The Quantitative TDT

(Quantitative Transmission–Disequilibrium Test)

Warren J. Ewens

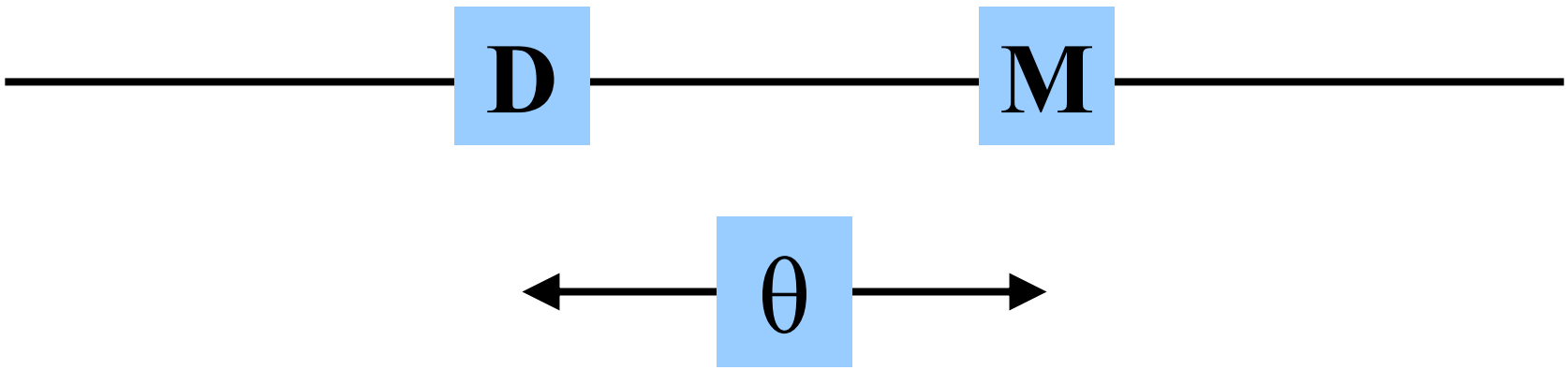
NUS, Singapore

10 June, 2009

The initial aim of the (**QUALITATIVE) TDT** was to test for *linkage* between a marker locus *M* and a disease locus *D*.

The aim of a QTDT is also to test for *linkage* between a marker locus *M* (with alleles M_1 and M_2) and a loci involved with a quantitative character.

Notation for the qualitative TDT ……



D_1 M_1

D_2 M_2

θ is the recombination fraction between disease and marker loci.

The null hypothesis (for linkage): $\theta = ½$

(Disease and marker loci unlinked. i.e. on different chromosomes)

The alternative hypothesis: $\theta < ½$

(Disease and marker loci linked, i.e (usually) close on the same chromosome)

So if we reject the null hypothesis, we get information on the location of the disease locus.

δ = Coefficient of linkage disequilibrium (LD)

= population frequency of D_1M_1 - population frequency of D_1 x population frequency of M_1

This is better called the coefficient of association.

It is a purely statistical concept. Linkage is a genetic concept. However, tests of linkage are often conducted by tests of association. Why?????

The TDT (transmission-disequilibrium test) of linkage is a family-based test. It thus avoids the problems of population stratification that can arise with population-based association (two-by-two table) tests, which are tests of the null hypothesis $\delta = 0$.

δ = Coefficient of linkage disequilibrium (LD)

= population frequency of D_1M_1 - population frequency of D_1 x population frequency of M_1

This is better called the coefficient of association.

It is a purely statistical concept. Linkage is a genetic concept. Tests of linkage are often conducted by testing whether $\delta = 0$. Why?????

The basic unit is the family trio of “mother, father and affected child”.

(More complicated cases, including pedigrees, can also be considered. Here, for simplicity, we focus throughout only on family trios.)

Population stratification problems are overcome by basing the test on the “within-family” transmission numbers.

Only transmissions from heterozygous (M_1M_2) parents are informative, so we (normally) only consider these. Given that the child is affected, the probability that any such parent transmits M_1 is not necessarily the "Mendelian" value ½ when $\theta < ½$, i.e. disease and marker loci are linked.

Prob (M_1M_2 parent transmits M_1 to an affected child)

$$= \tfrac{1}{2} + \delta(1-2\theta)K.$$

Here K is a complicated constant depending on disease and marker allele frequencies, the nature of the disease (dominant, recessive, additive, etc).

Note that if $\theta = \tfrac{1}{2}$, the two probabilities given above are equal.

These two probabilities are also equal if $\delta = 0$. Thus

the test can be used (with care) as a test of the hypothesis $\delta = 0$.

The formula confirms the fact that the test has no power as a test of $\theta = \tfrac{1}{2}$ if $\delta = 0$.

Notation (suitable later for QTDTs):

For family trio # i, ($i = 1, 2, \ldots, n$)

w_i is the observed excess of M_1 genes transmitted to the child over the null hypothesis mean of this number.

W_i is the corresponding random variable.

These two quantities are central to all that follows.

Example 1.

Suppose that both parents are M_1M_2.

There are three possibilities for the child:

M_1M_1 : here $w = 1$

M_1M_2 : here $w = 0$

$M_2\,M_2$: here $w = -1$

It follows that for these matings, Variance (W) = ½ under the null hypothesis.

Example 2.

Suppose that one parent is M_xM_x (for $x = 1$ or 2) and the other parent is M_1M_2 .

There are now two possibilities for the child:

M_1M_x : here $w = ½$.

M_2M_x : here $w = -½$.

For these matings, Variance (W) = ¼ under the null hypothesis.

Then the standard TDT statistic, viz. $(n_1 - n_2)^2/m$ can be written as:

$$\frac{\left(\sum_{i=1}^{n} w_i\right)^2}{Var\left(\sum_{i=1}^{n} w_i\right) \mid H_0}$$

This is (for all practical purposes) distributed as chi-square with one degree of freedom under the null hypothesis if m is large.

Quantitative TDTs

Here we consider some quantitative measurement in the child in each trio (e.g. BMI), and not the qualitative state (affected / not affected). The approaches that we consider are those of:

- Allison (1997)
- Abecasis *et al.* (2000a, 2000b),
- Rabinowitz (1997),
- Monks-Kaplan (2000).

All the above are in the Abecasis QTDT package.

The Rabinowitz procedure is also in the FBAT package.

(There are several other approaches available that are not considered here. Also, the PLINK package uses the Abecasis methods.)

Again to keep things simple, we consider only the case of *n* family trios, each consisting of mother, father and child. We assume that we know the marker locus genotypes of all three members of each trio, and also the quantitative measurement *(y)*, (for example BMI) in each child.

The Allison and the Abecasis methods are "regression-based". The Rabinowitz and the Monks-Kaplan approaches are in a sense the "converse" of this – see later.

We consider first the Allison and Abecasis approaches.

Let Y_i be the continuous phenotype (e.g. BMI) of the child in family i, ($i = 1, 2, \ldots, n$). This is taken as a random variable (hence the upper case notation).

For both Allison and Abecasis, the model is

$$Y_i \sim N(\mu_i, \sigma^2)$$

For the Allison and Abecasis models there is a null hypothesis value and an alternative hypothesis value for μ_i.

Regression test notation

PMT = parental mating type

PMT 1: one parent is M_1M_1, the other M_1M_2

PMT 2: both parents are M_1M_2

PMT 3: one parent is M_1M_2, the other M_2M_2

Allison linear

For family (trio) *i*, the mean of the measured quantity in the child is:

PMT 1: $\mu + \beta w$

PMT 2: $\mu + \alpha_1 + \beta w$

PMT 3: $\mu + \alpha_2 + \beta w$

The null hypothesis $\theta = ½$ becomes $\beta = 0$.

The alternative hypothesis makes no specification about β.

(Why linear? Linear in *w*.)

Abecasis "within only"

The mean of the measured quantity in the child is:

$\mu + \beta w$ for all three mating types.

In this model the null hypothesis $\theta = ½$ becomes

$\beta = 0.$

The alternative hypothesis makes no specification about β.

Abecasis "orthogonal"

The mean of the measured quantity in the child is:

PMT 1: $\mu \quad + \beta w$

PMT 2: $\mu + \alpha \; + \beta w$

PMT 3: $\mu + 2\alpha + \beta w$

The null hypothesis $\theta = ½$ again becomes $\beta = 0$.

The alternative hypothesis again makes no specification about β.

Allison quadratic

The mean of the measured quantity in the child is:

PMT 1: $\mu + \beta_1 w + \beta_2 w^2$

PMT 2: $\mu + \alpha_1 + \beta_1 w + \beta_2 w^2$

PMT 3: $\mu + \alpha_2 + \beta_1 w + \beta_2 w^2$

The null hypothesis $\theta = ½$ becomes $\beta_1 = \beta_2 = 0$. (This is the same null hypothesis as for the Allison linear model.)

The alternative hypothesis makes no specification about β_1 and β_2.

Abecasis "dominance"

The mean of the measured quantity in the child is:

PMT 1: $\mu + \beta w + \gamma d$

PMT 2: $\mu + \alpha + \beta w + \gamma d$

PMT 3: $\mu + 2\alpha + \beta w + \gamma d$

Here $d = -1$ if the child is homozygous (M_1M_1 or M_2M_2) and $d = +1$ if the child is heterozygous (M_1M_2).

In this model the null hypothesis is $\beta = \gamma = 0$. (Why??)

These models are all thus regression models. They take w as the independent variable. Also, some of them are nested within others.

In all models there is a proportion (R_1^2) of the total sum of squares removed under the null hypothesis and a (larger) proportion (R_2^2) of the total sum of squares removed under the alternative hypothesis.

This leads to standard regression hypothesis testing procedures.

Some examples of the test statistic:

Allison linear: $$F = \frac{(R_2^2 - R_1^2)}{(1 - R_2^2)/(n-4)}$$

Abecasis orthogonal: $$F = \frac{(R_2^2 - R_1^2)}{(1 - R_1^2)/(n-3)}$$

Abecasis within: $$F = \frac{(R_2^2 - R_1^2)}{(1 - R_1^2)/(n-2)}$$

Thus these procedures make an assumption of normality of *Y*.

Note also the degrees of freedom.

The Rabinowitz Approach

The main thing to remember about this approach (which is the basic FBAT approach) is that the quantitative measurements Y_i are taken as given, (that is, these are the independent variables, and thus denoted by y_i), and the transmission information w_i is then the dependent variable (and is thus denoted W_i).

THUS THE MEANING OF THE RANDOM VARIABLES IS REVERSED COMPARED TO THE ALLISON AND ABECASIS REGRESSION MODELS.

But: having W_i as the random variable is in line with the assumptions of the original qualitative TDT.

The numerator component of the test statistic is:

$$S = \Sigma_i \ (y_i - \bar{y}) \ W_i$$

Here W_i is, as before, the (random) difference between the number of genes that the child in trio *i* has and its null hypothesis mean.

Under the null hypothesis (no linkage between disease and marker loci) *S* has mean zero and variance

$$V = \Sigma_i (y_i - \bar{y})^2 \ \mathrm{Var}(W_i)$$

Recall:

$$\begin{aligned} \mathrm{Var}(W_i) &= \tfrac{1}{2} \text{ for } M_1M_2 \times M_1M_2 \text{ matings,} \\ &= \tfrac{1}{4} \text{ for all other matings} \end{aligned}$$

Thus the variance V is easily computed.

The test statistic is then $S / \sqrt{V}$, approx N(0,1) under H_0.

The numerator component of the Monks-Kaplan test statistic is

$$S = \Sigma_i \ (y_i - \bar{y}\ )\ W_i$$

That is, it is identical to the numerator of the Rabinowitz statistic.

The null hypothesis (no linkage between disease and marker) variance of S is estimated by

$$V^* = \Sigma_i \left[(y_i - \bar{y}\ )\ w_i\right]^2$$

The test statistic is then $S/\sqrt{V^*}$, approx N(0,1) under H_0.

The "reversed-role regression".

The Allison and Abecasis procedures are regressions of Y on w. What about a "reversed role" regression of W on y?

This regression model can be written as

$$W_i = \alpha + \beta(y_i - \bar{y}) + E.$$

The estimate of the slope in this regression is

$$\sum_i w_i (y_i - \bar{y}) / \sum_i (y_i - \bar{y})^2.$$

The numerator is the same as the numerator in the Rabinowitz z statistic. Using standard regression methods, we would test H_0 via a t statistic, defined as t = numerator / s.

(s = standard regression SD estimate of the numerator.)

That is, we test for non-zero slope of this regression line.

But the original TDT is a test of a non-zero intercept of this line!!!!

What is going on??? Isn't this very weird??

What are the properties, good and bad, of these procedures?

Main property of all of them: they do not test for the *absolute* values of the w_i. What they test is for *changes* in these values as a function of y, the phenotype in the child. This is obvious in the regression procedures, but is true also of the Rabinowitz procedure.

ALL these procedures would be unchanged if any arbitrary constant were added to the transmission values w.

This is in complete contrast to the aim of the qualitative TDT, which tests for the ABSOLUTE w values.

Thus quantitative TDTs test quite a different null hypothesis than do qualitative TDTs.

The aim of using the transmission approach is to overcome problems of population stratification. Do these procedures do this?

No – the Abecasis procedures do not do this if mating type is associated with population strata.

For the Abecasis "within" test. If mating type is associated with strata, the mean of the numerator in the Abecasis F ratio is

$$\sigma^2 + \text{positive term}$$

The mean of the denominator is

$$\sigma^2 + \text{a different positive term}$$

The Rabinowitz and Monks-Kaplan procedures, and the Allison regression procedures, ARE immune to population stratification.

There are many further considerations: power, dominance, using uninformative mating types (for example $M_1M_1 \times M_2M_2$) etc.

The take-home message: use QTDT packages with extreme caution. (More details are available in a handout.)