1 Nonparametric methods

In the context of linkage analysis, a nonparametric method is a technique for establishing linkage that does not require specification of the mode of transmission. This is in contrast to the previous methods in which we had to specify if the disease was recessive or dominant. A common feature of nonparametric methods is the use of small, nuclear families as opposed to the large multigeneration families used in parametric linkage analysis.

1.1 Sib-pair method

In the 1930s and 1940s, the most popular method of linkage analysis was based on the sib-pair method of Penrose (1935). While subsequent developments have supplanted this method, it is provides a natural introduction to modern affected sib-pair methods. In addition, this method uncovered the first case of linkage in humans.

The method is quite simple. Suppose the phenotype for each member of a sib-pair (i.e. a pair of siblings) is known. We suppose each phenotype corresponds to a genotype at some locus. Each sib-pair can be classified as having the same or not the same phenotype for each of the 2 phenotypes. We organize this data into a 2 by 2 table with a dimension for each phenotype, and a row or column for same or different. Each sib-pair contributes one entry to this table.

	same on phen. 2	differ on phen. 2
same on phen. 1		
differ on phen. 1		

If there is linkage between the loci corresponding to the phenotypes, the concordant cells of the table should fill up, i.e. sib-pairs will either have the same allele at the 2 loci or different alleles. We can then conduct a goodness of fit test, like Pearson's χ^2 test, to test for linkage.

While this method is simple to implement and easy to understand, it has some weaknesses. First, since it ignores the parental phenotypes, many non-informative matings are considered in the data, and this acts to dilute any effect we may otherwise see. Secondly, incomplete penetrance and phenocopies can disguise effects since we don't model these factors.

1.2 Affected sib-pair (ASP) methods

A simple modification to the idea of the basic sib-pair method greatly enhances the power of the method. Instead of using any sib-pair, we use *affected* sib-pairs (some methods also use affected-unaffected sib-pairs). While we may still consider non-informative matings (since the parents may be homozygous for the marker), we don't have to worry about incomplete penetrance since both of the siblings are affected (although phenocopies still pose a problem).

To understand the basis for these methods, and other methods of nonparametric linkage analysis, we need to introduce the idea of 2 markers being *identical by descent* (IBD). Consider 2 offspring and the alleles they have at some locus. If the 2 received the allele from the same parent, then the alleles are said to be identical by descent. If they have the same value for an allele but didn't receive it from the same parent, we say the alleles are *identical by state* (IBS). As an example, suppose the parents have 4 different marker alleles at some locus. Two of their children will either have 0, 1, or 2 alleles in common at this locus. If the marker locus is unlinked to the disease locus (recall both children are affected), the probability they share 0 or 2 alleles is 1/4 and the probability they share 1 allele is 1/2. If the marker is tightly linked

to the disease locus, then you would expect children to share alleles at the marker locus more frequently than not (with the surplus depending on the mode of genetic transmission). This notion of allele sharing has widespread application in statistical genetics, for example, it is what we meant by allele sharing in the discussion of quantitative traits. We can also express the pedigree likelihood in terms of IBD indicators rather than genotypes.

A variety of tests have been designed to test for linkage by comparing IBD proportions to their expected values under the hypothesis of no linkage. Define

$$z_i = P(i \text{ alleles are IBD for an ASP}).$$

As examples illustrate, if a marker is tightly linked to a recessive disease gene, we expect $z_0 = z_1 = 0$ and $z_2 = 1$, while for a dominant disorder $z_0 = 0$, and $z_1 = z_2 = 1/2$. Since the values of these probabilities under no linkage are $z_0 = 1/4$, $z_1 = 1/2$ and $z_2 = 1/4$, ASP methods are better at detecting linkage of a marker to a recessive trait.

The basic idea of any test using ASPs is to estimate the z_i with the sample proportions, then construct a test statistic based on these proportions. One could construct a χ^2 goodness of fit test to the proportions (a test with 2 degrees of freedom), but this test will have low power compared to tests that have a specific alternative. An asymptotically equivalent test, is to use the likelihood ratio test of the hypothesis of no linkage. If one specifies a mode of transmission in the alternative hypothesis, then one can construct a more powerful test than the χ^2 goodness of fit test. As is typical for non-parametric methods, one can devise a test that doesn't need a specification of the model, but one can devise a more powerful test if one proposes a model under the alternative. For this reason, there are 3 popular methods for testing the hypothesis of no linkage in ASP analysis. One specifies the alternative of a recessive mode of inheritance, one specifies a dominant mode of inheritance, and one tries to find a balance between the other two. Not surprisingly, the test that tries to strike a balance is not as powerful as either of the other 2 tests if the mode of inheritance is known.

1.2.1 The mean test

The mean test is designed to provide a powerful test under the alternative of a dominant mode of transmission. The test statistic is

$$T_{\text{mean}} = \hat{z}_2 + \frac{1}{2}\hat{z}_1.$$

One compares this to the null value of $1/2$, and the distribution of the statistic can be worked out based on the properties of sample proportions from multinomial samples. That is, the variance of a sample proportion is well known, but here we have to consider the correlation between $\hat{z}_2$ and $\hat{z}_1$.

1.2.2 The proportion test

The proportion test is designed to provide a powerful test when the alternative is a recessive mode of transmission. The test statistic is

$$T_{\text{proportion}} = \hat{z}_2,$$

and one compares this to its null value of $1/4$. The variance of this statistic is just the variance of a sample proportion. Recall, for recessive traits we expect $z_0 = z_1$, so we use neither of these in the test.

1.2.3 Minmax test of Whittemore and Tu

In a recent article, Whittemore and Tu (1998) propose a test of no linkage using the idea of minimax tests. Both the mean and proportion tests can be viewed as a special case of the more general test statistic

$$T_w = w_0\hat{z}_0 + w_1\hat{z}_1 + w_2\hat{z}_2.$$

This test statistic will be asymptotically normal, so if we use the square of T_w we get a test statistic that is asymptotically χ^2 with 1 degree of freedom (as opposed to the 2 degrees of freedom in the likelihood ratio test). This test statistic is invariant under linear transformations of the weights, so typically we set $w_0 = 0$ and $w_2 = 1$. This implies the properties of the test depend solely on the value of w_1. The idea of Whittemore and Tu is to select w_1 so that we minimize the maximal cost of making a mistake (hence the minmax terminology). An interesting result of Whittemore's in this connection is that any allowable value for w_1 must lie between 0 and 1/2 (the values used in the proportion and mean tests respectively). They show that the minmax value of w_1 is 0.275, slightly closer to the mean test (that is dominant inheritance)

1.3 Linkage disequilibrium

It is reasonable to suppose that some genetic disorders arise due to a mutation at some gene. A random mutation can lead to an organism that doesn't produce the protein necessary for some biological function. Consider some individual who has undergone a mutation that leads to a dominant disease. If we use D to represent the mutant disease allele, and single digit numbers to represent marker alleles, then his genotype on say chromosome 4 is something like

chromosome 4, copy 1

$$1, 4, 7, 3, 8, D, 3, 7, 3, 6, 2$$

chromosome 4, copy 2

$$4, 2, 6, 2, 5, d, 2, 6, 1, 7, 8.$$

He will transmit one of these chromosomes to each of his offspring. Consider an offspring who is affected. Such an offspring must have received copy 1, but perhaps there was a crossover, so this individual's chromosome 4 from the affected parent would be something like

$$4, 2, 6, 3, 8, D, 3, 7, 3, 6, 2.$$

As generations come and go, one of the chromosomes in all of the affected individuals must have the disease gene. Those with the disease gene will also have the markers near the disease loci that the original family member with the disease had unless there is some recombination that results in another marker taking its place. That is, we would expect to find alleles 8 and 3 at the marker loci next to disease allele in affected offspring in this family, a state we refer to as *linkage disequilibrium*.

We define the *linkage disequilibrium parameter*, δ, for the marker with allele value 8 in the above example as

$$\delta = P(D, 8) - P(D)P(8).$$

This parameter is not so useful as a measure of linkage disequilibrium because its minimal and maximal value depend on the allele frequencies. If there is no linkage disequilibrium

$$P(D, 8) = P(D)P(8),$$

and so we expect δ to be near zero (otherwise we expect it to be positive). Over time, the disequilibrium would decay at a rate depending on the recombination fraction between the disease and marker loci. If δ_t is the linkage disequilibrium at time t, then

$$\delta_t = (1-\theta)^t \delta_0,$$

where $\delta_0 = P(D,8) - P(D)P(8)$ for the original mutant, i.e. $\delta_0 = 1 - P(8)$. If one knows the recombination fraction and $P(8)$, one can thus deduce the time of the original mutation.

The linkage disequilibrium parameter can be estimated by using sample proportions. This also indicates how to construct confidence intervals for δ. To test the hypothesis

$$H_0 : \delta = 0$$

one tests if the disease haplotype (i.e. markers associated with the disease phenotype) occurs in significantly different proportions in the affected versus the non-affected populations. If haplotypes are not so easily constructed, one must average over phase. These tests fall into the category of association tests since they look for an association between haplotypes and disease phenotypes.

Linkage disequilibrium can arise for reasons other than a linkage of a marker to a loci where a mutation took place. For example, if you use a set of unrelated individuals it may appear that there is linkage disequilibrium. Given this, linkage disequilibrium will be useful for mapping when the population under study is a small isolated group with extensive intermarrying. In particular, this method will not be useful for analyzing heterogeneous populations with individuals coming from a variety of sources. When conducting association studies with heterogeneous populations one must use a test statistic that tests for association and linkage at the same time.

1.4 Haplotype relative risk

One of the major problems with tests of linkage disequilibrium is the unaffecteds are unrelated, hence the haplotypes of these individuals make for poor "control" haplotypes. A better way to define a control population would be to use unaffected individuals from the same family. For now we will suppose we have trios of parents and an affected child.

Consider the problem of whether some allele, denoted A, at a marker is in linkage disequilibrium with with some disease. *For each family*, define the disease genotype to be the genotype at the marker in question for the affected child. This disease genotype consists of 2 alleles. The control genotype consists of the 2 alleles not transmitted to the affected child. For example, if the parental genotypes are $A|B$ and $C|D$ and the child has genotype $A|D$, then the disease genotype for this family is $A|D$ and the control genotype is $B|C$. The *haplotype relative risk* is defined as

$$HRR = \frac{P(A|\text{disease genotype})}{P(A|\text{control genotype})}.$$

The test for linkage of the allele A to the disease loci in this context is

$$H_0 : HRR = 1.$$

We can test this hypothesis by setting up a certain 2 by 2 table and using Pearson's χ^2 test. To understand the table, realize we have cases (i.e. the disease genotypes) and controls (i.e. the genotypes not transmitted). Each case and control either has the disease allele in his or her genotype or does not, so we can form a 2 by 2 table as with any other case control study.

	has allele A	doesn't have allele A
case genotype		
control genotype		

If there are n families, then there are n affected children, hence there are n disease genotypes and n control genotypes (so there are $2n$ entries in the table). If there is linkage disequilibrium between the allele and the disease gene, we would expect the concordant cells of the table to fill up. Moreover, under the assumption of linkage equilibrium, alleles segregate independently of one another, so we can think of this table as representing data from $2n$ independent observations. Thus we can apply Pearson's χ^2 test to the table. For the previous example, we put an observation in the cell representing the A is in the disease genotype, and an observation in the cell corresponding to a not A in the control genotype. If there is linkage disequilibrium between allele A and the disease gene, we would expect to see observations pile up in the cells in which allele A is in the disease genotype.

1.5 The transmission disequilibrium test

Another test for testing the hypothesis of no linkage via linkage disequilibrium is known as the *transmission disequilibrium test*, TDT. This test is much like the HRR, but it ignores families where one parent is homozygous for the allele suspected to be in association with the disease gene (since these obscure our view of linkage disequilibrium). We think of each parent as contributing either the allele in disequilibrium with the disease genotype (i.e. the case allele) or not (the control allele). Since the parents are matched, the appropriate test is McNemar's χ^2 test for the table in which we have rows for the non-transmitted allele and columns for the transmitted allele.

	case allele A	case allele not A
control allele A	a	b
control allele not A	c	d

We expect the numbers in the off-diagonal of the table to fill up under the null so we compare these. When we work out the sample variance of this difference, we arrive at the test statistic, namely

$$\frac{(c-b)^2}{c+b}.$$

It can be shown that

$$\mathrm{E}(c-b) = 2n(1-2\theta)\delta/p_D,$$

where p_D is the probability of observing the disease allele and n is the sample size. Hence the expected value of the difference is zero if $\delta = 0$ or $\theta = \frac{1}{2}$, i.e. if there is linkage disequilibrium or if there is linkage (or if both are true). For this reason, it is recommended that linkage disequilibrium first be established using other methods (which is easy since we just have to detect association) then one tests for linkage using the TDT. In such a context, it has been shown that the TDT is a powerful test for establishing linkage. In contrast, it can be shown that the HRR tests $\delta = 0$ or $\theta = 0$ or both. Usually the null hypothesis of no linkage disequilibrium or complete linkage is not of interest.

Exercises for week 5

1. If the phenocopy rate for a recessive trait is 0.3, then what are the IBD probabilities for a marker tightly linked to the disease gene for 2 affected sibs? (Assume the genotypes of the parents are A, B and C, D, both parents are heterozygous at the disease locus.)

2. Find estimates of the standard deviation for the mean and proportion tests. Use these estimates to construct test statistics that have standard normal distributions under the relevant null hypothesis.

References

Penrose, L., (1935), "The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage", *Ann Eugen*, 6:133-138.

Whittemore, A. and Tu, I. (1998), "Simple, robust linkage tests for affected sibs", *American Journal of Human Genetics*, 62:1228–1242.