Statistics and Its Interface Volume 4 (2011) 1–10

A robust test for multi-ordered $2 \times J$ ordinal contingency tables

Min Yuan, Jinfeng Xu, Yaning Yang, and Gang Zheng*

Pearson's chi-square test and the Cochran-Armitage trend test are commonly used in the analysis of $2 \times J$ contingency tables. When the J columns are nominal, Pearson's test should be considered. On the other hand, when the J columns are ordinal and the ordering is well defined, the trend test should be used. In practice, however, the columns are often ordered but the ordering may not be uniquely defined, especially the J categories may be ordered in multiple ways according to several different factors. We assume that the columns could be either singly or multi-ordered, either being scientifically plausible, and consequently different scores could be assigned to the columns by the different ordering systems. Then the trend test, if applied, may lose substantial power when the orderings are misspecified. To guard against misspecifications of the scores for the columns, we propose a robust test by combining strengths of both Pearson's test and the trend test. In the trend test, we allow several different score specifications according to different ordering criteria and the scores are chosen to be robust enough. Extensive simulation studies demonstrate the efficiency robustness of the proposed approach. The proposed method is applied to two data sets from the Genetic Association Workshop 15 and an experiment on the use of sulfones and streptomycin drugs in the treatment of leprosy.

Keywords and phrases: Efficiency robustness, Minimum p-values, Ordered categorical data, Scoring, Trend tests, Two-locus model.

1. INTRODUCTION

In categorical data analysis, data are often summarized in a $2 \times J$ contingency table with the rows corresponding to two groups under comparison and the columns to J different ordinal categories. In many biomedical studies, the ordering of the J categories may not be uniquely defined and multiple ordering systems may exist. Some motivating examples are given below.

Example 1. In a genetic case-control association study, one is interested in testing association between a disease and a diallelic marker with three genotypes ($J = 3$). It is usually postulated that one of the two alleles is the risk allele, so that an individual's risk of having disease increases with the number of the risk allele and as such, the Cochran-Armitage's trend test should be applied [22].

Example 2. Also in the genetic association study, but a marker has three alleles. Then two of the three alleles may be regarded as risk alleles, and a total of $J = 6$ genotypes (combinations of two alleles) can be naturally ordered by the number of either alleles if they have additive effects on the disease. A trend test assigning proper scores to the two orderings for two risk alleles can be used to test for genotype-disease association and use of only one score in the trend test may lead to power loss since only the partial ordering with respect to one specific allele is used [4].

Example 3. In a two-locus case-control genetics association study, ordering of the loci may only be well defined marginally according to the genetic models at each locus but the interactions between the loci may be difficult to characterize. For example, if the two marker loci are diallelic (hence, there are totally $= 9$ genotype combinations), then two different orderings with respect to the two risk alleles can be assigned to the 9 columns. By using the two scoring systems, the between-locus interaction may be appropriately defined and then the trend test with the marginal scores and interaction terms can be applied.

Example 4. This example is similar to Example 3, but not in the genetics context. In studying association of smoking status and alcohol consumption to breast cancer, the smoking status is recorded as none (S0), less than a pack/day (S1) and more than a pack/day (S2) and the drinking status is recorded as none (N0), social drinker (N1), moderate drinking (N2) and heavy drinking (N3). There are 12 different combinations of smoking and drinking behaviors ($J = 12$). The cancer and cancer-free subjects are classified into these 12 categories to form a 2×12 table. When applying a trend test to it, ordered scores have to be assigned to each of the 12 columns. It is easier to assign increasing scores to smoking or drinking marginally, but it might not be easy to assign increasing scores jointly on the two risk factors (e.g., if $\mathrm{S0} + \mathrm{N2} < \mathrm{S1} + \mathrm{N1}$?). In other words, wrong scores are most likely to be assigned jointly than marginally. We also call this 2×12 table as a multi-ordered contingency table, because two orderings for smoking and drinking are involved.

*Corresponding author.

For a general multi-ordered contingency table, similar to the method of Czika and Weir [4], we can assign proper scores to all K possible linear independent orderings to obtain the trend test with a $J \times K$ score matrix. We still refer to this extended test as a trend test. The trend test asymptotically follows a chi-square distribution with K degree of freedoms (df) under the null hypothesis of no association. Obviously, the smaller the K, the more efficiency the trend test would achieve. When J is large and K is small, Pearson's test, with a larger degree of freedom, is usually less powerful compared to the trend test. Notice that if one assigns $K = J - 1$ linearly independent score vectors to the J categories, the trend test turns out to be the Pearson's chi-square test and consequently the choice of scores is irrelevant. But in practice, K is usually much less than $J-1$.

All test statistics that we consider here have the correct sizes under the null hypothesis H_0, because they all have asymptotic chi-squared distributions with different degrees of freedom. Hence our goal is to compare their efficiency and robustness under the alternative hypothesis H_1 due to model uncertainty. A model refers to a parametric distribution under which the observed data are generated. If the model is correctly specified (e.g., the normal distribution in the classical two-sample problem), an asymptotically optimum test can be applied (e.g., the T-test). If the actual distribution underlying the data is Cauchy distribution, the T-test has no power to detect a location shift (or it is not efficient). The power is used as a function of the sample size, while the efficiency is used in terms of a local H_1. Hence the T-test is not robust when the true model is misspecified. In fact, a more robust test is Wilcoxon rank test, which is less powerful than the T-test under the normal distribution, but retains high power when the distribution is Cauchy. Thus, it is much less sensitive to the misspecification of an underlying distribution than the T-test. The efficiency and robustness are a trade-off. We use the term efficiency robust test to refer to a test which seeks a balance between the efficiency and robustness [6]. A similar idea to efficiency robust tests in estimation is a minimax estimator. However, the efficiency robust test is used for hypothesis testing.

The power of the trend test is quite sensitive to the choices of scores and is therefore not robust to the misspecification of the scores (models) [8], the trade-off between efficiency and robustness for the trend test has been a recent debate in the literature [11, 23, 28]. When the model is known, scores are known. When the model is uncertain, scores are still needed to be specified in order to use the trend test, but subject to errors, which would affect the power performance of the trend test. In genetic case-control association studies, the scores are determined by the underlying mode of inheritance (i.e., the genetic model). For a single locus study, common genetic models include the recessive, additive, multiplicative and dominant models. The overdominant and underdominant models are less common but are also possible. The definitions of different genetic models are given later. For a multi-locus study, however, mode of inheritance becomes much more complicated when genetic interactions exist [16, 20, 27]. The true genetic model is usually unknown in practice, which is particularly true for many common and complex diseases as those studied by Wellcome Trust Case-Control Consortium (WTCCC) [26] in which a robust test to combine the strength of Pearson's test and the 1-df trend test for analysis of a 2×3 contingency table in genome-wide scans was proposed. Here the robustness refers to the power of a test being not largely affected even if the scoring vector or matrix is incorrectly specified. The robust test of WTCCC [26], referred to as Min2 by Joo et al. [13], takes the minimum of the p-values of Pearson's test and the trend test with the scores locally optimal for the additive model. Joo et al. [13] used this minimum p-value to rank about 400,000 genetic markers (SNPs) after quality control. The top ranked SNPs can be used for further focused analysis. Joo et al. [13] derived the asymptotic null distribution of Min2 for 2×3 tables, and studied some of its properties for genome-wide association studies.

For general J and $K < J-1$, similar to Joo et al. [13], we propose to combine the $(J-1)$-df Pearson's chi-square test and the K-df trend test by taking the minimum of their p-values, hereafter referred to as Min2 test. The scores (vector or matrix) of the trend test may be subject to misspecification yet are chosen to be robust enough. For example, in genetic association analysis, an additive score (equally spaced) for each risk allele can be used if the underlying genetic model is unknown [22]. The proposed method is shown to have both the robustness of the Pearson's test and the efficiency of the trend test. As a special case, our robust test can also be applied to a singly ordered contingency table ($K = 1$). Moreover, our approach extends the Min2 test of Joo et al. [13] and can be used in multi-locus association analysis with interactions.

The rest of this article is organized as follows. In Section 2, we introduce the proposed Min2 test for ordered $2 \times J$ contingency tables and study its properties. In Section 3, we evaluate the performance of the proposed test for two-locus genetic association studies via simulations. In Section 4, we apply our method to two subsets of Genetic Analysis Workshop 15 (GAW15) data set (http://www.gaworkshop.org/index.html) and an experiment on the use of sulfones and streptomycin drugs in the treatment of leprosy. Some concluding remarks are given in Section 5.

2. METHODS

2.1 Choice of scores

Consider a general $2 \times J$ contingency table with a binary outcome and J exposures. Without loss of generality, we consider data from a case-control design, but the developed methods are readily applicable to prospective or cross-sectional studies. Data including the underlying probabilities and cell counts are displayed in Table 1, where $y = 1$

Table 1. Probabilities and observed counts of a $2 \times J$ contingency table

Probabilities (Observed counts)					
	x_1	x_2	$\cdots$	x_J	
$y=1$	$p_1(r_1)$	$p_2(r_2)$	$\cdots$	$p_J(r_J)$	(R)
$y=0$	$q_1(s_1)$	$q_2(s_2)$	$\cdots$	$q_J(s_J)$	(S)
Total	(n_1)	(n_2)	$\cdots$	(n_J)	(N)

or 0 stands for the case or control group and $r_i(s_i)$ is the observed count in the case (control) group, $i = 1, \ldots, J$. Denote row margins by R and S and column margins n_i, $i = 1, \ldots, J$.

There might be different ordering methods for the J categories. For example, in genetic case-control association studies with I alleles, among which there are $K \leq I-1$ risk alleles. The $J = I(I+1)/2$ genotypes may be ordered by the number (or presence/absence) of any of the K risk alleles. Coding of these orderings depends on the underlying genetic model. Recessive, additive/multiplicative and dominant models are four commonly used genetic models. Under a recessive model, traits are only expressed if an individual has two risk alleles while under a dominant model, traits are expressed if an individual has at leat one risk allele. An individual's risk of expressing the trait increases with the number of risk alleles under an additive or multiplicative models. For example, when $I = 2$ ($J = 3$), the scores (0,0,1), (0,1/2,1), and (0,1,1) are used for the recessive, additive and dominant models, respectively [22, 7, 30]. When $I = 3$ ($J = 6$), we can assign two sets of scores according to the two risk alleles. Specifically, if the three alleles are A_1, A_2 and A_3 and among which A_1 and A_2 are the risk alleles, then the 6 possible genotypes (A_1A_1, A_1A_2, A_1A_3, A_2A_2, A_2A_3, A_3A_3) can be scored twice by the allele A_1 and A_2, respectively. For example, if both the risk alleles obey additive model, then we assign two sets of scores $(1, 0.5, 0.5, 0, 0, 0)$ and $(0, 0.5, 0, 1, 0.5, 0)$ to the 6 genotypes by one half of the number of A_1 and A_2, respectively. If, on the other hand, A_1 is dominant and A_2 is recessive, then the two sets of scores are $(1, 1, 1, 0, 0, 0)$ and $(0, 0, 0, 1, 0, 0)$.

A relevant but different example is the score specification in two-locus association analysis. Two-locus models with different interaction mechanisms are of great importance and were discussed in literature [16, 20, 27]. There are mainly two classes of two-locus models, namely, epistasis models and heterogeneity models [24]. All models here assume that the two loci are diallelic and unlinked. Let the two loci be diallelic and the alleles be A/a and B/b and A and B are risk alleles. Hence, there are $J = 9$ genotype combinations ($AABB$, $AABb$, $AAbb$, $AaBB$, $AaBb$, $Aabb$, $aaBB$, $aaBb$, $aabb$). We consider seven two-locus models. These two-locus models are summarized in Table 2 in terms of penetrance (penetrance is defined as the possibility of an individual being a case given a certain genotype), defined as follows.

Table 2. Penetrance tables for two disease loci ($g < f$)

	$D \cap D$			$D \cap R$			$R \cap R$		
	BB	Bb	bb	BB	Bb	bb	BB	Bb	bb
AA	f	f	g	f	g	g	f	g	g
Aa	f	f	g	f	g	g	g	g	g
aa	g	g	g	g	g	g	g	g	g
	$D \cup D$			$D \cup R$			$R \cup R$		
AA	f	f	f	f	f	f	f	f	f
Aa	f	f	f	f	f	f	f	g	g
aa	f	f	g	f	g	g	f	g	g
				threshold					
AA				g	g	g			
Aa				g	g	f			
aa				g	f	f			

The epistasis models include the intersection of dominant and dominant ($D \cap D$): the intersection of recessive and dominant ($D \cap R$), the intersection of recessive and recessive ($R \cap R$). For the $D \cap D$ model, a dominant disease allele must be present at both loci in order to contract the disease. Consequently, only those genotypes containing the two risk alleles will have the disease. This model has a good application in determining the color of the flowers produced by certain strains of pea plants [25]. For the $D \cap R$ model, the disease allele is dominant at locus 1 and recessive at locus 2. Therefore, the only vulnerable genotypes are AABB and AaBB. People can find an example of this model from the plumage color of chickens [15]. For the third epistasis model, traits manifest itself only if a recessive allele exists at both loci. Thus only individuals with genotype AABB are at risk.

The heterogeneity models include the union of dominant and dominant ($D \cup D$), the union of dominant and recessive ($D \cup R$) and the union of recessive and recessive ($R \cup R$). The $D \cup D$ model requires that at least one risk allele is at either locus 1 or locus 2. Thus, all genotypes are at risk except aabb. We can find the genetic model in the animal study for the control of leg feathers on chickens [15]. For the $D \cup R$ model, traits are determined by either a dominant form of the disease allele at the first loci or a recessive form at the second. For the $R \cup R$ model, those with a recessive allele at one of the two loci are at risk. The application of this genetic model can also be found in Lerner et al. [15].

Another commonly used interaction model is the threshold interaction model [18]. In a threshold model, trait manifests itself only when the total number of risk alleles at these two loci is greater than 1.

We show in details the scoring method for these models in what follows. For the nine genotypes, one can assign two sets of scores according to the marginal genetic model and one more score defined from the two scores to characterize the between-locus interaction. For example, suppose the two loci are both marginally dominant, then we can assign scores

$(1,1,1,1,1,1,0,0,0)$ and $(1,1,0,1,1,0,1,1,0)$ to the 9 genotypes based on the dominant model for each locus. A third score $(1,1,0,1,1,0,0,0,0)$, which is the element-wise intersection (minimum), can be used to represent the epistasis model or $(1,1,1,1,1,1,1,1,0)$, which the element-wise union (maximum), can be used to represent heterogeneity interaction. Formally, score matrices for the aforementioned epistasis and heterogeneity interaction models are summarized in (1)–(7), in which the 9 rows represent the genotypes and the first two columns of each matrix are the marginal scores of the two loci and the third column is the minimum/maximum of the first two columns for epistasis/heterogeneity model, representing the between-locus interaction.

$$(1)\qquad \mathbf{x}_{D\cap D}=\begin{pmatrix}1&1&1&1&1&1&0&0&0\\1&1&0&1&1&0&1&1&0\\1&1&0&1&1&0&0&0&0\end{pmatrix}^T$$

$$(2)\qquad \mathbf{x}_{D\cap R}=\begin{pmatrix}1&1&1&1&1&1&0&0&0\\1&0&0&1&0&0&1&0&0\\1&0&0&1&0&0&0&0&0\end{pmatrix}^T$$

$$(3)\qquad \mathbf{x}_{R\cap R}=\begin{pmatrix}1&1&1&0&0&0&0&0&0\\1&0&0&1&0&0&1&0&0\\1&0&0&0&0&0&0&0&0\end{pmatrix}^T$$

$$(4)\qquad \mathbf{x}_{D\cup D}=\begin{pmatrix}1&1&1&1&1&1&0&0&0\\1&1&0&1&1&0&1&1&0\\1&1&1&1&1&1&1&1&0\end{pmatrix}^T$$

$$(5)\qquad \mathbf{x}_{D\cup R}=\begin{pmatrix}1&1&1&1&1&1&0&0&0\\1&0&0&1&0&0&1&0&0\\1&1&1&1&1&1&1&0&0\end{pmatrix}^T$$

$$(6)\qquad \mathbf{x}_{R\cup R}=\begin{pmatrix}1&1&1&0&0&0&0&0&0\\1&0&0&1&0&0&1&0&0\\1&1&1&1&0&0&1&0&0\end{pmatrix}^T$$

$$(7)\qquad \mathbf{x}_{Thres}=\begin{pmatrix}1&1&1&0.5&0.5&0.5&0&0&0\\1&0.5&0&1&0.5&0&1&0.5&0\\1&1&0&1&0&0&0&0&0\end{pmatrix}^T,$$

where the first two rows are additive scores for the two loci and each entry of the last column is 1 if the number of risk alleles A and B in the two-locus genotype is greater than 2 (i.e., sum of the scores in the first two columns exceeds 1), otherwise it is 0.

Note that there are many other possible two-locus models underlying the case-control data. Because interaction mechanism may be so complicated that no simple scores can model it. A common approach in genetic study is to use the doubly additive model by assuming additive model for each locus, namely, the score matrix defined by A and B is taken to be

$$(8)\qquad \mathbf{x}_a=\begin{pmatrix}1&1&1&0.5&0.5&0.5&0&0&0\\1&0.5&0&1&0.5&0&1&0.5&0\end{pmatrix}^T$$

This model is the "multiplicative model within and between loci" defined in Marchini et al. [18]. This set of scores is a robust choice when the underlying two-locus model is unknown and we will use the trend test with this set of scores as our trend test in constructing Min2 test.

For the analysis of a general $2\times J$ contingency table, some guidelines of choices of scores in practice were suggested by Graubard and Korn [8]: (i) Use natural column scores which are based on substantive meaning of column categories; (ii) Use equally spaced column scores; and (iii) Always check the mid-rank scores if it is appropriate. In any case, the trend test is subject to misspecification of the scores and may be consequently less robust.

2.2 The trend test

For K different linearly independent scoring of the J categories, we have a $J\times K$ $(1\le K\le J-1)$ score matrix which is denoted as $\mathbf{x}=(\widetilde{x}_1,\ldots,\widetilde{x}_J)^T$ with $\widetilde{x}_j$ being the $K\times 1$ score vector assigned to the j-th category. Elements of $\mathbf{x}$ are normalized to lie in the interval $[0,1]$ in this article. The score matrix $\mathbf{x}$ is of full rank K because of the independency among the K sets of scores. Denote the jth category by E_j and $p_j=Pr(y=1|E_j)$. Then the logistic regression model has the form

$$\log\left(\frac{p_j}{1-p_j}\right)=\alpha+\beta^T\widetilde{x}_j,$$

where α is the baseline log odds, and β is the log odds ratio. The above prospective logistic regression model can also be used to fit a retrospective study [21].

Denote $\phi=R/N$, and

$$\mathbf{r}=(r_1,\ldots,r_J)^T,\ \mathbf{s}=(s_1,\ldots,s_J)^T,\ \mathbf{n}=(n_1,\ldots,n_J)^T.$$

The likelihood function $L(\alpha,\beta)$ is proportional to

$$\prod_{j=1}^{J}p_j^{r_j}(1-p_j)^{s_j}.$$

The score test for $H_0:\beta=0$ can be written as

$$(9)\qquad T_1(\mathbf{x})=U^T\left\{\mathrm{var}(U)\right\}^{-1}U,$$

where $U=\mathbf{x}^T\{(1-\phi)\mathbf{r}-\phi\mathbf{s}\}$ and

$$\mathrm{var}(U)=\{R(N-R)/N^3\}\mathbf{x}^T\left\{N\mathrm{diag}(\mathbf{n})-\mathbf{n}\mathbf{n}^T\right\}\mathbf{x}.$$

The score statistic $T_1(\mathbf{x})$ in (9) is called the trend test with scores $\mathbf{x}$. When the scores $\mathbf{x}$ are fixed (prespecified), $T_1(\mathbf{x})$ asymptotically follows a K-df chi-square distribution under H_0.

The robust tests can be derived from the trend test [6]. For example, the MAX test [28] can be written as $\text{MAX} = \max_{\mathbf{x}} T_1(\mathbf{x})$. Pearson's test, can also be obtained from the trend test when $K = 1$ and the scores are data-driven, namely $\hat{\mathbf{x}} = (r_1/n_1, \ldots, r_J/n_J)^T$, then $T_1(\hat{\mathbf{x}})$ is identical to Pearson's test which takes the form [31]

$$T_2 = \sum_{i=1}^{J} \frac{(r_i - n_i R/N)^2}{n_i R/N} + \sum_{i=1}^{J} \frac{(s_i - n_i S/N)^2}{n_i S/N}.$$

Under H_0, T_2 follows asymptotically a chi-square distribution with $J-1$ df. When $J = 3$ and $K = 1$, $T_1(\mathbf{x})$ becomes the trend test of Sasieni [22].

2.3 The proposed robust test

Let $\mathbf{x}$ be a prespecified $J \times K$ score matrix and the corresponding trend test be $T_1(\mathbf{x})$. This test is most powerful if the scores are correctly specified. But, as what we have argued before, $\mathbf{x}$ is subject to misspecification and therefore $T_1(\mathbf{x})$ may lose power. Our strategy is to compensate for the power loss by testing the hypothesis using both the Pearson's test and the trend test with some scores that are representative in practice (for example, additive scores for genetic association analysis, as discussed in Section 2.1). Denote p-values of $T_1(\mathbf{x})$ and T_2 by P_1 and P_2, respectively. Here we assume that the trend test follows a K-df chi-square distribution asymptotically. The robust test that we propose, also denoted by Min2, is given by

$$\text{Min2} = \min(P_1, P_2), \tag{10}$$

where $1 \le K \le J-1$ and $J \ge 3$. The null hypothesis is rejected when a smaller Min2 is observed. When $K = 1$ and $J = 3$, Min2 reduces to the special case considered in WTCCC [26] and Joo et al. [13]. Here, we not only extend their robust test from a 2×3 table to a general $2 \times J$ table but also allow the trend test with general degrees of freedom other than 1.

Denote the random variable with a chi-square distribution with f df by χ^2_f. Let $F_1(\cdot)$ and $F_2(\cdot)$ be the cumulative distribution functions of χ^2_K and χ^2_{J-1}, respectively. Then $P_1 = 1 - F_1(T_1(\mathbf{x}))$ and $P_2 = 1 - F_2(T_2)$.

Let $Y_1 = T_1(\mathbf{x})/T_2$ and $Y_2 = T_2$. Then, from Hogg and Craig [9] (Example 5, p. 138), Y_1 and Y_2 are asymptotically independent and Y_1 asymptotically has a beta distribution $Beta(K/2, (J-1-K)/2)$ under H_0. A simple proof for $K = 1$ was also given in Zheng et al. [28]. The joint density function of (Y_1, Y_2), $f_{Y_1,Y_2}(y_1, y_2)$, can be written as

$$f_{Y_1,Y_2}(y_1, y_2) = C y_1^{\frac{K-2}{2}} y_2^{\frac{J-3}{2}} (1-y_1)^{\frac{J-4}{2}} e^{-\frac{y_2}{2}},$$

where $C = \Gamma(\frac{K+J-2}{2})/\{2^{\frac{J-1}{2}}\Gamma(\frac{K}{2})\Gamma(\frac{J-1}{2})\Gamma(\frac{J-2}{2})\}$ and $0 \le y_1 \le 1$, $y_2 > 0$. By a transformation, the joint density of $T_1 = T_1(\mathbf{x})$ and T_2, denoted by $f_{T_1,T_2}(t_1, t_2)$, can be written as

$$f_{T_1,T_2}(t_1, t_2) = C t_1^{\frac{K-2}{2}} t_2^{\frac{1-K}{2}} (t_2 - t_1)^{\frac{J-4}{2}} e^{-\frac{t_2}{2}},$$

with $0 \le t_1 \le t_2$. For a given $c > 0$, noticing that $F_1^{-1}(1-c) \le F_2^{-1}(1-c)$, then under the null

$$\begin{aligned} Pr_{H_0}(\text{Min2} > c) &= Pr_{H_0}(P_1 > c, P_2 > c) \\ &= Pr_{H_0}(1 - F_1(T_1) > c, 1 - F_2(T_2) > c) \\ &= Pr_{H_0}\big(T_1 < F_1^{-1}(1-c), T_2 < F_2^{-1}(1-c)\big) \\ &= \int_0^{F_1^{-1}(1-c)} \int_{t_1}^{F_2^{-1}(1-c)} f_{T_1,T_2}(t_1, t_2) dt_1 dt_2. \end{aligned}$$

When $\text{Min2} = t$ is observed, the p-value of Min2 can be obtained using the following formula:

$$\begin{aligned} p_{\text{Min2}} &= p_{\text{Min2}}(t) = Pr_{H_0}(\text{Min2} \le t) \\ &= 1 - \int_0^{F_1^{-1}(1-t)} \int_{t_1}^{F_2^{-1}(1-t)} f_{T_1,T_2}(t_1, t_2) dt_1 dt_2. \end{aligned} \tag{11}$$

The above equations can be used to determine threshold values and p-values of the Min2 test. As an example, for a 2×6 table, the analytical threshold is 0.0306 when significance level is $\alpha = 0.05$. Therefore the threshold 0.025, if the Bonferroni correction is applied, is obviously too conservative. On the other hand, if $\text{Min2} = 0.05$ is observed for a 2×6 table, the p-value of it is then 0.0664 from the asymptotic distribution.

Note that the asymptotic null distribution of Min2 does not depend on the parameters of the contingency table. In genetic association studies, these parameters include allele frequency, relative risks, and penetrances. Furthermore, $\partial p_{\text{Min2}}/\partial t > 0$ (see Appendix). In genetic association studies, the above results imply that the asymptotic null distribution of Min2 is independent of the allele frequency of the marker of interest, and that the p-value of Min2 and the value of Min2 can both be used to rank SNPs in genome-wide association studies, which would result in the same ranking orders. The importance of this property in genome-wide scans is discussed in Li et al. [17].

In applications, sparse categorical data are likely obtained, e.g., genetic association studies with a minor allele frequency. It is known that the asymptotic null distribution of Pearson's test is sensitive to small or even zero cell counts. Thus it is expected that the asymptotic distribution of Min2 would also be affected. If small cell or empty cell counts are observed in the data, p-value of Min2 can be obtained by permutation methods. Some of our numerical results presented later are obtained by the permutation method.

Table 3. Empirical power comparison among Trend, Pearson's and Min2 tests for $\beta_1 = \beta_2 = 0.2$

β_3	Test	Null	$D \cup D$	$R \cup R$	$D \cup R$	$D \cap D$	$R \cap R$	$D \cap R$	Threshold
0.1	Trend	0.051	0.888	0.381	0.805	0.912	0.214	0.570	0.599
	Pearson	0.045	0.801	0.447	0.733	0.841	0.229	0.506	0.378
	Min2	0.048	0.877	0.452	0.801	0.903	0.240	0.577	0.537
0	Trend	0.050	0.731	0.198	0.500	0.725	0.197	0.491	0.432
	Pearson	0.049	0.585	0.205	0.404	0.581	0.203	0.393	0.244
	Min2	0.049	0.705	0.222	0.485	0.695	0.217	0.475	0.376
−0.1	Trend	0.052	0.483	0.088	0.200	0.455	0.176	0.427	0.295
	Pearson	0.048	0.361	0.083	0.157	0.340	0.192	0.328	0.176
	Min2	0.047	0.458	0.089	0.190	0.428	0.199	0.408	0.256

Table 4. Empirical power comparison among Trend, Pearson's and Min2 tests for $\beta_1 = \beta_2 = 0$

β_3	Test	Null	$D \cup D$	$R \cup R$	$D \cup R$	$D \cap D$	$R \cap R$	$D \cap R$	Threshold
0.4	Trend	0.051	0.650	0.565	0.922	0.789	0.052	0.156	0.316
	Pearson	0.047	0.750	0.700	0.904	0.841	0.069	0.232	0.404
	Min2	0.046	0.748	0.695	0.932	0.853	0.064	0.218	0.395
−0.4	Trend	0.046	0.774	0.431	0.936	0.690	0.054	0.110	0.215
	Pearson	0.048	0.821	0.595	0.910	0.773	0.066	0.180	0.307
	Min2	0.048	0.837	0.578	0.942	0.780	0.065	0.160	0.286

3. SIMULATION STUDIES

In this section, we present simulation results in terms of two-locus genetic case-control association analysis. We compare power of the Pearson's test, the trend test $T_1(\mathbf{x}_a)$ with additive scores $\mathbf{x}_a$ in equation (8) and the Min2 test. We denote the markers of interest by G and H. The alleles of marker G are A and a and those of marker H are B and b. We use $G_i(H_i)$ to represent the genotypes AA, Aa and aa (BB, Bb and bb). Therefore, there are up to 9 genotypes G_iH_j with $Pr(G_iH_j) = g_{ij}$, $i, j = 1, 2, 3$. Alleles A and B are assumed to be the risk alleles.

We use two-locus genetic models in (1)-(7) to generate data. Let the allele frequencies be p_1, p_2 for allele A and B. We assume Hardy-Weinberg equilibrium proportions hold in the population and the two markers are not in linkage disequilibrium in the population so that $g_{ij} = Pr(G_i)Pr(H_j)$ with $Pr(G_1) = p_1^2$, $Pr(G_2) = 2p_1(1-p_1)$, $Pr(G_3) = (1-p_1)^2$ and $Pr(H_1) = p_2^2$, $Pr(H_2) = 2p_2(1-p_2)$, $Pr(H_3) = (1-p_2)^2$. Let $k = Pr(\text{case})$ be the disease prevalence and $f_{ij} = Pr(\text{case}|G_i, H_j)$ be the penetrance for $i, j = 1, 2, 3$.

Let $y = 1$ and $y = 0$ for case and control individuals respectively. For a given score matrix $\mathbf{x}$, we use a logistic regression model

$$f_{ij} = Pr(y = 1|G_iH_j) = \frac{\exp(\alpha_0 + \tilde{x}_{l1}\beta_1 + \tilde{x}_{l2}\beta_2 + \tilde{x}_{l3}\beta_3)}{1 + \exp(\alpha_0 + \tilde{x}_{l1}\beta_1 + \tilde{x}_{l2}\beta_2 + \tilde{x}_{l3}\beta_3)}$$

where $l = j + 3(i-1)$, $\tilde{x}_{l1}, \tilde{x}_{l2}$ correspond to marginal model and $\tilde{x}_{l3}$ to interaction model; coefficients β_1 and β_2 represent the marginal effect with respect to the two markers of interest and β_3 reflects the interaction effect. Using Bayes formula, we have

$$Pr(G_iH_j|y = 1) = \frac{Pr(y = 1|G_iH_j)g_{ij}}{k},$$
$$Pr(G_iH_j|y = 0) = \frac{Pr(y = 0|G_iH_j)g_{ij}}{1 - k}.$$

Data of 1,000 cases and 1,000 controls are generated from these two distributions using the scores in equations (1)–(7).

For power comparison, allele frequencies of A and B are both set to be 0.3 and the significance level is 0.05. Results for other choices of allele frequencies are similar and are not reported here. We consider two settings: (a) Both marginal and interaction effects exist. Under this premise, we take $\beta_1 = \beta_2 = 0.2$ and β_3 equals to 0.1, 0 and -0.1 allowing that alleles could have predisposing or protective effects to disease through different interactions; (b) Only interaction effects exist. In this setting, we did simulation for $\beta_1 = \beta_2 = 0$ and $\beta_3 = 0.4$ or $\beta_3 = -0.4$. All simulations are conducted with 10,000 replicates. The powers of Pearson's test, trend test with score matrix $\mathbf{x}_a$ and their corresponding Min2 are reported in Tables 3–4. In these tables, the first column (Null) are the type I errors under the null model ($\beta_1 = \beta_2 = \beta_3 = 0$).

In Table 3, Pearson's test is better than the trend test only under the $R \cap R$ or $R \cup R$ model, while the trend test has better performance under the D-related models. All results show that the Min2 is always more powerful than at least one of the Pearson's test and trend test and its power is close to the more powerful one. For example, when $\beta_3 = 0$ in Table 3, the maximal relative power gain of Min2 over Pearson or trend test is 54% and the maximal relative power loss of

Table 5. Two-locus distribution of genotypes of rs2277726 and rs4414555

Genotype	00	01	02	10	11	12	20	21	22
Case	5	3	2	15	42	37	84	172	99
Control	1	2	2	32	54	33	59	160	117

Table 6. Two-locus distribution of genotypes of rs585491 and rs4245254

Genotype	00	01	02	10	11	12	20	21	22
Case	17	67	53	42	119	79	17	34	31
Control	17	51	50	47	110	66	25	47	47

Min2 is only 13%. When there exists interaction ($\beta_3 = 0.1$), the maximal relative power gain is 42% and maximal relative power loss is 10%.

Table 4 shows results for the situation when the main effects are absent ($\beta_1 = \beta_2 = 0$). From this table, the Pearson's test is generally more powerful than the additive trend test except, for example, the $D \cup R$ model when $\beta_3 = 0.4$. Similar to results in Table 3, the power of Min2 is close to or greater than the larger one between the Pearson's test and the trend test and hence shows robust efficiency.

In summary, Min2 has neither uniformly the worst nor the best power performance among the three tests across the seven models. When Min2 loses power compared with the Pearson's test and the trend test, it loses little; when it gains power it gains much. These features of Min2 demonstrate its efficiency and its robustness across various genetic models.

4. ILLUSTRATION

4.1 Two-locus interaction examples

We first apply Min2 test to two two-locus subsets from Problem 2 of the GAW15, which studies the genetic causes of rheumatoid arthritis. There are 459 cases and 460 controls genotyped at 2719 SNPs. The first subset is for SNPs rs2277726 and rs4414555 and the second one is for SNPs rs585491 and rs4245254. These two pairs of SNPs are randomly chosen from all 2719 SNPs to be in linkage equilibrium. We denote "0","1" and "2" the three genotypes of each locus, thus nine combined genotypes can be expressed as "00", "01", "02", ..., "22". We analyze these two data sets to illustrate our approach. Distributions of genotype frequencies of these two groups of SNPs are displayed in two 2×9 contingency tables (Tables 5 and 6).

For the first data set, if we use Pearson's test and trend test for each locus as in Joo et al. [13], i.e., marginal one-locus model, both tests are not significant at significance level 0.05 for SNPs rs2277726 (p-values of Pearson's test and trend test are 0.077 and 0.308 respectively) and rs4414555 (p-values of Pearson's test and trend test are 0.494 and 0.235 respectively). For the two-locus analysis, we choose to use the additive score matrix for the trend test, i.e., the score matrix $\mathbf{x}_a$. The Min2 test is constructed based on the trend test with score $\mathbf{x}_a$ and the Pearson's test. As Pearson's test may be invalid in the situation that some cell counts are very small, we use permutation to calculate p-values by randomly switching "case" or "control" labels with all margins being fixed. When two-locus model is considered, p-values of Pearson's test and trend test for the first data set are reported as 0.030, 0.287 respectively. It shows that the Pearson's test is significant at level 0.05 but not for the trend test. For the second data set, p-values of Pearson's test and the trend test for SNP rs585491 are 0.012 and 0.008 respectively that both show significance at level 0.05; p-values of Pearson's test and the trend test for SNP rs4245254 are 0.507 and 0.540 respectively that both show no significance at level 0.05. P-values of Pearson's test and the trend test under two-locus model are reported as 0.203, 0.026. Now the trend test is significant but not the Pearson's test.

In practice, one cannot determine a priori which test to use, so one may apply both tests and correct for multiple testing by Bonferroni correction. The corrected p-values (of p-values less than 0.05) for the first and second data sets are $2 \times 0.030 = 0.060$ and $2 \times 0.026 = 0.052$, respectively. Both corrected p-values are not significant at 0.05 level. On the other hand, both the p-values of the Min2 test for the two data sets are 0.046. These two examples show that Min2 (with score $\mathbf{x}_a$) have p-values that is very close to the minimum of the p-values of the Pearson's test and the trend tests and could be significant when Pearson's test or trend test is not.

Table 7. Change in clinical condition by degree of infiltration

	Degree of Infiltration	
Clinical Change	High	Low
Worse	1	11
Stationary	13	53
Slight improvement	16	42
Moderate improvement	15	27
Marked improvement	7	11

4.2 Example from clinical treatment

We also illustrated the efficiency robustness of the Min2 test by a singly-ordered contingency table. Table 7 refers to the experiment on the use of sulfones and streptomycin drugs in the treatment of leprosy [1]. The degree of infiltration at the start of the experiment measures a type of skin damage. The response is the change in the overall clinical condition of the patient after 48 weeks of treatment. Table 7 could be viewed as 2×5 contingency table and we can define orderings for different clinic changes. We reanalyze this data here with Pearson's test and the trend test with scores (following that in Graubard and Korn [8]) listed in Table 8. Min2 is constructed based on the trend test with

Table 8. Alternative scoring systems for column categories with exact one-sided p-values. All p-values are obtained based on permutations

	Clinical change by degree of infiltration				
	1	2	3	4	5
Mid ranks*	6	45	107	157	187
Standardized	−1.25	−0.73	0.09	0.75	1.15
P-value	Trend: 0.010, Pearson: 0.142, Min2: 0.016				
Equal space	−1	0	1	2	3
Standardized	−1.26	−0.63	0	0.63	1.26
P-value	Trend: 0.008, Pearson: 0.141, Min2: 0.016				
Un-equal space	1	2	3	4	99
Standardized	−0.90	−0.58	−0.26	0.06	1.67
P-value	Trend: 0.160, Pearson: 0.141, Min2: 0.016				

an equally spaced score and Pearson's test. We compute p-values using 10,000 permutations by fixing all the margins. Results in Table 8 show that, as expected, the trend test is sensitive to the choices of scores and different scores may lead to quite different conclusions. For example, the trend test is significant except that one uses the un-equal space score in Table 8. Min2 and Pearson's test are more robust to the score choices. Pearson's test is not significant while Min2 is significant for this data set at level 0.05. On the other hand, if the trend test is significant, Min2 has much smaller p-value than Pearson's test, indicating an improvement of Pearson's test due to incorporating the linear trend of the categories.

5. DISCUSSION

In this article, we propose a robust test, Min2, to a multi-ordered $2\times J$ ordered contingency table. Our results generalize those of WTCCC [26] and Joo et al. [13], which only considered a 2×3 table with only one score vector (a singly ordered contingency table) for case-control genetic association studies. In Joo et al. [13], Min2 is based on two chi-square tests with 1 degree of freedom and 2 degrees of freedom respectively. Our test allows chi-square tests with more general degrees of freedom by allowing different score specifications when the categories can be ordered in multiple ways.

One important application of our method is to test a two-locus association when the genotypes are ordered according to the two risk alleles on the two loci. We demonstrate the robustness of our proposed test by simulating two-locus genetic association studies, in which the Min2 is constructed from the Pearson's test and the additive trend test. The additive trend test is presumably powerful when the two-locus interaction is absent, but when an interaction exists, the Pearson's test in Min2 can compensate for the power loss of the additive trend test.

When the true model is unknown while a family of scientifically plausible models is available, Min2 is never the least powerful test while the trend test and Pearson's test could be the least powerful test under certain models. The trend test can gain power over the Pearson's test if the score is correctly specified, but lose much if it is not. The proposed Min2 test, however, combines their strengths and is always robust and efficient. When the score matrix is correct or closely so, Min2 behaves similarly to the trend test; when the score matrix is far from correct, Min2 is much closer to Pearson's test in power. Notably, Bonferroni correction can be used to correct for multiple testing when Pearson's test and the trend test are both applied, but it is too conservative due to relatedness of the two tests. The proposed Min2 test incorporates their correlation and is more powerful than applying the Bonferroni correction method.

In practice, it is difficult to define the orderings precisely, especially for multi-locus interaction models. In two-locus genetic association analysis, we proposed to use the additive scores in Min2 because the additive was shown to be more robust than recessive or dominant models and widely used in practice. For a general contingency table when the scores of ordered categories are difficult to determine, we proposed to use equally spaced scores. These choices may incur substantial power loss in the trend test if they are far from the true underlying model. Fortunately, however, the Pearson's test picks the power when the scores are incorrectly specified in the trend test.

Similar to a single-locus study, it is also promising to consider an adaptive score in the trend test as in Zheng and Ng [29]. This adaptive procedure will need an analogue of the Hardy-Weinberg disequilibrium test to determine the two-locus interaction models which is still not available in the literature. Besides, ICPT proposed by Song et al. [24] has a wide application for testing association by considering monotone penetrances along the genotypes. However it is not easy to find its asymptotic distribution. Besides motivated by the idea of Min2, it might be interesting to construct the test as the minimum of the Pearson's test, trend test and ICPT. As one of the referees pointed out, the sum test statistic is another choice to combine the Pearson's test and the trend test [10]. We will investigate these issues in future.

The proposed robust methods are useful in initial exploratory analysis, where a large number of ordered tables with a lot of uncertainty about the underlying distributions need to be examined and tested. The genome-wide scan conducted by the WTCCC is a good example of applying the proposed robust tests. Although we focused on the applications of our proposed methods to genetic epidemiology studies, the proposed robust tests, as shown by one example, can be generally applied to the analysis of any ordered $2\times J$ contingency tables.

ACKNOWLEDGMENT

The work of Y. Yang is supported by China NSF Grant. The work of J. Xu is supported by NUS grant R155-000-075-112. We would like to thank both reviewers, and the editors

for helpful comments and suggestions which improved our presentation. The use of GAW data was approved by the GAW grant, R01 GM031575.

APPENDIX

By the properties of derivation of inverse function and compound function, we have that $\partial p_{\text{Min2}}/\partial t$ is proportional to

$$g_1(t) \times \int_{F_1^{-1}(1-t)}^{F_2^{-1}(1-t)} \frac{(t_2 - F_1^{-1}(1-t))^{\frac{J-4}{2}}}{t_2^{\frac{K-1}{2}} e^{t_2/2}} dt_2$$

$$+ g_2(t) \times \int_{0}^{F_1^{-1}(1-t)} \frac{(F_2^{-1}(1-t) - t_1)^{\frac{J-4}{2}}}{t_1^{\frac{2-K}{2}}} dt_1,$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are the density functions of chi-square distributions with K df and $J-1$ df, respectively, and $g_1(t) = [F_1^{-1}(1-t)]^{\frac{K-2}{2}}/f_1(F_1^{-1}(1-t))$, $g_2(t) = [F_2^{-1}(1-t)]^{\frac{1-K}{2}}/\{e^{F_2^{-1}(1-t)/2} f_2(F_2^{-1}(1-t))\}$. For a given c, $F_l^{-1}(\cdot)$ is an increasing function of integer l, so $\partial p_{\text{Min2}}/\partial t > 0$ for any $t \in [0,1]$, which proves that p_{Min2} is a strictly increasing function of the Min2 statistic t.

Received 16 April 2009

REFERENCES

[1] Armitage, P. (1955). Tests for linear trends in proportions and frequencies *Biometrics*. **11**, 375–386.

[2] Agresti, A. (1996). An introduction to Categorical Data Analysis. *John Wiley and Sons, Inc*, New York. MR1394195

[3] Cochran, W.G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics*. **10**, 417–451. MR0067428

[4] Czika, W. and Weir, B. S. (2004). Properties of the multiallelic trend test. *Biometrics* **60**, 69–74. MR2043620

[5] Dewan, A. et al. (2006). HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989–992.

[6] Freidlin, B., Podgor, M. J., and Gastwirth, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics* **55**, 883–886. MR1705680

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

[8] Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics* **43**, 471–476. MR0897415

[9] Hogg R. V. and Craig A. T. (1978). Introduction to Mathematical Statistics (the 4th edition). *Macmillan publishing co., Inc*, New York. MR0467974

[10] Hoh, J., Wille, A., and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res* **11**, 2115–2119.

[11] Ivanova, A. and Berger, V. W. (2001). Drawbacks to integer scoring for ordered categorical data. *Biometrics* **57**, 567–570. MR1855693

[12] Isabelle Schrauwen, Megan Ealy, Matthew J. Huentelman, Melissa Thys, Nils Homer, Kathleen Vanderstraeten, Erik Fransen, Jason J. Corneveaux, David W. Craig, Mireille Claustres, Cor W. R. J. Cremers, Ingeborg Dhooge, Paul Van de Heyning, Robert Vincent, Erwin Offeciers, Richard J. H. Smith, and Guy Van Camp (2009). A Genome-wide Analysis Identifies Genetic Variants in the RELN Gene Associated with Otosclerosis. *The American Journal of Human Genetics* **84**, 328–338.

[13] Joo, J., Kwak, M., Ahn, K., and Zheng, G. (2009). A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium. *Biometrics*. **65**, 1115-1122.

[14] Klein, R. J. et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*. **308**, 385–389.

[15] Lerner, I. M. (1968). Heredity, Evolution, and Society. *San Francisco: W.H. Freeman*.

[16] Li, W. and Reich, J. (2000). A complete enumeration and classification of two locus disease models. *Hum Hered* **50**, 334–349.

[17] Li, Q., Yu, K., Li, Z., and Zheng, G. (2008). MAX-rank: a simple and robust genome-wide scan for case-control association studies. *Hum. Genet.* **123**, 617–623.

[18] Marchini, J., Donnelly, P., and Cardon, L.R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413–417.

[19] Mills, J. L. and Graubard, B. I. (1987). Is moderate drinking during pregnancy associated with an increased risk for malformations? *Pediatrics* **80**, 309–314.

[20] Neuman, R. J. and Rice J. P. (1992). Two-locus models of disease. *Genet Epidemiol* **9**, 347–365.

[21] Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411. MR0556730

[22] Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics* **53**, 1253–1261. MR1614374

[23] Senn, S. (2007). Drawbacks to noninteger scoring for ordered categorical data. *Biometrics* **63**, 296–298. MR2345600

[24] Song, M. and Nicolae, D. L. (2009). Restricted parameter space models for testing gene-gene interaction. *Genet. Epidemiol.* **33**, 386–393. MR2627365

[25] Suzuki, D. T., Griffiths, A. J. F., Miller, J. H., and Lewontin, R. C. (1986). An Introduction to Genetic Analysis (3rd ed). *San Francisco: W.H. Freeman*.

[26] The Wellcome Trust Case Control Consortium (WTCCC) (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

[27] Zhao, J., Jin, L., and Xiong, M. (2006). Test for interaction between two unlinked loci. *Am J Hum Genet.* **79**, 831–845.

[28] Zheng, G. (2008). Analysis of ordered categorical data: Two score-independent approaches *Biometrics* **64**, 1276–1279.

[29] Zheng, G. and Ng, H. K. T. (2008). Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics*, **9**, 391–399.

[30] Zheng, G., Freidlin, B., Li, Z., and Gastwirth, J. L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical J.* **45**, 335–348. MR1973305

[31] Zheng, G., Joo, J., and Yang, Y. (2009). Pearson's test, trend tests and MAX are all trend tests with different types of scores. *Ann. Hum. Genet.* **73**, 133–140.

Min Yuan
School of Public Health,
Fudan University
Shanghai 200032, China
E-mail address: myuan@fudan.edu.cn

Jinfeng Xu
Department of Statistics and Applied Probability
National University of Singapore
117546, Singapore
E-mail address: staxj@nus.edu.sg

Yaning Yang
Department of Statistics and Finance,
University of Science and Technology of China
Hefei 230026, China
E-mail address: ynyang@ustc.edu.cn

Gang Zheng
Office of Biostatistics Research, DPPS
National Heart, Lung and Blood Institute
Bethesda, MD 20892, USA
E-mail address: zhengg@nhlbi.nih.gov