BIOSTAT551 Statistical Genetics II: Quantitative Traits — Winter 2004

Homework 4

General Instructions: For all problems, show how you determined your answer. Be sure to define all variables that you use and justify your logic. If you have used software to obtain answers, output should be attached, but this is not a substitute for showing the steps of your analysis! The required data sets are available from the course web-site.

4.1 The apolipoprotein E (APOE) gene is associated with late-onset familial and sporadic Alzheimer Disease and early-onset sporadic Alzheimer Disease. There are three primary variants as listed below (with estimated allele frequencies):

Allele	Frequency
E2	0.08
E3	0.77
E4	0.15

Due to the late-onset of Alzheimer disease, collection of parents is not typically feasible. However, it is reasonable to collect unaffected siblings. Suppose that you have a sample of discordant sib-pairs (one affected and one unaffected) and that you wish to test for association using the E4 variant. Define ***X*** to be the number of APOE-E4 alleles in the affected sib and ***p*** to be the frequency of the APOE-E4 variant in a given population. (collapse E2 and E3 into one class of variants, i.e. look at APOE as a locus with 2 alleles)

a) List the 6 possible genotype configurations for a discordant sibpair along with the value of ***X*** for the given sib-pair. (5 pts)

b) Under the null hypothesis, what is the expectation of ***X*** (conditional on the pair of genotypes obtained per sibship)? Note that this conditioning does not pair a genotype with disease status. (10pts)

c) Under the null hypothesis, what is the variance of ***X***? (10pts)

d) Propose a test statistic for testing the null hypothesis of no linkage or no association given a sample of ***N*** discordant sib-pairs. (5pts)

4.2 Suppose that you are interested in testing whether or not a marker is linked to a susceptibility locus, i.e. the trait of interest is dichotomous (z=1 for diseased and 0 otherwise). Suppose that the data consist of full-sib pairs and that Haseman-Elston regression is to be used to test for linkage.

a) Denote the squared trait difference for the j^{th} sib-pair, $(z_{1j}-z_{2j})^2$, by Y_j. What values of Y_j are possible? To what do these values correspond, i.e. what are the types of sib-pairs?

b) Restate the null and alternative hypotheses for the Haseman-Elston regression in terms of the classes of sib-pairs designated in part a).

c) The file HE.dat, found on the exam link of the class web-page, contains the results of the Haseman-Elston test applied to 300 markers located across 6 chromosomes. Give your conclusions and recommendations for further genetic work from this partial genome scan. Use an individual significance level of 0.05/300=0.00016667 for each marker.

d) Suppose that you tested each marker at 0.05. What would your conclusions be then? (I'll tell you where the susceptibility genes are located once you turn in the final).

e) What are the risks to the genetic study if the testing procedure from part c) is used? If the testing procedure from part d) is used?

4.3 LDL cholesterol (LDL-C) concentration has been found to be a risk factor for cardiovascular disease. Significant linkage for one measure of LDL-C was found for a region on chromosome 4. The following are allele frequencies for marker D4S1647 which is centrally located in this region:

$\Pr(M_1)= 0.2000$
$\Pr(M_2)= 0.0143$
$\Pr(M_3)= 0.2000$
$\Pr(M_4)= 0.3143$
$\Pr(M_5)= 0.1571$
$\Pr(M_6)= 0.0714$

$\Pr(M_7) = 0.0429$

a) Suppose that we have some evidence that marker allele M_5 represents a functional polymorphism and that we wish to perform the transmission/disequilibrium test for this specific allele (after dichotomizing LDL-C into high and low). What fraction of the parents will be informative for the TDT for this marker allele?

b) The following table summarizes transmissions from parents to children with high LDL-C for 104 parent-child trios. Compute the TDT statistic. What can you **CONCLUDE** from this test. State your **INTERPRETATION** of the results.

	NOT TRANSMITTED	
TRANSMITTED	M_5	**Other**
$\mathbf{M_5}$	8	39
Other	19	142