Published in final edited form as:

Ann Hum Genet. 2018 September ; 82(5): 318–324. doi:10.1111/ahg.12257.

Decomposing Pearson's χ^2 test: a linear regression and its departure from linearity

ZHENGYANG ZHOU^{1,2,4}, HUNG-CHIH KU⁵, GUAN XING⁶, and CHAO XING^{1,2,3}

¹McDermott Center of Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX 75390

²Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390

³Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390

⁴Department of Statistical Science, Southern Methodist University, Dallas, 75275

⁵Department of Mathematical Sciences, DePaul University, Chicago, IL 60604

⁶Gilead Sciences, Inc., 199 East Blaine Street, Seattle, WA 98102

Summary

In case-control genetic association studies, a standard practice is to perform the Cochran-Armitage (CA) trend test under the assumption of the additive model due to its robustness. We could even identify situations it outperformed the analysis model consistent with the underlying inheritance mode. In this paper we analytically reveal the statistical basis that leads to the phenomenon. By elucidating the origin of the CA trend test as a linear regression model, we decompose Pearson's χ^2 test statistic into two components—one is the CA trend test statistic that measures the goodness of fit of the linear regression model, the other measures the discrepancy between data and the linear regression model. Under this framework we show the additive coding scheme, as well as the multiplicative coding scheme, increases the coefficient of determination of the regression model by increasing the spread of data points. We also obtain the conditions under which the CA trend test statistic equals the MAX statistic and Pearson's χ^2 test statistic.

Keywords

trend test; Pearson's chi-squared test; ordinary least squares; linear regression

In genetic association studies, there is no single best test for all situations. If the true genetic model were known, the association test reflecting this model would have optimal power. In practice the Cochran-Armitage (CA) trend test (Cochran, 1954; Armitage, 1955) under the

Corresponding author: Chao Xing, Ph.D., MC 8591, UT Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390, Phone: 214-648-1695, Fax: 214-648-1666, chao.xing@utsouthwestern.edu.

Conflict of Interest

None.

assumption of the additive model, i.e., a dose-response model, is a standard approach in case-control genome-wide association screens (GWAS) for its robustness, as shown in numerous simulation studies (e.g., Gonzalez *et al.*, 2008; Kuo & Feingold, 2010; Li *et al.*, 2009; Loley *et al.*, 2013). We happened to find by simulation that the CA trend test either under the additive assumption or under the multiplicative assumption outperformed that under the recessive assumption even if the true genetic model is recessive when the minor allele frequency is low and the sample size is not large enough (Figure 1). This anomalous phenomenon met the “man bites dog” criterion to be newsworthy (Elston, 1989) and motivated us to investigate the problem analytically. By elucidating the origin of the CA trend test as a test to examine whether there is a trend in binomial proportions across levels of an ordinal variable by a linear regression model, we decompose Pearson’s χ^2 test statistic (Pearson, 1900) by the ordinary least squares (OLS). Under this theoretical framework, the anomalous behavior of the CA trend test can be explained; further, we prove some results previously published in the *Journal* by other investigators (Zheng *et al.*, 2009).

Consider a diallelic locus with the major and minor alleles denoted as a and A , respectively, in a case-control study (Table 1). Denote by r_i and s_i the numbers of cases and controls, respectively, in genotype category G_i , where $i \in \{0, 1, 2\}$ reflects the number of A alleles a subject has. Thus G_0 , G_1 , and G_2 correspond to genotypes aa , Aa , and AA , respectively. Denote by R , S , and n_i the marginal sums such that $R = \sum_{i=0}^2 r_i$, $S = \sum_{i=0}^2 s_i$, and $n_i = r_i + s_i$, and denote by N the total sample size such that $N = R + S = \sum_{i=0}^2 n_i$. Assume r_i ’s are drawn from binomial distributions $\text{Bin}(n_i, \pi_i)$. The null hypothesis of no association between the disease and genotype is $H_0: \pi_0 = \pi_1 = \pi_2$. Note that π_i is the probability of a subject with genotype G_i being affected conditional on the marginal counts n_i . R and S , different from the penetrance function $f_i = P(\text{affected}|G_i)$ defined in a general population. Equivalently, we can also assume (r_0, r_1, r_2) follow a trinomial distribution with parameters R and (τ_0, τ_1, τ_2) , and (s_0, s_1, s_2) follow a trinomial distribution with parameters S and (ν_0, ν_1, ν_2) . The null hypothesis is then $H_0: \tau_i = \nu_i$.

Assuming G_0 , G_1 , and G_2 are three ordered categories, a more restricted alternative hypothesis is $H_1: \pi_0 \leq \pi_1 \leq \pi_2$ or $\pi_0 \geq \pi_1 \geq \pi_2$ with at least one strict inequality. Assign a set of scores (x_0, x_1, x_2) to G_0 , G_1 , and G_2 , respectively, with the constraints $x_0 \leq x_1 \leq x_2$ and $x_0 < x_2$. The CA trend test examines whether there is a linear relationship between π_i ’s and x_i ’s by fitting a linear regression model $\pi_i = \alpha + \beta x_i + e_i(1)$, where e_i is a random error term with mean equal to zero and variance equal to σ^2 . The null hypothesis is $H_0: \beta = 0$. The

CA trend test statistic is $T_{CA} = \frac{N(N \sum_{i=0}^2 r_i x_i - R \sum_{i=0}^2 n_i x_i)^2}{RS[N \sum_{i=0}^2 n_i x_i^2 - (\sum_{i=0}^2 x_i n_i)^2]}$. Under the null hypothesis,

T_{CA} follows a χ^2 distribution with 1 degree-of-freedom (d.f.). In practice score sets of $(0, 0, 1)$, $(0, 0, 5, 1)$, and $(0, 1, 1)$ are typically used to represent recessive, additive, and dominant models, respectively.

Above we review the data structure of a genetic association study and the CA trend test used for analysis; below we first describe the simulation models and results, and then derive analytical results to interpret the simulation results. Set the genotype group G_0 , i.e., aa , as

the reference group. Define genotype relative risk as $\lambda_i = f_i/f_0$, thus $\lambda_i = 1$. For simulation purpose we adopt a reparametrization strategy (Zheng *et al.*, 2009) by defining $\lambda_1 = 1 + \lambda \cos \theta$ and $\lambda_2 = 1 + \lambda \sin \theta$, where $\lambda \geq 0$ is the distance between point $P = (\lambda_1, \lambda_2)$ and point $O = (1, 1)$, and $\theta \in [\pi/4, \pi/2]$ is the angle between OP and the horizontal line in a two-dimensional space. Thus, θ determines the genetic model and λ determines how far the genetic model is from the null. The null hypothesis can be rewritten as $H_0: \lambda = 0$. In terms of genetic models, $\theta = \pi/4$, $\arctan 2$, and $\pi/2$ correspond to dominant, additive, and recessive models, respectively. We performed simulations under the following alternative settings. Assume a disease prevalence (K) of 0.1 and the minor allele A frequencies $q \in (0.05, 0.1, 0.15, 0.2)$. Fix the alternative hypothesis as $\lambda = 1$ and vary the genetic models by setting $\theta' = \theta/\pi$ from $\arctan 2$ to $1/2$, i.e., from the additive model to the recessive model, with an increment of 0.01. Under each genetic model, penetrances are determined by $f_0 = K/[(1 - q)^2] + 2\lambda_1 q(1 - q) + \lambda_2 q^2$ and $f_i = \lambda_i f_0$. The probabilities of the two trinomial distributions for cases and controls are then $\tau_i = P(G_i) f_i / K$ and $v_i = P(G_i) (1 - f_i) / (1 - K)$, respectively. Assume a balanced design, i.e., $R = S$ with the total sample size $N \in (300, 500, 800, 1000)$. For each model, 10,000 replicates are simulated and each dataset is examined by the CA trend test under the additive model with the score set $(0, 0, 5, 1)$, denoted as T_{CA}^{Add} , under the recessive model with the score set $(0, 1, 1)$, denoted as T_{CA}^{Rec} , and under the multiplicative model with the score set $(0, \frac{1}{\sqrt{\lambda_2} + 1}, 1)$, denoted as T_{CA}^{Mul} . Note that there are unlimited ways to assign the multiplicative model scores. As $\lambda_2 > \lambda_1$ when $\theta \in [\pi/4, \pi/2]$ under the alternative hypothesis, we fix the upper limit λ_2 , which leads to a multiplicative model score set $(1, \sqrt{\lambda_2}, \lambda_2)$. It can be further transformed into $(0, \frac{1}{\sqrt{\lambda_2} + 1}, 1)$ because the CA trend test is invariant to the linear transformation of the scores (Tarone & Gart, 1980). The empirical power at the 0.05 level is calculated as the proportion of the 10,000 replicates for which the P-value is less than or equal to 0.05.

First, in most of the parameter space T_{CA}^{Add} and T_{CA}^{Mul} have comparable power except that T_{CA}^{Mul} is more powerful when the underlying model is close to recessive (Figure S1). It can be predicted because $\frac{1}{\sqrt{\lambda_2} + 1}$ is constrained between approximately 0.414, when $\theta = \pi/4$, and 0.421, when $\theta = \arctan 2$ according to the simulation setting. The multiplicative model lies between the recessive and additive models but in close proximity to the additive end. Therefore overall T_{CA}^{Add} and T_{CA}^{Mul} have comparable power, but under near recessive models the latter is closer to the truth, and thus outperforms the former. Regardless of the simulation setting, the upper limit of $\frac{1}{\sqrt{\lambda_2} + 1}$ is 0.5 because $\lambda_2 > 1$, and $\frac{1}{\sqrt{\lambda_2} + 1} = 0.4$ when $\lambda_2 = 2.25$, which is greater than most GWAS hits (see e.g., Hodge & Greenberg, 2016). Therefore, the similarity between T_{CA}^{Add} and T_{CA}^{Mul} in terms of power exists in general.

Second, in most of the parameter space the results are as expected that T_{CA}^{Add} and T_{CA}^{Mul} are more powerful when the underlying genetic model is close to additive, whereas T_{CA}^{Rec} is more

powerful when the underlying model is close to recessive (Figure S1). However, T_{CA}^{Add} and T_{CA}^{Mul} are more powerful than T_{CA}^{Rec} at the recessive end when the minor allele frequency is low and the sample size is small (Figures 1 & S1), although T_{CA}^{Add} and T_{CA}^{Mul} are just above the nominal test level and T_{CA}^{Rec} is below the nominal level. In other words, the wrong analysis model outperforms the true model that is used for data simulation in those situations. When the sample sizes are 300 or 500, the power of T_{CA}^{Add} and T_{CA}^{Mul} are greater than that of T_{CA}^{Rec} ; with the increase of sample sizes, the power of T_{CA}^{Rec} increases to surpass that of T_{CA}^{Add} and T_{CA}^{Mul} .

Here we decompose Pearson's χ^2 test statistic by the theory of OLS. Unlike the CA trend test, Pearson's χ^2 test is an omnibus test for independence in a contingency table without pre-assigning scores. In terms of genetics, it makes no assumption about the genetic model. It tests the alternative hypothesis of not all π_i 's are equal. The test statistic is

$$T_P = \sum_{i=0}^2 \frac{(r_i - n_i R/N)^2}{n_i R/N} + \sum_{i=0}^2 \frac{(s_i - n_i S/N)^2}{n_i S/N}. \text{ Under the null hypothesis, } T_P \text{ follows a } \chi^2$$

distribution with 2 d.f.. Pearson's χ^2 test statistic T_P can be decomposed into two components—one is the CA trend test statistic T_{CA} that measures the goodness of fit of the linear regression model (1), the other, denoted as T_{SSE} , measures the discrepancy between the data and model (1). The derivation details are described in Appendix I, and here we summarize the results. Denote by $p_i (= r_i/n_i)$ the observed value of π_i , which has a weight of n_i , by $p = R/N$ the overall proportion of cases, and by $\bar{x} = \sum_{i=0}^2 n_i x_i / N$ the average score. By

$$\text{OLS we can estimate for model (1) that } \hat{\beta} = \frac{\sum_{i=0}^2 n_i (p_i - p)(x_i - \bar{x})}{\sum_{i=0}^2 n_i (x_i - \bar{x})^2} \quad (2) \text{ and } \hat{\pi}_i = p + \hat{\beta}(x_i - \bar{x})$$

(3). It can be shown

$$T_P = \frac{1}{p(1-p)} \sum_{i=0}^2 n_i (p_i - p)^2 = \frac{1}{p(1-p)} \sum_{i=0}^2 n_i (\hat{\pi}_i - p)^2 + \frac{1}{p(1-p)} \sum_{i=0}^2 n_i (p_i - \hat{\pi}_i)^2 \quad (4) \text{ and}$$

$$T_{CA} = \frac{1}{p(1-p)} \sum_{i=0}^2 n_i (\hat{\pi}_i - p)^2. \text{ Note that } \sum_{i=0}^2 n_i (p_i - p)^2, \sum_{i=0}^2 n_i (\hat{\pi}_i - p)^2, \text{ and}$$

$\sum_{i=0}^2 n_i (p_i - \hat{\pi}_i)^2$, are total sum of squares (SSTO), regression sum of squares (SSR), and error sum of squares (SSE), respectively, for model (1); thus we define

$$T_{SSE} = \frac{1}{p(1-p)} \sum_{i=0}^2 n_i (p_i - \hat{\pi}_i)^2. \text{ In summary, equation (4) is identical to } T_P = T_{CA} + T_{SSE},$$

and also $\frac{SSTO}{p(1-p)} = \frac{SSR}{p(1-p)} + \frac{SSE}{p(1-p)} \quad (5)$, which decompose T_P into a component measuring the goodness of fit of the linear regression model (1) and a component measuring the departure of data from linearity. Note that $p(1-p)$ can be regarded as the pooled variance of π_i 's; by Cochran's theorem (Cochran, 1934) T_P , T_{CA} , and T_{SSE} follow χ^2 distributions with 2, 1, and 1 d.f., respectively.

According to the analytical work above on decomposing Pearson's χ^2 test statistic T_P we hypothesize that the additive and multiplicative coding scheme in the linear regression model (1) fits data better than the recessive coding scheme in case of low allele frequencies and relatively small sample sizes, even though the true mode of inheritance is recessive,

which explains why T_{CA}^{Add} and T_{CA}^{Mul} are more powerful than T_{CA}^{Rec} . We provide a numerical example to illustrate the point (Figure 2). A 2×3 table with $(r_0, r_1, r_2) = (443, 52, 5)$ and $(s_0, s_1, s_2) = (465, 34, 1)$ was generated by simulation under the recessive model with $K = 0.1$, $q = 0.05$, and $R = S = 500$. When analyzing under the additive model, the coefficient of determination $R^2 (= SSR/SSTO)$, which measures the goodness of fit of the linear regression model (1), equals 0.0067; $R^2 = 0.0069$ when analyzing under the multiplicative model; in contrast $R^2 = 0.0027$ when analyzing under the recessive model. Therefore, both the additive and the multiplicative models fit the data better than the recessive model. Correspondingly, the association is significant under the additive model (P -value=0.0095) and under the multiplicative model (P -value=0.0086) but insignificant under the recessive model (P -value=0.1014).

Further we show that the additive and multiplicative coding schemes lead to better fit of the linear regression model (1) than the recessive coding scheme partially by increasing the spread of the data points. Denote the sample variance of the scores by s_x^2 , which equals $\sum_{i=0}^2 n_i(x_i - \bar{x})^2 / (N - 1)$. By OLS we know $Var(\hat{\beta}) = \sigma^2 / [(N - 1)s_x^2]$ and equation (3). Given a dataset $SSTO$ is fixed. Thus,

$$E(R^2) = E(SSR)/SSTO = E(\sum_{i=0}^2 n_i(\hat{\pi}_i - p)^2)/SSTO = (N - 1)s_x^2 E(\hat{\beta}^2)/SSTO = [\sigma^2 + (N - 1)s_x^2 \beta^2] / SSTO$$

which shows $E(R^2)$ is a monotonically increasing function of s_x^2 . Therefore, the more spread out the scores are, the better a liner regression model could potentially fit the data. However, note that $E(R^2)$ is also a monotonically increasing function of β^2 , which will be different with different score sets. The overall goodness of fit depends on $s_x^2 \beta^2$. For the same numerical example above, under the additive model $s_x^2 = 0.025$, $E(\hat{\beta}) = 0.259$, and thus $s_x^2 [E(\hat{\beta})]^2 = 0.00168$; under the multiplicative model $s_x^2 = 0.019$, $E(\hat{\beta}) = 0.301$, and thus $s_x^2 [E(\hat{\beta})]^2 = 0.00173$; under the recessive model, $s_x^2 = 0.006$, $E(\hat{\beta}) = 0.335$, and $s_x^2 [E(\hat{\beta})]^2 = 0.00067$.

With the analytical work elucidating the CA trend test and decomposing Pearson's χ^2 test statistic T_B we can easily prove some results previously obtained by other investigators.

Zheng et al. (Zheng *et al.*, 2009) showed that the CA trend test with the score set

$$\left(0, \frac{p_1 - p_0}{p_2 - p_0}, 1\right), \text{ denoted as } T_{CA}^{Best}, \text{ equals } T_B; \text{ moreover, if } 0 < x_1 < 1, T_{CA}^{Best} \text{ equals the so-called}$$

MAX statistic (Freidlin *et al.*, 2002), which is defined as $\max_{x_1 \in [0, 1]} T_{CA}^{(0, x_1, 1)}$. By OLS we

have equations (2) and (3). When the score set is $\left(0, \frac{p_1 - p_0}{p_2 - p_0}, 1\right)$, it can be shown that equation (2) leads to $\hat{\beta} = p_2 - p_0$, and equation (3) leads to $\hat{\pi}_i = p_i$. It indicates the three points $(0, p_0)$, $\left(\frac{p_1 - p_0}{p_2 - p_0}, p_1\right)$, and $(1, p_2)$, are right on the regression line, and the fitted values equal the

observed values (Figure 2D). As such, $SSE = 0$, and thus $T_P = T_{CA}^{Best}$. Note that this equation is valid regardless of the range of x_1 . Therefore $\max_{x_1 \in [0, 1]} T_{CA}^{(0, x_1, 1)} \leq T_{CA}^{Best}$, and the equality holds when $0 \leq x_1 \leq 1$. For the same numerical example above, if the score set $\left(0, \frac{p_1 - p_0}{p_2 - p_0}, 1\right)$ could be used, the association is most significant (P -value=0.0083) among all tested models. Moreover, $s_x^2 = 0.015$, which is smaller than that under either the additive model or the multiplicative model; however, $E(\hat{\beta}) = 0.345$, and thus $s_x^2[E(\hat{\beta})]^2 = 0.00174$, which is the greatest among all the tested models.

We previously showed Pearson's χ^2 test is more powerful than T_{CA}^{Add} in case of the heterozygote advantage model and provided a theoretical argument from the distribution perspective that Pearson's χ^2 test examines both the location and dispersion shifts, whereas T_{CA}^{Add} only examines the location shift (Zhou *et al.*, 2017). Under the heterozygote advantage model the dispersion information overshadows the location information, therefore Pearson's χ^2 test is more powerful. The analytical work in this paper provides justification from another perspective that T_{CA}^{Add} only tests the goodness of fit of the linear model (1), whereas T_P also includes a component measuring the departure of data from linearity. Under the heterozygote advantage model $\lambda_1 < 1$ and $\lambda_2 > 1$, i.e., severe departure from linearity, therefore Pearson's χ^2 test is more powerful. Although we started this project with an observation that the additive coding scheme outperforms the recessive coding scheme even the true genetic model is recessive, note from previous simulation studies (e.g. Gonzalez *et al.*, 2008; Kuo & Feingold, 2010; Li *et al.*, 2009; Loley *et al.*, 2013) that T_{CA}^{Add} overall performs worse when the true genetic model is recessive than when the model is dominant or additive. The underlying reason is that under a recessive model there is less information on mean score difference between cases and controls, which the CA trend test measures.

The CA trend test under the additive model shows robustness across various genetic models. Intuitively it makes sense in that it lies in the middle of the parameter space between the two extreme ends. By the analytical work in this paper we speculate a numerical reason is that the additive coding scheme overall leads to a good fit of the linear regression model (1) by increasing the spread of the data points. Note that the anomalous behavior of the CA trend test observed in this paper happens only to binary traits, but not to continuous traits (data not shown). It can also be deduced from the linear regression nature of the CA trend test. Compared to the additive coding scheme, the recessive coding scheme not only leads to less spread of the data points in terms of scores, in case of binary traits it also leads to heteroscedasticity because subjects with the same score 0 consist of two groups π_0 and π_1 (Figure 2C). On the contrary this issue does not arise in case of continuous traits.

Following Sasieni's seminal paper (Sasieni, 1997) T_{CA}^{Add} became the standard practice in GWAS. In recent years, the generalized linear model, in particular, logistic regression for a binary trait, is advocated for its flexibility in modeling different types of traits and including

covariates, as well as providing estimates of parameters of interest such as odds ratio (Dizier *et al.*, 2017; Loley *et al.*, 2013; So & Sham, 2011; Wellek & Ziegler, 2012). One study closely related to our work is on the general regression model (GRM) (Dizier *et al.*, 2017). It tests for both additive effect and deviation from additive effect, which is analogous to our decomposing T_P into a component measuring the goodness of fit of the linear regression model and a component measuring the departure of data from linearity. Here we focus on elucidating the characteristics of the CA trend test to explain the anomalous behavior of the test in a certain circumstance. The superior power of GRM in GWAS warrants further investigation on how to modify the CA trend test to increase power based on the current analytical work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the reviewers for their constructive comments that tremendously improve the manuscript. CX was partially supported by the NIH grant UL1TR001105. Author contributions— Study design ZZ, GX, CX; Theoretical analysis: all authors; Simulation: ZZ, HK, CX; Manuscript preparation: ZZ, GX, CX.

References

- Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics*. 1955; 11:375–386.
- Cochran WG. The distribution of quadratic forms in a normal system with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*. 1934; 30:178–191.
- Cochran WG. Some methods for strengthening the common chi-square tests. *Biometrics*. 1954; 10:417–451.
- Dizier MH, Demenais F, Mathieu F. Gain of power of the general regression model compared to Cochran-Armitage Trend tests: simulation study and application to bipolar disorder. *BMC Genet*. 2017; 18:24. [PubMed: 28283021]
- Elston RC. Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. *Am J Med Genet*. 1989; 34:487–8. [PubMed: 2624257]
- Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered*. 2002; 53:146–52. [PubMed: 12145550]
- Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V. Maximizing association statistics over genetic models. *Genet Epidemiol*. 2008; 32:246–54. [PubMed: 18228557]
- Hodge SE, Greenberg DA. How can we explain very low odds ratios in GWAS? I. polygenic models. *Hum Hered*. 2016; 81:173–180. [PubMed: 28171865]
- Kuo CL, Feingold E. What's the best statistic for a simple test of genetic association in a case-control study? *Genet Epidemiol*. 2010; 34:246–253. [PubMed: 20025064]
- Li Q, Zheng G, Liang X, Yu K. Robust tests for single-marker analysis in case-control genetic association studies. *Ann Hum Genet*. 2009; 73:245–52. [PubMed: 19208106]
- Loley C, König IR, Hothorn L, Ziegler A. A unifying framework for robust association testing, estimation, and genetic model selection using the generalized linear model. *Eur J Hum Genet*. 2013; 21:1442–8. [PubMed: 23572026]
- Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*. 1900; 50:157–175.
- Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics*. 1997; 53:1253–61. [PubMed: 9423247]

- So HC, Sham PC. Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates. *Behav Genet.* 2011; 41:768–75. [PubMed: 21305351]
- Tarone RE, Gart JJ. On the robustness of combined tests for trends in proportions. *Journal of the American Statistical Association.* 1980; 75:110–116.
- Wellek S, Ziegler A. Cochran-Armitage test versus logistic regression in the analysis of genetic association studies. *Hum Hered.* 2012; 73:14–7. [PubMed: 22212245]
- Zheng G, Joo J, Yang Y. Pearson's test, trend test, and MAX are all trend tests with different types of scores. *Ann Hum Genet.* 2009; 73:133–40. [PubMed: 19183346]
- Zhou Z, Ku HC, Huang Z, Xing G, Xing C. Differentiating the Cochran-Armitage trend test and Pearson's χ^2 test: location and dispersion. *Ann Hum Genet.* 2017; 81:184–189. [PubMed: 28653322]

Appendix

Proof of $T_P = T_{CA} + T_{SSE}$

Pearson's χ^2 statistic

$$T_P = \sum_{i=0}^2 \frac{(r_i - n_i R/N)^2}{n_i R/N} + \sum_{i=0}^2 \frac{(s_i - n_i S/N)^2}{n_i S/N} = \sum_{i=0}^2 \frac{n_i(p_i - p)^2}{p} + \sum_{i=0}^2 \frac{n_i(p_i - p)^2}{1-p}. \text{ Given } \hat{\pi}_i$$

$$= \frac{1}{p(1-p)} \sum_{i=0}^2 n_i(p_i - p)^2$$

$$= p + \hat{\beta}(x_i - \bar{x}) \text{ and } \hat{\beta} = \frac{\sum_{i=0}^2 n_i(p_i - p)(x_i - \bar{x})}{\sum_{i=0}^2 n_i(x_i - \bar{x})^2}, \text{ by some algebra manipulations, it can be}$$

shown $\sum_{i=0}^2 n_i(p_i - \hat{\pi}_i)(\hat{\pi}_i - p) = 0$. Therefore,

$$T_P = \frac{1}{p(1-p)} \sum_{i=0}^2 n_i(p_i - \hat{\pi}_i + \hat{\pi}_i - p)^2 = \frac{1}{p(1-p)} \sum_{i=0}^2 n_i[(p_i - \hat{\pi}_i)^2 + (\hat{\pi}_i - p)^2 + 2(p_i - \hat{\pi}_i)(\hat{\pi}_i - p)]$$

$$= \frac{1}{p(1-p)} \sum_{i=0}^2 n_i(\hat{\pi}_i - p)^2 + \frac{1}{p(1-p)} \sum_{i=0}^2 n_i(p_i - \hat{\pi}_i)^2$$

(6). Again by some algebra manipulations, it can be shown

$$\frac{1}{p(1-p)} \sum_{i=0}^2 n_i(\hat{\pi}_i - p)^2 = \frac{N(N \sum_{i=0}^2 r_i x_i - R \sum_{i=0}^2 n_i x_i)^2}{RS[N \sum_{i=0}^2 n_i x_i^2 - (\sum_{i=0}^2 n_i x_i)^2]}, \text{ which is exactly the CA trend test}$$

statistic T_{CA} . Define $T_{SSE} = \frac{1}{p(1-p)} \sum_{i=0}^2 n_i(p_i - \hat{\pi}_i)^2$. Therefore equation (6) is identical to

$$T_P = T_{CA} + T_{SSE}$$

Figure 1. Power comparison of the Cochran-Armitage trend test by the additive, multiplicative, and recessive models under recessive modes of inheritance for low-frequency variants
On the simulation models, the relative risks for Aa and AA are defined as $\lambda_1 = 1 + \cos\theta$ and $\lambda_2 = 1 + \sin\theta$, respectively. Therefore, $\theta \in (0.47\pi, 0.48\pi, 0.49\pi, 0.50\pi)$ ranges from near-recessive models to the recessive model; The labels “Additive”, “Multiplicative”, and “Recessive” indicate the analysis models. The y-axis is the empirical power based on 10,000 replicates at the 0.05 level. The disease prevalence equals 0.1; the minor allele frequency equals 0.05; and the total sample size $N \in (300, 500, 800, 1000)$ with equal numbers of cases and controls.

Figure 2.

Comparison of the Cochran-Armitage trend test by different analysis models under the recessive mode of inheritance: a numerical example

A 2×3 table with $(r_0, r_1, r_2) = (443, 52, 5)$ and $(s_0, s_1, s_2) = (465, 34, 1)$ was generated by simulation under the recessive model with $K = 0.1$, $q = 0.05$, and $R = S 500$. It was analyzed using the CA trend test assuming (A) the additive model ($x_0 = 0, x_1 = 0.5, x_2 = 1$), (B) the

multiplicative model $\left(x_0 = 0, x_1 = \frac{1}{\sqrt{\lambda_2} + 1} = \sqrt{2} - 1, x_2 = 1\right)$, (C) the recessive model ($x_0 = 0,$

$x_1 = 0, x_2 = 1$), and (D) the "best" model $\left(x_0 = 0, x_1 = \frac{p_1 - p_0}{p_2 - p_0} \cong 0.338, x_2 = 1\right)$. The x -axis is

the scores used for the genotype groups in the CA trend test; the y -axis is the proportion of cases in each genotype group; the area of the circles is proportional to the counts; the straight line is the fitted regression line based on model (1); and R^2 is the corresponding coefficient of determination.

Table 1

Genotype distribution at a diallelic marker in a case-control study

Phenotype	Genotype			Total
	<i>aa</i>	<i>Aa</i>	<i>AA</i>	
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript