

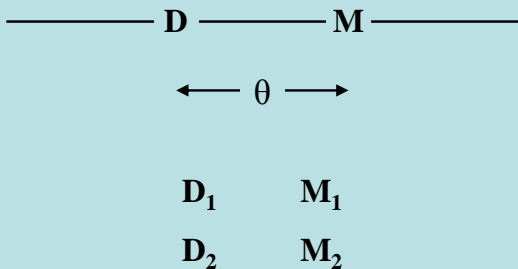
The qualitative and the quantitative TDT

(Transmission – Disequilibrium Test)

The initial aim of the *qualitative* TDT was to test for *linkage* between a marker locus M (with alleles M_1 and M_2) and a disease locus D (with alleles D_1 and D_2). Assume that D_1 is the disease-predisposing allele.

Qualitative = affected or not affected.

Using the TDT as a test of *association* between the alleles at M and the alleles at D came later. So we start with linkage.



Null hypothesis (for linkage): $\theta = \frac{1}{2}$
(Disease and marker loci unlinked)

Alternative hypothesis: $\theta < \frac{1}{2}$
(Disease and marker loci linked)

How can we test this null hypothesis? There are two main approaches in the qualitative (affected / not affected) case.

1. Population-based tests. Of these the oldest, and most popular, is the case-control study.
2. Family-based tests.

CASE / CONTROL ANALYSIS

Gene counts in R_1 cases and R_2 controls

	M_1	M_2	Total
AFFECTED (CASES)	n_{11}	n_{12}	$2R_1$
NOT AFFECTED (CONTROLS)	n_{21}	n_{22}	$2R_2$

The case-control test is a test of the null hypothesis $\delta = 0$, where

$$\delta = \text{freq}(D_1M_1) - \text{freq}(D_1) \times \text{freq}(M_1)$$

The main problem with the case-control approach is that it is a test of association, not directly of linkage. So why is it used as a (surrogate) test of linkage?

Historical. The disease allele D_1 is assumed to have arisen comparatively recently. Suppose that it arose by mutation from D_2 on an M_1 - bearing gamete. Then at the time of the mutation there was 100% association between D_1 and M_1 . If disease and marker loci are closely linked, this association might persist to some extent today.

Unfortunately, association can arise through population stratification as well as through linkage.

Family-based tests are used to overcome this problem.

The TDT

The TDT (transmission-disequilibrium test) of linkage is a family-based test – to avoid problems of population stratification. It relies however for its operation on there being association between marker and disease loci.

How does it work?

The simplest basic unit is the family trio of “mother, father and affected child”. So we only consider this case in detail.

Only transmissions from heterozygous (M_1M_2) parents are informative, so we only consider these. Given that the child is affected, the probability that any such parent transmits M_1 is not necessarily $\frac{1}{2}$ if $\theta < \frac{1}{2}$.

Prob (het. parent transmits M_1) - Prob (het. parent transmits M_2)

$$= k(1-2\theta)\delta$$

Where k is a complicated constant depending on disease and marker allele frequencies.

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

For family trio # i , we define w_i as the excess of the number of M_1 genes transmitted to the child over its null hypothesis mean.

The TDT test is based on the sum of the w_i values, summed over all trios in the data.

Example

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_2M_2 : here $w = -1$

The TDT statistic is $4(\sum_i w_i)^2/m$, where m is the total number of heterozygous parents in the data set.

Under the null hypothesis this has approximately a chi-square distribution with one degree of freedom.

The TDT statistic is more commonly written as

$$(n_1 - n_2)^2/m,$$

where n_1 is the total number of transmissions of M_1 and n_2 is the total number of transmissions of M_2 from all the heterozygous parents in the data set.

This is the TDT statistic. A property of the TDT procedure:-

When $H_0 : \theta = \frac{1}{2}$ is true, transmissions of marker alleles to two or more affect sibs are independent.

Therefore the TDT may be used as a test of this null hypothesis when the data contain families with two or more affected children.

Another property of the TDT is that it has *increased* power when association is higher. This is shown by the probabilities considered above when there is no stratification. The larger is δ the larger is the value of $(1-2\theta)\delta$, and thus the larger is the power to test the null hypothesis $\theta=\frac{1}{2}$.

The TDT as a test of association

- The TDT is now often, even mainly, used as a test of association.
- Often this happens to confirm linkage in a situation where association has been found via a two-by-two population-based test.
- Here H_0 is $\delta = 0$.

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).

There are several approaches to TDTs in the literature, and we do not discuss them all. Instead, we discuss some of those in the QTD and the FBAT packages (more details later).

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, for example BMI, in each child.

For models where this trait is treated as a random variable, it is denoted in upper case as Y . For models where this trait is treated as a non-random, it is denoted in lower case as y .

The null hypothesis is that the marker locus is not linked to any locus affecting the quantitative trait. This is equivalent to saying that y does not depend on w .

Example (remember?)

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_2M_2 : here $w = -1$

Regression-based models

The first approaches that we consider are regression-based. Here the quantitative measurement is taken as a random variable, and thus is denoted Y . The independent variable is w . Y is (usually) also assumed to depend on the parental mating type of each trio (more details later).

In more detail, the value Y_i of the measurement of the child in family i , ($i = 1, 2, \dots, n$), is a random variable with mean possibly depending on w_i and the family mating type, and with variance σ^2 . This measurement is also assumed to have a normal distribution.

Allison linear:

$$Y = \mu + \beta w + E \quad \text{if one parent is } M_1M_1, \text{ the other is } M_1M_2$$

$$Y = \mu + \alpha_1 + \beta w + E \quad \text{if both parents are } M_1M_2$$

$$Y = \mu + \alpha_2 + \beta w + E \quad \text{if one parent is } M_1M_2, \text{ the other is } M_2M_2$$

Abecasis orthogonal:

$$Y = \mu + \beta w + E \quad \text{if one parent is } M_1M_1, \text{ the other is } M_1M_2$$

$$Y = \mu + \alpha + \beta w + E \quad \text{if both parents are } M_1M_2$$

$$Y = \mu + 2\alpha + \beta w + E \quad \text{if one parent is } M_1M_2, \text{ the other is } M_2M_2$$

Abecasis within:

$$Y = \mu + \beta w + E \quad \text{for all three parental mating types.}$$

In all three models the null hypothesis is $\beta = 0$.

These models are nested, with Allison's being the most general.

In all three 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.

Test statistic:

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

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

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

Thus these procedures use standard regression methods.

The Rabinowitz approach

The main thing to remember about this approach (which is the basic FBAT approach) is that the quantitative measurements Y_i in family i is 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.

The Monks-Kaplan approach (also in the QTD package) also takes W_i as the random variable.

The numerator component of the Rabinowitz test statistic is

$$\sum_i y_i \cdot w_i$$

Here w_i is, as before, the difference between the number of M_1 genes that the child in trio i has and its null hypothesis mean, and y_i is the difference between y_i and the average of the n values of y in the data set.

Under the null hypothesis (no linkage between disease and marker loci) this numerator component has mean zero and variance

$$V = \sum_i (y_i)^2 \sigma_i^2.$$

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

What are the properties of these procedures?

Main property of all of them: they do not test for the *absolute* W_i values. 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.

This is in contrast to the aim of the qualitative TDT.

In fact the test of the *intercept*, and not the slope, of the "role reversed" regression is equivalent to the original qualitative TDT test.

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.

The Allison procedures are immune to this problem, as is the Rabinowitz procedure.

There are many further considerations: power, dominance, using uninformative mating types (for example $M_1M_1 \times M_2M_2$) etc. Some of these properties are not yet known.

The take-home message: use the Abecasis QTD T package with extreme caution. (More details are available in a handout.) If in doubt, use the Allison or the Rabinowitz methods. Think about your data and which procedure is best suited to it.