STAT 550 Howework 6 Anton Amirov

1. This question relates to the same study you saw in Homework-4, by Dr. Arno Motulsky and coworkers, and published in Thompson et al. (1988; Am.J.Hum.Genet, 42, 113-124). There were three population samples (all from around Seattle), (Caucasian, African American, and Japanese American), and three tightly linked diallelic loci, designated M, P and S.
In the Japanese subsample, the 68 individuals (136 chromosomes) were scored from all three loci. As in the Homework-4 question, there were few double-heterozygotes for any pair of the loci. So you can treat the estimated sample haplotype frequencies as though they were observed.
(a) For loci S and M, the estimated haplotype frequencies are 0.551, 0.082, 0.008, 0.360. Is there evidence for disequilibrium between loci S and M?

In 2x2 contingency table

$$\chi^2 = n\frac{(q_{11}q_{22} - q_{12}q_{21})^2}{(q_{11}+q_{12})(q_{11}+q_{21})(q_{12}+q_{22})(q_{21}+q_{22})} = 92.36$$

For 5% confidence χ^2 with 1 degree of freedom should exceed 3.841, so there is clear evidence for disequilibrium between S and M.

(b) For loci P and M, the estimated haplotype frequencies are 0.514, 0.398, 0.045, 0.043. Is there evidence for disequilibrium between loci P and M?

By the same formula

$$\chi^2 = 0.12$$

It is too small to infer disequilibrium.

(c) For loci S and P, the estimated haplotype frequencies are 0.592, 0.041, 0.320 0.047. Is there evidence for disequilibrium between loci S and P?

$$\chi^2 = 1.57$$

Again, it is not enough to infer disequlibrium.

2. In a simple backcross experiment between two inbred lines, hybrid AB/ab individuals are crossed with ab/ab individuals, and the numbers of recombinant offspring are counted. Among a total of 120 offspring in which the hybrid individual was female, 50 were of the recombinant types.
Among 90 offspring in which the hybrid individual was male, 38 were of the recombinant types.
(a) Is there evidence for linkage, using only the data on the offspring of hybrid males?

For males ρ has MLE of 38/90=0.4222, and its base-10 lod score is

$$lod(\hat{\rho}) = t\log(\hat{\rho}) + (n-t)\log(1-\hat{\rho}) + n\log 2 = 0.4748$$

This score is too low to infer linkage

(b) Is there evidence for linkage, using only the data on the offspring of hybrid females?

For females ρ has MLE of 50/120=0.4166, and its base-10 lod score is

$$lod(\hat{\rho}) = t\log(\hat{\rho}) + (n-t)\log(1-\hat{\rho}) + n\log 2 = 0.7272$$

This score is too low to infer linkage

(c) Assuming male and female recombination frequencies are the same, is there evidence for linkage?

For both ρ has MLE of 88/210=0.4190, and its base-10 lod score is

$$lod(\hat{\rho}) = t\log(\hat{\rho}) + (n-t)\log(1-\hat{\rho}) + n\log 2 = 1.2006$$

This score is still too low to infer linkage

(d) Is there evidence from this experiment that male and female recombination frequencies differ?

There is almost no difference in MLE for ρ between males and females, so nothing suggests that recombination frequencies differ. Difference in lod scores is explained by the fact that there is more data on females than on males.

3. **A couple, each of whom is unaffected, have two kids affected by a recessive disease. At a linked marker, the parents have four distinct alleles. One parent is ab, and the other is cd. Both the affected kids are bd at the marker, and the recombination probability between trait and marker is r. A fetus is typed as bc. Show that the probability the baby will be affected is**
$r(1-r)((1-r)^4 + r(1-r)\,R + r^4)/R^2$
where $R = r^2 + (1-r)^2$.

Both children should be either recombinant or non recombinant is meioses from both parents. So the probability that they are both recombinant in meiosis

$$\frac{r^2}{r^2 + (1-r)^2}$$

To be affected, for meioses from parent ab fetus need also to be recombinant if children are recombinant and non-recombinant otherwise, probability of that event is

$$\frac{r^2}{r^2 + (1-r)^2} r + \frac{(1-r)^2}{r^2 + (1-r)^2}(1-r) = \frac{r^3 + (1-r)^3}{r^2 + (1-r)^2}$$

For meioses from parent cd fetus need also to be non-recombinant if children are recombinant and recombinant otherwise, probability is

$$\frac{r^2}{r^2 + (1-r)^2}(1-r) + \frac{(1-r)^2}{r^2 + (1-r)^2} r = \frac{r(1-r)}{r^2 + (1-r)^2}$$

So the probability that fetus is affected is the product of those probabilities (as events are independent) and is

$$\frac{r^3 + (1-r)^3}{r^2 + (1-r)^2} \frac{r(1-r)}{r^2 + (1-r)^2} = r(1-r)\frac{r^3 + (1-r)^3}{(r^2 + (1-r)^2)^2} = r(1-r)\frac{r^3 + (1-r)^3}{R^2}$$

We can see that

$$\begin{aligned}(1-r)^4 + r(1-r)R + r^4 &= (1-r)^4 + r^3(1-r) + r(1-r)^3 + r^4 \\ &= r^3(1-r+r) + (1-r)^3(1-r+r) = r^3 + (1-r)^3\end{aligned}$$

Therefore probability that fetus is affected can also be written as

$$r(1-r)\frac{(1-r)^4 + r(1-r)R + r^4}{R^2}$$

4. Suppose we sample families with two offspring, in which one parent is heterozygous for a dominant genetic disease allele, and marker type ab, while the unaffected spouse has marker type aa. The familes are divided into two groups:
Group 1: the two offspring share the same disease status, or the same marker type, but not both
Group II: all other combinations.
Suppose the recombination frequency between the disease locus and the marker locus is r.

(a) Show that the probability a given family is of Group-I type is s = 2 r (1-r).

If both children have the same marker (probability ½) to have different disease status one child needs to recombinant in meiosis from the affected parent, and another one non-recombinant (probability 2r(1-r)). If they have different markers (probability ½) to have the same disease status again one child has to be recombinant and another one -non-recombinant. So the total probability is

$$\frac{1}{2}2r(1-r)+\frac{1}{2}2r(1-r)=2r(1-r)$$

(b) Suppose n families are sampled, and k are of Group-I type. Show that the log-likelihood is, up to an additive constant,
k log s + (n-k) log (1-s) or k log(2 r (1-r)) + (n-k) log(1 - 2 r (1-r))

Probabilty of such event is proportional to

$$\big(2r(1-r)\big)^k\big(1-2r(1-r)\big)^{n-k}$$

Therefore logarithm of probability, up to a constant is

$$\log\Big(\big(2r(1-r)\big)^k\big(1-2r(1-r)\big)^{n-k}\Big)$$
$$= k\log\big(2r(1-r)\big)$$
$$+ (n-k)\log\big(1-2r(1-r)\big) = k\log s + (n-k)\log(1-s)$$

(c) Show that, provided k is not bigger than n/2 the maximum likelihood estimate of r is (1 - sqrt(1 - 2k/n))/2 .

It easy to see that s=k/n maximizes $k \log s + (n-k)\log(1-s)$: at that point

$$\frac{d}{ds}\left(k \log s + (n-k)\log(1-s)\right) = \frac{k}{s} - \frac{n-k}{1-s} = n - n = 0$$

$$\frac{d^2}{dx^2}\left(k \log s + (n-k)\log(1-s)\right) = -\frac{k}{s^2} - \frac{n-k}{(1-s)^2} < 0$$

So the log likelihood is optimized if

$$s = 2r(1-r) = \frac{k}{n} \ or$$

$$r^2 - r + \frac{k}{2n} = 0 \ \rightarrow r = \frac{1}{2}\left(1 \ \pm \sqrt{1 - \frac{2k}{n}}\right)$$

Since recombination probability is always <1/2, the only possible value is

$$r = \frac{1}{2}\left(1 - \sqrt{1 - \frac{2k}{n}}\right)$$

(d) It can be shown that the Fisher information for estimating r is (2 n (1-2 r)^2)/ (r (1-r) (1 - 2 r (1-r)))
You do not need to show this.
What does this tell you about estimating r, when in fact r is close to 1/2 ?

As r approaches ½ Fisher information approaches 0, therefore it becomes more difficult to estimate r, i.e. more experimental data (larger n) is required to make estimation of the same accuracy (or to infer linkage).