Distribution of the Admixture test for the Detection of Linkage under Heterogeneity

Julian J. Faraway
Department of Statistics, University of Michigan,
Ann Arbor, Michigan 48109, USA.
Phone: (734)-763-5238

Running Title: Admixture Test

Abstract

The admixture test for the detection of linkage under heterogeneity is considered. We show that the null distribution of this test statistic has half its weight concentrated on zero and the other half on a complicated distribution that can be approximated by $\max(X_1, X_2)$ where X_1 and X_2 are independent χ^2_1 variables. We also give exact critical values for small samples and show that the power of this test to detect linkage is generally greater than the standard test that assumes homogeneity.

1 INTRODUCTION

We consider the detection of linkage when linkage heterogeneity exists, that is when only a fraction of sibships may be linked to a given genetic marker. Smith(63) introduced the admixture model based on the recombination fraction and the proportion of linked families. Ott(83,85) and Risch(88) consider tests for *heterogeneity* based on this model, whereas Hodge et. al.(83) and Risch(89) consider tests for *linkage* based on this same model. The latter is discussed here. Martinez & Goldin (89) discuss sample sizes needed for such tests.

The test for linkage is one-sided since recombination fractions greater than one half make no biological sense and should its estimated value be greater than one half, one would not take this as evidence of linkage. Hence, the true null distribution of the admixture statistic has half its weight concentrated at zero and the other half on some other distribution which is the subject of our interest here. Because of the symmetry of the problem, its convenient and notationally simpler to just compute the null distribution for the two-sided test statistic to discover the aforementioned distribution. Bear in mind that, although we shall be concerned with the two-sided test statistic in what follows, the true null distribution is as above.

Hodge claimed that the asymptotic (as the number of sibships becomes large) null distribution of the (two-sided) admixture test statistic was χ^2_1 but Risch conjectured it was χ^2_2. We claim here that neither is correct and that the true asymptotic distribution is quite complicated but can be adequately approximated by the $\max(X_1, X_2)$ where X_1 and X_2 are independent χ^2_1 variables. This distribution lies somewhere between the two previous claims and thus this result is of more than just technical interest given the popularity of the test. Ghosh & Sen(85) study the asymptotic distribution of the likelihood ratio test statistic for a mixture model that is similar to the one here and obtained a result similar in form to ours.

2 DISTRIBUTION OF THE TEST STATISTIC

Let the recombination fraction be θ, the proportion of linked sibships be α and the sibship size be s. Let the number of sibships be n and let X_i be the number of recombinant gametes out of s for sibship i.

Thus the likelihood for this set of sibships would be

$$L(\theta,\alpha) = \prod_{i=1}^{n} \left[\alpha\theta^{X_i}(1-\theta)^{s-X_i} + (1-\alpha)(1/2)^s\right]$$

Note that if we map $X_i \mapsto s - X_i$ (producing an outcome that has equal probability under the hypothesis of no linkage) and $\theta \mapsto 1-\theta$ then the likelihood stays the same. This symmetry allows us to consider the two-sided test statistic in our computation of the actual one-sided admixture test. If we wish to test for linkage, the natural null and alternative hypotheses are

$$H_0 : \theta = 1/2 \qquad H_A : \theta < 1/2$$

and the maximum likelihood-ratio test statistic is

$$T = 2\log(L(\hat{\theta},\hat{\alpha})/L(1/2,\tilde{\alpha}))$$

where $\hat{\theta}$ and $\hat{\alpha}$ are the maximum likelihood estimates(m.l.e) under the alternative hypothesis and $\tilde{\alpha}$ is the m.l.e. under the null. Note that when $\theta = 1/2$, α is unidentifiable i.e any value of α produces the same likelihood so the actual value of $\tilde{\alpha}$ is immaterial, although this unidentifiability is the source of the difficulty in determining the distribution of T. We use natural logs here for statistical convenience; lod scores will be discussed later. So

$$T = 2\max_{\alpha,\theta} T(\alpha,\theta) = 2\max_{\alpha,\theta} \sum_{i=1}^{n} \log\left[\alpha(2^s\theta^{X_i}(1-\theta)^{s-X_i}-1)+1\right]$$

where $0 \leq \alpha \leq 1$, $0 \leq \theta \leq 1$

Unfortunately, the asymptotic distribution under the null is not simply χ^2_1 as it would be if the usual theory were applicable. This is because a regularity condition regarding the identifiability of the parameters is not satisfied; see Wald(1949). This means the asymptotic distribution of T must be derived by other means. We give a heuristic justification of our result and verify it by simulation.

Since α is unidentifiable at the null, the likelihood will be rather flat in the α direction and since the range of α is restricted, the value of α maximizing T will tend to occur at the boundary of the range of α for large n. To see this, expand T in θ about 1/2, with α bounded away from 0,

$$T(\alpha,\theta) \approx -8\alpha\sum_{i=1}^{n}(X_i - s/2)(\theta - \frac{1}{2}) + 4\alpha[(1-\alpha)\sum_{i=1}^{n}(2X_i - s)^2 - ns](\theta - \frac{1}{2})^2$$

(where $\approx$ means approximately) Maximizing over θ gives

$$\max_{\theta} T(\alpha,\theta) \approx \frac{4\alpha[\sum_{i=1}^{n}(X_i - s/2)]^2}{ns - 4(1-\alpha)\sum_{i=1}^{n}(X_i - s/2)^2} \quad (\dagger)$$

Let $Z = \sum_{i=1}^{n}(X_i - s/2)$ and $S^2 = \sum_{i=1}^{n}(X_i - s/2)^2$ and now differentiating with respect to α

$$\frac{d}{d\alpha}\max_{\theta} T(\alpha,\theta) \approx \frac{-8Z^2(4S^2 - ns)}{(ns - 4S^2(1-\alpha))^2}$$

which will be positive or negative depending on whether S^2 is less or more than $ns/4$, independent of the value of α so for n sufficiently large T will be maximized at $\alpha = 1$ or for α small ($T = 0$ when $\alpha = 0$). Since $S^2 \to ns/4$ as $n \to \infty$, both cases will be roughly equally likely. So we consider the distribution of $\max_\theta T(\alpha,\theta)$ for $\alpha = 1$ and for α small.

When $\alpha = 1$, $\max_\theta T(1,\theta) \approx \frac{4}{ns}[\sum_{i=1}^{n}(X_i - s/2)]^2$ using $(\dagger)$. Since $EX_i = s/2$ and $Var\, X_i = s/4$, $\max_\theta T(1,\theta)$ is asymptotically χ^2_1, just applying the central limit theorem.

However, when α is small the distribution of T is not so clear:

Write k_j= {number of $X_i = j$} for $j = 0, 1, \ldots s$ then

$$T = 2\max_{\alpha,\theta}\sum_{i=0}^{s} k_i \log\left[\alpha(2^s\theta^i(1-\theta)^{s-i}-1)+1\right]$$

Now since α is small, we can expand log in terms of α ($\log(1+x) \approx x - x^2/2$):

$$T \approx 2\max_{\alpha,\theta}\sum_{i=0}^{s} k_i\left[\alpha(2^s\theta^i(1-\theta)^{s-i}-1) - \frac{1}{2}\alpha^2(2^s\theta^i(1-\theta)^{s-i}-1)^2\right]$$

and maximizing this over α gives

$$T \approx \max_{\theta} \frac{\left[\sum_{i=0}^{s} k_i (2^s \theta^i (1-\theta)^{s-i} - 1)\right]^2}{\sum_{i=0}^{s} k_i \left[2^s \theta^i (1-\theta)^{s-i} - 1\right]^2}$$

Now under the null $\theta = 1/2$, and so $Ek_i = n\binom{s}{i}2^{-s}$ so replacing k_i by it's expectation and then by applying the binomial theorem, we see that the numerator is approximately

$$\sum_{i=0}^{s} n\binom{s}{i} 2^{-s} \left[2^s \hat{\theta}^i (1-\hat{\theta})^{s-i} - 1\right]^2 = n\left[(\theta^2 + (1-\theta)^2)^s 2^s - 1\right]$$

Hence

$$T \approx \max_{\theta} \left[\sum_{i=0}^{s} k_i c_i(\theta)\right]^2$$

where

$$c_i(\theta) = \frac{(2^s \theta^i (1-\theta)^{s-i} - 1)}{\sqrt{n\left[(\theta^2 + (1-\theta)^2)^s 2^s - 1\right]}}$$

The distribution of this cannot be explicitly stated for general s, but given that k_i is asymptotically normal as $n \to \infty$, $\sum_{i=0}^{s} k_i c_i(\theta)$ is asymptotically a weighted sum of normals and is hence normal for given θ. This might suggest a χ_1^2 as a possible approximation and simulation shows that this is indeed a good fit but it should be emphasized that this is not the exact distribution.

Now when T is maximized for α small, T is a weighted sum of the k_i and when maximized for $\alpha = 1$, T is a function of the sample mean, so the maximizing values at these two points will be tend to be independent especially for large s. This suggests a distribution for T as the maximum of two independently distributed χ_1^2 variables. Again, this is not an exact result but simulation indicates that it is a good approximation. The true asymptotic distribution a function of the maximum of a particular Gaussian process but since this cannot be explicitly calculated, the suggested approximation will be of more practical utility.

To check the validity of this suggested approximating distribution, consider the following simulation results: With sibship size s set to 3 and the number of sibships set to 100, 100,000 datasets were generated with the true $\theta = 1/2$. The likelihood was maximized by first transforming α and θ to a logit scale ($x \mapsto \log(x/(1-x))$) so that the constraints on α and θ can be removed and then using the Nelder-Mead simplex method described in Press et al. (1988) to find the maximum. The maximum at $\alpha = 1$ and α small as indicated in the discussion above was also calculated.

The quantiles of the suggested distribution of T may be simply calculated by noting that

$$P(T < q) = P(\chi_1^2 < q)^2$$

In figure 1, we show a quantile-quantile plot of the simulated test-statistic against it's claimed distribution. We have converted to a lod scale and focused only on the upper tail (the largest 284 observations) of the distribution since this is the area of most interest and the fit is good for the rest of the distribution anyway.

Figure 1

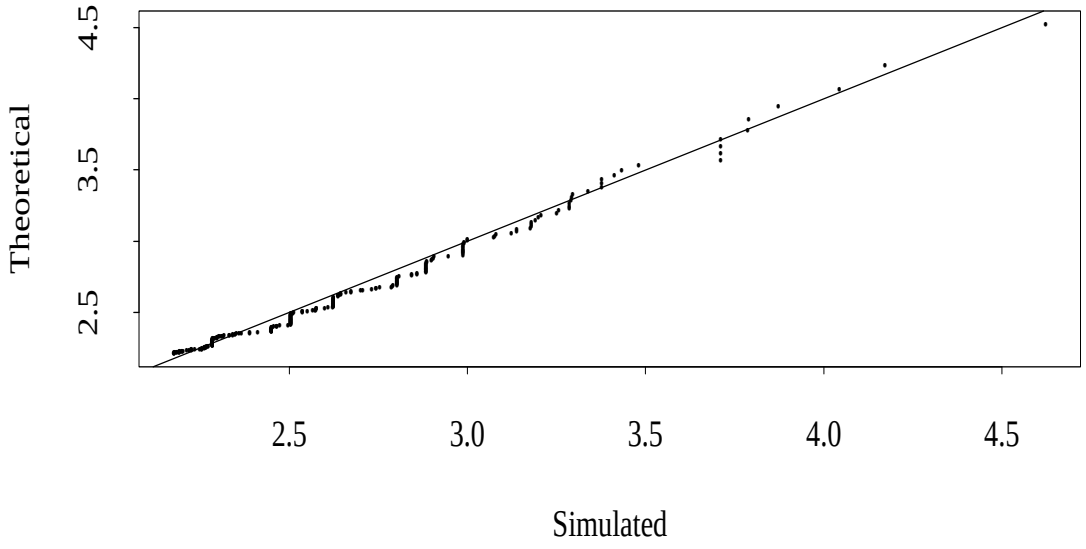

Figure 1: A Q-Q plot showing the upper tails of the distributions of the simulated and theoretical T on a lod scale

The agreement between the simulated and theoretical distributions is good. Any divergence from a perfect match can reasonably be attributed to simulation sampling error and that for finite n, T is discrete. Similar results have been observed for other small values of s and the fit improves as n gets larger.

3 CRITICAL VALUES

Recall that if $\hat{\theta} \geq 0.5$ we have no evidence for linkage, otherwise we can determine the significance of the observed T by referring to the approximate null distribution that we have calculated. If lod scores are preferred, one would use

$$T' = 2\log(10)T$$

If the same level of test is desired as for a χ^2_1 distributed statistic, lod scores of 2,3 and 5 correspond to scores for T' of 2.28,3.28 and 5.27 respectively. (Compare the values given by Risch(89) of 2.62,3.70 and 5.80 respectively).

This result is asymptotic in nature and may not be good for the small samples used in practice. It should be noted that it is computationally feasible to calculate exact critical values for small samples. To guarantee at least the same level of test corresponding to using a lod score of 3 as a criterion (a significance level of approximately 0.02144%), the null hypothesis should be rejected when T' exceeds the critical values given in the following table:

Table I									
sibship	Number of families								
size	2	3	4	5	6	7	8	9	10
2	-	-	-	-	-	4.05	3.59	3.58	3.29
3	-	-	-	3.98	3.40	3.45	3.44	3.22	3.16
4	-	-	3.59	3.50	3.30	3.25	3.20	3.45	3.29
5	-	3.98	3.55	3.32	3.16	3.06	3.35	3.40	3.18
6	-	2.79	3.30	3.21	3.43	3.37	3.25	3.31	3.23

Table I: Critical lod scores for the admixture statistic corresponding to a nominal lod score of 3

No entry (-) indicates that the maximum possible value of T' has a probability exceeding the stated significance level and so under these conditions there is insufficient data - the null hypothesis will never be rejected. The critical values fluctuate but approach the expected 3.28 as n gets larger. Note that there is one value less than 3, which may seem odd, but remember these are exact values, and this happens to be the critical value corresponding to the stated level of significance.

4 POWER

Risch(89) compared the power of the heterogeneity test against the standard test where homogeneity is assumed ($\alpha = 1$), and concluded that the homogenous test was more powerful in most circumstances. Contrary to this, we demonstrate here, by using the correct critical value and computing the power exactly, that the heterogenous test is generally preferable.

Exact critical values for a significance level corresponding to a lod of 3 for both tests were computed and the exact power to detect linkage was calculated for a range of values of α from 0 to 1 and of θ from 0 to 1/2. Figure 2 shows the power of the heterogenous test minus that of the homogenous test. The lines show contours of equal difference in power (probability expressed as a percentage) and "=" denotes the region where there is a less than a 0.01 difference in the power.

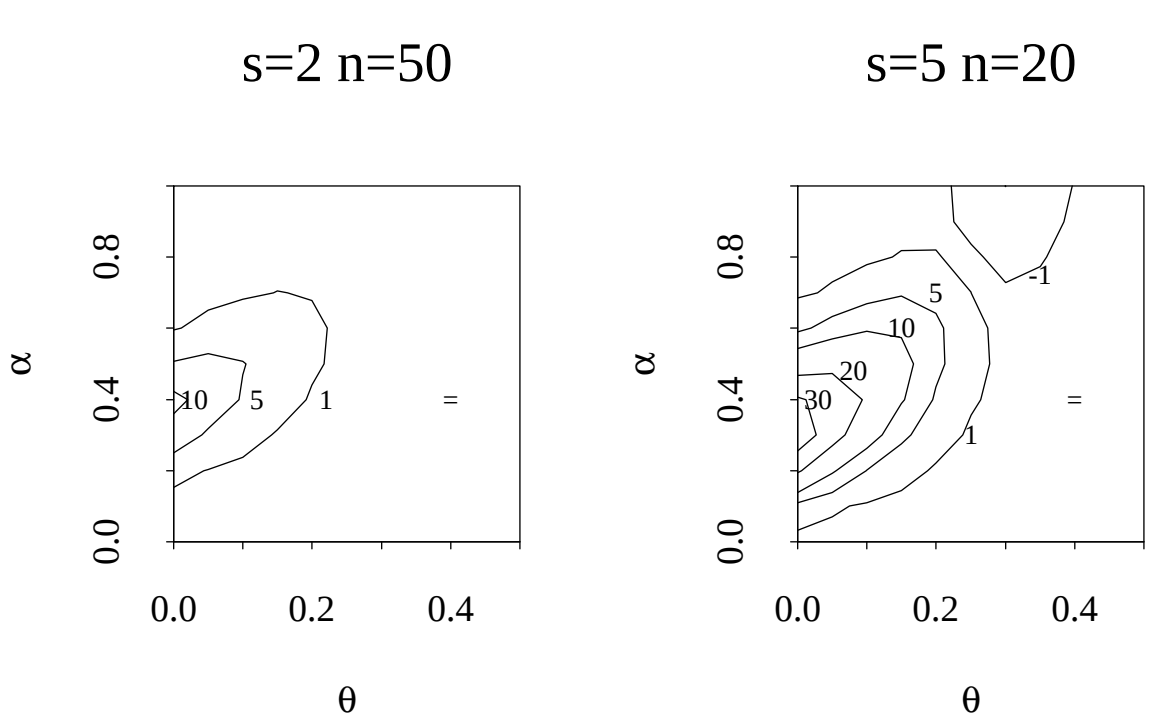

Figure 2: Contour plots showing the difference in power between the heterogenous test and the homogeneous test

In the case of sibship size 2 and 50 sibships, the homogenous test exceeds the power of heterogenous test by no more than 0.01 and can be 0.1 less powerful when α=0.4 and θ=0. When the sibship size is 5 and with 20 sibships, the heterogenous test exceeds the power of homogenous test by 0.37 when α=0.3 and θ=0. The region where the homogenous test is mildly preferable is confined to an area of low mixing and moderate linkage. Other comparisons show that the region where the heterogenous test is clearly preferable expands with sibship size and number of sibships. Even when there is no mixing the homogenous test is only mildly more (0.05-0.1 at best) powerful than the heterogenous test, but if there is some mixing the heterogenous test can be substantially more powerful.

5 DISCUSSION

We have approximated the null distribution of the admixture test for the detection of linkage and demonstrated that if the possibility of heterogeneity exists, this admixture test is generally more powerful than the usual test which takes no account of heterogeneity.

We have considered constant sibship size here for simplicity of the exposition but this is not crucial and the same asymptotic result would follow even if the sibship size were allowed to vary. Furthermore, the same result holds even when the meioses are not completely informative. For the least informative, phase unknown, case, the test statistic is

$$T = \max_{\alpha,\theta} \sum_{i=1}^{n} \log\left[\alpha(2^{s-1}\{\theta^{X_i}(1-\theta)^{s-X_i} + \theta^{s-X_i}(1-\theta)^{X_i}\} - 1) + 1\right]$$

and a similar reasoning to the one above may be used to get the same result.

ACKNOWLEDGEMENT

Thanks to Michael Boehnke of the Department of Biostatistics, University of Michigan for bringing my attention to this problem and offering helpful comments and thanks also to two referees for improving the initial draft.

REFERENCES

Ghosh J.K. & Sen P.K. (1986) "On the Asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results" *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer Volume II* Wadsworth

Hodge SE, Anderson CE, Neiswanger K, Sparkes RS, Rimoin DL (1983) "The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): Linkage studies, two-locus models, and genetic heterogeneity" *Am J Hum Genet* **35** 1139-55

Martinez M. & Goldin L. (1989) "The detection of linkage and heterogeneity in nuclear families for complex disorders: One versus two marker loci" *Am J Hum Genet* **44** 552-559

Ott J (1983) "Linkage analysis and family classification under heterogeneity" *Ann Hum Genet* **47** 80-96

Ott J (1985) "Analysis of Human Genetic Linkage" *Baltimore, The John Hopkins University Press*

Press W.H., Flannery B.P., Teukolsky S.A. & Vetterling W.T. (1988) Numerical Recipes. *Cambridge University Press*

Risch N (1988) "A new statistical test for linkage heterogeneity" *Am J Hum Genet* **42** 353-364

Risch N (1989) "Linkage detection tests under heterogeneity" *Genet Epidem.* **6** 473-480

Smith CAB (1963) "Testing for heterogeneity of recombination values in human genetics" *Ann Hum Genet* **27** 175-182

Wald A. (1949) Note on the consistency of the maximum likelihood estimate" *Ann Math Statist* **20** 595-601