Exact Conditional Efficiency Robust P-values from an Arbitrary Ranking of a Sample Space: an Application to Genome-wide Association Studied

Max Moldovan[1] and Mette Langaas[2]

[1] *Australian Institute of Health Innovation, University of New South Wales, Level 1 AGSM Building, Sydney NSW 2052, Australia;*

[2] *Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*

Abstract

We introduce a general method for computation of exact conditional efficiency robust enumeration p-values for detection of genotype-phenotype associations at a single bi-allelic genetic locus. Our method can be based on any arbitrary ranking test statistics, such as efficiency robust test statistics or asymptotic p-values. The resulting p-values are exact conditional enumeration p-values and satisfy the basic statistical validity property $Pr(P \leq \alpha|H_0) \leq \alpha$ for all parameters under the null hypothesis and all significance levels α. Practically, the method allows performing statistically valid significance testing in genomic analyses with unknown modes of inheritance at individual bi-allelic genetic loci - the situation typical in genome-wide association studies. We provide an open-source R code implementing the method.

Keywords: mode of genetic inheritance; efficiency robust statistics; exact conditional inference; enumeration; genome-wide association study.

1 Introduction

Genome-wide association studies (GWAS) consider hundreds of thousands of single nu-cleotide polymorphisms (SNPs) covering the entire human genome. Each SNP is nor-mally represented by a bi-allelic locus and assessed for association with a speci?c genetic trait, usually in the context of a case-control study. Complex diseases such as asthma, diabetes and multiple sclerosis, among many others, are generally targeted by GWAS in order to identify common genetic variations as potential disease risk factors. The number of genetic markers in a particular GWAS can vary from several hundred thousands to several millions, depending on the platform used for genotyping and the type of genomes to be studied. For example, more SNPs are required for GWAS that utilize African pop-ulations than for GWAS involving European populations since the former are older, and thus being exposed to random gene recombinations for more genorations. See Hirschhorn and Daly[1] and Manolio[2] for the interesting well illustrated introductions to GWAS.

At a given genetic locus, there is a pair of markers, called alleles, inherited from each of two parents. Given a trait is passed through this locus, there are

several modes of inheritance can be in effect. The dominant mode of inheritance requires the presence of a single "disease" allele from one of parents for a trait to be inherited. The recessive mode of inheritance requires "disease" alleles from both parents to be passed to the offspring for a trait to express. In case of the additive mode of inheritance, a trait is expressed only partly if a single disease allele is inherited, but express in full if both "disease" alleles are in place. There are a few more modes of inheritance can be specified depending on the degree to which a trait is expressed in an offspring, see Visscher et al.[3] for the discussion of heritability concepts.

When the mode of inheritance at a genetic locus is known, higher power of a test for genotype-phenotype association can be achieved through using the Cochran-Armitage trend test (CATT) under the explicit assumption of the specific genetic model, see Lettre et al.[4] and Gonzlez et al.[5]. In practice, however, the mode of inheritance is usually unknown. Under this typical scenario, the so-called efficiency robust tests (see Podgor et al.[6]) can be used, the tests which remain sensitive to detection of genotype-phenotype associations even though the genetic model is either unknown or misspecified.

There are several efficiency robust testing strategies. For example, the MAX test, first suggested by Freidlin et al.[7], has been recommended by the several authors, see Zheng and Gastwirth [8] and Gonzlez et al.[5]. This testing approach is implemented as aa sequential application of several statistical tests optimal for alternative genetic models with retaining the most significant result. The traditional version of the MAX test, normally referred to as MAX3, is based on the three CATTs with scores motivated by dominant, recessive and additive genetic models. Alternatively, Persons chi-square test (χ^2) can be included within the same MAX testing strategy, leading to MAX4, see Li et al.[9]. Zheng et al.[10] demonstrated that χ^2 test can be considered as a type of a trend test and also noted that this test is sensitive to detection of overdominant (underdominant) modes of inheritance. MIN2 is one more variant of the MAX test implemented as a combination of the additive CATT and χ^2 , see Joo et al.[11]. A slightly different efficiency robust testing strategy is known as MERT and is a weighted version of CATT optimal for recessive and dominant models, see Gastwirth[12] and Freidlin et al.[7]. There are several more efficiency robust testing approaches can be specified and some authors even suggest applying a combination of different versions of efficiency robust tests within a single testing procedure, see Joo et al.[11].

In finite sample settings, many of the currently known and used efficiency robust tests are not guaranteed to lead to statistically valid inference. This is because the under-lying computational procedures are based either on random sampling or on asymptotic distributions of efficiency robust statistics (Gonzlez

et al.[5], Joo et al.[13], So anda Sham [14]). For the methods that use random permutations (see Sladek et al.[15]), the statistical inference will be valid, but a very high number of random permutations is needed to achieve the required precision for traditionally low GWAS type significance levels (often in the order of 10^{-8}). Recently, Loley et al.[16] attempted to unify the efficiency robust testing approaches by proposing a framework also leading to inference of unknown statistical validity.

In the current paper, we introduce a computational procedure that takes as an input the ordering of a sample space imposed by any of test statistics or p-values, including the ones introduced above. The procedure outputs exact conditional enumeration p-values that satisfy the basic validity property $Pr(P \leq \alpha | H_0) \leq \alpha$, for all parameters under the null hypothesis and all significance levels α.

2 Notation and the Method

Let the information on a single SNP be represented by the 23 contingency table given by Table 1, where x_i and y_i are the counts of observed genotypes for n_1 cases and n_2 controls, respectively, with $n = n_1 + n_2$. We denote this empirically observed table by $s * (x_1, x_2 | m_1, m_2, n_1, n_2)$, because all the other entries of the table can be calculated from these numbers. Note that for given n_1 , n_2 , m_1 and m_2 , there is a finite number of possible contingency tables, called a reference set (Verbeek[17]) and denoted here by (m_1, m_2, n_1, n_2). Next let T be an arbitrary ranking statistic with the value t corresponding to the empirically observed table $s * (x_1, x_2 | m_1, m_2, n_1, n_2)$. Given a general hypothesis of 'H_0 : no association between genotypes and the case-control status of the subjects' tested against 'H_A : there is association between genotypes and the case-control status of the subjects', and larger values of T being more hostile to the null H_0 , the set of tables ranked lower or equal than the observed table $s * (x_1, x_2 | m_1, m_2, n_1, n_2)$ is given by the critical set:

$$R(x_1, x_2 | m_1, m_2, n_1, n_2) := \{ s(i, j | m_1, m_2, n_1, n_2) : T \geq t \} \tag{1}$$

By definition, a p-value is the probability of obtaining the outcome as extreme or

Table 1 Genotype counts at a bi-allelic locus.

	AA	AB	BB	Total
Case	x_1	x_2	x_3	n_1
Control	y_1	y_2	y_3	n_2
Total	m_1	m_2	m_3	n

worse than the empirically observed outcome $s * (\cdot)$ under the null, which is just the probability of the critical set $R(\cdot)$. Under the null and based on the assumed

underlying hypergeometric sampling scheme (see Lehmann[18] for descriptions of alternative sampling schemes), the probability of obtaining each individual table $s(i,j|m_1,m_2,n_1,n_2)$ within the reference set can be computed as follows:

$$f(i,j|m_1,m_2,n_1,n_2) = \frac{\begin{pmatrix} m_1 \\ i \end{pmatrix}\begin{pmatrix} m_2 \\ j \end{pmatrix}\begin{pmatrix} n-m_1-m_2 \\ n-i-j \end{pmatrix}}{\begin{pmatrix} n \\ n_1 \end{pmatrix}} \tag{2}$$

See Lloyd[19] for the generalization of the central multivariate hypergeometric probability function given by (2). The p-value $p_s*;T$ corresponding to the empirically observed table $s*(x_1,x_2|m_1,m_2,n_1,n_2)$ is the probability of the critical set $R(\cdot)$ given by (1):

$$ps*,T(x_1,x_2|m_1,m_2,n_1,n_2) = \sum_{s\in R} f(i,j|m_1,m_2,n_1,n_2) \tag{3}$$

Note that p_s*,T is a Fisher-type conditional p-value by construction, inheriting positive(e.g. statistical validity and empirical relevance) as well as negative (e.g. conservatism and computational challenges) aspects of Fisher's p-values.

3 Numerical Illustration

Denote statistics obtained from CATTs optimal for dominant, recessive and additive models, respectively, by T_D, T_R and T_A (Sasieni[20]):

$$T_D = \frac{\mathrm{n(nx_1 - n_1m_1)^2}}{n_1m_1(n-n_1)(n-m_1)}$$

$$T_R = \frac{\mathrm{n(nx_3 - n_1m_3)^2}}{n_1(n-n_1)(nm_3-m_3^2)}$$

$$T_A = \frac{\mathrm{n(n(x_2+2x_3) - n_1(m_2+2m_3))^2}}{\mathrm{n_1(n-n_1)(n(m_2+4m_3) - (m_2+2m_3)^2)}}$$

All three statistics asymptotically follow the chi-square distribution with one degree of freedom. The MAX3 test statistic is given by $T_{MAX3} = max(T_D,T_R,T_A)$ with the observed value $t_{MAX3} = max(t_D,t_R,t_A)$. For the empirically observed table $s*(0,2|3,4,4,5)$,the p-value p_s*,T_{MAX3} can be computed as shown in Table 2. Specifically, there are 11 tables in the reference set $S(3,4,4,5)$ and only three tables in the critical set $R(0,2|3,4,4,5)$ since only the tables with the values of T_{MAX3} statistics equally or more extreme then observed are included in the critical set, i.e. $T_{MAX3} \geq t_{MAX3}$. The resulted exact conditional efficiency robust p-value $p_s*,T(0,2|3,4,4,5) = 0:0952$ and is the sum of $f(\cdot|m_1,m_2,n_1,n_2)$ given

by (2) of the three tables in $R(0,2|3,4,4,5)$.
Table 2 The illustrative example is based on $(m_1, m_2, n_1, n_2) = (3,4,4,5)$ with an observed value $(x_1, x_2) = (0,2)$. The critical region R is given by the lower part of the table under the horizontal line.

Table 2 Genotype counts at a bi-allelic locus.

x_1	x_2	T_D	T_R	T_A	T_{MAX3}	$f(x_1, x_2 \mid m_1, m_2)$
1	2	0.2250	0.0321	0.1636	0.2250	0.2857
2	1	0.9000	0.0321	0.2557	0.9000	0.1905
1	3	0.2250	2.0571	0.2557	2.0571	0.0952
2	2	0.9000	2.0571	2.0045	2.0571	0.1429
1	1	0.2250	3.2143	1.7284	3.2143	0.0952
2	0	0.9000	3.2143	0.1636	3.2143	0.0238
0	3	3.6000	0.0321	1.7284	3.6000	0.0635
0	4	3.6000	2.0571	0.1636	3.6000	0.0079
0	2	3.6000	3.2143	4.9500	4.9500	0.0476
3	0	5.6250	0.0321	2.0045	5.6250	0.0159
3	1	5.6250	2.0571	5.4102	5.6250	0.0317
					$p_s*, T =$	0.0952

4 Conclusion

The method we suggested above is by no means new. The initial idea can be traced back to Fisher [22] and PS;T given by (3) is based on the combinatorial results known for many decades, see Freeman and Halton [23]. Our contribution to the original Fisher's methodology is the idea of ordering the sample space, given by the reference set S, based on any arbitrary chosen ranking statistics, the efficiency robust test statistics in our case. We have borrowed this approach from the unconditional exact testing literature, see Barnard [24] and Lloyd and Moldovan [25] for the origination of the unconditional inference philosophy and one of the initial attempts to combine the conditional and unconditional types of exact inference, respectively.

To conclude, it should be pointed out that only the basic form of the adjustment procedure has been given above. In practice, more special cases can arise, such as the presence of covariates (e.g. additional SNPs, environmental factors or baseline factors) or involvement of additional shifted parameters (e.g. in power studies). While this is clearly the limitation of the presented procedure, the basic general exact conditional method introduced above gives a solid basis for further investigations to these and possibly several more theoretical and applied research directions. We provide an open-source R code to encourage and facilitate such

investigations. The R code is available upon request from the authors.

References

[1] Hirschhorn, J.N., and Daly, M.J. (2005), "Genome-wide association studies for common diseases and complex traits", *Nature Reviews Genetics*, 6, 95-108.

[2] Manolio, T.A. (2010), "Genomewide association studies and assessment of the risk of disease", *New England Journal of Medicine*, 363, 166-176.

[3] Visscher, P.M., Hill, W.G., and Wray, N.R. (2008), "Heritability in the genomics era-concepts and misconceptions.", *Nature Reviews Genetics*, 9, 255-266.

[4] Lettre, G., Lange, C., and Hirschhorn, J.N. (2007), "Genetic model testing and statistical power in population-based association studies of quantitative traits" *Genetic Epidemiology*, 31, 358-362..

[5] Gonzalez, J.R., Carrasco, J.L., Dudbridge, F., Armengol, L., Estivill, X., and Moreno, V. (2008), "Maximizing association statistics over genetic models" *Genetic Epidemiology*, 32, 246-254..

[6] Podgor, M.J, Gastwirth, J.L., and Mehta C.R (1996), "Efficiency robust tests of independence in contingency tables with ordered classi

cations." *Statistics in Medicine*, 15, 2095-2105.

[7] Freidlin, B., Zheng, G., Li, Z. and Gastwirth, J.L. (2002), "Trend tests for casecontrol studies of genetic markers: Power, sample size and robustness", *Human Heredity*, 53, 146-152.

[8] Zheng, G., and Gastwirth, J.L. (2006), "On estimation of the variance in Cochran- Armitage trend tests for genetic association using case-control studies." *Statistics in Medicine*, 25, 3150-3159.

[9] Li, Q., Zheng, G., and Yu, K. (2009), "Robust tests for single-marker analysis in case-control genetic association studies." *Annals of Human Genetics*, 73, 245-252.

[10] Zheng, G., Joo, J., and Yang, Y. (2009), "Pearson's test, trend test, and MAX are all trend tests with different types of scores." *Annals of Human Genetics*, 73, 133-140.

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

[12] Gastwirth, J.L. (1985), "The use of maximin efficiency robust tests in combining contingency tables and survival analysis", *Journal of the American Statistical Association*, 80, 380-384.

[13] Joo, J., Kwak, M., and Zheng, G. (2010), "Improving power for testing genetic association in case-control studies by reducing the alternative space", *Biometrics*, 66, 266-276.

[14] So, H.C., and Sham, P.C. (2011), "Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates", *Behavior Genetics*, 41, 768-775.

[15] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C., and Froguel, P. (2007), "A genome-wide association study identifies novel risk loci for type 2 diabetes", *Nature*, 445, 881-885.

[16] Loley, C.,, I.R., Hothorn, L., and Ziegler, A. (2013), "A unifying framework for robust association testing, estimation, and genetic model selection using the generalized linear model", *European Journal of Human Genetics, to appear*.

[17] Verbeek, A. (1985), "A survey of algorithms for exact distributions of test statistics in $r \times c$ contingency tables with fixed margins", *Computational Statistics and Data Analysis*, 3, pp.159-185.

[18] Lehmann, E.L. (1986), *Testing statistical hypotheses*, 2nd ed., New York: Wiley.

[19] Lloyd, C.J. (1999), *Statistical Analysis of Categorical Data*, New York: Wiley

[20] Sasieni, P.D. (1997), "From genotype to genes: doubling the sample size", *Biomet- rics*, 53, pp.1253-1261.

[21] Devlin, B., and Roeder, K. (1999), "Genomic control for association studies", *Bio- metrics*, 55, pp.997-1004.

[22] Fisher, R.A. (1935), "The logic of inductive inference (with discussion)", *Journal of the Royal Statistical Society*, 98, pp.39-54.

[23] Freeman, G.H., and Halton J.H. (1951), "Note on an exact treatment of contingency, goodness of fit and other problems of significance", *Biometrika*, 38, pp.141-149.

[24] Barnard, G.A. (1947), "Significance tests for 2×2 tables", *Biometrika*, 34, pp.123-138.

[25] Lloyd, C.J., and Moldovan, M. (2007), "Unconditional efficient one-sided confidence limits for the odds ratio based on conditional likelihood", *Statistics in Medicine*, 26, pp.5136-5146.

Corresponding Author

Authors can be contacted at: max.moldovan@gmail.com.