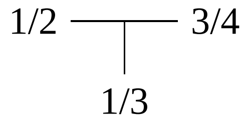
Practical exercises on transmission/disequilibrium (1): the TDT and extensions

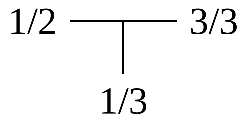
David Clayton and Joanna Howson

1 Informative transmissions

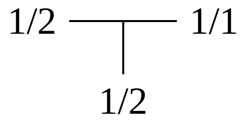
The following represent trios of an affected offspring and both parents. Alleles are coded 1–4 and unknown genotypes are denoted by ?/?.



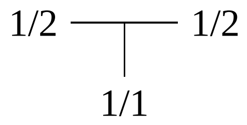
(1)



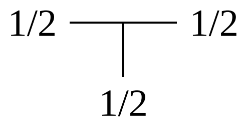
(2)



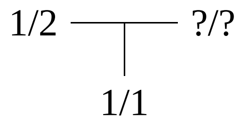
(3)



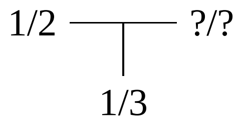
(4)



(5)



(6)



(7)

Determine, in each case, how many informative transmissions are provided by the family. Excluding uninformative transmissions, what are the observed and "expected" numbers of transmissions of each allele?

2 Preparing computer files

Prepare the data for the above families in the standard "pre-ped" format used in the Linkage package.[1] You can create the data file either in Stata's own data editor (started from the toolbar), or by preparing a standard text file named, for example, `exercise.pre`. [2] In either case, the variables to be entered should appear in the following order:

pedigree code, member id, father's id, mother's id, sex, affected, allele1, allele2 .

[1] If you wish to skip this part of the exercise, a data file is supplied. You can read it in by typing "`use solution`"

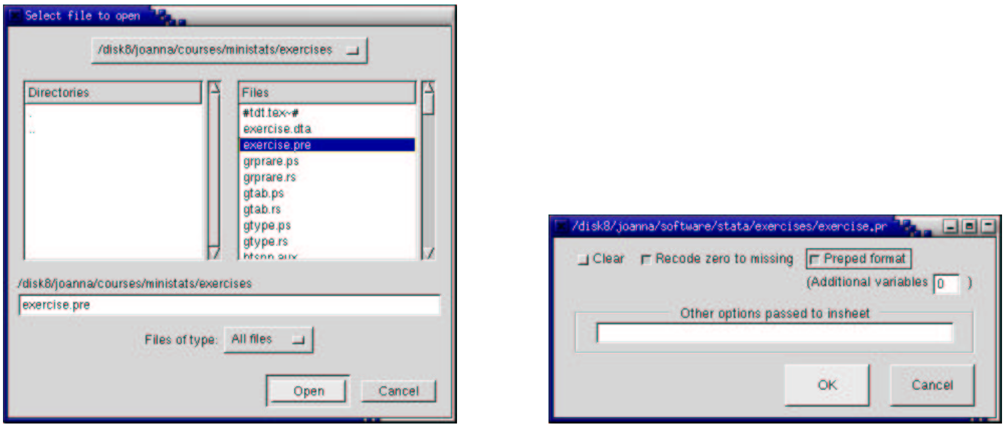
[2] If you enter the data using the Windows "Notepad" tool, remember to select file type as "all" when you save the file; otherwise the file will be named exercise.ped.txt!

In this example, the file contains data on a single marker genotype — further markers contribute additional pairs of columns. Disease status ("affected") should be coded 1 for unaffected and 2 for affected. Sex is coded 1 for Male and 2 for Female (although these data are not given here). Identifiers for pedigrees and for members within pedigrees are usually integers. You should code offspring as "affected" (2). It does not matter how you code parents since this information is not used in the simple analyses described here.

If you are using an editor to create a text file, fields should be separated by tab characters and missing items should be entered as `0` (zero). Each subject's data should form a single line, and you should make sure that the last line has been properly terminated so that the end of file is at the start of the next line (otherwise the last line will not be read). Begin by generating the menu file with the command,

```
. gamenu
```

You can now read the preped file by selecting **Read Spreadsheet**, from the **Data management** sub menu of the **GenAssoc** menu. Select your file from those listed in the window. Note, to see your file, you may have to select 'Files of type: `All files`', see Figure 1. Click open. In the new window, click the 'Preped format' option ('Recode zero to missing' should already be selected). See Figure 1.



Select file to open

/disk8/joanna/courses/ministats/exercises

Directories: . ..

Files: #tdt.tex~#, exercise.dta, exercise.pre, grprare.ps, grprare.rs, gtab.ps, gtab.rs, gtype.ps, gtype.rs, htspp.aux

/disk8/joanna/courses/ministats/exercises

exercise.pre

Files of type: All files

/disk8/joanna/software/stata/exercises/exercise.pr

□ Clear ☑ Recode zero to missing ☑ Preped format

(Additional variables 0)

Other options passed to insheet

Figure 1: Importing data into Stata.

The equivalent command line is,

```
. ginsheet using exercise.pre, zmiss preped
```

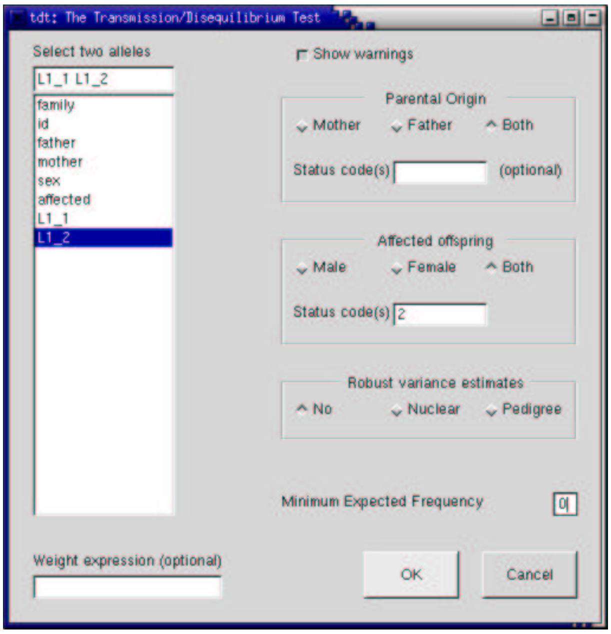
Note that the command `ginsheet` needs the option `zmiss`, to convert the missing data code zero, into Stata's own internal code for missing, denoted by `.` (period). The locus is named `L1` by default and its two alleles denoted by `_1` and `_2` respectively.

Whichever way you chose to enter the data, you should save the data as a Stata `.dta` file for later use:

```
. save exercise
```

If you have already saved an earlier version of the data, you must type

```
. save exercise, replace
```



tdt: The Transmission/Disequilibrium Test

Select two alleles: L1_1 L1_2

family
id
father
mother
sex
affected
L1_1
L1_2

□ Show warnings

Parental Origin: ○ Mother ○ Father ● Both
Status code(s): (optional)

Affected offspring: ○ Male ○ Female ● Both
Status code(s): 2

Robust variance estimates: ● No ○ Nuclear ○ Pedigree

Minimum Expected Frequency: 0

Weight expression (optional):

Figure 2: TDT analysis.

After you have entered the data, run the Stata program `tdt` by selecting **TDT** from the **GenAssoc** menu. In the window that opens, select `L1_1` and `L1_2`, and set the 'Minimum Expected Frequency' to `0`, as in Figure 2.

Alternatively you may wish to use the command line,

```
. tdt L1_1 L1_2, emin(0)
```

(Note that `L1_1 L1_2` may be replaced by `L1_*`). Check that these results agree with the answers you worked out yourself. It is quite likely that they won't. Think very hard about pedigree (5)!

3 IDDM and three SNP markers in the MHC class 3 region

The next exercise concerns insulin dependent diabetes mellitus (IDDM) and a set of three closely spaced markers in the MHC class 3 region. The data are a very small subset of a much larger study. The markers `bat2`, `bat3` and `ng36` are separated by 20 kb and 260 kb respectively. These cover a region strongly implicated in IDDM by linkage analysis. Read the file in as follows: select **Read Stata data file**, from the **Data management** sub menu under the **GenAssoc** menu, and read in the mhc3iddm.dta file. Alternatively you can use the command,

```
. use mhc3iddm
```

Now find out the name of the markers with

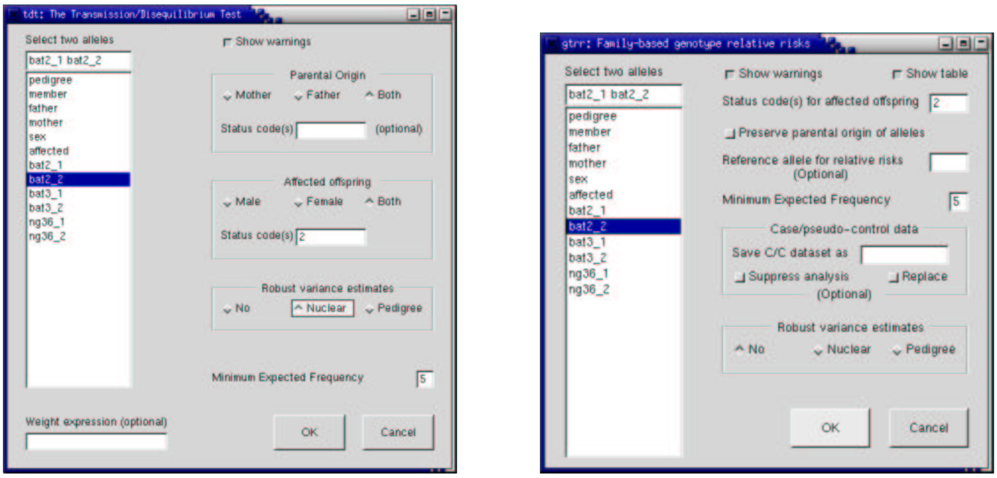
```
. describe
```

Use the `tdt` command to test for association between disease and each marker in turn.

Note that these data concern affected sib pairs and there is very strong linkage in the region. Investigate the effect of using the `robust` option with the `tdt` program, e.g.

```
. tdt bat2_*, robust
```

Alternatively, if you wish to use the menus, select **TDT** from the menu as before. Select the two alleles, `bat2_1` and `bat2_2`; keep the 'Minimum Expected Frequency' as the default `5`, but click on `Nuclear` in the 'Robust variance estimates' box, in Figure 3.



tdt: Select two alleles: bat2_1 bat2_2; Affected offspring Status code(s): 2; Robust variance estimates: Nuclear; Minimum Expected Frequency: 5

gtrr: Select two alleles: bat2_1 bat2_2; Status code(s) for affected offspring: 2; Minimum Expected Frequency: 5; Robust variance estimates: No

Figure 3: The left Figure shows TDT with the robust option. GTRR is shown in the Figure on the right.

The command `gtrr` estimates genotype relative risks. Select **Genotype RR** from the **TDT etc** sub menu of **GenAssoc**. Choose alleles, `bat2_1` and `bat2_2`, as in Figure 3.

Alternatively you can use the command line,

```
. gtrr bat2_*
```

You should again investigate the effect of using the `robust` option. Does this analysis suggest anything about the mode of inheritance (dominant or recessive) of the gene in this region?

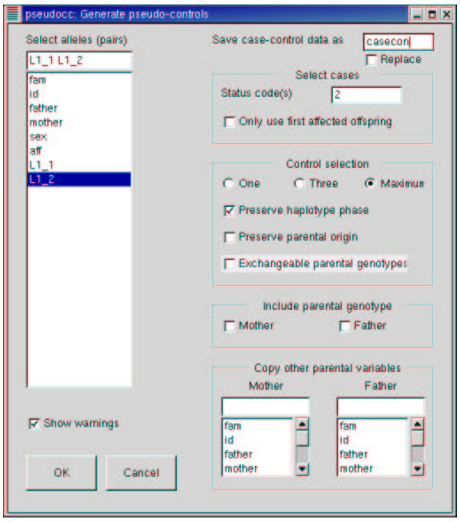
4 Case/pseudo–control studies

The command `gtrr` works by creating a case/pseudo–control study consisting of sets comprising (a) the genotype of the affected offspring (the "case"), and (b) the other three genotypes which the subject might have received from his or her parents (the "controls").[3] The relative risks can then be estimated using conditional logistic regression. Optionally, `gtrr` will save this file for later analysis, but the more general command `pseudocc` can also be used to create a case/pseudo–control data-set.

Read in the data file you created in the first exercise. Assuming that the variables holding the two alleles are called `L1_1` and `L1_2`, the case-control dataset can be created

[3] In some cases, the case or some of the controls are inadmissible since they would not allow inference of the parental genotypes. In such cases the set is either missing or may contain fewer controls.

as follows. Choose **Pseudo-CC** from the **TDT etc** sub menu of **GenAssoc**. Fill in the fields as in Figure 4: 'Select alleles' `L1_1` and `L1_2`; fill in the 'Save case-control data as' box with, `casecon`; run the command by clicking OK.



pseudocc: Generate pseudo-controls

Select alleles (pairs): L1_1 L1_2

Save case-control data as: casecon

□ Replace

Select cases

Status code(s): 2

□ Only use first affected offspring

Control selection

○ One ○ Three ● Maximum

☑ Preserve haplotype phase

□ Preserve parental origin

□ Exchangeable parental genotypes

Include parental genotype

□ Mother □ Father

Copy other parental variables

☑ Show warnings

Figure 4: The pseudo-cc window.

The command line for this is,

```
. pseudocc L1_*, saving(casecon)
```

Delete the data set from Stata memory with

```
 . clear
```

Read the case-control data into Stata either using the menu **Read Stata data set** and clicking on casecon.dta, or with the command line,

```
. use casecon
```

Use the data browser to inspect the file contents (thats the blue spreadsheet button with the magnifying glass, just below **GenAssoc**). Has the program produced the case-control sets you would expect?

The same results you obtained with `gtrr` could be obtained by creating a case/pseudo–control study in this way, and analysing it by conditional logistic regression (the `clogit` command in Stata). While `gtrr` is more convenient for simple analyses, this approach is very useful since, as we shall see, it allows more difficult questions to be addressed.

5 Several loci

It is possible to extend case/pseudo–control analysis with several loci. There are then two ways we can make the case-control study, according to whether or not we require *haplotype phase* to be known for cases and controls. Delete the pseudo-cc data, using `clear` and read in the mhc3iddm.dta file once more. Again select the **Pseudo-CC** menu. This time choose alleles, `bat2_1`, `bat2_2`, `bat3_1`, `bat3_2`. Click the 'Preserve haplotype phase' button off and save the case control data set as `mhccc`. To create case/pseudo–control data where you wish haplotype phase to be known, do the same except, click the 'Preserve haplotype phase' button *on*, and save the dataset as `mhcccph`. We can also make these files for the two markers `bat2`, `bat3`, with the commands:

```
. pseudocc bat*, saving(mhccc)
. pseudocc bat*, saving(mhcccph) phase
```

As might have been expected, the phased case-control study is rather smaller.

Our initial analyses will only consider models in which phase need not be known, so we use the first file:

```
. use mhccc, clear
```

Or `clear` the mhc3iddm dataset, and use the **Read Stata dataset** menu to read in the mhccc.dta file you created. We shall first generate the necessary indicator variables. Choose **Allele frequencies** from the **Tabulate** sub menu. Click on `bat2_1` and `bat2_2`. Put `B2_` in the 'Variable prefix string' of the 'Generate indicator variables' box. Do the same for bat3 but use `B3_` as the indicator variable prefix. Alternatively,

```
. quietly gtab bat2_*, gen(B2_)
. quietly gtab bat3_*, gen(B3_)
```

(The word `quietly` suppresses the output from the command, but the indicator variables will still be generated.) The commands to carry out the single locus analyses are,

```
. clogit case B2_*, group(set)
. clogit case B3_*, group(set)
```

Within the menu system, you can select **Fit** from the **Regression** sub menu. Click on `clogit` as the 'Regression command', `case` as the response variable, and `B2_1` and `B2_2`, the 'Metric' explanatory variables. Use the option, `group(set)`. Do the same for B3, except select the `B3_1` and `B3_2`.

One possible two-locus analysis is, having chosen the **Fit** menu, select `clogit` as the regression command, `case` as the response variable, have `group(set)` as the option but select `B2_1`, `B2_2`, `B3_1`, `B3_2` as the metric explanatory variables. Or use the command line,

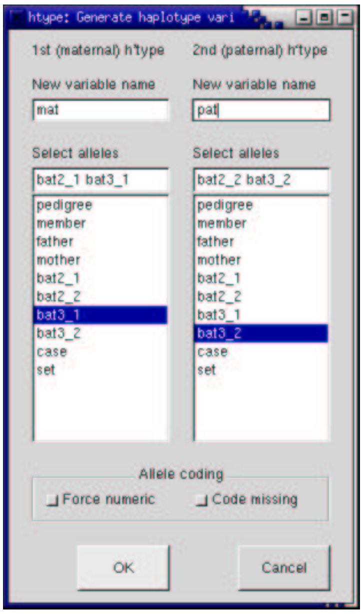
```
. clogit case B2_* B3_*, group(set)
```

How would you interpret this result? You should investigate the effect of using "robust" standard error estimates by adding the `cluster(pedigree)` option. [4]

Our next analysis looks at haplotypes for the two loci. We first read in the data, having used `clear` to delete the previous data set, and create two variables to hold the maternal and paternal *haplotypes*. Read in the Stata data set you created, mhcccph.dta, with either the **Read Stata data set** menu, or with the command,

```
. use mhcccph
```

Select, **Create haplotype variables** from the **Recode** sub menu. Fill in the boxes as in Figure 5. Alternatively you can use the command lines,



htype: Generate haplotype vari

1st (maternal) h'type	2nd (paternal) h'type
New variable name: mat	New variable name: pat
Select alleles: bat2_1 bat3_1	Select alleles: bat2_2 bat3_2

Allele coding

□ Force numeric □ Code missing

Figure 5: Generating haplotype variables.

[4] This option is not available in version 7 of Stata. It should become available in later versions. In the meantime, an alternative unofficial version, `rclogit`, is distributed with this package.

```
. egen mat = htype(bat*_1)
. egen pat = htype(bat*_2), co(mat)
```

(Note that `bat*_1` will expand to `bat2_1 bat3_1`, which hold the maternal copies of the `bat2` and `bat3` loci. The `co` option forces the paternal haplotype to be coded identically to the maternal one). Indicator variables counting occurrences of each haplotype can now be created. So select **Allele Frequencies** from the **Tabulate** sub menu. Choose the haplotypes, `mat` and `pat` in the 'Select two alleles' field, and `hap` as the indicator variable prefix string. This is the same as,

```
. quietly gtab mat pat, gen(hap_)
```

and we can estimate relative risks with the command

```
. clogit case hap_*, group(set)
```

This can be achieved within the **Fit** menu by choosing `clogit` as the regression command, `group(set)` as the option, `case` as the response variable, and `hap_1`, `hap_2`, `hap_3`, `hap_4` as the metric explanatory variables.