NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Increasing power with the unconditional maximization enumeration test in small samples – a detailed study of the MAX3 test statistic

by

Mette Langaas and Øyvind Bakke

PREPRINT
STATISTICS NO. 1/2013

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2013/S1-2013.pdf
Mette Langaas has homepage: http://www.math.ntnu.no/∼mettela
E-mail: Mette.Langaas@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

Increasing power with the unconditional maximization enumeration test in small samples – a detailed study of the MAX3 test statistic

Mette Langaas Øyvind Bakke

Abstract

We present and compare statistical methods for calculating p-values for discrete distributions in the presence of nuisance parameters in small samples. The methods we consider are asymptotic, conditional, and unconditional, and combinations thereof, where the p-value from one method is used in sequel as a test statistic in another method. We consider tests were for a given significance level we reject the null hypothesis if the p-value is not greater than the significance level. It is well known that the unconditional maximization method yields a valid p-value. This implies that when the unconditional maximization method is applied in sequel to a p-value that is not valid, the new maximization p-value will always be valid. Statistical methods for calculating valid p-values can be ranked according to the power of the corresponding test. When the unconditional maximization method is applied in sequel to a p-value that is valid (refer to the corresponding test as the original test), the new maximization p-value will always be as least as small as the starting p-value. This is true for all possible outcomes, and will thus give a new test with uniformely at least as high power as the original test. The unconditional maximization method can therefore be seen as a post processing tool to increase the power of an existing test. We elaborate on these general findings in a detailed study of the robust MAX3 test statistics for testing associations between a genotype and a phenotype in case–control data for sample sizes on the order of tens. 2×3 contingency table; Cochrane–Armitage trend test; P-value; Robust test statistic; Polymorphism.

1 Introduction

We live in the millennium of *big data*, where scientists routinely collect hundreds to thousands of samples and study tens of thousands of genetic markers, such as in genomewide association (GWA) studies, presented in Sladek *and others* (2007), Djurovic *and others* (2010), and Martinelli-Boneschi *and others* (2012). However, the first step on the road to the important discoveries is often a targeted pilot study involving small sample sizes. For randomized clinical studies, intervention studies and studies of rare diseases, these pilot studies might involve samples that are only on the order of tens, and only one or a few genetic markers (Serra *and others*, 2011). When data are scarce it is of great importance to apply the most powerful statistical method of analysis available.

Many biological and medical studies use statistical hypothesis testing to find an association between a phenotype and a genotype. Different test statistics may be applied and tailored to different alternative hypotheses such as different genetic models as one example. Instead of performing multiple separate hypothesis tests and regarding this as a multiple testing problem, it is possible to combine one test statistic for each alternative hypothesis into a robust test statistic by taking the maximum over the different test statistics, for example. This may result in a robust test statistic with an unknown parametric distribution, like the MAX3 test statistic of Freidlin *and others* (2002) that is the subject of presentation.

There are a number of comparative studies that can help researchers choose the most powerful test statistic available, such as for the association between a dichotomous phenotype and genotype (see Zheng *and others*, 2006; Joo *and others*, 2009), but few studies compare different methods for calculating p-values based on one test statistic. The methods for calculating p-values that we discuss are asymptotic, conditional and unconditional methods, all based on the same test statistic.

There is a large body of literature on hypothesis testing in 2×2 contingency tables, where conditional tests are often found to be less powerful than unconditional alternatives, as described by Mehrotra *and others* (2003) and Lydersen *and others* (2009). Our focus is on genotype–phenotype association for a biallelic marker in a case–control setting, which means that the data are presented in a 2×3 contingency table. Due to the less discrete nature of higher order $(r \times c)$ contingency tables, some researchers have found that the power advantage of the unconditional test over the conditional test tends to be less pronounced than for the 2×2 contingency table (Mehta and Hilton, 1993).

Our presentation is organized as follows. In Section 2 we present the MAX3 test statistic, background and genetic notation. In Section 3 we look at how to calculate exact power, and stress the concept of validity of a p-value. We then present four general methods for calculating a p-value, and look at how these methods can be combined in Section 4. Section 5 presents the results from conducting a large study on method validity and power, we discuss and conclude in Sections 6 and 7.

2 Association between genotype and phenotype in case–control studies using the robust MAX3 test statistic

Assume that genotype and phenotype data are collected in a case–control study, and that the genotype data come from one biallelic genetic marker. Index the three genotypes aa, aA, AA, where A is the high risk allele, by 0, 1, 2, respectively. Phenotype (case or control), and genotype data (three categories) can be presented in a 2×3 contingency table (Table 1). The number of cases and controls with genotype i is denoted by x_i and y_i, respectively, and the total number of cases and controls with genotype i by $m_i = x_i + y_i$, $i = 0, 1, 2$. Let $n_1 = x_0 + x_1 + x_2$ denote the total number of cases, $n_2 = y_0 + y_1 + y_2$ the total number of controls, and let $N = n_1 + n_2 = m_0 + m_1 + m_2$.

Denote by k the prevalence of the disease, i.e. the probability that a randomly drawn individual from the population under study has the disease, and by g_i the probability

Table 1: Notation for 2×3 table.

	Genotype			
	aa	aA	AA	Total
Case	x_0	x_1	x_2	n_1
Control	y_0	y_1	y_2	n_2
Total	m_0	m_1	m_2	N

that an individual has genotype i, $i = 0, 1, 2$. Let f_i be the penetrance, i.e. the conditional probability of disease given genotype i. When we test for association between a phenotype and a genotype, the null hypothesis is

$$f_0 = f_1 = f_2.$$

The above probabilities cannot be estimated in a case–control study when the disease prevalence of the population are unknown. Therefore it is convenient to express the null hypothesis in terms of conditional probabilities of genotypes given disease status. Let p_i and q_i denote the conditional probabilities of an individual having genotype i given that the individual is diseased and not diseased, respectively. Then $f_i = kp_i/g_i$ and $1 - f_i = (1 - k)q_i/g_i$ by elementary probability laws. If the null hypothesis $f_0 = f_1 = f_2$ is true, then all p_i/g_i are equal, and, likewise, all q_i/g_i are equal, $i = 0, 1, 2$. By dividing the former by the latter, it is seen that all p_i/q_i are equal. Since both the p_i and the q_i add to 1, $p_i = q_i$ for all i. Conversely, $p_i = f_i g_i/k$ and $q_i = (1 - f_i)g_i/(1 - k)$. If $p_i = q_i$ for all i, then $f_i/k = (1 - f_i)/(1 - k)$, giving $f_i = k$ for all i, thus all the f_i are equal. We now have an equivalent formulation of the null hypothesis,

$$p_0 = q_0, \qquad p_1 = q_1, \qquad p_2 = q_2. \tag{1}$$

Tests of this null hypothesis against the alternative that not all $p_i = q_i$ include the classical test for homogeneity using Pearson's chi-square statistic, which asymptotically has the chi-square distribution with two degrees of freedom under the null hypothesis, and generalizations of Fisher's exact test for 2×2 contingency tables to 2×3 tables.

However, genetic models provide more specific alternative hypotheses, and the recessive, dominant and additive models are often investigated, each associated with a different alternative hypothesis. Introducing genotype relative risks (GRRs), $\lambda_1 = f_1/f_0$ and $\lambda_2 = f_2/f_0$, the alternative hypotheses for the three models formulated in terms of penetrances, the GRRs or the p_i and q_i (the formulations in terms of the latter can be found by arguments similar to the one for the null hypothesis) are:

$$\begin{array}{lllll} \text{Recessive:} & f_0 = f_1 < f_2, & 1 = \lambda_1 < \lambda_2, & p_0/q_0 = p_1/q_1 < p_2/q_2 \\ \text{Monotone:} & f_0 < f_1 < f_2, & 1 < \lambda_1 < \lambda_2, & p_0/q_0 < p_1/q_1 < p_2/q_2 \\ \text{Dominant:} & f_0 < f_1 = f_2, & 1 < \lambda_1 = \lambda_2, & p_0/q_0 < p_1/q_1 = p_2/q_2 \end{array} \tag{2}$$

The monotone case with $f_1 = (f_0 + f_2)/2$ is called *additive*. The Cochran–Armitage test for trend (CATT) (Armitage, 1955; Cochran, 1954; Sasieni, 1997; Slager and Schaid, 2001) is often used to test the null hypothesis (1) towards one of the genetics models

(alternative hypoteses) in (2), partly motivated by the formulation in terms of the p_i and q_i. It is based on the statistic $\sum_{i=0}^{2} s_i(x_i/n_1 - y_i/n_2)$, where s_0, s_1, s_2 are scores appropriate for the alternative hypothesis in question. Standardizing and replacing unknown parameters p_i, q_i by estimators m_i/N, we obtain the CATT test statistic,

$$\mathrm{CATT} = \frac{\sum_{i=0}^{2} s_i(n_2 x_i - n_1 y_i)}{\sqrt{n_1 n_2 \left(\sum_{i=0}^{2} s_i^2 m_i - \frac{1}{N}\left(\sum_{i=0}^{2} s_i m_i\right)^2\right)}},$$

which asymptotically has a standard normal distribution under the null hypothesis. The absolute value of CATT is invariant to linear transformations of the scores, so they are chosen $(s_0, s_1, s_2) = (0, s, 1)$, and we use the notation CATT_s. The value of s is chosen as $s = 0$, $\frac{1}{2}$, 1 for the recessive, additive and dominant model of (2), respectively (Zheng *and others*, 2003). The index s thus denotes which genetic model (alternative hypothesis) is used. A large value for CATT_s indicates rejection of the null hypothesis.

When the genetic model is unknown, a popular strategy is to form one combined alternative hypothesis by taking the union of the three alternative hypotheses in (2). A test statistics tailored towards this combined alternative hypothesis is the MAX3 test statistic, which is the maximum of the three CATT test statistics for the recessive, additive and dominant models (Freidlin *and others*, 2002), $\max(\mathrm{CATT}_0, \mathrm{CATT}_{1/2}, \mathrm{CATT}_1)$. If the the potential high risk allele is unknown, the combined alternative hypothesis is defined by the union of the three alternative hypotheses in (2) and these three alternative hypotheses with the inequalies reversed. This may be written as

$$\begin{array}{ll} (1) & p_0/q_0 \neq p_2/q_2, \\ (2) & p_1/q_1 \text{ lies in the closed interval joining } p_0/q_0 \text{ and } p_2/q_2, \end{array} \tag{3}$$

and will cover all six combinations of genetic models and which allele is the high risk one. To test (1) against (3) the MAX3 test statistics is used, defined as

$$\mathrm{MAX3} = \max(|\mathrm{CATT}_0|, |\mathrm{CATT}_{1/2}|, |\mathrm{CATT}_1|). \tag{4}$$

The exact parametric distribution of the MAX3 statistic is unknown.

The MAX3 test statistic is just one out of many possible so-called robust test statistics that can be used to test for an association between genotype and phenotype in case–control studies. Other choices include MIN2, MERT and CLRT. Evaluations and comparisons have been done by Zheng *and others* (2006) and Joo *and others* (2009). We will focus on the MAX3 test statistic in this presentation, but will not provide a comparison of various robust test statistics.

3 P-values and power

Assume that the outcome Z of an experiment is used to choose between a null hypothesis $H_0\colon \theta \in \Theta_0$ and an alternative hypothesis $H_1\colon \theta \notin \Theta_0$, where the probability distribution of Z depends on a parameter (vector) θ. Following Casella and Berger (2001, Section 8.3.4), we define a p-value as a test statistic $p(Z)$ satisfying $0 \leq p(z) \leq 1$ for all possible outcomes z. We consider tests, where for a significance level α, H_0 is rejected

when $p(Z) \leq \alpha$. A p-value is valid if $\Pr_\theta(p(Z) \leq \alpha) \leq \alpha$ for all α and $\theta \in \Theta_0$. In this case, the test preserves its nominal level in the sense that the probability of a type I error is at most α.

In Section 4 we will consider several choices of p-value $p(Z)$ for testing association between genotype and phenotype in a case–control study (1, 3). Desireable properties of a p-value are validity and high power (the probability to reject H_0). If the sample space is discrete, the power at θ of a test defined by $p(Z)$ for a given α is

$$\gamma(\theta) = \Pr_\theta(p(Z) \leq \alpha) = \sum_{p(z) \leq \alpha} \Pr_\theta(Z = z), \tag{5}$$

where the probabilites depend on θ, and the summation is over all outcomes z having a p-value not greater than α. The test size is $\sup_{\theta \in \Theta_0} \gamma(\theta)$.

If $p(Z)$ and $p'(Z)$ are two p-values giving power functions γ and γ', respectively, it follows from (5) that if $p(z) \leq p'(z)$ for all outcomes z, then $\gamma(\theta) \geq \gamma'(\theta)$ for all θ. So uniformly smaller p-values over the sample space give uniformly higher power over the parameter space.

In our case, it is reasonable to assume that the numbers (x_0, x_1, x_2) of cases having genotype 0, 1, 2, respectively, are trinomially distributed with parameters $(n_1; p_0, p_1, p_2)$, and independent of the numbers (y_0, y_1, y_2) of controls, which are trinomially distributed with parameters $(n_2; q_0, q_1, q_2)$. We consider n_1 and n_2 being part of the experimental design, so that the sample space consists of $\binom{n_1+2}{2}\binom{n_2+2}{2}$ outcomes, giving a bound of the number of summands in (5). Power can then be calculated by the sum at the right of (5), where the joint probability of $z = (x_0, x_1, x_2, y_0, y_1, y_2)$ with $\theta = (p_0, p_1, p_2, q_0, q_1, q_2)$ is

$$\Pr_\theta(Z = z) = \frac{n_1!}{x_0!x_1!x_2!} p_0^{x_0} p_1^{x_1} p_2^{x_2} \cdot \frac{n_2!}{y_0!y_1!y_2!} q_0^{y_0} q_1^{y_1} q_2^{y_2}. \tag{6}$$

4 Methods for calculating p-values

A standard way of obtaining a p-value is by means of a test statistic $T(Z)$ (for example MAX3). Assume that large values of $T(Z)$ give evidence of H_1. If H_0 is simple, i.e. Θ_0 contains only one point θ,

$$p(z) = \Pr_\theta(T(Z) \geq T(z)) = \sum_{T(z') \geq T(z)} \Pr_\theta(Z = z') \tag{7}$$

is a valid p-value. For the latter equality, a discrete sample space is assumed. Note that only the ordering of the sample space provided by $T(Z)$ matters: If $T'(Z)$ is another statistic, and $T(z_1) \leq T(z_2)$ if and only if $T'(z_1) \leq T'(z_2)$ for all outcomes z_1, z_2, then $T(Z)$ and $T'(Z)$ define the same p-value.

Equation (7) is not directly applicable in our case, since H_0 is composite, stating that genotype frequencies are equal for cases and controls, consequently being equal to the unconditional frequencies, $p_i = q_i = g_i$, so that Θ_0 consists of all $\theta = (g_0, g_1, g_2, g_0, g_1, g_2)$. We now review various approaches to calculating p-values in other ways.

4.1 Maximization approach (M)

For the maximization approach, (7) is replaced with

$$p(z) = \sup_{\theta \in \Theta_0} \Pr(T(Z) \geq T(z)) = \sup_{\theta \in \Theta_0} \sum_{T(z') \geq T(z)} \Pr(Z = z'). \quad (8)$$

The ordering of the outcomes z defined by decreasing $T(z)$ is the same as the ordering defined by increasing $p(z)$ as seen from (8).

If a $1 - \gamma$ confidence region for Θ_0 is available, the maximization of (8) restricted to this region, and with a penalty of γ added to the p-value, also results in a valid p-value (Berger and Boos, 1994). We are, however, unaware of such confidence regions for the present 2×3 case.

4.2 Estimation approach (E)

The E p-value is defined by (7), but in place of the unknown parameters $g_i = p_i = q_i$ of (6), the maximum likelihood estimates under the null hypothesis are used, $\hat{g}_i = m_i/N$.

This approach will not give valid p-values in general, since there is no guarantee that the true parameter vector g would not give a greater p-value than the E p-value (compare (8)). However, the estimators converge in probability to the true parameters as N grows (as do maximum likelihood estimators under some regularity conditions in general), so the E p-values, being sums of continuous functions of the form (6), are asymptotically valid as N tends to infinity.

The E p-value can be estimated by simulation by drawing random samples from the probability distribution (6) inserted $\hat{g}_i$ in place of $p_i = q_i$. This is parametric bootstrap simulation (retaining the disease status of each individual and drawing genotypes according to $\hat{g}_0$, $\hat{g}_1$, $\hat{g}_2$). In this case it is the same as nonparametric bootstrap simulation (retaining the disease status of each of the N individuals and drawing genotypes with replacement, since the probability of drawing genotype i then will be $\hat{g}_i$).

4.3 Conditioning on sufficient statistics (C)

When an outcome $z = (x_0, x_1, x_2, y_0, y_1, y_2)$, is presented as a contingency table (Table 1) the column margins are $m_0 = x_0 + y_0$, $m_1 = x_1 + y_1$ and $m_2 = x_2 + y_2$. When we condition on the column margins $M(z) = (m_0, m_1, m_2)$, the probability under the null hypothesis of an outcome z is a trivariate hypergeometric probability

$$\Pr(Z = z \mid M(Z) = M(z)) = \frac{\binom{m_0}{x_0}\binom{m_1}{x_1}\binom{m_2}{x_2}}{\binom{N}{n_1}}, \quad (9)$$

showing that the column margins are sufficient statistics for the genotype frequencies. Conditioning on column margins is also done in Fisher's exact test for testing equality of two binomial proportions.

The p-value of an outcome z conditioned on its column margins $M(z)$, named the C p-value, is calcuated by the sum

$$p(z) = \Pr(T(Z) \geq T(z) \mid M(Z) = M(z)) = \sum_{T(z') \geq T(z)} \Pr(Z = z' \mid M(Z) = M(z)). \quad (10)$$

The C p-value is valid conditional on every possible column margin vector $m = (m_0, m_1, m_2)$, i.e. if H_0 is rejected when the p-value does not exceed α, then $\Pr(\text{rejection} \mid M(Z) = m)) \leq \alpha$ under H_0 for all m. It is, however, also valid considered as a p-value for the original unconditional experiment, since by the law of total probability

$$\begin{aligned}\Pr(\text{rejection}) &= \sum_m \Pr(\text{rejection} \mid M(Z) = m) \Pr(M(Z) = m) \\ &\leq \sum_m \alpha \Pr(M(Z) = m) = \alpha \sum_m \Pr(M(Z) = m) = \alpha,\end{aligned}$$

where the sum is over all possible column margin vectors $m = (m_0, m_1, m_2)$ and in our situation $\Pr(M(Z) = m)$ is trinomial under the null hypothesis (1). The number of nonzero summands in (10) is much smaller than it would have been without conditioning, making summation also feasible for relatively large studies. Bakke and Langaas (2012) found a formula for the maximal number of nonzero summands in (10), and numerical examples are given in the C column of Table 3.

Tian *and others* (2009) have devised an efficient algorithm to calculate exact MAX3 probabilities by adding bivariate hypergeometric probabilities of sample points conditioned on marginal sums m_i, n_i . Other authors, Sladek *and others* (2007), have instead used an approximate version of the C p-value using permutation testing.

We have seen that the outcome of an experiment can be presented as a contingency table $z = (x_0, x_1, x_2, y_0, y_1, y_2)$. The outcome may alternatively be given on the individual level as two vectors of length N, one giving the disease status and one giving genotype status. Thus, entry k in the disease vector gives the disease status of individual k and entry k in the genotype vector gives the coded genotype of individual k. In permutation testing we generate B new outcomes of our experiment by permuting (shuffling) the genotypes vector, while keeping the disease vector fixed. This gives B new contingency tables with the same margins as the observed contingency table. The permutation p-value is given as the proportion of the $B + 1$ outcomes (the original outcome and the B permutation outcomes) having a value of the test statistic T greater than or equal to that of the original outcome. The permutation p-value is valid (Phipson and Smyth, 2010). When B tends to infinity the permutation p-value equals the C p-value. This can be seen by the fact that the permutation procedure is a trivariate hypergeometric experiment, drawing genotypes of the n_1 cases from the m_0, m_1, m_2 of each genotype.

We recommend using the C p-value, and not the permutation p-value, based on the following arguments. If the permutation algorithm is run more than once for the same observed outcome, this may result in a different permutation p-value for each run, which for a given significance level may lead to different hypothesis testing decisions. For GWA studies a significance level of $5 \cdot 10^{-8}$ is routinely used. To be able to arrive at a p-value below this significance level B must at least be $2 \cdot 10^7$. Using permutation with very large values of B is very inefficient compared to using (10) directly.

4.4 Asymptotic approach (A)

Another way of dealing with nuisance parameters and to avoid summation over a large set of outcomes, is to use of the asymptotic distribution of $T(Z)$ under the null hypothesis (1). For large samples, the approximate distribution of $(\text{CATT}_0, \text{CATT}_{1/2}, \text{CATT}_1)$ under

the null hypothesis is trivariate normal, and the correlation coefficients can be estimated consistently (Freidlin *and others*, 2002; Zheng and Gastwirth, 2006). Asymptotic p-values can be calculated from $\Pr(\text{MAX3} < T(z)) = \Pr(|\text{CATT}_0| < T(z), |\text{CATT}_{1/2}| < T(z), |\text{CATT}_1| < T(z))$, when $T(z)$ is the observed value of the MAX3 test statistic for an outcome z. González *and others* (2008) and Zang *and others* (2010) used numerical integration to calculate asymptotic p-values for the MAX3 test statistic. Asymptotic p-values are in general not valid.

4.5 Combination of methods

Instead of applying one of the methods mentioned above (M, E, C, A) to the original test statistic T, it is possible to let the negative of the p-value of one method serve as a test statistic for another. This may be repeated, and it is possible to have sequences of methods, the p-value of one serving as a test statistic for the next. We will denote by E∘M the p-value obtained by first applying E to T and then M to the resulting p-value, E∘E∘M if E is applied twice and then M, and so on.

The idea of applying the p-value of one test as the test statistic for another is used in Fisher–Boschloo's test for equality of two binomial proportions (Boschloo, 1970; McDonald *and others*, 1977). Here, the p-value of Fisher's exact test, which may be seen as a C method, was used as a test statistic for the M method, resulting in a C∘M p-value. Silva Mato and Martín Andrés (1997) gave a review of methods for 2×2 tables, including M, C∘M and E∘M. Lloyd (2008) explored combinations of E and M in general, including iterated applications of E before M.

Applying the C method twice leaves the C p-value unaltered, that is, the C p-value equals the CC p-value, since conditional on the column margins the ordering of the test statistics and the ordering of the C p-values will be the same. Applying the M method twice also leaves the M p-value unaltered, since the ordering of the M p-values are by construction the same as the ordering of the test statistics, over all possible outcomes.

The application of the M method to an existing p-value – valid or not – always leaves a valid p-value, since M gives a valid p-value by construction. Also, it is interesting to note that the application of M to an existing, valid p-value gives a p-value that is at least as small as the original p-value. This is true for all possible outcomes, and thus gives a test with uniformly (for all α and θ) at least as high power as that of the original test. Röhmel and Mansmann (1999) showed this for the 2×2 contingency table, but the proof extends to a general situation.

Theorem 1. *Let $p(Z)$ be a p-value and let $p'(Z)$ be the M p-value defined by $p(Z)$, i.e., $p'(z) = \sup_{\theta \in \Theta_0} \Pr_\theta(p(Z) \leq p(z))$ for all outcomes z.*

(a) $p'(Z)$ is a valid p-value.

(b) If $p(Z)$ is a valid p-value, then $p'(z) \leq p(z)$ for all outcomes z.

Proof. (a) This is true in general for any statistic $p(Z)$ (see Casella and Berger, 2001, Theorem 8.3.27).

(b) If $p(Z)$ is a valid p-value, $\Pr_\theta(p(Z) \leq \alpha) \leq \alpha$ for all significance levels α and all $\theta \in \Theta_0$. In particular, setting $\alpha = p(z)$, we get $\Pr_\theta(p(Z) \leq p(z)) \leq p(z)$ for all $p(z)$ and all $\theta \in \Theta_0$. Thus $p'(z) = \sup_{\theta \in \Theta_0} \Pr_\theta(p(Z) \leq p(z)) \leq p(z)$ for all z. □

Table 2: A hypothetical case–control study with $n_1 = 40$ cases and $n_2 = 40$ controls.

	Genotype			
	aa	aA	AA	Total
Case	6	4	30	40
Control	13	7	20	40
Total	19	11	50	80

Because of Theorem 1 we will only consider combinations of methods that ends with an M step.

Both M and C give valid p-values, but one does not in general give uniformly (for all α and θ) better power than the other. By the above, C∘M always gives power that is uniformly (for all α and θ) as high as C. The reason that M is not routinely applied, is due to the computational effort required.

Note that the M p-value can be calculated based on the C p-values of all possible outcomes, since by the law of total probability,

$$p(z) = \sup_{\theta \in \Theta_0} \sum_m \Pr(T(Z) \geq T(z) \mid M(Z) = m) \Pr(M(Z) = m), \tag{11}$$

where the sum is over all possible values of $m = (m_0, m_1, m_2)$ giving sum N. Further, the probability $\Pr(T(Z) \geq T(z) \mid M(Z) = m)$ equals the C p-value of the outcome with column margins m having the smallest test statistic which is at least as large as the observed $T(z)$. If there are no test statistics at least at large as $T(z)$ for the m in question the probability is zero. Further, the distribution of $M(Z)$ is trinomial under H_0 (1). The C∘M p-value can be calculated using (11), with the negative of the C p-value in place of the test statistic $T(Z)$, since the C and the CC p-values are identical.

4.6 The 2×3 case illustrated with a hypothetical example

Assume we observe Table 2 in a hypothetical case–control study with $n_1 = 40$ cases and $n_2 = 40$ controls, and that we do not know the mode of inheritance of the biallelic marker under study or which allele is the potential high risk one. We would like to test the null hypothesis (1) against the alternative hypothesis (3), and we base our calculations on MAX3 defined by (4).

The MAX3-test statistic for our observed table is 2.31. Of the $\binom{42}{2}\binom{42}{2} = 741321$ possible outcomes, 562184 have a MAX3 statistic not less than that of the observed table. If we knew the explicit probability of each outcome under the null hypothesis, we would, according to (7), add the probabilities of the 562184 outcomes to get the p-value.

M: In the maximization approach (M), we approximate (8) by maximizing over a discrete grid for the nuisance parameters (g_0, g_1, g_2) with increments 0.01, giving $\binom{102}{2} = 5151$ values of the nuisance parameters. This means computing 5151 sums of 562184 terms each, giving a maximum of 0.0490 at $(g_0, g_1, g_2) = (0.08, 0.84, 0.08)$, so that the M p-value is 0.0490 (Figure 1).

E: For the estimation approach (E), we insert the maximum likelihood estimates of the nuisance parameters under the null hypothesis, $(\hat{g}_0, \hat{g}_1, \hat{g}_2) = (19/80, 11/80, 50/80) =$

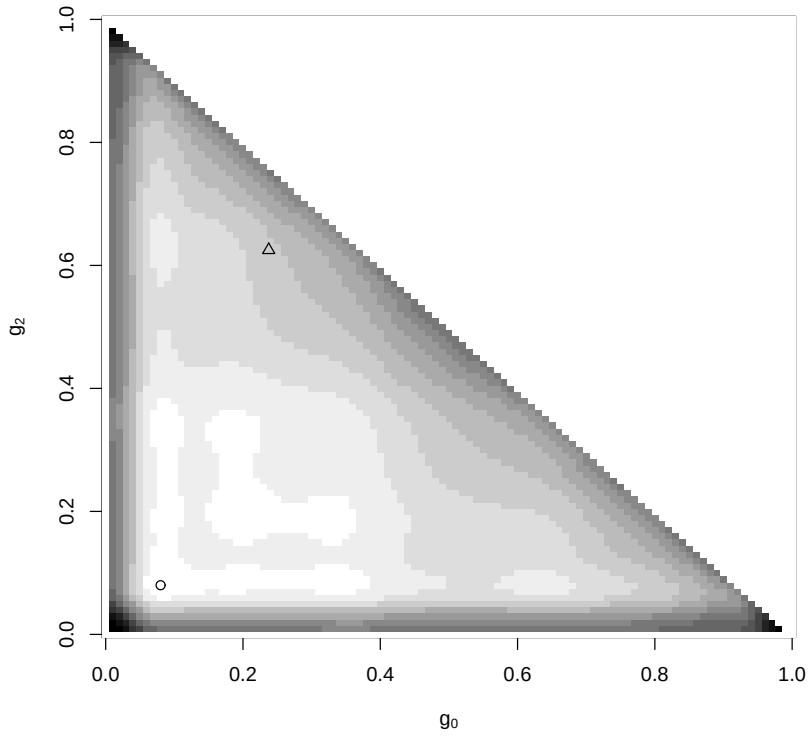

Figure 1: Illustration of the E and M p-values for the hypothetical example in Section 4.6. The M p-value of 0.0490 is found at $(g_0, g_1, g_2) = (0.08, 0.84, 0.08)$ (marked with a circle), and the E p-value of 0.0379 at $(\hat{g}_0, \hat{g}_1, \hat{g}_2) = (0.2375, 0.1375, 0.6250)$ (marked with a triangle). Light shades indicate large p-values, black denotes zero.

$(0.2375, 0.1375, 0.6250)$, into (6) to calculate the probability of each outcome, by (7) giving a p-value of 0.0379 (Figure 1). Note that E p-values need not be valid for small samples.

E∘M: The E p-values are considered a test statistic (instead of the original MAX3), thus the E p-value needs to be calculated for all 741321 tables. This changes which outcomes are equal to or more extreme than the observed outcome. In this scenario there are 556926 outcomes with an E p-value not larger than the observed 0.0379. Applying the M method to the E p-values gives a maximum over 5151 values of nuisance at $(g_0, g_1, g_2) = (0.00, 0.07, 0.93)$ and a p-value of 0.0411. Note that this p-value is valid, and that in this case the E∘M p-value is smaller than the M p-value.

C: To calculate the conditional p-value we only consider outcomes having the same column margins as the observed table. There are 240 tables having column margins $(m_0, m_1, m_2) = (19, 11, 50)$ (see Bakke and Langaas, 2012). Of these, 174 have a MAX3 test statistic not less than our observed table. The p-value according to (10) is 0.0590.

C∘M: The C p-values are considered a test statistic, thus the C p-value needs to be calculated for all 741321 tables (conditioned on the column margins for each outcome, the maximal number of tables having the column margins of any of these tables is 574,

see Bakke and Langaas, 2012). Now there are 567464 outcomes in the tail. Applying the M method to the C p-values gives a maximum over 5151 values of nuisance at $(0.38, 0.35, 0.27)$ and a p-value of 0.0499, which is, in accordance with Section 4.5, indeed smaller than the C p-value.

A: We use the asymptotic distribution as described in (González *and others*, 2008; Zang *and others*, 2010), which gives a p-value of 0.03732, using the `asy` method of the Rassoc R package (Zang *and others*, 2010). Note that A p-values are not valid for small samples.

A∘M: Considering A p-values as a test statistic changes which outcomes are equal to or more extreme than the observed outcome. The A p-values for all outcomes need to be calculated, and the `asy` method of the Rassoc R package (Zang *and others*, 2010) adds 0.5 to all table cells of tables in which at least one cell entry is zero. Using the A p-value as the test statistic there are 551696 outcomes in the tail. Applying the M method to the A p-values gives maximum over 5151 values of nuisance at $(0.4, 0.05, 0.55)$ and a p-value of 0.0387. Note that the A∘M p-value is valid.

To sum up, the smallest p-values are achieved for the A (0.03732) and E (0.0379) methods, but these methods are not in general valid. Of the valid methods the C (0.0590) method gives the largest p-value, followed by C∘M (0.0499), M (0.0490), E∘M (0.0411), and finally A∘M (0.0387). We compare the methods further in the next section.

5 Power study

To study the validity of the E and A methods, and compare the power of all the methods presented in Section 4, we have conducted an extensive power study.

We have studied balanced and slightly unbalanced situations with respect to the row margins n_1 (smaller sample) and n_2 (larger sample). A selection of these combinations are presented in Table 3 along with the number of possible outcomes with these row margins, and the maximum number of outcomes conditional also on column margins (Bakke and Langaas, 2012).

For the smaller sample, we have considered ten values, $n_1 = 5$, 10, 15, ..., 50, and for the larger sample, we have considered $n_2 = n_1$, $n_1 + 5$, $n_1 + 10$, $n_1 + 15$, and $n_1 + 20$. This gives a total of 50 combinations (n_1, n_2). For each of these combinations we have calculated p-values for all possible outcomes and for all the seven methods considered in Section 4.6: the conditional method, C, the asymptotic method, A, the unconditional methods E and M, and the combined unconditional methods E∘M, C∘M and A∘M. The test size and power were calculated for grids in p_i and q_i (as explained below) using Equation (5). Thus, these are exact calculations and no simulations are involved.

Test size calculation set-up The null hypothesis (1) states that $p_i = q_i = g_i$. The test size was calculated at values of (g_0, g_1, g_2) placed in a regular grid, which we denote the H_0 grid. The grid had increments of 0.01, giving $\binom{102}{2} = 5151$ points. In biological analyses, Hardy–Weinberg equilibrium (HWE) is sometimes assumed, and the test size under HWE was calculated for values of the disease allele frequency q from 0.01 to 0.99 with increments of 0.0005, giving 1961 triplets $(g_0, g_1, g_2) = ((1-q)^2, 2q(1-q), q^2)$. We denote this the H_0–HWE grid.

Table 3: The number of outcomes with given row margins (U) and the maximum number of outcomes conditional on column margins (C) for selected sample sizes from the study in Section 5.

n_1	n_2	U	C
10	10	4356	44
20	20	53361	154
30	30	246016	331
40	40	741321	574
50	50	1758276	884
10	30	32736	66
20	40	198891	231
30	50	657696	474
40	60	1628151	784
50	70	3389256	1161

Results for test size The methods E and A were not found to keep the nominal level, which was expected as explained in Section 4. The percentage of points in the H_0 grid at which a level of 0.05 was violated was between 0 and 56% ($n_1 = 10,\ n_2 = 15$) for the E method, with a mean of 29% over all $50 \cdot 5151$ combinations of sample sizes and grid points. For the H_0–HWE grid the percentages were between 0 and 75% ($n_1 = 25$, $n_2 = 30$), with a mean of 32%. Maximum test size for E was 0.064 at one gridpoint for $n_1 = 10$ and $n_2 = 15$ in the H_0 grid and 0.056 at one gridpoint for $n_1 = 25$ and $n_2 = 30$ in the H_0–HWE grid. The points of violation for the E method when $n_1 = n_2 = 40$ are shown in Figure 2.

For the A method there were violations of between 0 and 42% ($n_1 = 40,\ n_2 = 60$) of the points of the H_0 grid, with a mean of 30%. For the H_0–HWE grid the percentages were between 0 and 37% ($n_1 = 40$, $n_2 = 50$), with a mean of 9%. Maximum test size for A was 0.072 for $n_1 = 5$ and $n_2 = 25$ in the H_0 grid and 0.053 for $n_1 = 40$ and $n_2 = 50$ in the H_0–HWE grid. The points of violation for the A method with $n_1 = 50$ and $n_2 = 70$ are shown in Figure 2.

For E, given n_1, there were the fewest violations when $n_2 = n_1$. For A, surprisingly, the trend was for more violations when n_1 grows.

Power calculation set-up The power study was inspired by Joo *and others* (2010). We used a disease prevalence of $k = 0.1$, and disease allele frequencies $q = 0.1, 0.2, \ldots, 0.5$. We assumed HWE, so that genotype frequencies were $g_0 = (1-q)^2,\ g_1 = 2q(1-q)$ and $g_2 = q^2$. This means that we only studied five genotype frequency vectors (g_0, g_1, g_2).

The genotype relative risks, λ_1 and λ_2, were chosen based on four genetic models. Two sets of λ_2 were considered: A moderate effect set 1.2, 1.4, ..., 5.0 and a large effect set 1.5, 2.5, ..., 20.5, each containing 20 values for λ_2. For the recessive model, $\lambda_1 = 1$, for the dominant model $\lambda_1 = \lambda_2$, for the additive model $\lambda_1 = (1 + \lambda_2)/2$, and for the overdominant model $\lambda_1 = 1.2\lambda_2$.

Combining the four genetic models, the 20 values for λ_2 and the five values of

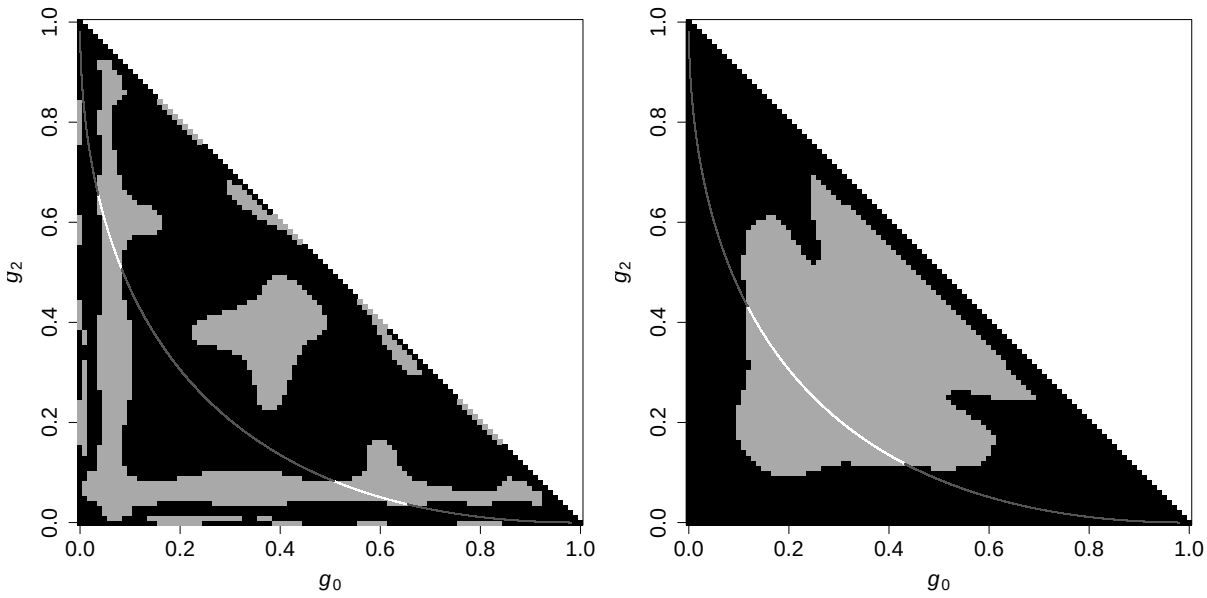

Figure 2: Genotype frequencies for which test size is greater than a significance level of 0.05 (grey grid points). Left: The E method with $n_1 = n_2 = 40$. Right: The A method with $n_1 = 50$ and $n_2 = 70$. The curve indicates points of Hardy–Weinberg equilibrium.

(g_0, g_1, g_2), we get 400 parameter sets under H_1 for each of the moderate and large effect sets. This would give a total of 800 parameter sets, but 15 of these gave invalid values for f_2, leaving us with 785 parameter value sets to investigate.

Results for power For each combination of 50 pairs of sample sizes and seven methods (M, E, E∘M, C, C∘M, A, A∘M), we obtained power at the 785 parameter sets under H_1. In Figure 3, we show graphs of power as a function of genotype relative risk λ_2 for disease allele frequency $q = 0.1$ for sample sizes $(n_1, n_2) = (10, 20)$ (very small sample sizes), $(40, 40)$ (equal sample size for cases and controls) and $(50, 70)$ (40% more controls than cases), and additionally for $q = 0.3$ in the $(40, 40)$ case. For each graph test size under the null ($\lambda_2 = 1$) is added for comparison. In Table 4 we present power results for a selection of 24 parameter sets, together with 3 parameter sets under the null for comparison for sample size $(40, 40)$.

For each of the 785 parameter sets and the 50 sample sizes (in total 39250 scenarios) we have compared the performance of the five methods that gave valid p-values – A∘M, C, C∘M, E∘M and M – with respect to power at the 0.05 significance level in a pairwise set-up. Overall the ranking of the methods were C∘M, E∘M, M, C and A∘M, with C∘M as the best method. Results for pairwise comparison between methods are presented in Table 5, for all scenarios, and for subsets of the parameter sets from each of the recessive, additive, dominant and over-dominant genetic models, and for subsets of the sample sizes: larger ($n_2 \geq 30$), smaller ($n_2 < 30$), balanced ($n_2 - n_1 < 10$) and unbalanced ($n_2 - n_1 \geq 10$) sample sizes.

To summarize, C∘M wins over (is more powerful than) E∘M in 60.4% of the scenarios, while E∘M wins over C∘M in 34.6% of the scenarios. C∘M wins over M in 72.2% of the scenarios, over C in 97.4% and over A∘M in 83.7% of the scenarios. Furthermore, E∘M

Table 4: Power (%) for various methods at a significance level of 0.05, $n_1 = n_2 = 40$. Hardy–Weinberg equilibrium is assumed, and the parameters considered are disease allele frequency q and genotype relative risks λ_1 and λ_2, where the model determines λ_1 in terms of λ_2. The prevalence k is 0.1.

Parameters			Method						
Model	λ_2	q	A	A∘M	C	C∘M	E	E∘M	M
H_0	1	0.1	1.7	1.5	3.0	3.6	5.0	4.5	2.4
		0.3	4.0	3.6	3.9	4.4	4.9	4.6	4.4
		0.5	5.1	4.5	4.0	4.8	4.8	4.6	4.6
Recessive	2	0.1	1.9	1.6	3.3	3.9	5.2	4.7	2.7
		0.3	15.3	13.5	13.2	14.3	15.8	14.4	14.2
		0.5	30.1	28.0	26.7	29.4	29.4	28.8	28.8
	5	0.1	4.2	3.7	6.4	7.4	10.0	9.3	7.3
		0.3	83.7	82.4	80.4	82.9	83.6	82.7	82.5
		0.5	96.0	95.1	94.9	95.7	95.8	95.4	95.4
Additive	2	0.1	7.2	6.4	10.5	12.0	14.1	13.0	8.8
		0.3	18.9	17.3	17.7	19.3	19.7	18.7	18.4
		0.5	19.5	18.2	17.3	19.2	19.0	18.6	18.7
	5	0.1	61.4	58.7	68.4	71.2	73.4	72.0	64.5
		0.3	79.1	77.1	76.8	79.0	79.3	78.4	78.3
		0.5	64.7	63.3	62.1	64.6	64.7	64.1	64.2
Dominant	2	0.1	20.2	18.2	25.8	28.4	31.6	29.9	22.2
		0.3	31.0	28.7	29.4	31.7	32.2	30.7	30.3
		0.5	20.6	19.6	17.9	19.9	20.2	19.9	19.9
	5	0.1	93.7	92.7	95.4	96.1	96.7	96.4	94.3
		0.3	91.5	90.6	90.0	91.4	91.8	91.2	91.2
		0.5	63.4	63.0	60.3	62.4	63.6	63.3	63.4
Over-	2	0.1	34.0	31.4	40.6	43.7	47.3	45.4	36.1
dominant		0.3	42.5	39.7	40.4	43.1	44.0	42.1	41.6
		0.5	24.6	23.8	21.3	23.4	24.3	23.9	23.9
	5	0.1	97.6	97.1	98.4	98.7	98.9	98.8	97.9
		0.3	94.6	94.1	93.5	94.5	94.9	94.4	94.4
		0.5	66.7	66.4	63.4	65.4	66.9	66.7	66.7

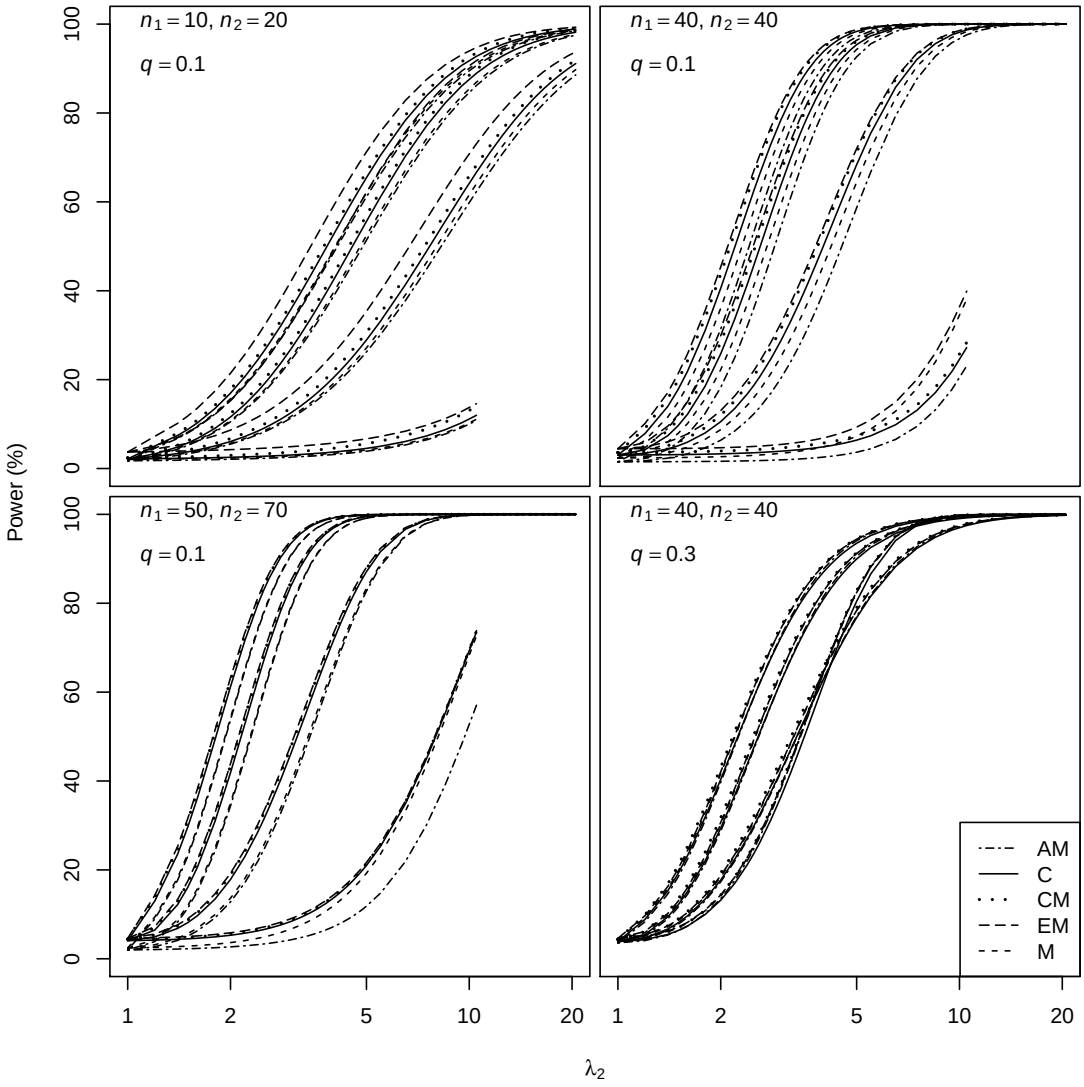

Figure 3: Power (%) at a significance level of 0.05 as a function of genotype relative risk λ_2 (on a logarithmic scale) for various methods (A∘M, C, C∘M, E∘M, M; see legend) and sample sizes. Hardy–Weinberg equilibrium is assumed; the prevalence k is 0.1. Disease allele frequency q is 0.3 (lower right figure) and 0.1 (other figures). The genotype relative risk λ_1 is determined by the model. On each figure, the upper five graphs are for the overdominant model, the next five graphs for the dominant model, the next five graphs for the additive model, and the lower (left) five graphs for the recessive model.

wins over M in 59.4% of the scenarios, over C in 78.0% and over A∘M in 69.8% of the scenarios. M wins over C in 60.8% and over A∘M in 73.8% of the scenarios. For these comparisons the number of draws varies from 2.5% to 5.1%. The comparison between the A∘M and C methods is not that clear. We find that A∘M wins over C in 48.4% of the scenarios, while C wins over A∘M in 49.1% of the scenarios. The average difference in power is given in Table 5.

For the comparisons between the E∘M and C∘M methods, C∘M wins in the majority of the scenarios within each genetic model, for larger sample sizes and for both balanced and unbalanced samples, but E∘M is better than C∘M for smaller sample sizes. We have also looked in more detail into the comparison of C and M, since the discussion on conditional versus unconditional methods has been given much emphasis in the literature, mainly for the 2×2 case. For all genetic models, M wins over C in the majority of scenarios, with the strongest win for the recessive model, while C (compared to M) performs it's best for the additive model. Further, M shows the greatest wins for smaller and balanced sample sizes, while C (compared to M) performs it's best for the larger and unbalanced sample sizes.

6 Discussion

Ordering of the sample space Consider an 2×3 case–control experiment with sample size (n_1, n_2). The $\binom{n_1+2}{2}\binom{n_2+2}{2}$ possible outcomes can be ordered according to decreasing test statistics, e.g. the MAX3, or to increasing p-value obtained by any of the seven methods (M, E, E∘M, C, C∘M, A, A∘M) based on the same test statistic. From the definition of the M p-value we have seen (in Section 4.1) that the ordering of the outcomes based on the (negative of the) p-value will be the same as the ordering based on the test statistic. Any test statistic giving the same ordering of the sample space as the MAX3 test statistic will give exactly the same M p-value as the MAX3 test statistic. This also means that the C∘M p-value will give the same ordering of the possible outcomes as the C p-value, the same for E and E∘M, A and A∘M, and so on. On the other hand, the C and E p-value provides a different ordering of the sample space than the MAX3 test statistic. Combining this fact with the finding (from Section 5) that the E∘M and C∘M methods perform better than the M method, we question the quality of the MAX3 test statistics for the genetic models considered (recessive, additive, dominant and over-dominant) in our power study, and find that it might be of interest to search for a better test statistic. (This conclusion also holds if we remove the over-dominant model from the power study.) Alternatively, the C∘M and E∘M methods provide an alternative way of improving the chosen test statistic by reordering the outcomes of the sample space.

Including covariates In genotype–phenotype association studies, covariates (age, sex, smoking status, body mass index, among others) may be present, and researchers may want to take these covariates into account in the statistical analyses. So and Sham (2011) present an asymptotic MAX3 method with an adjustment for covariates. The method is based on the asymptotic normality of score statistics in a multiple logistic regression, and is suggested for use in GWA study sample sizes on the order of hundreds or thousands. For small sample sizes, conditional (exact) versions of multiple logistic

Table 5: A pairwise comparison of power for the five methods C∘M, E∘M, M, C and A∘M, for various subsets of the 785 parameter sets and the 50 sample sizes studied in Section 5. For each subset and comparison two numbers are given, the percentage of times the method of the upper column header is at least as powerful as the method of the lower column header (first row for each subset), and the average difference in power between these methods (second row for each subset). *All* refers to all 39259 scenarios, *Recessive, Additive, Dominant* and *Over-dominant* refers to all 50 sample sizes for the parameter sets with the indicated genetic model (185 parameter sets for recessive and 200 for the others). For the remaining categories all 785 parameter sets are used, but subsets of sample sizes are used. *Smaller* refers to sample sizes where $n_2 < 30$ and *Larger* where $n_2 \geq 30$, *Balanced* has $n_2 - n_1 < 10$ and *Unbalanced* has $n_2 - n_1 \geq 10$ (number of sample sizes ranges from 15 to 35).

	C∘M E∘M	C∘M M	C∘M C	C∘M A∘M	E∘M M	E∘M C	E∘M A∘M	M C	M A∘M	C A∘M
All	65.4	76.7	100.0	86.2	64.5	81.1	72.4	63.5	77.2	51.6
	0.1	1.2	1.6	2.1	1.1	1.5	2.0	0.4	0.9	0.5
Recessive	57.7	74.1	100.0	72.5	67.6	85.0	60.9	71.8	59.7	36.8
	−0.3	0.5	1.3	0.9	0.8	1.6	1.2	0.9	0.5	−0.4
Additive	73.7	87.5	100.0	93.9	62.3	76.2	70.9	59.0	78.9	54.1
	0.4	1.7	1.7	2.6	1.3	1.3	2.2	0.1	0.9	0.8
Dominant	66.2	74.3	100.0	89.3	63.6	81.4	77.1	61.7	83.6	56.5
	0.2	1.4	1.7	2.4	1.2	1.5	2.2	0.3	1.1	0.8
Over-dominant	63.5	70.8	100.0	88.0	64.6	81.9	79.9	62.2	85.4	57.8
	0.1	1.3	1.7	2.4	1.2	1.6	2.3	0.4	1.1	0.7
Smaller	47.0	79.0	100.0	90.2	82.4	98.1	90.9	70.0	79.3	54.2
	−0.6	1.6	3.0	3.4	2.2	3.6	4.0	1.4	1.8	0.4
Larger	73.3	75.7	100.0	84.5	56.8	73.8	64.5	60.8	76.3	50.4
	0.4	1.0	1.0	1.5	0.6	0.6	1.1	0.0	0.5	0.5
Balanced	60.6	76.4	100.0	91.2	69.9	89.4	80.2	71.9	92.0	54.6
	−0.2	1.1	2.0	2.9	1.3	2.2	3.1	0.9	1.7	0.8
Unbalanced	68.6	76.9	100.0	82.9	60.8	75.5	67.3	58.0	67.4	49.6
	0.3	1.3	1.3	1.6	0.9	1.0	1.3	0.1	0.3	0.3

regression are available (Mehta and Patel, 1995), but to our knowledge MAX3 methods with covariates are not available for conditional (exact) inference.

An alternative strategy to including covariates in a logistic regression (with the assumption that covariates have a multiplicative effect on the odds ratio) is to divide the data into partial tables, e.g. based on sex and age categories, and compute the MAX3 test statistic and p-values with e.g. E∘M or C∘M for each partial table. This might be used when the assumption of multiplicative effects of covariates is violated, or if for example there is only an association between genotype and phenotype for one of the sexes or one of the age groups. If the effects of genotype are similar between the partial tables, it might be possible to combine the MAX3 test statistics for each partial table in the fashion of the CochranMantelHaenszel statistic, or to combine p-values from the partial tables using Fisher's (1950) method.

Population substructure Violation of the assumption of independence between individual observations in a study may occur if population substructure is present in the data, and may be dealt with using the method of scaling with a variance inflation factor (Devlin and Roeder, 2004) as is done in Sladek *and others* (2007). This may easily be done for the C and M methods presented here, by externally estimating one scaling factor for each genetic model to be applied to each CATT statistic before the maximum is taken. However, this requires that data for more than one biallelic maker are present, and that the sample size is large enough so that the scaling factors can be estimated consistently.

Larger sample sizes In GWA studies sample sizes are on the order of thousands of cases and controls, and ten thousands of genetic markers are studied. For these sample sizes the unconditional methods (E and M) will be too computationally intensive. Even for 100 cases and 100 controls, the number of possible outcomes is more than 27 million, and without smart algorithms and parallel programming this would be infeasible for use. However, conditional methods are easily applicable even for very large sample size (see Bakke and Langaas, 2012).

7 Conclusions

We have presented conditional (C), unconditional (E and M) and asymptotic (A) methods, and combinations thereof, and have used the robust MAX3 test statistic as a model statistic. It is well known that the A and E methods may produce invalid p-values, while the C and M methods will produce valid p-values. This applies to any test statistic. Specifically we have found that for small sample sizes for the MAX3 statistic (for all sample sizes studied in this presentation, that is, on the order of tens) the E and A method should not be used since they give invalid p-values. For the MAX3 test statistic we advocate using the E∘M and C∘M methods, which are more powerful than the M, C and A∘M methods in the scenarios we have investigated (recessive, additive, dominant and overdominant genetic models).

Furthermore, we in general advocate the use of an M step as a post processing step. The M step does not change the order of the test statistic over the sample space, makes an invalid p-value valid, and always improves a valid p-value (or leaves it unchanged).

The M step has been criticized for giving a conservative p-value, however, viewed as a post processing step this is not true. Applying the M step to a valid p-value will increase power, as shown in Theorem 1.

In the literature on 2×2 tables there has been an emphasis on comparing the conditional (C) and unconditional (M) method. However, if there are computational resources available to perform an M step, the C∘M method will always be preferable to the C method, so that the comparison should then be between the M and the C∘M methods. If a power study has been conducted and the C∘M method is found to be more powerful than the M method this will be a motivation to look for a better test statistic.

Acknowledgments

The authors would like to thank Bo Lindqvist for valuable comments.

Conflict of Interest: None declared.

References

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

Bakke, Ø. and Langaas, M. (2012). The number of $2 \times c$ tables with given margins. *Technical Report*, Department of Mathematical Sciences, Norwegian University of Science and Technology. Preprint in Statistics, 11/2012.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89**, 1012–1016.

Boschloo, R. D. (1970). Raised conditional level of significance for the $2\times$ 2-table when testing the equality of two probabilities. *Statistica Neerlandica* **24**, 1–9.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*, 2nd edition. Pacific Grove, CA: Duxbury.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10**, 417–451.

Devlin, B. and Roeder, K. (2004). Genomic control for association studies. *Biometrics* **55**, 997–1004.

Djurovic, S., Gustafsson, O., Mattingsdal, M., Athanasiu, L., Bjella, T., Tesli, M., Agartz, I., Lorentzen, S., Melle, I., Morken, G. *and others.* (2010). A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. *Journal of Affective Disorders* **126**, 312–316.

Fisher, R. A. (1950). *Statistical Methods for Research Workers*, 11th edition. Edinburgh: Oliver & Boyd.

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. *Human Heredity* **53**, 146–152.

González, 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.

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.

Joo, J., Kwak, M., Chen, Z. and Zheng, G. (2010). Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. *Statistics in Medicine* **29**, 158–180.

Lloyd, C. J. (2008). Exact p-values for discrete models obtained by estimation and maximization. *Australian & New Zealand Journal of Statistics* **50**, 329–345.

Lydersen, S., Fagerland, M. W. and Laake, P. (2009). Recommended tests for association in 2×2 tables. *Statistics in Medicine* **28**, 1159–75.

Martinelli-Boneschi, F., Esposito, F., Brambilla, P., Lindström, E., Lavorgna, G., Stankovich, J., Rodegher, M., Capra, R., Ghezzi, A., Coniglio, G., Colombo, B., Sorosina, M., Martinelli, V., Booth, D., Bang Oturai, A., Stewart, G., Harbo, H. F., Kilpatrick, T. J., Hillert, J., Rubio, J. P., Abderrahim, H., Wojcik, J. *and others.* (2012). A genome-wide association study in progressive multiple sclerosis. *Multiple Sclerosis Journal* (in press). http://msj.sagepub.com/content/early/2012/03/28/1352458512439118.

McDonald, L. L., Davis, B. M. and Milliken, G. A. (1977). A nonrandomized unconditional test for comparing two proportions in 2× 2 contingency tables. *Technometrics* **19**, 145–157.

Mehrotra, D. V., Chan, D. S. F. and Berger, R. L. (2003). A cautionary note on exact unconditional inference for a difference between two independent bionomial proportions. *Biometrics* **59**, 441–450.

Mehta, C. R. and Hilton, J. F. (1993). Exact power of conditional and unconditional tests: Going beyond the 2×2 contingency table. *The American Statistician* **47**, 91–98.

Mehta, C. R. and Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine* **14**, 2143–60.

Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero. *Statistical Applications in Genetics and Molecular Biology* **9**.

Röhmel, J. and Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal* **41**, 149–170.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

Serra, A., Schuchardt, K., Genuneit, J., Leriche, C. and Fitze, G. (2011). Genomic variants in the coding region of *neuronal nitric oxide synthase* (*NOS1*) in infantile hypertrophic pyloric stenosis. *Journal of Pediatric Surgery* **46**, 1903–1908.

Silva Mato, A. and Martín Andrés, A. (1997). Simplifying the calculation of the p-value for Barnard's test and its derivatives. *Statistics and Computing* **7**, 137–143.

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 others.* (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885.

Slager, S. L. and Schaid, D. J. (2001). Case–control studies of genetic markers: Power and sample size approximations for Armitages test for trend. *Human Heredity* **52**, 149–153.

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**, 768775.

Tian, J., Xu, C., Zhang, H. and Yang, Y. (2009). Exact max test in case–control association analysis using R: Package MaXact. Unpublished manuscript.

Zang, Y., Fung, W. K. and Zheng, G. (2010). Simple Algorithms to Calculate Asymptotic Null Distributions of Robust Tests in Case-Control Genetic Association Studies in R. *Journal of Statistical Software* **33**, 1–24.

Zheng, G., Freidlin, B. and Gastwirth, J. L. (2006). Comparison of robust tests for genetic association using case–control studies. In: Rojo, J. (editor), *Optimality: The Second Erich L. Lehmann Symposium*, Volume 49, Lecture Notes—Monograph Series. Beachwood, OH: Institute of Mathematical Statistics. pp. 253–265.

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 Journal* **45**, 335–348.

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.