Brazilian Journal of Probability and Statistics
2009, Vol. 23, No. 1, 18–35
DOI: 10.1214/08-BJPS001

Screening method for genetic linkage analysis: Case of the transmission disequilibrium test

Smail Mahdi

University of the West Indies

Abstract. In this paper we investigate the use of a two-stage case-control design to test for linkage disequilibrium in a large sample with a large number of null makers and one potential candidate marker. The scores, or signals, obtained at the markers, at the same stage, are assumed to be independent. The aim is to reduce the cost due to the number of laboratory analyses. In the first stage, the test is carried out at all markers of a randomly selected proportion λ of the sample at hand. Then the markers showing a score over a specified threshold, say, the median score, along with an average random proportion p of the makers with scores below the median are selected for the second stage of the study. Combined scores are then computed at the second stage and these cross-stage scores are not assumed to be necessarily additive or independent. This, partially, extends Satagopan et al. (*Biometrics* **58** (2002) 163–170) analysis in the case of independent marker outcomes. The aim is to identify optimal values for p and λ that maximize the probability to detect association in the case of association. The transmission-disequilibrium test is considered in the analysis and analytical formulas for the underlying probabilities are derived throughout. Furthermore, simulation results on the performance of the two designs are presented.

1 Introduction

Although a lot of statistical methods have been developed to test for association between markers and diseases in case-control studies, such as those of Spielman and Ewens (1998), Spielman, McGinnis and Ewens (1993), Terwilliger (1995) and Choulakian and Mahdi (2000), it remains that little attention has been paid to the design of such studies, as pointed out in Satagopan et al. (2002). In association studies for detecting the possible loci, conferring a risk for certain diseases, often a large number of marker loci is needed to be genotyped for each individual of a large sample. The cost of genotyping all marker loci for the whole sample of individuals could be then very costly. Therefore, an optimal design of the study, in order to reduce the cost of testing all markers and all individuals, would be useful. To this end, Satagopan et al. (2002) proposed a two-stage setup when the primary limitation on resources is the number of gene evaluations performed, rather than the total number of individuals. They describe the cost constraint as it occurs

Key words and phrases. Screening method, two-stage test, transmission-disequilibrium test.
Received May 2007; accepted January 2008.

in the analysis of association between a BRCA mutation and a single nucleotide polymorphisms (SNP), based on approximately 1500 anonymous SNP's and several other known markers, in a total of 2000 cancer cases. In the first stage of the analysis, all genes are evaluated on a subset of n_1 individuals and then only the most promising ones, say the top ith proportion, are evaluated on additional n_2 individuals. The value of n_2 depends on the remaining resources allocated to the study. The computed test statistic for each subject is assumed to have an asymptotic normal distribution with mean μ and variance σ^2. The statistic is afterward rescaled to lead to a Gaussian test statistic with variance 1. In the absence of association, that is, at the null markers, the mean is $\mu = 0$. The computed statistic on the first n_1 subjects is assumed to have a normal distribution with mean $n_1\mu$ and variance n_1^2. This argument suggests that the statistic, based on the n_1 subjects, is given by the sum of the n_1 independent statistics for each individual. The used assumption helps, therefore, to compute the mean and variance of the combined statistic upon the $n_1 + n_2$ individuals as $(n_1 + n_2)\mu$ and $(n_1 + n_2)^2$, respectively. Furthermore, the covariance between these two statistics is equal to n_1. It is also concluded that the joint distribution is binormal. Note that this is just a simplifying assumption since univariate normal distributions do not necessarily have joint normal distributions, see, for instance, Johnson and Wichern (2002); only the converse is true. The question is then to find the values n_1 and i optimizing the probability that the true gene is selected at the end of the study.

The drawback of this setup is that some potential marker candidates may provide, at first, signals slightly below the top ith proportion. Such markers would be disregarded because of possible sampling errors that might be due to n_1 and more so to i, especially, when a small value like $i = 10$, the one recommended in Satagopan et al. (2002), is used. Furthermore, the inference is solely reduced to the investigation of the mean of the distribution after the scaling of the test statistic. It is worth noting that this scaling is not always possible when, for instance, the variance of the statistic is a function of the mean as it is in the case of the transmission disequilibrium statistic (TDT) which is based upon a binomial variable. We recall that the TDT statistic was proposed by Spielman, McGinnis and Ewens (1993) as a family-based association test to test for the presence of genetic linkage between a genetic marker and a trait.

Following the methodology of Satagopan et al. (2002), we propose an alternative and safer setup in which all markers are tested on a subsample of size n_1 in the first step and all makers giving a signal over or equal to the sample median Ξ_{XY} of the whole sample, made up with the signal of the true gene marker X and the null markers $Y^{(i)}, i = 1, \dots, m$, will be further analyzed. Moreover, an average small random proportion p of the markers with signal below Ξ_{XY} will also be considered for the second stage of the analysis. This type of model, where only one locus is assumed to be responsible of the trait (say, disease), is useful in linkage disequilibrium analysis for localizing young mutations in isolated populations (see, e.g., Terwilliger (1995)).

We recall that linkage disequilibrium (LD) between a marker locus and a disease locus indicate that the two loci are closely linked. Indeed, when a new disease-causing mutation occurs in a population, it necessarily occurs at some locus of some haplotype. The disease allele is then associated with the allele, say A, carried at a very close locus of the same haplotype. Since these loci are tightly linked, the two alleles will co-segregate for many subsequent generations since recombinations are rare in such DNA short sequences. This will lead to a significant difference in frequency of allele A in the disease population with respect to the control population, provided that the populations are fairly stable and isolated to major mixing. Note that associations decay in the long run, mainly because of the recombination process.

It is worth noting that not all real applications require a high cost, including the screening of a large number of markers and individuals. In the case of noncomplex disorders, some association studies are based on moderate values such as $n = 100$ to $n = 500$ and $m = 10$ to $m = 50$. This even happens in studies on common complex diseases; see, for instance, the analysis carried out in Fallin et al. (2001) on the relationship between SNP's within the APOE gene region and Alzheimer's disease (AD). Their case/control sample consisted of only 210 AD cases and 159 nondemented elderly controls. Furthermore, they have just used 8 SNP's in a 205-Kb region of chromosome 19 that contains the APOE gene and a set of 5 SNP's in a 200-Kb region on chromosome 13. However, finding genes predisposing to very complex phenotypes is far from simple, and is likely to require a large high cost, that is, multi-disciplinary research groups accumulating hundreds, and probably thousands, of DNA samples from rigorously phenotyped subjects, as pointed out in Buckland (2001).

We organize this paper as follows. In Section 2, following the Introduction, we present the one-stage model and in Section 3, the two-stage model. The type one error along with the power of the tests, based on the one-stage and the two-stage models, are derived in Section 4 and 5, respectively. In Section 6, we develop the screening procedure using the TDT statistic and in Section 7 we analyze the sensitivity of the test performance with respect to the value of p (the proportion of markers to be taken among those with scores below the median). Finally, we display and discuss the obtained simulation results, on the performance of the two methods with the TDT test, in Section 8.

2 Single stage-model

We consider throughout the case of m independent null markers and one potential true associated marker in a fairly large sample of n individuals. Let $Y^{(i)}, i = 1, \ldots, m$, and X denote the signals (outcomes) of the null markers and the true marker, respectively. In the case of one-stage model, the probability P that the true gene, represented by X, is found at the end of the study is given by

$$P = P\Big(X_n > \max_{i=1}^{m} Y_n^{(i)}\Big) = \int_{-\infty}^{\infty} (1 - F_{X_n}(y_n^*)) g_n(y_n^*)\, dy_n^*,$$

where F_{X_n} and g_n represent the distribution functions of the scores X_n and $\max Y_n^{(j)}$, $j = 1, \ldots, m$, respectively, both evaluated at $\max_j(Y_n^{(j)}) = y_n^*$. The subscript n stands for the used sample size. We consider below the two-stage case.

3 Two-stage model

The steps of the model are outlined as follows.

Step 1. At this first stage, all the $m + 1$ markers are tested on n_1 individuals as in Satagopan et al. (2002). Let X_{n_1} and $Y_{n_1}^1, \ldots, Y_{n_1}^m$ denote the obtained scores at this stage. The setup consists in selecting all markers with a score above or equal to the sample median Ξ_{XY} for further testing. We also add to the second sage, on average, a random proportion p of the markers that provided a score below Ξ_{XY}. Let p_1 denote the probability that the X marker is selected for stage 2. This probability is given by

$$\begin{aligned} p_1 &= P\big[(X_{n_1} \geq \Xi_{XY}) \cup \big((X_{n_1} < \Xi_{XY}) \cap (S = 1)\big)\big] \\ &= P[(X_{n_1} \geq \Xi_{XY})] + pP[(X_{n_1} < \Xi_{XY})], \end{aligned}$$

where S is a Bernoulli variable with probability of success p. We use the following lemma to evaluate p_1.

Lemma 1. *For continuous distributions of X and Y, we have*

$$P[X \geq \Xi_{XY}] = P[X \geq \Xi_Y],$$

where Ξ_Y is the sample median of the Y scores.

Proof. The random variable X satisfies the equality

$$\begin{aligned} P[X \geq \Xi_{XY}] = {} & P[X \geq \Xi_{XY} | X > \Xi_Y] P[X > \Xi_Y] \\ & + P[X \geq \Xi_{XY} | X \leq \Xi_Y] P[X \leq \Xi_Y]. \end{aligned}$$

Note that under the condition $X > \Xi_Y$, we necessarily have $\Xi_Y < \Xi_{XY}$ and if $X \leq \Xi_Y$, then $\Xi_Y \geq \Xi_{XY}$. To evaluate $P[X \geq \Xi_{XY} | X > \Xi_Y]$, we consider first the case where there is no Y value between X and Ξ_Y. In such a case, we will necessarily have $X \geq \Xi_{XY} \geq \Xi_Y$. Thus, $P[X \geq \Xi_{XY} | X > \Xi_Y] = 1$ in such case. On the other hand, if there is a Y value between X and Ξ_Y, say, $\tilde{Y}$, then $\Xi_Y \leq \Xi_{XY} \leq \tilde{Y}$ and therefore, $X \geq \Xi_{XY}$. Thus, $P[X \geq \Xi_{XY} | X > \Xi_Y] = 1$ as well. Thus, $P[X \geq \Xi_{XY} | X > \Xi_Y] = 1$ in both cases. Let us now consider the situation $X \leq \Xi_Y$. First, if there is no Y value between X and Ξ_Y then, we must have $X \leq \Xi_{XY} \leq \Xi_Y$, and thus, $P[X \geq \Xi_{XY} | X \leq \Xi_Y]P[X \leq \Xi_Y] = PX = \Xi_{XY}] = 0$ since X and Y are assumed to be continuous variables. Furthermore, if there is a Y value between X and Ξ_Y, then $X \leq \Xi_{XY} \leq \Xi_Y$; thus, $P[X \geq \Xi_{XY} | X \leq \Xi_Y]P[X \leq \Xi_Y] = PX = \Xi_{XY}] = 0$. We conclude then the equality $P[X \geq \Xi_{XY}] = P[X \geq \Xi_Y]$. □

Thus, we have

$$p_1 = 1 + (p-1)\int_{-\infty}^{\infty} F_{X_{n_1}}(\xi)\psi(\xi)\,d\xi, \tag{3.1}$$

where ψ represents the probability density function of the sample median Ξ_Y and $F_{X_{n_1}}(\xi)$ is the distribution function of X_{n_1} evaluated at $\Xi_Y = \xi$. The distribution of the variable Ξ_Y can be approximated as

$$\psi(\xi) \sim \exp\left[-m\left(2f(\xi_{\text{med}})(\xi - \xi_{\text{med}})\right)^2\right],$$

where ξ_{med} is the true median of Y_{n_1} and f is the probability density function of Y_{n_1}; see, for example, Lupton (1993).

Evaluation of p_2

The quantity p_2 represents the probability that the marker X score outperforms the Y scores at the second stage, given that the marker X is selected in phase 1. Let η denote the random number of selected markers among whose with scores below the median at the first stage. Then the total number of selected markers for phase 2 is $m_2 = k + \eta$ where $k = \frac{m+4}{2}$ if m is even and $\frac{m+3}{2}$ if m is odd and η is a binomial variable with parameters $m - k$ and p. The probability p_2 is given by

$$\begin{aligned} p_2 &= \sum_{r=0}^{m-k+1} p_2(r)Pr[\eta = r] \\ &= \sum_{r=0}^{m-k+1} p_2(r)\binom{m-k+1}{r} p^r (1-p)^{m-k-r+1}, \end{aligned}$$

where

$$p_2(r) = Pr[A(r)|B(r)]$$

and

$$\begin{aligned} A(r) &= \left\{X_{n_2} > Y_{n_2}^{(1)}, \ldots, X_{n_2} > Y_{n_2}^{(k)}, X_{n_2} > Y_{n_2}^{(k+1)}, \ldots, X_{n_2} > Y_{n_2}^{(k+r)}\right\}; \\ B(r) &= \left[Y_{n_1}^{(1)} \geq \Xi_Y, \ldots, Y_{n_1}^{(k)} \geq \Xi_Y\right] \\ &\quad \cap \left[\bigcap_{l=1}^{r} (Y_{n_1}^{(k+l)} < \Xi_Y \cap S = 1)\right] \cap [X_{n_1} \geq \Xi_Y \cup [X_{n_1} < \Xi_Y \cap S = 1]], \end{aligned}$$

and

$$
\begin{aligned}
p_2(r) = {} & \int_{x_2=-\infty}^{\infty} \{P[Y_{n_2} \leq x_2 | Y_{n_1} \geq \Xi_Y]\}^k \times \{P[Y_{n_2} \leq x_2 | Y_{n_1} < \Xi_Y]\}^r \\
& \qquad \times \frac{P[X_{n_1} \geq \Xi_Y]}{P[X_{n_1} \geq \Xi_Y] + pP[X_{n_1} < \Xi_Y]} \times dP[X_{n_2} \leq x_2 | X_{N_1} \geq \Xi_Y] \\
& + \int_{x_2=-\infty}^{\infty} \{P[Y_{n_2} \leq x_2 | Y_{n_1} \geq \Xi_Y]\}^k \times \{P[Y_{n_2} \leq x_2 | Y_{n_1} < \Xi_Y]\}^r \\
& \qquad \times \frac{pP[X_{n_1} < \Xi_Y]}{P[X_{n_1} \geq \Xi_Y] + pP[X_{n_1} < \Xi_Y]} \\
& \qquad \times dP[X_{n_2} \leq x_2 | X_{n_1} < \Xi_Y].
\end{aligned}
$$

In order to compute the above integral, we need to evaluate separately the following terms.

(i) The first term is given by

$$
\begin{aligned}
& P[Y_{n_2} \leq x_2 | Y_{n_1} \geq \Xi_Y] \\
& \quad = \frac{\int_{-\infty}^{\infty} P[Y_{n_2} \leq x_2, Y_{n_1} \geq \xi_Y] \psi(\xi_Y)\, d\xi_Y}{\int_{-\infty}^{\infty} P[X_{n_1} \geq \xi_Y] \psi(\xi_Y)\, d\xi_Y} \\
& \quad = \frac{\int_{-\infty}^{\infty} P[Y_{n_2} \leq x_2 | Y_{n_1} \geq \xi_Y] P[Y_{n_1} \geq \xi_Y] \psi(\xi_Y)\, d\xi_Y}{\int_{-\infty}^{\infty} P[X_{n_1} \geq \xi_Y] \psi(\xi_Y)\, d\xi_Y} \\
& \quad = \frac{\int_{y_1=\xi_Y}^{\infty} \int_{-\infty}^{\infty} P[Y_{n_2} \leq x_2 | Y_{n_1} = y_1] f_{Y_1}(y_1) \psi(\xi_Y)\, d\xi_Y\, dy_1}{\int_{-\infty}^{\infty} P[X_{n_1} \geq \xi_Y] \psi(\xi_Y)\, d\xi_Y} \\
& \quad = \frac{\int_{-\infty}^{\infty} \psi(\xi_Y) [\int_{\xi_Y}^{\infty} \Psi_{Y_{n_2}|y_1}(x_2) f_{Y_1}(y_1)\, dy_1]\, d\xi_Y}{\int_{-\infty}^{\infty} [1 - W_{X_1}(\xi_Y)] \psi(\xi_Y)\, d\xi_Y},
\end{aligned}
$$

where ψ is the pdf of Ξ_Y, f_{Y_1} probability density function of Y_{n_1}, Ψ the conditional distribution function of $Y_{n_2}|Y_{n_1}$ and W the distribution function of Y_{n_1}.

(ii) Similarly, we have

$$
P[Y_{n_2} \leq x_2 | Y_{n_1} < \Xi_Y] = \frac{\int_{-\infty}^{\infty} \psi(\xi_Y) [\int_{-\infty}^{\xi_Y} \Psi_{Y_{n_2}|y_1}(x_2) f_{Y_1}(y_1)\, dy_1]\, d\xi_Y}{\int_{-\infty}^{\infty} W_{X_1}(\xi_Y) \psi(\xi_Y)\, d\xi_Y},
$$

(iii)

$$
P[X_{n_1} < \Xi_Y] = 1 - P[X_{n_1} \geq \Xi_Y] = \int_{-\infty}^{\infty} W_{X_1}(\xi_Y) \psi(\xi_Y)\, d\xi_Y
$$

and

(iv)

$$dP[X_{n_2} \le x_2 | X_{n_1} > \Xi_Y]$$
$$= \frac{d}{dx_2}\left\{\frac{\int_{-\infty}^{\infty}\psi(\xi_Y)[\int_{\xi_Y}^{\infty}\psi_{X_{n_2}|x_1}(x_2)\,f_{x_1}(x_1)\,dx_1]\,d\xi_Y}{\int_{-\infty}^{\infty}[1-W_{X_1}(\xi_Y)]\psi(\xi_Y)\,d\xi_Y}\right\}dx_2,$$

where f_{x_1} is the probability density function of X_1 and $\Psi_{X_{n_2}|x_1}$, the conditional distribution function of $X_{n_2}|X_{n_1}$.

(v) The last term is given by

$$dP[X_{n_2} \le x_2 | X_{n_1} \le \Xi_Y]$$
$$= \frac{d}{dx_2}\left\{\frac{\int_{-\infty}^{\infty}\psi(\xi_Y)[\int_{-\infty}^{\xi_Y}\psi_{X_{n_2}|x_1}(x_2)\,f_{X_1}(x_1)\,dx_1]\,d\xi_Y}{\int_{-\infty}^{\infty}W_{X_1}(\xi_Y)\psi(\xi_Y)\,d\xi_Y}\right\}dx_2.$$

4 Type one error

4.1 One-stage model

We recall that the type one error is given by the probability to reject the null hypothesis H_0 of no association under H_0. Under the null hypothesis, the scores X_n and $Y_n^{(i)}$ for $i = 1, \dots, m$ have the same distribution as, say, F_y. Therefore, for a significance level α, we reject H_0 under H_0 if $\max_{i=1}^{m+1} Y_n^{(i)} > K_0$ where $K_0 = F_Y^{-1}(1-\alpha)^{1/(m+1)}$. This derives from the following equation:

$$P[Reject H_0|H_0] = P\Big[\max_{i=1}^{m+1} Y_n^{(i)} > K_0\Big]$$
$$= 1 - P\Big[\max_{i=1}^{m+1} Y_n^{(i)} \le K_0\Big] = 1 - [F_Y(K_0)]^{m+1} = \alpha.$$

4.2 Performance and power

We measure the performance of the one-stage model by the probability to conclude association given that there is indeed association. This performance measure is given by the following formula:

$$P\big[\max\big(Y_n^{(1)}, \dots, Y_n^{(m)}, X_n\big) > K_0\big] = 1 - [F_Y(K_0)]^m F_{X_n}(K_0).$$

Furthermore, the probability that the X marker will outperform the Y's markers and show association is referred to as the power of the testing procedure, which is given by

$$Po_1 = P\Big[X_n > \max_{i=1}^{m} Y_n^{(i)} \cap X_n > K_0\Big]$$
$$= P\Big[X_n > \max_{i=1}^{m} Y_n^{(i)}\Big]P\Big[K_0 \le \max_{i=1}^{m} Y_n^{(i)}\Big]$$
$$+ P[X_n > K_0]P\Big[K_0 > \max_{i=1}^{m} Y_n^{(i)}\Big]$$

$$
\begin{aligned}
&= [1 - G_n(K_0)] \times \int_{-\infty}^{\infty} [1 - F_{X_n}(y_n^*)] g(y_n^*)\, dy_n^* \\
&\quad + [1 - F_{X_n}(K_0)] G_n(K_0),
\end{aligned}
$$

where G_n represents the cumulative function of $\max Y_n^{(i)}$ for $i = 1, \ldots, m$.

5 Two-stage model

We derive below the type one error and the power of the two-stage model.

5.1 Type one error

For the same significance level α, as above, the corresponding two-stage procedure critical value K_1 for rejecting wrongly the null hypothesis H_0 is given by the solution of the equation

$$
\begin{aligned}
\alpha &= \sum_{j=0}^{m-k+1} P\Big[\max_{i=1}^{k+j} Y_N^{(i)} > K_1 | \eta = j\Big] P[\eta = j] \\
&= \sum_{j=0}^{m-k+1} [1 - F_Y(K_1)]^{k+j} \binom{m-k+1}{j} p^j (1-p)^{m-k-j+1} \\
&= 1 - \sum_{j=0}^{m-k+1} F_Y(K_1)^{k+j} \binom{m-k+1}{j} p^j (1-p)^{m-k-j+1} \\
&= 1 - F_Y(K_1)^k [pF_Y(K_1) + 1 - p]^{m-k+1}.
\end{aligned}
\tag{5.1}
$$

From a partial check of the above formula, we can state the following remarks.

Remark 1. When $p = 1$, we have $K_1 = K_0$. This agrees with the one-stage model since all markers are selected for step 2 in such situations.

Remark 2. For a fixed value α, the right-hand side of equation (5.1) increases as p increases, and so does the value K_1, which reaches its maximum at $K_1 = K_0$. Indeed, we have

$$
\frac{d\alpha(p)}{dp} = (m - k + 1) F_Y(K_1)^k [pF_Y(K_1) + 1 - p]^{m-k} \big(1 - F_Y(K_1)\big) \geq 0
$$

since $pF_Y(K_1) + 1 - p \geq 0$ and $F_Y(K_1) \leq 1$.

5.2 Power

To evaluate the power, we consider first the case where $k+r$ markers are selected for the second phase. In this case, the probability for the selected set of markers to show association, given that the true marker X is included in the subset, is

$$\begin{aligned}\mathcal{P}(r) &= P\big[\max\big(Y_N^{(1)},\ldots,Y_N^{(k+r-1)},X_N\big) > K_1|\eta=r\big]\\ &= 1-[F_Y(K_1)]^{k+r-1}F_{X_N}(K_1).\end{aligned}$$

Therefore, the performance of the procedure, that is, the probability to observe association, with any possible size of the selected markers set, is obtained as

$$\wp = p_1\sum_{r=0}^{m-k+1}\binom{m-k+1}{r}p^r(1-p)^{m-k-r+1}\mathcal{P}(r),$$

where p_1 is given in equation (3.1). Furthermore, the probability that the marker X will outperform the Y markers and show association in the case where $k+r$ markers, including the true marker, are selected for the final step is given by

$$\begin{aligned}\mathcal{P}'(k+r) &= P\Big[X_N > \max_{i=1}^{k+r-1} Y_N^{(i)} \cap X_N > K_1|\eta=r\Big]\\ &= P\Big[X_N > \max_{i=1}^{k+r-1} Y_N^{(i)}\Big]P\Big[K_1 \le \max_{i=1}^{k+r-1} Y_N^{(i)}\Big]\\ &\quad + P[X_N > K_1]P\Big[K_1 > \max_{i=1}^{k+r-1} Y_N^{(i)}\Big]\\ &= \big[1-G_{(k+r-1)}(K_1)\big]\times\left\{\int_{-\infty}^{\infty}[1-F_X(\xi_Y)]g_{k+r-1}(\xi_Y)\,d\xi_Y\right\}\\ &\quad + [1-F_X(K_1)]G_{k+r-1}(K_1),\end{aligned}$$

where G_{k+r-1} and g_{k+r-1} are the distribution and probability density functions of $\max_{i=1}^{k+r-1} Y_i$, respectively. Thus, the power of the two-stage test is given by

$$Po_2 = p_1\sum_{r=0}^{m-k+1}\binom{m-k+1}{r}p^r(1-p)^{m-k-r+1}\mathcal{P}'(k+r).$$

Remark 3. For $p=1$, $p_1=1$, $Po_2=\mathcal{P}'(m+1)=Po_1$ since $K_1=K_0$ in such a case. This agrees with the power of the one-stage procedure.

We illustrate below the screening methodology through the X-linkage transmission disequilibrium test (TDT). Following the terminology of Spielman and Ewens (1998), as reported in Ho and Bailey-Wilson (2000), this test uses the total number of transmissions T of allele A_1 in n pairs of heterozygous mothers (A_1, A_2) and their affected children to test for associations disease-gene. Mothers with multiple affected children contribute multiple pairs to n. The used test statistic is the Z-score test with Yates's continuity correction.

6 Screening with the TDT statistic

We consider here the transmission disequilibrium statistic evaluated at the X marker as well as at the Y_i's markers. When this statistic is based on n_1 individuals, that is, n_1 pairs of heterozygous mothers (A_1, A_2) and their affected children, it is given by

$$X_{n_1} = \frac{|\zeta^x_{n_1} - n_1/2| - 0.5}{\sqrt{n_1/4}}$$

and

$$Y_{n_1} = \frac{|\zeta^y_{n_1} - n_1/2| - 0.5}{\sqrt{n_1/4}}$$

at the true marker X and the null marker Y, respectively. The variables $\zeta^y_{n_1}$ and $\zeta^x_{n_1}$ have independent binomial distributions with parameters $(n_1, \frac{1}{2})$ and $(n_1, \tilde{p})$, respectively. When the full sample size $n = n_1 + n_2$ is used, we similarly have

$$X_{n_1+n_2} = \frac{|\zeta^x_{n_1} + \zeta^x_{n_2} - (n_1+n_2)/2| - 0.5}{\sqrt{(n_1+n_2)/4}}$$

and

$$Y_{n_1+n_2} = \frac{|\zeta^y_{n_1} + \zeta^y_{n_2} - (n_1+n_2)/2| - 0.5}{\sqrt{(n_1+n_2)/4}}.$$

Note that, in the above statistics, the variable $X_{n_1+n_2}$ cannot be algebraically decomposed into a sum of X_{n_1} and X_{n_2} as in Satagopan et al. (2002). This is also true for $Y_{n_1+n_2}$ statistic as well.

We compute below the mean, variance and covariance of the above statistics. The covariance between the scores Y_{n_1} and $Y_{n_1+n_2}$ is obtained as

$$\begin{aligned} &\operatorname{cov}(Y_{n_1}, Y_{n_1+n_2}) \\ &\quad = \frac{4}{\sqrt{n_1(n_1+n_2)}} E[|\zeta^y_{n_1} - n_1/2| \times |(\zeta^y_{n_1} - n_1/2) + (\zeta^y_{n_2} - n_2/2)|] \qquad (6.1) \\ &\qquad - E[|\zeta^y_{n_1} - n_1/2|]E[|(\zeta^y_{n_1} - n_1/2) + (\zeta^y_{n_2} - n_2/2)|]. \end{aligned}$$

Using the classical normal approximation, $\zeta^y_l \sim N(\frac{l}{2}, \frac{l}{4})$ for l is large, we get after integration and numerical approximation, the following formulas

$$E[|\zeta^y_l - l/2|] \simeq 0.2\sqrt{l} \qquad (6.2)$$

and

$$E[|\zeta^y_{n_1} - n_1/2||(\zeta^y_{n_1} - n_1/2) + (\zeta^y_{n_2} - n_2/2)|] \simeq (0.73 + 0.34\tau)\frac{\tau n_2^2}{16}, \qquad (6.3)$$

where l can take the values n_1 or $n = n_1 + n_2$. After substitution of equations (6.2) and (6.3) into equation (6.1), we get

$$\operatorname{cov}(Y_{n_1}, Y_{n_1+n_2}) = \frac{0.73 + 0.34\tau}{\sqrt{\tau(1+\tau)}} n - 0.16,$$

where $\tau = \frac{n_1}{n_2}$ and $n = n_1 + n_2$. Using similar techniques, one can easily derive the following quantities:

$$E(Y_{n_1}) \simeq 0.8 - \frac{1}{\sqrt{n_1}}$$

and

$$\mathrm{Var}(Y_{n_1}) \simeq 1 - \left[0.8 - \frac{1}{\sqrt{n_1}}\right]^2.$$

Furthermore, if we assume that the vector $(Y_{n_1}, Y_{n_1+n_2})$ has an approximate binormal distribution, then the conditional distribution of $Y_{n_1+n_2}|Y_{n_1}$ is also normal. The mean and variance of this conditional distribution can easily be deducted from the parent joint distribution.

Similarly, using the normal approximation $\zeta^x_{n_1} \sim N(n\tilde{p}, n\tilde{p}(1-\tilde{p}))$, we get

$$\begin{aligned} E(X_{n_1}) = \int_{-\infty}^{\infty} &\left\{\frac{|\zeta^x_{n_1} - n_1/2| - 0.5}{\sqrt{n_1/4}}\right\} \\ &\times \frac{1}{\sqrt{2\pi}\sqrt{n_1\tilde{p}(1-\tilde{p})}} \exp\left[-\frac{1}{2}\left[\frac{\zeta^x_{n_1} - n_1\tilde{p}}{\sqrt{n_1\tilde{p}(1-\tilde{p})}}\right]^2\right] d\zeta^x_{n_1}. \end{aligned}$$

Making now the change of variable, $z = \frac{\zeta^x_{n_1} - n_1\tilde{p}}{\sqrt{n_1\tilde{p}(1-\tilde{p})}}$, yields after integration,

$$E(X_{n_1}) = 2\sqrt{\tilde{p}(1-\tilde{p})}\left[\sqrt{\frac{2}{\pi}}\exp\left(-\frac{K^2}{2}\right) + K\big(1 - 2\Phi(-K)\big)\right] - \frac{1}{\sqrt{n_1}},$$

where $K = (\tilde{p} - \frac{1}{2})\sqrt{\frac{n_1}{\tilde{p}(1-\tilde{p})}}$. Moreover, with the use of the same techniques, we obtain the following approximation:

$$\begin{aligned} E(X^2_{n_1}) = \int_{-\infty}^{\infty} &\left\{\frac{|\zeta^x_{n_1} - n_1/2| - 0.5}{\sqrt{n_1/4}}\right\}^2 \\ &\times \frac{1}{\sqrt{2\pi}\sqrt{n_1\tilde{p}(1-\tilde{p})}} \exp\left[-\frac{1}{2}\left[\frac{\zeta^x_{n_1} - n_1\tilde{p}}{\sqrt{n_1\tilde{p}(1-\tilde{p})}}\right]^2\right] d\zeta^x_{n_1} \\ = n_1 &+ 4\tilde{p}(1-\tilde{p})(1+n_1) \\ &- 4\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n_1}}\left[\sqrt{\frac{2}{\pi}}\exp\left(-\frac{K^2}{2}\right) + K\big(1 - 2\Phi(-K)\big)\right] + \frac{3}{n_1} \\ = n_1 &+ 4\tilde{p}(1-\tilde{p})(1+n_1) + O(n^{-1/2}). \end{aligned}$$

Note that in the case $\tilde{p} = \frac{1}{2}$, we get $E(X_{n_1}) = \sqrt{\frac{2}{\pi}} - \frac{1}{\sqrt{n_1}}$ and $E(X^2_{n_1}) = 1 - 2\sqrt{\frac{2}{\pi n_1}}$ since $K = 0$. This yields, $\mathrm{Var}(X_{n_1}) = 1 - \frac{2}{\pi} - \frac{1}{n_1} \simeq 1 - \frac{2}{\pi} \simeq (0.6)^2$ for large n_1. These values agree with the obtained values at the null marker Y_{n_1}. On the other hand, for $\tilde{p} > \frac{1}{2}$, we get $E(X_{n_1}) = (2\tilde{p} - 1)\sqrt{n_1} + O(n_1^{-1/2})$ and $E(X_{n_1}^2) = 4\tilde{p}(1 - \tilde{p}) + (2\tilde{p} - 1)^2 n_1 + O(n_1^{-1/2})$ which are function of n_1 and $\tilde{p}$. This yields $\text{Var}(X_{n_1}) = 4\tilde{p}(1 - \tilde{p}) + O(n_1^{-1/2})$. Therefore, the asymptotic value of $\text{Var}(X_{n_1})$ which is $4\tilde{p}(1 - \tilde{p})$ is not the same for every value of $\tilde{p}$.

The covariance between the statistics X_{n_1} and $X_{n_1+n_2}$, evaluated on a set of n_1 individual pairs and on a set of $n_1 + n_2$ individual pairs that include the first n_1 ones, is given by

$$\begin{aligned}
&\text{cov}(X_{n_1}, X_{n_1+n_2})\\
&\quad= \frac{4}{\sqrt{n_1(n_1+n_2)}} E[|\zeta_{n_1}^x - n_1/2||(\zeta_{n_1}^x - n_1/2) + (\zeta_{n_2}^x - n_2/2)|]\\
&\qquad - E[|\zeta_{n_1}^x - n_1/2|]E[|(\zeta_{n_1}^x - n_1/2) + (\zeta_{n_2}^x - n_2/2)|]\\
&\quad\simeq \chi \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\zeta_{n_1}^x - n_1/2|\\
&\qquad\qquad \times |(\zeta_{n_1}^x - n_1/2) + (\zeta_{n_2}^x - n_2/2)|\delta(\zeta_{n_1}^x, \zeta_{n_2}^x)\, d\zeta_{n_1}^x\, d\zeta_{n_2}^x\\
&\qquad - \int_{-\infty}^{\infty} \left\{\frac{|\zeta_{n_1}^x - n_1/2| - 0.5}{\sqrt{n_1/4}}\right\}^2\\
&\qquad\qquad \times \frac{1}{\sqrt{2\pi}\sqrt{n_1\tilde{p}(1-\tilde{p})}} \exp\left[-\frac{1}{2}\left[\frac{\zeta_{n_1}^x - n_1\tilde{p}}{\sqrt{n_1\tilde{p}(1-\tilde{p})}}\right]^2\right] d\zeta_{n_1}^x\\
&\qquad\quad \times \int_{-\infty}^{\infty} \left\{\frac{|\zeta_{n_1+n_2}^x - (n_1+n_2)/2| - 0.5}{\sqrt{(n_1+n_2)/4}}\right\}^2 \frac{1}{\sqrt{2\pi}\sqrt{(n_1+n_2)\tilde{p}(1-\tilde{p})}}\\
&\qquad\qquad \times \exp\left[-\frac{1}{2}\left[\frac{\zeta_{n_1+n_2}^x - (n_1+n_2)\tilde{p}}{\sqrt{(n_1+n_2)\tilde{p}(1-\tilde{p})}}\right]^2\right] d\zeta_{n_1+n_2}^x,
\end{aligned}$$

where the δ is the joint normal density function of the independent variables $\frac{\zeta_{n_1}^x - n_1\tilde{p}}{\sqrt{(n_1\tilde{p}(1-\tilde{p})}}, \frac{\zeta_{n_2}^x - n_2\tilde{p}}{\sqrt{(n_2\tilde{p}(1-\tilde{p})}}$ and $\chi = \frac{4}{\sqrt{n_1(n_1+n_2)}}$. The use of the normal approximation and the Delta method, yields, after computation and simplification, the following result:

$$\begin{aligned}
&\text{cov}(X_{n_1}, X_{n_1+n_2})\\
&\quad\simeq \frac{4}{\sqrt{n_1(n_1+n_2)}}[n_1(n_1+n_2)|\tilde{p} - 1/2|]\\
&\qquad - \left(2\sqrt{\tilde{p}(1-\tilde{p})}\left[\sqrt{\frac{2}{\pi}}\exp\left(-\frac{K^2}{2}\right) + K(1 - 2\Phi(-K))\right] - \frac{1}{\sqrt{n_1}}\right)
\end{aligned}$$

$$\times\left(2\sqrt{\tilde{p}(1-\tilde{p})}\left[\sqrt{\frac{2}{\pi}}\exp\left(-\frac{K'^2}{2}\right)+K\left(1-2\Phi(-K')\right)\right]\right.$$
$$\left.-\frac{1}{\sqrt{n_1+n_2}}\right),$$

where $K' = (\tilde{p}-\frac{1}{2})\sqrt{\frac{n_1+n_2}{\tilde{p}(1-\tilde{p})}}$ and Φ is the distribution function of the standard normal variable.

7 Sensitivity to markers with scores below median

To check on the sensitivity of the analysis with respect to the markers that provide scores below the median at the first stage, we evaluate the probability that the true marker's score X is among that markers. This probability is approximately given by

$$P[X_{n_1} < \Xi_Y] = \int_{-\infty}^{\infty}\Phi_0\left(\frac{\xi-\mu_{X_{n_1}}}{\sigma_{X_{n_1})}}\right)\psi(\xi)\,d\xi,$$

where Φ_0 is the distribution function of the standard normal variable and ψ the probability density function of the sample median Ξ_Y which is approximately Gaussian with mean ξ_{med} and variance $\sigma^2_{\Xi_Y} = \frac{1}{4m[f(\xi_{\text{med}})]^2}$; see Cramér (1946). Using the second-order approximation of the function $\Phi_0(\frac{\xi-\mu_{X_{n_1}}}{\sigma_{X_{n_1}}})$ at the point $\xi = \xi_{\text{med}}$, we get after integration and simplification,

$$P[X_{n_1} < \Xi_Y] \simeq \Phi_0\left(\frac{\xi_{\text{med}}-\mu_{X_{n_1}}}{\sigma_{X_{n_1}}}\right)$$
$$-\frac{\xi_{\text{med}}-\mu_{X_{n_1}}}{2\sqrt{2\pi}\sigma^2_{X_{n_1}}}\left\{\exp\left[-\frac{1}{2}\left(\frac{\xi_{\text{med}}-\mu_{X_{n_1}}}{\sigma_{X_{n_1}}}\right)^2\right]\right\}\sigma^2_{\Xi_Y}.$$

Furthermore, by using the normal approximation for the distribution of Y_{n_1}, that holds when n_1 is sufficiently large, we obtain $E(Y_{n1}) = \xi_{\text{med}}$; thus,

$$\begin{aligned}
&P[X_{n_1} < \Xi_Y]\\
&\quad\simeq \Phi_0\left(\frac{\xi_{\text{med}}-\mu_{X_{n_1}}}{\sigma_{X_{n_1}}}\right)\\
&\qquad-\frac{(\xi_{\text{med}}-\mu_{X_{n_1}})\pi\,\text{Var}(Y_{n_1})}{4m\sqrt{2\pi}\sigma^2_{X_{n_1}}}\left\{\exp\left[-\frac{1}{2}\left(\frac{\xi_{\text{med}}-\mu_{X_{n_1}}}{\sigma_{X_{n_1}}}\right)^2\right]\right\}.
\end{aligned}\tag{7.1}$$

7.1 Partial check

To partially check the accuracy of the derived expression for $P[X_{n1} < \Xi_Y]$, we consider again the Satagopan et al. (2002) two-stage design test and the transmission disequilibrium test. In Satagopan et al. (2002) case, we have $\mu_{X_{n_1}} = n_1\mu$, $\xi_{\text{med}} = 0$, $\mu_{Y_{n_1}} = 0$, $\text{Var}(Y_{n_1}) = n_1$ and $\sigma^2_{X_{n_1}} = n_1$. The substitution of these values into equation (7.1) yields

$$P[X_{n_1} < \Xi_Y] \simeq \Phi_0(-\sqrt{n_1}\mu) + \frac{\pi n_1 \mu}{4\sqrt{2\pi m}} \exp\left[-\frac{n_1\mu^2}{2}\right].$$

Note that if we substitute $\mu = 0$ in the above equation, we get $P[X_{n_1} < \Xi_Y] = \Phi_0(0) = \frac{1}{2}$ as it should be since the score variables X and Y have the same distribution in such a case; thus, the probability to observe a value below the median is $\frac{1}{2}$. On the other hand, for the studied case $\mu > 0$ in the simulation of Satagopan et al. (2002), $P[X_{n_1} < \Xi_Y]$ decreases as n_1 or m increases and tends toward zero when $n_1 \to \infty$. This agrees with the sign of μ. On the other hand, for $\mu < 0$ we have that $P[X_{n_1} < \Xi_Y]$ tends toward 1 as $n_1 \to \infty$. This agrees as well with the fact that μ is negative.

Case of the transmission disequilibrium test. In this case, we have $\mu_{X_{n1}} = 2\sqrt{\tilde{p}(1-\tilde{p})}(\tilde{p} - 0.5)\sqrt{n_1} + O(n^{-1/2})$, $\sigma^2_{X_{n_1}} = 4\tilde{p}(1-\tilde{p}) + O(n^{-1/2})$, $\xi_{\text{med}} \simeq \mu_{Y_{n_1}} = 0.8 + O(n^{-1/2})$ and $\text{Var}(Y_{n_1}) = 0.36 + O(n^{-1/2})$. After substitution of the above values into equation (7.1), we get

$$P[X_{n_1} < \Xi_Y] \simeq \Phi_0\left(\frac{(0.8 - 2\sqrt{\tilde{p}(1-\tilde{p})}(\tilde{p}-0.5)\sqrt{n_1})}{2\sqrt{\tilde{p}(1-\tilde{p})}}\right) - \Delta\Delta',$$

where

$$\Delta = \exp\left[-\frac{1}{2}\left(\frac{(0.8 - 2\sqrt{\tilde{p}(1-\tilde{p})}(\tilde{p}-0.5)\sqrt{n_1})}{2\sqrt{\tilde{p}(1-\tilde{p})}}\right)^2\right]$$

and

$$\Delta' = \frac{0.36\pi[0.8 - 2\sqrt{\tilde{p}(1-\tilde{p})}(\tilde{p}-0.5)\sqrt{n_1}]}{16\sqrt{2\pi m}\tilde{p}(1-\tilde{p})}.$$

From above, we see that $P[X_{n_1} < \Xi_Y] \to \Phi_0(-\infty) = 0$ as $n_1 \to \infty$ for $\tilde{p} > \frac{1}{2}$, as it should be since the X value is almost surely above the Y values. Note that the case $\tilde{p} > \frac{1}{2}$ is the most relevant one for the analysis presented here. Similarly, for $\tilde{p} < \frac{1}{2}$ we obtain $P[X_{n_1} < \Xi_Y] \to 1$ when n_1 gets sufficiently large; this is also expected because the X value will tend to be much larger than the Y values in this situation. Finally, if $\tilde{p} = \frac{1}{2}$ we get $P[X_{n_1} < \Xi_Y] = \Phi_0(0) = \frac{1}{2}$ which agrees with the fact that X and Y are identically distributed in such case.

These results lead to the following.

Remark 4. The probability that the X marker takes a value below the sample median decreases as the sample size or the number of markers increases. This probability becomes negligible when the first stage is based on fairly large sample sizes.

8 Simulation results and discussion

To compare the performance of the two methods with the TDT statistic, we carried out a simulation study for various values of n_1, n_2, m, p and $\tilde{p}$. To this end, for fixed values of $\lambda = \frac{n_1}{n_1+n_2} = 0.05, (0.1), 0.95$, $\tilde{p} = 0.50(0.05)1.0$ and $p = 0.05; 0.10$, we have estimated the probability that the associated marker is selected at the end of the study using both one-stage and two-stage methods. The notation used $\tilde{p} = 0.50(0.05)1.0$ is to state that $\tilde{p}$ starts at 0.50 and increases steadily by 0.05 until reaching the value 1.0. For each configuration of the parametric space, we have used 5000 samples. Note that we could not use much larger values for n and m because of computing power limitation. Nevertheless, our considered parametric values are sufficiently informative and provide us with a good guideline. Overall, the analysis showed that the two-stage model performs as the one-stage model for $\lambda \geq 0.25$, and that the results are not significantly dependent of p since m and n are fairly large. This agrees with the findings in Satagopan et al. (2002) and also confirms the Remark 4 about the sensitivity of the performance to the value of p. Therefore, we suggest taking a small value such as $p = 5\%$. Note that the asymptotic variance of the TDT statistic at the true marker is different from the one at the null marker. In this case, the two statistics cannot be scaled in order to have a common unit variance as in Satagopan et al. (2002). We display in Table 5 empirical values for the asymptotic mean and variance of the TDT statistic in the case $n_1 = 10000$ and $\tilde{p} = 0.5(0.1)0.9$. These empirical values agree with the derived corresponding formulas. Tables 1–4 displayed below, illustrate the obtained results in the cases of $p = 0.05, n = n_1 + n_2 = 500, 1000, 5000$ and $m = 100, 500$. The last rows of the tables give the corresponding empirical value

Table 1 *Empirical values for the probability that the X marker outperforms the Y markers as a function of λ and $\tilde{p}$ in the case of one-stage and two-stage test procedures. The last row gives the corresponding values obtained with the one-stage test procedure. $n = 500$ and $m = 100$*

$\tilde{p}$	0.50	0.55	0.60	0.65	0.70
λ					
0.05	0.01	0.30	0.76	0.90	0.96
0.15	0.01	0.30	0.85	0.97	0.99
0.25	0.02	0.34	0.93	0.98	1.00
0.35	0.01	0.35	0.95	1.00	1.00
0.45	0.01	0.34	0.94	1.00	1.00
0.55	0.01	0.34	0.95	1.00	1.00
0.65	0.01	0.32	0.95	1.00	1.00
0.75	0.01	0.30	0.95	1.00	1.00
0.85	0.01	0.32	0.95	1.00	1.00
0.95	0.01	0.32	0.94	1.00	1.00
1.00	**0.013**	**0.31**	**0.94**	**1.00**	**1.00**

Table 2 *Empirical values for the probability that the X marker outperforms the Y markers as a function of λ and $\tilde{p}$ in the case of one-stage and two-stage test procedures. The last row gives the corresponding values obtained with the one-stage test procedure.* $n = 1000$ *and* $m = 100$

$\tilde{p}$	0.50	0.55	0.60	0.65	0.70
λ					
0.05	0.01	0.51	0.82	0.96	0.98
0.15	0.01	0.59	0.96	1.00	0.99
0.25	0.01	0.62	0.99	1.00	1.00
0.35	0.01	0.62	0.99	1.00	1.00
0.45	0.01	0.62	0.99	1.00	1.00
0.55	0.01	0.62	1.00	1.00	1.00
0.65	0.01	0.63	1.00	1.00	1.00
0.75	0.01	0.63	0.99	1.00	1.00
0.85	0.01	0.64	1.00	1.00	1.00
0.95	0.01	0.64	1.00	1.00	1.00
1.00	**0.01**	**0.62**	**0.99**	**1.00**	**1.00**

Table 3 *Empirical values for the probability that the X marker outperforms the Y markers as a function of λ and $\tilde{p}$ in the case of one-stage and two-stage test procedures. The last row gives the corresponding values obtained with the one-stage test procedure.* $n = 5000$ *and* $m = 100$

$\tilde{p}$	0.50	0.55	0.60	0.65	0.70
λ					
0.05	0.01	0.51	0.82	0.96	0.97
0.15	0.01	0.58	0.96	0.98	1.00
0.25	0.02	0.66	0.99	1.00	1.00
0.35	0.01	0.67	1.00	1.00	1.00
0.45	0.01	0.65	1.00	1.00	1.00
0.55	0.01	0.66	1.00	1.00	1.00
0.65	0.01	0.68	1.00	1.00	1.00
0.75	0.01	0.66	1.00	1.00	1.00
0.85	0.01	0.66	1.00	1.00	1.00
0.95	0.01	0.68	1.00	1.00	1.00
1.00	**0.02**	**0.69**	**1.00**	**1.00**	**1.00**

obtained with the one-stage test procedure. Note that from Table 4, we see that the probability values are smaller than their corresponding values given in the other tables; this suggests that larger samples are needed when both m and n are of the same large magnitude.

The case of correlated markers will be analyzed in a future paper using properties of a covariance matrix based upon the recombination rates between loci; see, for example, Lessard and Mahdi (1995) and Mahdi and Lessard (1996).

Table 4 *Empirical values for the probability that the X marker outperforms the Y markers as a function of* λ *and* $\tilde{p}$ *in the case of one-stage and two-stage test procedures. The last row gives the corresponding values obtained with the one-stage test procedure.* $n = 500$ *and* $m = 500$

$\tilde{p}$	0.50	0.55	0.60	0.65	0.70
λ					
0.05	0.00	0.19	0.75	0.84	0.92
0.15	0.00	0.20	0.82	0.99	0.98
0.25	0.00	0.20	0.88	0.99	1.00
0.35	0.00	0.21	0.89	1.00	1.00
0.45	0.00	0.21	0.89	1.00	1.00
0.55	0.00	0.15	0.88	1.00	1.00
0.65	0.00	0.17	0.85	1.00	1.00
0.75	0.00	0.17	0.85	1.00	1.00
0.85	0.00	0.15	0.84	1.00	1.00
0.95	0.00	0.15	0.85	1.00	1.00
1.00	**0.00**	**0.17**	**0.89**	**1.00**	**1.00**

Table 5 *Empirical values for the asymptotic mean* (μ_{TDT}) *and variance* (σ^2_{TDT}) *of the TDT statistic computed with* $n_1 = 10000$, $m = 1000$ *and* $\tilde{p} = 0.5(0.1)0.9$

$\tilde{p}$	μ_{TDT}	σ^2_{TDT}
0.5	0.7948	0.3610
0.6	63.2501	0.9666
0.7	126.4886	0.8393
0.8	189.7367	0.6420
0.9	252.9794	0.3594

Acknowledgments

The author is grateful to the Editors and Referees for their help and useful comments. The financial support of UWI is also gratefully acknowledged.

References

Buckland, P. R. (2001). Genetic association studies of alcoholism-problem with the candidate gene approach. *Alcohol and Alcoholism* **36** 99–103.

Choulakian, V. and Mahdi, S. (2000). A new statistic for the analysis linkage between trait and polymorphic marker loci. *Mathematical Biosciences* **164** 139–145. MR1751268

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press. MR0016588

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D. and Schork, N. J. (2001). Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE Locus variation and Alzheimer's disease. *Genome Research* **11** 143–151.

Ho, G. Y. F. and Bailey-Wilson, J. E. (2000). The transmission/disequilibrium test for linkage on the X chromosome. *American Journal of Human Genetics* **66** 1158–1160.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey.

Lessard, S. and Mahdi, S. (1995). Convergence de la variabilité dans les modèles polygéniques gaussiens. *Genetics Selection Evolution* **27** 395–421.

Lupton, R. (1993). *Statistics in Theory and Practice*. Princeton Univ. Press. MR1228644

Mahdi, S. and Lessard, S. (1996). Convergence of covariance structures in additive Gaussian polygenic models. *Biometrics* **52** 833–845. MR1411734

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E. and Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics* **58** 163–170. MR1891375

Spielman, R. S. and Ewens, W. J. (1998). A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. *American Journal of Human Genetics* **62** 450–458.

Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52** 506–516.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci at one or more polymorphic marker loci. *American Journal of Human Genetics* **56** 777–787.

Department of Computer Science,
Mathematics and Physics
University of the West Indies
Cave Hill Campus
Barbados
E-mail: smahdi@uwichill.edu.bb