrates the improvements in goodness-of-fitness achieved by the introduced bootstrap approach. The paper concludes with a concise summary in Section 5.

2. Cochran-Armitage trend test

The Cochran-Armitage's trend test is a widely used test for trend among binomial proportions which uses the genotype contingency table (Table 1) in a different manner than Pearson's test. Power is very often improved as long as the probability of having disease increases with the number of disease-associated alleles. In genetic association studies in which the underlying genetic model is unknown, the additive version of this test is most commonly used. In order to measure the effect of genotype i and to detect particular types of association, we introduce a weight w_i. The special choice $(w_0,w_1,w_2)=(0,1,2)$, represents the additive effect of allele A. (See Table 2.)

Table 1
Genotype counts distribution for the case-control studies.

	$w_0=0$	$w_1=1$	$w_2=2$	Total
Case	n_0	n_1	n_2	n
Control	m_0	m_1	m_2	m
Total	N_0	N_1	N_2	N

Let us consider a single-marker locus with two possible alleles which are commonly denoted by A and a. Thus, each individual has three possible genotypes AA, Aa, and aa. In the following we denote the two alleles by 0 and 1 instead of A and a and the genotypes by 0, 1, 2, the sum of the two allele indices involved. We assume a random sample of n cases and m unrelated controls. The case-control data can then be summarized according to genotypes as shown in Table 1.

Here, (n_0,n_1,n_2) are counts of the genotypes in cases and (m_0,m_1,m_2) are counts of the genotypes in controls, and (N_0,N_1,N_2) are counts of the genotypes in case-control samples. Let n and m be the total number of cases and controls, respectively, and the total sample size, $N=n+m$. As cases and controls are independently sampled the genotype counts for cases and controls follow independent multinomial distributions with parameters (p_0,p_1,p_2), and (p_0',p_1',p_2'), respectively, where p_i and p_i', $i=0,1,2$, are the genotype probabilities in cases and controls.

$$(n_0,n_1,n_2):\mathrm{Multi}(n,p_0,p_1,p_2),$$

$$(m_0,m_1,m_2):\mathrm{Multi}(m,p_0',p_1',p_2').$$

Under the null hypothesis of no association, $H_0: p_i=p_i'$ for $i=0,1,2$. The Cochran-Armitage's trend test statistic for the data in Table 1 is given by

$$T=\frac{N(N(n_1+2n_2)-n(N_1+2N_2))^2}{n(N-n)\left(N(N_1+4N_2)-(N_1+2N_2)^2\right)}. \quad (1)$$

The statistic in Eq. (1) follows the chi-square distribution with one degree of freedom (df), see (Armitage, 1955). Let us denote the Cochran-Armitage trend test as CA in the rest of work.

Agresti (2007) states CA in terms of the Pearson chi-squared statistic. Consider a contingency table $2\times J$ with ordered column, see Table 1. Let $n_j\sim bin(N_j,p_j)$, $j=0,\ldots,J-1$, it is of interest to test the following null hypothesis

$$\begin{aligned} &H_0: p_0=p_1=\ldots=p_{J-1},\\ &H_1: p_i\neq p_j,\ \exists i\neq j. \end{aligned} \quad (2)$$

It can be carried out by using a linear probability model

$$p_j=\alpha+\beta w_j. \quad (3)$$

Table 2
Frequency table.

score				
w_0	w_1	...	w_{J-1}	total
n_0	n_1	...	n_{J-1}	n
m_0	m_1	...	m_{J-1}	m
N_0	N_1	...	N_{J-1}	N