Stat 597D: Topics in bioinformatics — Spring 2001
Assignment 3 — Due March 27

Problem 1. The data below are taken from Morgan et al. (1935). This is the same Morgan for whom the unit of distance is named. He and his collaborators phenotyped 16136 flies at 9 loci on the Drosophila X chromosome. For each fly, the researchers were able to determine for one gamete whether that gamete was recombinant on each of the 8 interlocus intervals.

Table 1: Gamete Counts in the Morgan et al. Data

S	**n**	S	**n**	S	**n**	S	**n**
ϕ	6607	$\{2,4\}$	38	$\{6,7\}$	21	$\{2,5,6\}$	3
$\{1\}$	506	$\{2,5\}$	85	$\{6,8\}$	30	$\{2,5,7\}$	4
$\{2\}$	1049	$\{2,6\}$	237	$\{7,8\}$	2	$\{2,5,8\}$	1
$\{3\}$	855	$\{2,7\}$	123	$\{1,2,3\}$	1	$\{2,6,7\}$	2
$\{4\}$	1499	$\{2,8\}$	70	$\{1,2,6\}$	1	$\{2,6,8\}$	3
$\{5\}$	937	$\{3,4\}$	22	$\{1,3,5\}$	1	$\{2,7,8\}$	2
$\{6\}$	1647	$\{3,5\}$	55	$\{1,4,5\}$	1	$\{3,4,7\}$	2
$\{7\}$	683	$\{3,6\}$	177	$\{1,4,6\}$	1	$\{3,4,8\}$	1
$\{8\}$	379	$\{3,7\}$	88	$\{1,4,7\}$	2	$\{3,5,6\}$	1
$\{1,2\}$	3	$\{3,8\}$	38	$\{1,4,8\}$	1	$\{3,5,7\}$	2
$\{1,3\}$	6	$\{4,5\}$	41	$\{1,5,7\}$	2	$\{3,5,8\}$	3
$\{1,4\}$	41	$\{4,6\}$	198	$\{1,5,8\}$	1	$\{3,6,7\}$	1
$\{1,5\}$	55	$\{4,7\}$	159	$\{1,6,8\}$	1	$\{3,6,8\}$	1
$\{1,6\}$	118	$\{4,8\}$	91	$\{2,3,6\}$	1	$\{4,5,8\}$	1
$\{1,7\}$	54	$\{5,6\}$	35	$\{2,4,6\}$	4	$\{4,6,8\}$	4
$\{1,8\}$	34	$\{5,7\}$	49	$\{2,4,7\}$	5	$\{4,7,8\}$	1
$\{2,3\}$	3	$\{5,8\}$	40	$\{2,4,8\}$	6	$\{5,6,8\}$	1

For example, the number of gametes observed to be recombinant on the 2nd and the 6th intervals is 237. The number of gametes with no recombinations on any of the 8 intervals is 6607.

Suppose we fit Haldane's model to these data.

(a) Give maximum likelihood estimates of the eight interlocus interval lengths, expressed in centimorgans.

(b) Give the expected information matrix evaluated at the MLE and its inverse. The inverse can be used as an estimate of the covariance matrix for the MLE. Report the standard error for each of the 8 estimates in part (a)—These will be the square roots of the diagonal entries of the inverse information matrix, converted to centimorgans.

Hints: These questions are not as computationally intensive as they may look; they can even be done by hand (even the inversion of the 8×8 matrix!). Also, this example comes from Lange's book, pp. 221–222, but beware: the Haldane entries in Table 12.2 are incorrect.

Problem 2. In a study involving 16 Costa Rican children affected with the disease ataxia-telangiectasia (AT), 28 of their 32 parents are heterozygous at a particular marker locus. For these 28 heterozygous parent-affected child pairs, the summary table is given below:

	A	B	C	D	E	F	G	H	I	J	Total
Transmitted	3	0	22	0	1	0	0	0	0	2	28
Not transmitted	0	4	0	4	3	4	1	1	2	9	28
Total	3	4	22	4	4	4	1	1	2	11	56

The Pearson χ^2 statistic for the above table is 46.46; however, the usual Pearson χ^2 test is not valid here. Unfortunately, I could not obtain the original data from which this table comes. Therefore, I fabricated the following dataset, which may be seen to produce the same table as above:

Parent genotype; allele transmitted	Count
A/B ; A	2
A/E ; A	1
B/C ; C	2
C/D ; C	3
C/E ; C	2
C/F ; C	3
C/H ; C	1
C/I ; C	2
C/J ; C	9
D/E ; E	1
F/J ; J	1
G/J ; J	1
Total	28

Sample a large number of times (say, 10^5) from the null distribution of the χ^2 statistic for the null hypothesis of no linkage or linkage disequilibrium.

(a) Plot a histogram of the sample you obtain. Scale this histogram so that it represents a probability distribution and superimpose a plot of the chi square density on 9 degrees of freedom. Does the null distribution appear to be well-approximated by the chi-square?

(b) Estimate a p-value for the observed $\chi^2 = 46.46$.

(c) Define the statistic

$$Z_{\max} = \frac{n_{ij} - (n_j/2)}{\sqrt{n_j/4}},$$

where n_{ij} is the table entry in the ith row and the jth column, $1 \le i \le 2$, $1 \le j \le 10$, and n_j is the jth column sum. Repeat parts (a) and (b) for this new statistic (except you should not superimpose a chi-square distribution on the histogram—just produce the histogram and estimate the p-value).

Problem 3. Go to Beatrix's second lecture notes at

`http://www.stat.psu.edu/%7Etrix/teaching/genetics/`

and do what you are asked to do in the section marked **Homework** at the bottom of page 2. The data are all found at the web site above.