Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Fu et al., J Biom Biostat 2014, S12
DOI: 10.4172/2155-6180.S12-002

Research Article Open Access

# Sample Size Calculation for the Modified Likelihood Ratio Test in Genetic Linkage Analysis

**Yuejiao Fu*, Pengfei Li and Soowoon Chung**

*York University and University of Waterloo, Canada*

**Abstract**

Mixture models provide flexible means of handling heterogeneity in data. The possible latent structure suggested by mixture model analysis should be carefully examined using designed experiments. Sample size determination is an important and difficult step in design of experiments. We investigate the sample size calculation for the modified likelihood ratio test for binomial mixture models arising in genetic linkage analysis. We obtain limiting distributions for the modified likelihood ratio test under two sets of commonly used local alternatives. A simple sample-size formula is obtained and illustrated using both simulations and a genetic linkage study for schizophrenia.

**Keywords:** Contiguity theory; Genetic linkage analysis; Hypothesis testing; Local asymptotic power; Mixture models; Modified likelihood; Recombinant

## Introduction

Mixture models provide flexible means of handling observed or unobserved heterogeneity in data. The data analysis using mixture models could unveil the possible underlying or latent structure. Well-designed clinical trials and scientific experiments are usually needed to examine the validity of the suggested latent structure. Sample size determination is a major issue in those studies, see Chow et al. [1] and references therein. There is a vast literature covering sample size calculation for comparative research studies especially in medical context, for example, hypothesis testing for proportions in two groups.

Instead of considering simple designs such as a two-sample test, we consider calculating sample size for hypothesis tests in mixture model framework. More specifically we propose a formula for determining required sample size for performing a test of homogeneity. A test of homogeneity, which tests the null hypothesis of one component parametric model versus the alternative of a two-component mixture, is one of the most difficult and important problems in finite mixture models. There is some literature on power and sample size calculations for tests of homogeneity in finite mixture models. Hall and Stewart [2] provided theoretical analysis of power in a two-component normal mixture model and addressed the irregular feature of the problem. Recently, Chen et al. [3] addressed the issue of sample size calculation for tests of homogeneity using the EM-test and $C(\alpha)$ test. Instead of a general homogeneity test, we consider a special binomial mixture model arising in genetic linkage analysis. This particular binomial mixture model in pedigree studies has been studied in Lemdani and Pons [4], Liang and Rathouz [5], Fu et al [6]. showed that the modified likelihood ratio test (MLRT) which was proposed by Chen [7] and Chen et al. [8] has better power for detecting the aforementioned binomial mixture alternative than other methods discussed in their paper. Since sample size calculation is test-specific, for the homogeneity test of the special binomial mixture, we choose the MLRT as the basis for the sample size determination. Following Chen et al. [3], we investigate the power properties of the MLRT under two sets of commonly used local alternatives. A simple sample size formula is obtained and illustrated by both simulations and a genetic linkage study for schizophrenia.

The rest of the article is as follows. Section 2 presents the problem set up and gives the asymptotic distribution of the MLRT and sample-size formula for two local alternatives. Section 3 presents a real data example in genetic linkage study and simulation results are given in Section 4. The proof of the theorem is given in the Appendix.

## Main Results

The particular binomial mixture model in pedigree studies we consider is a two-component binomial mixture with one component distribution completely known. This model is commonly used to model the recombinant data in pedigree studies and known as phase known model. See Liang and Rathouz [5] and Fu et al. [6] for more details. Suppose we have a random sample $X_1,\ldots,X_n$ drawing from the following binomial mixture model

$$(1-\gamma)\mathrm{Bin}(m,0.5) + \gamma\mathrm{Bin}(m,\theta),$$

where $\gamma$ is the mixing proportion and $\theta \in [0,0.5]$ is the component parameter with a specified range. Our interest is to test homogeneity with the null hypothesis specified as

$$H_0 : \gamma\,(\theta - 0.5) = 0.$$

Note that there are two unusual features of the homogeneity test: (1) the null hypothesis lies on the boundary of the parameter space, and (2) the parameters $\gamma$ and $\theta$ are not identifiable under the null model. The log-likelihood function of $(\gamma,\theta)$ is

$$l_n(\gamma,\theta) = \sum_{i=1}^{n} \log\left\{(1-\gamma)\binom{m}{X_i}0.5^m + \gamma\binom{m}{X_i}\theta^{X_i}(1-\theta)^{m-X_i}\right\}.$$

The modified log-likelihood function is defined as

$$pl_n(\gamma,\theta) = l_n(\gamma,\theta) + C\log(\gamma)$$

with $C>0$. In this paper, we choose $C=1$ as suggested in Fu et al. [6]. The MLRT statistic is defined as

***Corresponding author:** Yuejiao Fu, York University and University of Waterloo, Canada, Tel: 416-736-2100 ext. 33772; Fax: 416-736-5757; E-mail: yuejiao@mathstat.yorku.ca

**Received** July 01, 2014; **Accepted** August 07, 2014; **Published** August 11, 2014

**Citation:** Fu Y, Li P, Chung S (2014) Sample Size Calculation for the Modified Likelihood Ratio Test in Genetic Linkage Analysis. J Biom Biostat S12: 002. doi:10.4172/2155-6180.S12-002

$$M_n = 2\left\{\max_{\gamma\in[0,1],\,\theta\in[0,0.5]} pl_n(\gamma,\theta) - pl_n(1,0.5)\right\}.$$

The limiting distribution of $M_n$ is $0.5\chi_0^2 + 0.5\chi_1^2$, where $\chi_0^2$ denotes a degenerate distribution with all its mass at zero. Given a significance level α<0.5, the MLRT rejects $H_0$ when $M_n > z_\alpha^2$, where $z_\alpha$ is the αth upper quantile of standard normal distribution.

The key step of sample size determination is to find the distributions of the test statistics under alternative hypotheses. However, such distributions are usually not available. In the context of homogeneity test, along the same line of Chen et al. [3], we consider power and sample size calculations under local alternative models. Among many possible deviations from the null model, we choose the following local alternatives which are contiguous to the null distribution see Le Cam and Yang [9]:

$$H_{A1}^n : \gamma = \gamma_0, \theta = 0.5 - n^{-1/2}\tau;$$

$$H_{A2}^n : \gamma = n^{-1/2}\eta; \theta = \theta_0 \quad (1)$$

where $\gamma_0$ and $\theta_0$ are constants within the parameter space. For testing homogeneity in finite mixture models, we usually encounter two types of loss of identifiability, which lead to the two specified local alternatives. $H_{A1}^n$ refers to the situation where two-component distributions are close to each other, and $H_{A2}^n$ refers to the situation where one mixing proportion is close to 0. In pedigree studies, $H_{A1}^n$ suggests even for the families with linkage the linkage is weak; while $H_{A2}^n$ suggests that there hardly exist any families with disease locus linked with the marker under consideration.

From Le Cam's contiguity theory, the limiting distribution of the MLRT statistic $M_n$ under two specified local alternatives $H_{A1}^n$ or $H_{A2}^n$ can be determined. The results are given in the following theorem and the proof is in the Appendix.

Theorem 1. Let $\delta=n^{1/2}\gamma(0.5-\theta)$. Under $H_{A1}^n$ or $H_{A2}^n$ we have as $n\to\infty$,

$$M_n \to \left\{(Z + 2\delta\sqrt{m})^+\right\}^2$$

in distribution, where $Z$ denotes a standard normal random variable.

Note that under the two specified local alternatives, $\delta$ does not depend on $n$. It is equal to $\gamma_0\tau$ under $H_{A1}^n$ and $\eta(0.5-\theta_0)$ under $H_{A2}^n$. We use the above asymptotic distribution of $M_n$ under the two specified local alternatives or $H_{A2}^n$ as the basis for power and sample size calculations. For a given alternative model $(1-\gamma)\text{Bin}(m,0.5) + \gamma\text{Bin}(m,\theta)$, the local power of the MLRT can be approximated by

$$\Phi(2\delta\sqrt{m} - z_\alpha) = \Phi\left(2n^{1/2}\gamma(0.5-\theta)\sqrt{m} - z_\alpha\right). \quad (2)$$

Note that the three basic components of sample size calculation are significance level α, target power $1-\beta$ and a potential alternative model. For the two sequences of local alternative model $H_{A1}^n$ or $H_{A2}^n$, if the target power is $1-\beta$ at a significance level α, the required sample size approximately satisfies the following equation:

$$2n^{1/2}\gamma(0.5-\theta)\sqrt{m} - z_\alpha = z_\beta.$$

In other words, the minimum sample size requirement is

$$n_{\alpha,\beta} = \left\{\frac{z_\alpha + z_\beta}{2\gamma(0.5-\theta)\sqrt{m}}\right\}^2.$$

The validity of the sample size formula is examined using a real data example and computer simulations which are given in the next two sections.

## Application

We applied the developed theory to a genetic linkage study for schizophrenia conducted at the Johns Hopkins School of Medicine. The details of the study design and data collection can be found in Pulver et al. [10] and Liang and Rathouz [5]. This study included 486 individuals from 54 families with at least two affected relatives. Here "affected" refers to someone who was diagnosed with either schizophrenia or schizoaffective disorder based on the DSM-III-R criteria. Based on previous studies, one is particularly interested in Marker D22S941 on chromosome 22. However, it is well known that schizophrenia is prone to heterogeneity. Research showed that the following two-component binomial mixture

$$0.6Bin(9,0.5) + 0.4Bin(9,0.06)$$

may fit the data well. Suppose our interest is to validate above mixture structure at the 0.5% level, which is typical in linkage studies, with at least 80% power. The approximate sample size is $n_{0.005,0.2} \approx 10$. We also used computer simulations to check: (1) whether the limiting distribution provides reasonable approximation to the finite sample distribution under the calculated sample size; (2) whether the MLRT statistic has the desired power to detect the heterogeneity. In the simulations, we set $C$=1 as recommended by Fu et al. [6]. The simulated type I error is 0.4%, and the power of $M_n$ is 87% based on 50,000 repetitions.

Similarly, we consider the situation where the significance level is 1%, and target power is 80%. The approximate sample size is $n_{0.01,0.2} \approx 9$. The simulated type I error and power of $M_n$ are around 1.4% and 86%, respectively.

## Simulation Study

We further examined the performance of the sample size calculation formula under other settings. We considered eight alternative models which are determined by the various combinations of $\gamma$=(0.05,0.3), $\theta$=(0.05,0.3), and $m$=(4,8).

| γ | θ | *m* | $n_{0.005,0.2}$ | Type I error | Power |
|---|---|---|---|---|---|
| 0.05 | 0.05 | 4 | 1442 | 0.46% | 96.8% |
| 0.05 | 0.05 | 8 | 721 | 0.53% | 100.0% |
| 0.05 | 0.3 | 4 | 7299 | 0.55% | 80.8% |
| 0.05 | 0.3 | 8 | 3650 | 0.47% | 85.9% |
| 0.3 | 0.05 | 4 | 40 | 0.55% | 89.5% |
| 0.3 | 0.05 | 8 | 20 | 0.49% | 92.3% |
| 0.3 | 0.3 | 4 | 203 | 0.55% | 81.3% |
| 0.3 | 0.3 | 8 | 101 | 0.49% | 82.4% |

**Table 1:** Simulated type I error and power of $M_n$ with $C$=1 under the null model and under the given alternative model $(1-\gamma)\text{Bin}(m,0.5) + \gamma\text{Bin}(m,\theta)$. The significance level is 0.5%.

| γ | θ | *m* | $n_{0.01,0.2}$ | Type I error | Power |
|---|---|---|---|---|---|
| 0.05 | 0.05 | 4 | 1239 | 1.10% | 95.8% |
| 0.05 | 0.05 | 8 | 620 | 1.05% | 100.0% |
| 0.05 | 0.3 | 4 | 6273 | 1.07% | 80.8% |
| 0.05 | 0.3 | 8 | 3136 | 1.01% | 85.1% |
| 0.3 | 0.05 | 4 | 34 | 1.11% | 87.5% |
| 0.3 | 0.05 | 8 | 17 | 1.24% | 92.4% |
| 0.3 | 0.3 | 4 | 174 | 1.13% | 80.9% |
| 0.3 | 0.3 | 8 | 87 | 1.16% | 82.8% |

**Table 2:** Simulated type I error and power of $M_n$ with $C$=1 under the null model and under the given alternative model $(1-\gamma)\text{Bin}(m,0.5) + \gamma\text{Bin}(m,\theta)$. The significance level is 1%.

We considered two significance levels 0.5% and 1%, with the same desired power 80%. For each alternative model, we calculated the required sample size, the simulated type I error rate, and power of $M_n$ with $C$=1 based on 50,000 repetitions. The results were summarized in Tables 1 and 2. From the tables we can see that the proposed sample size formula reliably achieves the desired power under different alternative models.

## Acknowledgment

The research was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank the editor, and two referees for their valuable suggestions and comments.

### References

1. Chow SC, Shao J, Wang H (2003) Sample size calculation in clinical research. New York: Marcel Dekker.
2. Hall P, Stewart M (2005) Theoretical analysis of power in a two-component normal mixture model. Journal of Statistical Planning and Inference 134: 158-179.
3. Chen J, Li P, Liu Y (2014) Sample-size calculation for tests of homogeneity. Submitted Manuscript.
4. Lemdani M, Pons O (1995) Tests for genetic linkage and homogeneity. Biometrics 51: 1033-1041.
5. Liang KY, Rathouz PJ (1999) Hypothesis testing under mixture models: application to genetic linkage analysis. Biometrics 55: 65-74.
6. Fu Y, Chen J, Kalbeisch JD (2006) Testing for homogeneity in genetic linkage analysis. Statistica Sinica 16: 805-823.
7. Chen J (1998) Penalized likelihood ratio test for finite mixture models with multinomial observations. Canadian Journal of Statistics 26: 583-599.
8. Chen H, Chen J, Kalbeisch JD (2001) A modified likelihood ratio test for homogeneity infinite mixture models. J R Statist Soc B 63: 19-29.
9. Le Cam L, Yang GL (1990) Asymptotics in Statistics; Some Basic Concepts. Springer-Verlag, New York.
10. Pulver AE, Karayiorgou M, Wolyniec PS, Lasseter VK, Kasch L, et al. (1994) Sequential strategy to identify a susceptibility gene for schizophrenia: report of potential linkage on chromosome 22q12-q13.1: Part 1. Am J Med Genet 54: 36-43.
11. van der Vaart AW (2000) Asymptotic statistics. Cambridge University Press.
12. Hajek J, Sidak Z (1967) Theory of Rank Tests. Academic Press, New York.