



Diagnostics for the Homogeneity of Capture  
Probabilities in Capture-Recapture Experiments

By CHANGXUAN MAO and BRUCE G. LINDSAY

Technical Report #01-07-05

2001

---

**Center for Likelihood Studies**  
DEPARTMENT OF STATISTICS  
THE PENNSYLVANIA STATE UNIVERSITY  
UNIVERSITY PARK, PA 16802

# Diagnosics for the homogeneity of capture probabilities in capture-recapture experiments

Changxuan Mao\*

Division of Biostatistics, School of Public Health, University of California at Berkeley, Earl Warren Hall 7360, Berkeley, CA 94720-7360, USA.

and

Bruce G. Lindsay

Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802, USA

July 5, 2001

**SUMMARY.** In capture-recapture studies, the capture probabilities often vary with individuals. Three graphic diagnostic procedures are developed to detect the existence of heterogeneity of capture probabilities. They are log ratio plot, residual plot and gradient plot. Confidence bands are available for all three plots. The sign sequence of the empirical residual function is also an indicator of heterogeneity. The dispersion score test is developed to test homogeneity of capture probabilities. Four examples from epidemiology and ecology are studied for illustration.

**KEY WORDS:** Population size; Number of species; Binomial mixture; Heterogeneity.

## 1. Introduction

Suppose there are  $N$  individuals in a population, where  $N$  is unknown. One important issue in the biological and ecological studies is the estimation of the population size  $N$  through capture-recapture experiments. Such an estimation problem can trace back to Laplace (1783). Chao (2001) provides the most recent references for a closed population. Chao (2001) also discusses applications in epidemiology and health sciences, software reliability, population census adjustment and sociological research.

---

\* *email:* cmao@stat.psu.edu

An important issue in building models for these experiments is the determination of whether heterogeneity of capture probabilities is present or absent. We evaluate several possibilities in this paper. We first introduce our notation for capture-recapture studies. Suppose there are  $M$  independent capture occasions. Each individual is captured or not independently at each occasion. Imagine that the individuals are indexed by  $1, 2, \dots, N$ . Let  $y_{ij}$  be the indicator of capture, that is,  $y_{ij}$  equals 1 if the  $i$ th individual is captured at the  $j$ th occasion, else  $y_{ij}$  equals 0, and let  $\pi_{ij}$  be the probability that the  $i$ th individual is captured at the  $j$ th occasion,  $i = 1, 2, \dots, N, j = 1, 2, \dots, M$ . The  $y_{ij}$ 's are independent and constitute a *multiple Bernoulli sample*. In this paper, each individual is assumed to have the same probability to be captured at all occasions, that is,

$$\pi_{ij} = \pi_i, \quad j = 1, 2, \dots, M, \quad i = 1, 2, \dots, N.$$

Let  $x_i$  be the number of occasions when the  $i$ th individual is captured. Then

$$x_i = \sum_{j=1}^M y_{ij}, \quad i = 1, 2, \dots, N.$$

So  $x_i$  is a binomial random variable with index  $M$  and capture probability  $\pi_i$ . The  $x_i$ 's are independent and constitute a *binomial sample*. However, if  $x_i = 0$ , so that the  $i$ th individual is never captured, the presence of the individual is unknown. Hence the observed data come from a zero-truncated binomial distribution.

In the *homogeneity model*, the  $\pi_i$ 's are assumed to be identical. In the alternative *heterogeneity model*, the  $\pi_i$ 's may be different and are often assumed to be a random sample arising from a *latent distribution*, which leads to a *mixture model*. When the latent distribution is degenerate, the homogeneity model is obtained, naturally called a *one-component model* in the mixture modeling literature. Testing homogeneity of individual capture probabilities can be put in a hypothesis testing problem as follows:

$$H_0: \text{a one-component model} \quad \text{versus} \quad H_a: \text{any mixture model.}$$

Some procedures have been developed to test homogeneity, for example, see Lloyd (1992), Cormack (1966) and Chapman (1952). In the computer package, CAPTURE, developed by Overton and Burnham, a goodness-of-fit  $\chi^2$  test is available (Otis et al. 1978). Lloyd (1992) proposed two kinds of procedures to test homogeneity and concluded that the procedures in Lloyd (1992) were more efficient and had higher power in a comparison with other

procedures, such as the goodness-of-fit  $\chi^2$  test in the program CAPTURE, Cormack's test for  $M = 3$ , and Chapman's test. However, Lloyd (1992) did not claim that the procedures had optimality properties.

In this paper, we will introduce a new powerful test which has some local optimality properties. Moreover, three new graphical diagnostics are presented in this paper, which are sensitive to mixture structure. In fact, the mixture alternative hypothesis deviates from the null hypothesis in very distinctive ways, a fact we will exploit.

However, in the original binomial sample, the unknown population size  $N$  plays the role of sample size. The diagnostics and the test have been developed previously only for the case that  $N$  is known. We will show that they can be extended to a zero-truncated binomial sample obtained from the original binomial sample. The zero-truncated binomial sample is defined and the justification for using the zero-truncated binomial sample is presented in Section 2.

In Section 3, some results about the zero-truncated binomial mixture distribution are introduced, which motivate the development and provide mathematical background for the graphical diagnostics. In Section 4, three graphical diagnostics are developed for the zero-truncated binomial sample. In Section 5, the dispersion score test is developed for the zero-truncated binomial sample. We will show that to perform these graphical diagnostics and the dispersion score test, it is not necessary to model the latent distribution or to estimate the population size  $N$ . In Section 6, four real examples are studied. In Section 7, the properties of the graphical diagnostics and the dispersion score test are discussed. A general confidence bands result about the log ratio plot for distributions from a one-parameter exponential family is given in **Appendix B**.

## 2. The zero-truncated binomial sample

We first show that testing homogeneity based on the original binomial sample can be adapted to testing homogeneity based on a zero-truncated binomial sample.

The odds ratio is used to parameterize the binomial density and the corresponding zero-truncated binomial density in this section. Let  $g(x; \lambda)$  be the density of a binomial distribution with index  $M$  and odds ratio  $\lambda$  with respect to a  $\sigma$ -finite measure on  $\mathcal{R}$ , called the *supporting measure* and denoted by  $d\gamma(x)$ . The density  $g(x; \lambda)$  is given by

$$g(x; \lambda) = (1 + \lambda)^{-M} \lambda^x,$$

where

$$d\gamma(x) = \frac{M!}{(M-x)!x!}, \quad x = 0, 1, \dots, M.$$

Suppose  $x$  is a binomial random variable given  $\lambda$  and  $\lambda$  itself is from a latent distribution  $P(\lambda)$ . The marginal density of  $x$  is a mixture density, denoted by  $g(x; P)$ , which can be written as

$$g(x; P) = \int g(x; \lambda) dP(\lambda), \quad x = 0, 1, \dots, M.$$

The zero-truncated mixed binomial therefore has the density given by

$$g(x; P)/[1 - g(0; P)], \quad x = 1, 2, \dots, M.$$

Our next goal is to show that the above “truncated mixed binomial density” can be re-written as a mixture of zero-truncated binomial distributions, albeit with a different latent distribution  $Q$ .

Define a re-weighted measure from  $P$  as follows:

$$dQ(\lambda) = \frac{[1 - (1 + \lambda)^M] dP(\lambda)}{\int [1 - (1 + \lambda)^M] dP(\lambda)}.$$

We can invert the above relation to obtain

$$dP(\lambda) = \frac{[1 - (1 + \lambda)^M]^{-1} dQ(\lambda)}{\int [1 - (1 + \lambda)^M]^{-1} dQ(\lambda)}.$$

So the latent distribution  $P$  is also a re-weighted version of  $Q$ . Thus the re-weighting is a bijection. Note that  $P$  is degenerate if and only if  $Q$  is degenerate.

Now let  $f(x; \lambda)$  be a zero-truncated binomial density of  $g(x; \lambda)$ . Then

$$f(x; \lambda) = [(1 + \lambda)^M - 1]^{-1} \lambda^x I(x > 0),$$

with respect to  $d\gamma(x)$ . Let  $f(x; Q)$  be the mixture density of the zero-truncated binomial distribution, that is,

$$f(x; Q) = \int f(x; \lambda) dQ(\lambda), \quad x = 1, 2, \dots, M.$$

It can be shown that

$$g(x; P)/[1 - g(0; P)] = f(x; Q), \quad x = 1, 2, \dots, M,$$

which was our goal.

We next define a zero-truncated binomial sample. Let  $n$  be the number of distinct individuals that are captured in the sample. Suppose the individuals are re-indexed such that the first  $n$  individuals are present in the sample. Since we do not know the number of individuals who have never appeared in the sample, we define a new dataset from those  $x_i$ 's larger than zero as follows:

$$z_i = x_i | x_i > 0, \quad i = 1, 2, \dots, n.$$

Then each  $z_i$  follows a zero-truncated binomial distribution with index  $M$  and capture probability  $\pi_i$ . The  $z_i$ 's constitute a *zero-truncated binomial sample*. The diagnostics and the dispersion score test will be based on the  $z_i$ 's because the  $x_i$ 's arise from  $g(x; P)$  if and only if the  $z_i$ 's arise from  $f(z; Q)$ .

### 3. The zero-truncated binomial distribution

Let  $z$  be a zero-truncated binomial random variable with an known parameter  $M$  and unknown parameter  $\pi$ , where  $M$  is the index and  $\pi$  is the capture probability of the original binomial distribution. Let  $\Omega$  be the *sample space* of  $z$ . Let  $\mu$  be the *mean value parameter* of  $z$  and  $\sigma^2$  be the variance of  $z$ . It is clear that  $\mu$  is in  $[1, M]$  and the mean value parameter space is bounded. Let  $f(z; \mu)$  be the density of the zero-truncated binomial distribution with respect to  $d\gamma(z)$ . Suppose  $\mu$  is a random variable from a latent distribution  $Q(\mu)$ , which results in a mixture density, denoted by  $f(z; Q)$ .

We now introduce some key diagnostic functions. Let

$$H(z) = \log f(z; Q) / d\gamma(z),$$

$$r(z) = f(z; Q) / f(z; \mu_0) - 1,$$

and

$$D(\mu) = \int r(z) f(z; \mu) d\gamma(z),$$

where  $\mu_0$  is the mean of the latent distribution and is given by

$$\mu_0 = \int \mu dQ(\mu).$$

The functions  $H(z)$ ,  $r(z)$  and  $D(\mu)$  are called the *log ratio function*, the *residual function* and the *gradient function* respectively. The gradient function  $D(\mu)$  is a smoothed version of  $r(z)$  with the kernel function  $f(z; \mu)$ .

The zero-truncated binomial distributions form a one-parameter exponential family. Our graphical diagnostics are based on the following results:

The log ratio function  $H(z)$  and the residual function  $r(z)$  are convex functions in  $z$ . The gradient function  $D(\mu)$  is a convex function in  $\mu$ . These functions are linear if and only if  $Q$  is degenerate. If  $Q$  is not degenerate, then the residual function  $r(z)$  has the sign sequence  $+, -, +$  when  $z$  traverses the real line.

The convexity of  $r(z)$  and  $H(z)$  are relatively easy to obtain, for detailed proof, see Lindsay (1995). The sign sequence was first discussed in Shaked (1980). Lindsay (1995) gave a simple general result about the relationship between the moments of two measures and the sign sequence of the difference of the two measures. The convexity of  $D(\mu)$  was proved in Lindsay and Roeder (1992), which is also easy but requires results about totally positive kernels from Karlin (1968).

#### 4. Graphical diagnostics

Graphical diagnostic procedures are often used in mixture modeling, for example, see Titterton, Smith and Makov (1985), Lindsay and Roeder (1992) and Lindsay (1995). They are used to show the presence of mixture structure, provide raw estimates for underlying parameters, such as the number of components in the mixture, and tell uniqueness of non-parametric maximum likelihood estimator for the latent distribution, etc. Here we consider using the results in Section 3 to construct graphical diagnostics to test homogeneity.

Let  $n_z$  be the number of individuals that are captured exactly  $z$  times, that is,

$$n_z = \sum_{i=1}^n I(z_i = z), \quad \forall z \in \Omega.$$

The  $n_z$ 's are the most important summary statistics in capture-recapture studies, called *frequency counts*. Let  $\bar{z}$  be the sample mean, that is

$$\bar{z} = \sum_{i=1}^n z_i/n = \sum_{z=1}^M zn_z/n.$$

The maximum likelihood estimator  $\hat{\mu}$  for  $\mu$  equals  $\bar{z}$  under the one-component model. In this section, we use the notation  $f(z; \mu)$  to represent the density of the zero-truncated binomial distribution with respect to counting measure on  $\mathcal{N}$ .

Define the *empirical log ratio function*, the *empirical residual function* and the *empirical gradient function* as follows:

$$\hat{H}(z) = \log n_z/[nd\gamma(z)],$$

$$\hat{r}(z) = n_z / [nf(z; \hat{\mu})] - 1,$$

and

$$\hat{D}(\mu) = \sum_{z=1}^M \hat{r}(z) f(z; \mu).$$

The functions  $\hat{H}(z)$ ,  $\hat{r}(z)$  and  $\hat{D}(\mu)$  are empirical counterparts of  $H(z)$ ,  $r(z)$  and  $D(\mu)$  respectively. When  $n$  is sufficient large, they can be good estimators for their theoretic counterparts. Suppose  $z$  follows  $f(z; Q)$ , then  $\hat{H}(z)$ ,  $\hat{r}(z)$  and  $\hat{D}(\mu)$  will converge pointwise to  $H(z)$ ,  $r(z)$  and  $D(\mu)$  respectively as  $n$  goes to infinity, see Lindsay and Roeder (1992).

We will call plots  $(z, \hat{H}(z))$ ,  $(z, \hat{r}(z))$  and  $(\mu, \hat{D}(\mu))$  the the *log ratio plot*, the *residual plot* and *gradient plot* respectively. The empirical gradient function  $\hat{D}(\mu)$  is a smoothed version of  $\hat{r}(z)$  with the kernel function  $f(z; \mu)$ .

The above discussions lead to the following conclusions:

*If the heterogeneity model is true, all these plots will show convexity and the empirical residual function will have a sign sequence +, -, +. The plots are approximately linear if and only if the homogeneity model is true. If neither the linearity nor convexity holds, the binomial model should be suspected.*

We next discuss the zero frequency counts problem. For levels of  $z$  with small probabilities of being observed, we expect to see  $n_z = 0$  frequently. There are also the levels of  $z$  for which the plots have large standard errors. The empirical residual function takes the lower limit  $-1$  with  $n_z = 0$ . The visual effect will be distorted at these points. The empirical log ratio function takes value  $-\infty$  with  $n_z = 0$ , which can not be plotted. Our strategy here will be to truncate the range of the plots. If  $n_{(z^*+1)} = 0$  but  $n_z \neq 0$  for  $z \leq z^*$ , then we will only plot the points with  $z$  from one to  $z^*$ . In order to show sufficient evidence for or against convexity, we may require that  $z^*$  is no less than a fixed natural number, say 4. For the gradient plot, if there exists  $z^*$  such that  $n_{z^*} \neq 0$  but  $n_z = 0$  for  $z > z^*$ , then the visual effect can be distorted after  $\mu = z^*$ .

An alternative to the simple range truncation used here would be to group data in the tails, as in a goodness-of-fit  $\chi^2$  test, and use the corresponding grouped density function.

In order to have interpretable plots, we will construct confidence bands for these plots when  $n$  is small. The confidence bands for the log ratio plot, the residual plot and the gradient plot are based on the following three results:

*The empirical log ratio function  $\hat{H}(z)$  is approximately normally distributed for each  $z$ . Let  $e_H(z)$  be the approximate pointwise standard error of  $\hat{H}(z)$*

either under the one-component model or any mixture model, which is given by

$$e_H^2(z) = 1/n_z - 1/n, \quad z \in \Omega, \quad n_z \neq 0.$$

For each  $z$ , the empirical residual function  $\hat{r}(z)$  is approximately normally distributed. The approximate pointwise standard error  $e_r(z)$  of the residual plot under the one-component model is given by

$$e_r^2(z) = n^{-1}[1/f(z; \hat{\mu}) - 1 - (z - \hat{\mu})^2/\sigma^2(\hat{\mu})].$$

The empirical gradient function  $\hat{D}(\mu)$  is approximately normally distributed for each  $\mu$ . The approximate pointwise standard error  $e_D(\mu)$  of the gradient plot under the one-component model is given by

$$e_D^2(\mu) = n^{-1}\{E_z[f(z; \mu)/f(z; \hat{\mu})]^2 - 1 - (\mu - \hat{\mu})^2/\sigma^2(\hat{\mu})\}.$$

The results about the residual plot and the gradient plot are direct applications of the theory in Lindsay and Roeder (1992). The proof for the log ratio plot is given in **Appendix B**. The empirical gradient plot is directly related to the theory of nonparametric maximum likelihood. In particular, the one-component model is the nonparametric maximum likelihood estimator for  $Q$  if and only if  $\hat{D}(\hat{\mu})$  is no greater than 0.

If the one-component model is true, the residual plot and the gradient plot should be horizontal lines through the origin while the log ratio plot should be a nonhorizontal line. Since it is easier to recognize a line if its slope is near to zero than a line with nonzero slope, we suggest adjusting the slope and intercept on the log ratio plot. Suppose

$$H(z) = a + bz + E(z), \quad a, b \in \mathcal{R}.$$

One kind of simple estimators for  $a$  and  $b$ , denoted by  $\hat{a}$  and  $\hat{b}$ , can be obtained using weighted least square estimation by regressing  $\hat{H}(z)$  on  $z$  with weights  $n_z$  for each  $z$ . These weights ensure that the slope and intercept are largely determined by those points having large  $n_z$ 's, which usually are those points having small  $z$ 's. Note that the statistic  $\hat{b}$  is the first order efficient estimator for the parameter  $\log \lambda$  in the homogeneity model, see Lindsay (1995). Let

$$\hat{E}(z) = \hat{H}(z) - \hat{a} - \hat{b}z.$$

The only difference between  $\hat{H}(z)$  and  $\hat{E}(z)$  is the intercept and slope. Since we can treat  $\hat{a}$  and  $\hat{b}$  as constants although they are obtained from the data,

the functions  $\widehat{H}(z)$  and  $\widehat{E}(z)$  share the same confidence bands. The strict convexity does not change if it exists.

Now we can define the confidence bands. Let  $Z_{\alpha/2}$  be the upper  $\alpha/2$  quantile of the standard normal distribution, we have  $\alpha$ -level pointwise confidence bands

$$(z, \widehat{E}(z) \pm Z_{\alpha/2} e_H(z)), (z, \widehat{r}(z) \pm Z_{\alpha/2} e_r(z)), \text{ and } (z, \widehat{D}(\mu) \pm Z_{\alpha/2} e_D(\mu))$$

for the log ratio plot, the residual plot and the gradient plot respectively. If there is no convex curve but a horizontal line contained in the confidence bands the homogeneity model may be the first choice due to the principle of parsimony. Moreover, the log ratio plot and the residual plot tend to be very similar in appearance. An advantage to the log ratio plot is that the assessment of convexity is more stringent. That is, if the log ratio plot is convex, then the residual plot must also be, but not vice versa.

## 5. The dispersion score test

The  $C(\alpha)$  test for homogeneity is a popular test procedure (Neyman and Scott 1966 and Lindsay 1995). We follow Lindsay (1995) to call it the *dispersion score test*. The dispersion score test is quite simple and tractable. It has an asymptotic normal distribution and has local optimality properties over a wide range of alternatives. There are several ways that arrive at the dispersion score test. The following derivation is due to Lindsay (1995). For detailed derivation and the motivation of this approach, the readers are referred to Lindsay (1995), where the original derivation of Neyman and Scott (1966) is also described.

Let  $\phi$  be the *natural parameter* of the zero-truncated binomial distribution, which is the log odds ratio of the original binomial distribution. Let  $f(z; \phi)$  be the density parameterized by  $\phi$  with respect to  $\gamma(z)$ . Define the *score function* to be

$$\nu_1(\phi, z) = [f(z; \phi)]^{-1} \frac{\partial f(z; \phi)}{\partial \phi}.$$

Then define the (Neyman) *dispersion score function* to be

$$\nu_2(\phi, z) = [f(z; \phi)]^{-1} \frac{\partial^2 f(z; \phi)}{\partial \phi^2} = \nu_1(\phi, z)^2 + \frac{\partial \nu_1(\phi, z)}{\partial \phi}.$$

Let  $\hat{\phi}$  be ML estimate for  $\phi$  under the one-component model. The test statistic is given by

$$V = \sum_{i=1}^n \nu_2(\hat{\phi}, z_i).$$

To standardize the test statistic  $V$ , define  $\tau(\phi)$  to be

$$\tau(\phi) = E_z[\nu_2(\phi, z)]^2 - \frac{\{E_z[\nu_2(\phi, z)\nu_1(\phi, z)]\}^2}{E_z[\nu_1(\phi, z)]^2}.$$

The dispersion score test is based on the following result:

*The standardized test statistic  $T = [n\tau(\hat{\phi})]^{-1/2}V$  is an asymptotically standard normal random variable under the one-component model.*

For detailed discussion, the readers are referred to Neyman and Scott (1966) and Lindsay (1995). The last proposition gives a simple test for homogeneity:

*Reject the homogeneity model at the level  $1 - \alpha$  if  $T > Z_\alpha$ .*

The dispersion score test is the least specific of our tools as it only examines one predictive feature of the mixture model, namely that the mixture model data is overdispersed relative to the one-component data. However, this makes it simple as well as more powerful in small samples. Since it is a test statistic, the magnitude of  $T$  should not be used as a measure of the degree of overdispersion; for a large sample, it could be quite large despite the presence of a small amount of heterogeneity.

Finally, the dispersion score test is related to the nonparametric maximum likelihood estimator for  $Q$  in that if the estimator for the latent distribution  $Q$  is the one-component model, then the sign of  $T$ , which is also the sign of  $\hat{D}(\hat{\mu})$ , must be the negative sign  $-$ .

## 6. Examples

### 6.1 *The Down's syndrome incidence data*

In the prevalence studies of epidemiology, exhaustive attempts to find diagnosed individuals and report all affected individuals are usually expensive. The number of missing cases can be estimated by the ascertained cases from various sources using capture-recapture methodology.

Names of Down's syndrome children in Massachusetts were obtained from five different record sources. The children were born between January 1, 1955 and December 31, 1959, and still alive on December 31, 1966. The five different record sources were: obstetric records, other hospital records, Massachusetts Department of Health, Massachusetts Department of Mental Health and schools. The dataset was analyzed by several researchers, such as Wittes (1970), Fienberg (1972) and Hook and Regal (1982). This dataset

is a multiple Bernoulli sample. We assume that each child had the same probability to appear in each of these five records, then the dataset becomes a binomial sample. The frequency counts from  $n_1$  to  $n_5$  are 248, 188, 81, 18 and 2.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

The residuals from  $r(1)$  to  $r(5)$  are 0.025,  $-0.046$ , 0.009, 0.101 and 0.502. There is a sign sequence  $+, -, +$ , which indicates that there is some mixing effect. In the log ratio plot and the residual plot, the curves themselves show slight convexity. There is a horizontal line through zero lying in the confidence bands of the residual plot and that of the gradient plot but the log ratio plot confidence bands does not contain a straight line. We obtain  $V = 19.399$ ,  $\tau(\hat{\phi}) = 0.686$ ,  $T = 1.011$  and P-value = 0.156. So the dispersion score test is not significant even at the level 0.10. This dataset displays only weak heterogeneity relative to the one-component model.

## 6.2 *Hepatitis A virus infection data*

From April to July 1995, an outbreak of the Hepatitis A virus infection occurred in and round a technical college in Taiwan with about 5000 students. The dataset concerns the infected students in that college; it was analyzed in Tsay and Chao (2001). There were three lists of records: (1) records based on a serum test taken by the Institute of Preventive Medicine, Department of Health of Taiwan, (2) records by doctors in local hospitals, and (3) records based on questionnaires conducted by epidemiologists, in which cases were either confirmed by serum test or identified by symptom combinations.

We assume that for each student, he/she had the same probability to be in each list. But the probabilities may be different among students. We have  $n_1 = 187$ ,  $n_2 = 56$  and  $n_3 = 28$ .

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

The residuals are  $r(1) = 0.0824$ ,  $r(2) = -0.3370$  and  $r(3) = 1.0339$ . We have the  $+, -, +$  sign sequence. Although  $M = 3$ , the graphical diagnostics give striking convexity. However, it seems that the gradient plot does not work better than the residual plot and the log ratio plot. The log ratio plot seems the best among these plots.

Here are statistics for the dispersion score test:  $V = 28.466$ ,  $\tau(\hat{\phi}) = 0.117$ ,  $T = 5.048$  and P-value= 0. The dispersion score test is highly significant.

### 6.3 *Meadow vole data*

The capture-recapture dataset of meadow vole was analyzed by Pollock et al. (1990) and Lee and Chao (1994). There were five consecutive trapping days and totally 102 distinct voles were captured. The frequency counts from  $n_1$  to  $n_5$  are 29, 15, 15, 16 and 27.

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

The residuals from  $r(1)$  to  $r(5)$  are 2.290,  $-0.401$ ,  $-0.579$ ,  $-0.368$  and 2.753. There is a sign sequence  $+, -, +$  to the residuals. All graphical diagnostics show strong convexity. We have  $V = 140.584$ ,  $\tau(\hat{\phi}) = 1.740$ ,  $T = 10.554$  and P-value= 0. So the dispersion score test is highly significant. Both the graphical diagnostics and the dispersion score test confirm the conclusion that there existed heterogeneity of capture probabilities that was made by Pollock et al. (1990) and Lee and Chao (1994).

### 6.4 *Hong Kong bird data*

This dataset was analyzed in Chao et al. (2000). The *World Wide Fund for Nature Hong Kong* hosts an annual *Big Bird Race*. There were 20 teams in the *Year 2000 Big Bird Race*. Each team had four members and they went around the whole city and recorded the bird species that they had observed. The total frequency-counts over all teams from  $n_1$  to  $n_{20}$  are 21, 16, 13, 10, 4, 13, 6, 4, 11, 1, 6, 5, 8, 3, 4, 6, 11, 15, 8 and 55.

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

[Figure 13 about here.]

The residuals from  $r(1)$  to  $r(20)$  are 116148.4, 6757.2, 662.9, 86.2, 6.906, 6.457, 0.248,  $-0.628$ ,  $-0.444$ ,  $-0.967$ ,  $-0.840$ ,  $-0.871$ ,  $-0.757$ ,  $-0.868$ ,  $-0.681$ , 0.112, 5.287, 36.3, 136.2 and 13682.3. There is a sign sequence  $+, -, +$  to the residuals. The log ratio plot shows a strong overall convexity with some departure in the midrange. The residual plot and the gradient plot show convexity but they are not so convincing due to the scaling problem. The gradient plot localized to an interval centered at  $\hat{\mu} = 11.59$  shows a strong convexity. We have  $V = 10591.02$ ,  $\tau(\hat{\phi}) = 45.126$ ,  $T = 106.296$  and P-value = 0. So the dispersion score test confirms the strong heterogeneity.

## 7. Discussion

The heterogeneity model introduces a larger variance than does the homogeneity model. Such a phenomenon is called “over-dispersion”, which is clear from the following equality:

$$\text{Var}(z; Q) = \text{Var}(E(z|\mu)) + E(\text{Var}(z|\mu)).$$

The convexity of the log ratio function, residual function and gradient function provides more refined assessment of this overdispersion. The conceptually simple graphical diagnostic procedures can display heterogeneity if it exists. The gradient plot shows convexity throughout the mean value parameter space which is smoother than the residual plot and the log ratio plot. The residual plot and the log ratio plot show “local perturbations in fit as well as an approximately convex configuration” (Lindsay and Roeder 1992). Based on our experiences with these plots, we recommend the log ratio plot as being the best single plot.

Over-dispersion is the basis for the dispersion score test, see Neyman and Scott (1966) and Lindsay (1995). The dispersion score test has some optimality properties stated in Neyman and Scott (1966). While graphical diagnostics procedures may provide satisfactory informal assessments, the dispersion score test provides a locally most powerful test for the existence of heterogeneity.

As a test, the dispersion score test is not specific for the mixture model, as it has power against any overdispersion alternative. On the other hand, it is usually simpler than likelihood ratio testing, which requires an iterative algorithm and lacks an explicit form for its asymptotic distribution, see Seidel, Mosler and Alker (2000) for other irregularities of likelihood ratio testing.

The model used here for population size estimation through capture-recapture experiments is the same as the model for the estimation of the

number of species in an ecological community using presence/absence information of each species. So the methodology presented here can be applied to test species evenness. The authors are investigating the diagnostics and the dispersion score test for the species estimation using abundance information. Similar results will be reported elsewhere.

#### ACKNOWLEDGEMENTS

Thanks for Dr. Anne Chao's suggestions to write this paper. Professor Lindsay's research was supported by National Science Foundation grant DMS-9870193.

#### REFERENCES

- Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge, Massachusetts: MIT Press.
- Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics* **6**, 138–155.
- Chao, A., Lin, S. P., Yang, H. C. and Yip, P. S. F. (2000). The analysis of Hong Kong Big Bird Race data for the Year of 2000. *Journal of Chinese Statistical Association* **38**, 231–241.
- Chapman, D. G. (1952). Inverse multiple and sequential sample census. *Biometrics* **8**, 286–306.
- Cormack, R. M. (1966). A test for equal catchability. *Biometrics* **22**, 330–342.
- Fienberg, S. (1972). The multiple recapture census for closed populations and incomplete  $2^k$  contingency tables. *Biometrika* **59**, 591–603.
- Hook, E. and Regal, R. (1982). Validity of bernoulli census, log-linear, and truncated binomial models for correcting under-estimates in prevalence studies. *American Journal of Epidemiology* **116**, 168–176.
- Karlin, S. (1968). *Total positivity*. Stanford University Press.
- Laplace, P. S. (1783). Sur les naissances, les mariages et les morts. *Paris* in Histoire de L'Académie Royale des Sciences.
- Lee, S. M. and Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* **50**, 88–97.
- Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. Institute of Mathematical Statistics.
- Lindsay, B. G. and Roeder, K. (1992). Residual diagnostics for mixture models. *Journal of the American Statistical Association* **87**, 785–794.

- Lloyd, C. J. (1992). Testing capture homogeneity in a recapture model. *Biometrika* **79**, 555–561.
- Neyman, J. and Scott, E. L. (1966). On the use of  $C(\alpha)$  optimal tests of composite hypothesis. *Bulletin of the Institute of International Statistics* **41 I**, 477–497.
- Otis, D. L., Burnham, K. P., White, G. C. and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife monograph* **62**, 1–135.
- Pollock, K. H., Nichols, J. D., Bronwie, C. and Hines, J. E. (1990). Statistical inference for capture-recapture experiments. *Wildlife monographs* **107**, 1–97.
- Seidel, W., Mosler, K. and Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics* **52**, 481–487.
- Shaked, M. (1980). On mixtures from exponential families. *Journal of the Royal Statistical Society, Series B* **42**, 192–198.
- Titterton, D. M., Smith, A. F. and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley.
- Tsay, P. K. and Chao, A. (2001). Population size estimation for capture-recapture models with applications to epidemiological data. *Journal of Applied Statistics* **28**, 25–26.
- Wittes, J. T. (1970). *Estimation of population size: the Bernoulli census*. PhD thesis, Harvard University, Department of Statistics.

## APPENDIX A

### *The formulas for computation*

Four parameters can be used to parameterize the binomial density and the zero-truncated binomial density: the capture probability  $\pi$ , the mean value  $\mu$ , the odds ratio  $\lambda$  and the log odds ratio  $\phi$ . There are bijections between each pair of the four parameters. In the development of the graphical diagnostics and the dispersion score test, we have used  $\mu$ ,  $\lambda$  and  $\phi$  to facilitate the narration of results. For computational convenience, we use the density parameterized by  $\pi$  and present all necessary expressions in terms of  $\pi$ .

Let  $\pi_*$  be  $1 - \pi$  and  $\hat{\pi}_*$  be  $1 - \hat{\pi}$ . Let  $\hat{\pi}$  be the maximum likelihood estimator of  $\pi$  under the one-component model. The density of the zero-truncated binomial distribution parameterized by  $\pi$  with respect to counting

measure on  $\mathcal{N}$  is given by

$$f(z; \pi) = \frac{M! \pi^z \pi_*^{M-z}}{z!(M-z)!(1-\pi_*^M)}, \quad z \in \Omega.$$

Let

$$\mu_i = E(z^i), \quad \kappa_i = (1 - \pi_*^M) \mu_i, \quad i = 1, 2, 3, 4.$$

Then

$$\begin{aligned} \kappa_1 &= M\pi, \\ \kappa_2 &= [M\pi + M(M-1)\pi^2], \\ \kappa_3 &= M\pi + 3M(M-1)\pi^2 + M(M-1)(M-2)\pi^3, \\ \kappa_4 &= M\pi + 7M(M-1)\pi^2 + 6 \prod_{j=0}^2 (M-j)\pi^3 + \prod_{j=0}^3 (M-j)\pi^4. \end{aligned}$$

Moreover, we have

$$\mu = \mu_1, \quad \sigma^2 = \mu_2 - \mu_1^2 = \frac{M\pi\pi_*}{(1-\pi_*^M)^2} (1 - \pi_*^M - M\pi\pi_*^{M-1}),$$

$$\nu_1(\phi, z) = z - \mu, \quad \nu_2(\phi, z) = (z - \mu)^2 - \sigma^2,$$

$$V = \sum_{i=1}^n (z_i - \hat{\mu})^2 - n\sigma^2(\hat{\mu}),$$

$$\tau(\phi) = \mu_4 + 8\mu_2\mu_1^2 - 4\mu_1\mu_3 - \mu_2^2 - 4\mu_1^4 - (\mu_2 - \mu_1^2)^{-1}(\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3)^2,$$

and

$$E \left[ \frac{f(z; \mu)}{f(z; \hat{\mu})} \right]^2 = \frac{\pi_*^{2M}(1 - \hat{\pi}_*^M)^2}{\hat{\pi}_*^{2M}(1 - \pi_*^M)^3} \left\{ \left[ \pi_* + \pi \left( \frac{\pi \hat{\pi}_*}{\pi_* \hat{\pi}} \right)^2 \right]^M - \pi_*^M \right\}.$$

## APPENDIX B

### *The pointwise normality of the log ratio plot*

Let  $z_i$  given  $\phi_i$  follow a discrete distribution  $f(z; \phi_i)$ ,  $i = 1, 2, \dots, n$ , where  $f(z; \phi_i)$  is the density with respect to some  $\sigma$ -finite measure in the form as follows:

$$f(z; \phi) = c(\phi)e^{z\phi}, \quad z \in \Omega,$$

where  $\Omega$  is the sample space. Some examples of such distributions are binomial distributions, Poisson distributions and all kinds of truncated versions of these distributions.

Suppose the  $\phi_i$ 's constitute a random sample from a latent distribution  $Q(\phi)$ . Let  $f(z; Q)$  be the mixture density with respect to counting measure. Then the  $z_i$ 's constitute a random sample from  $f(z; Q)$ . For each  $z$ , by the law of large numbers, we have

$$n_z/n \xrightarrow{P} f(z; Q) \text{ as } n \rightarrow +\infty.$$

By the central limit theorem, we have

$$n^{1/2}(n_z/n - f(z; Q)) \xrightarrow{d} N(0, f(z; Q)[1 - f(z; Q)]) \text{ as } n \rightarrow +\infty.$$

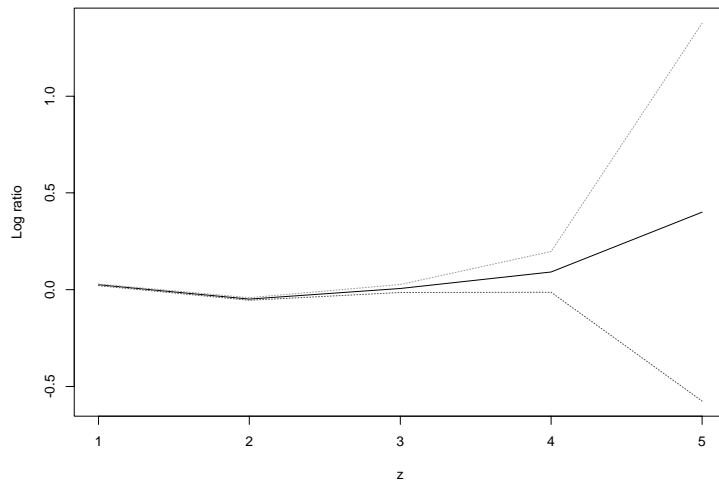
By the delta method, we obtain

$$n^{1/2}(\log n_z/n - \log f(z; Q)) \xrightarrow{d} N(0, [f(z; Q)]^{-1}[1 - f(z; Q)]),$$

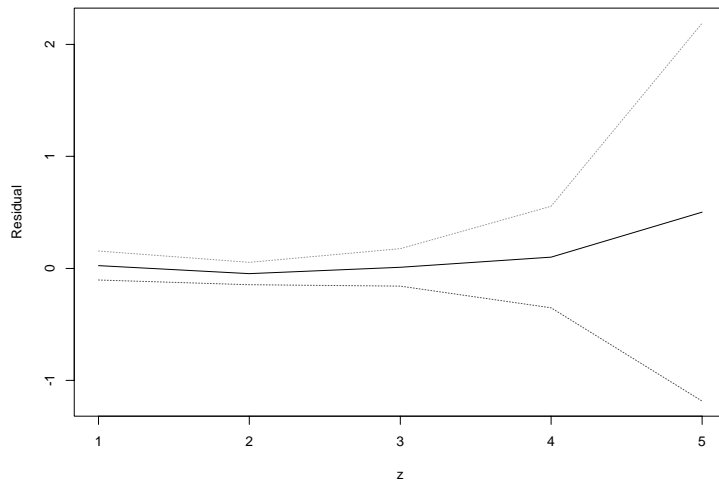
as  $n$  goes to infinity. That is, approximately we have

$$\log n_z/[nd\gamma(z)] \sim N(\log f(z; Q)/d\gamma(z), n_z^{-1} - n^{-1}),$$

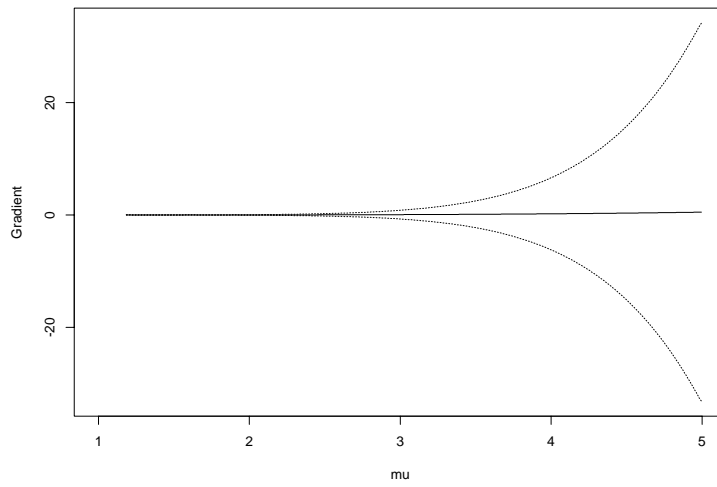
where the variance has been estimated, that is,  $f(z; Q)$  is replaced by its consistent estimator  $n_z/n$ .



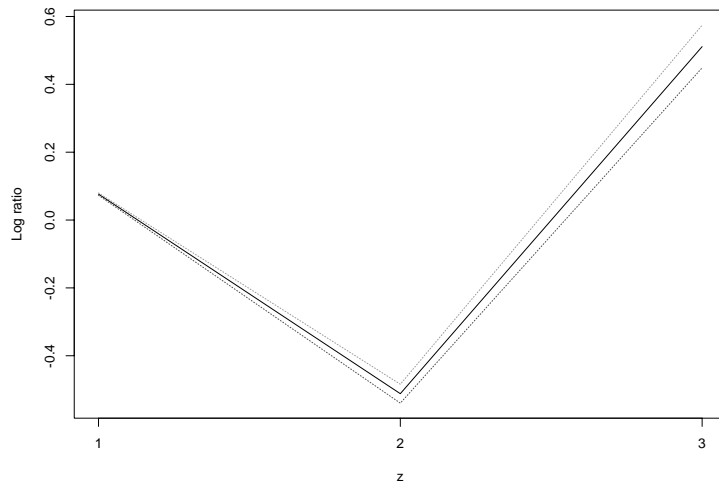
**Figure 1.** Log ratio plot: Down's syndrome incidence data



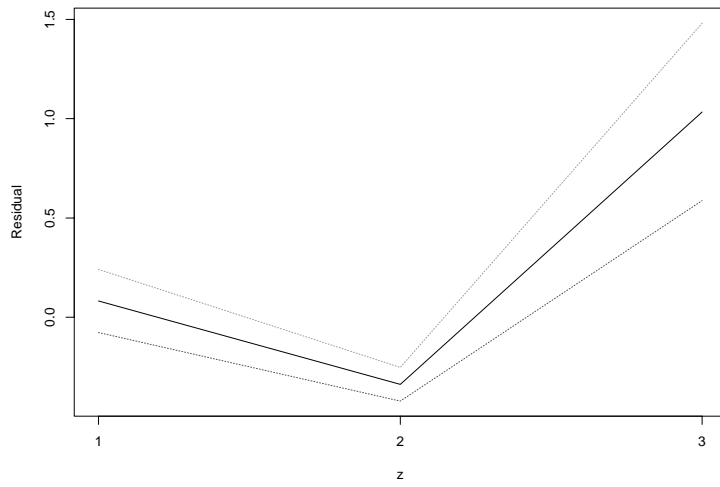
**Figure 2.** Residual plot: Down's syndrome incidence data



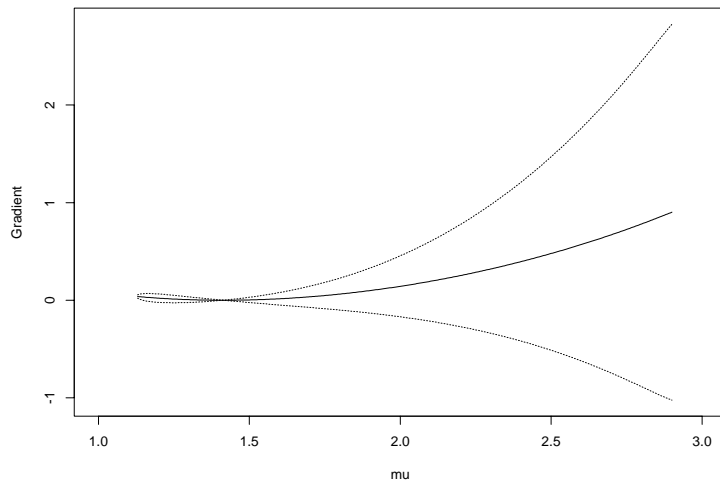
**Figure 3.** Gradient plot: Down's syndrome incidence data



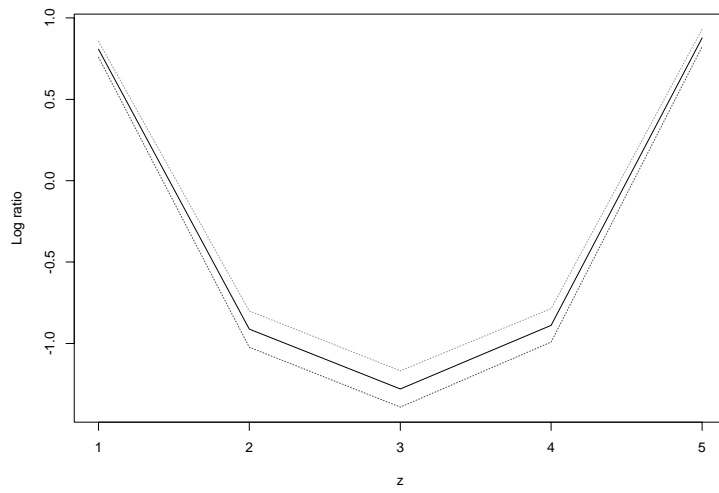
**Figure 4.** Log ratio plot: Hepatitis A virus infection data



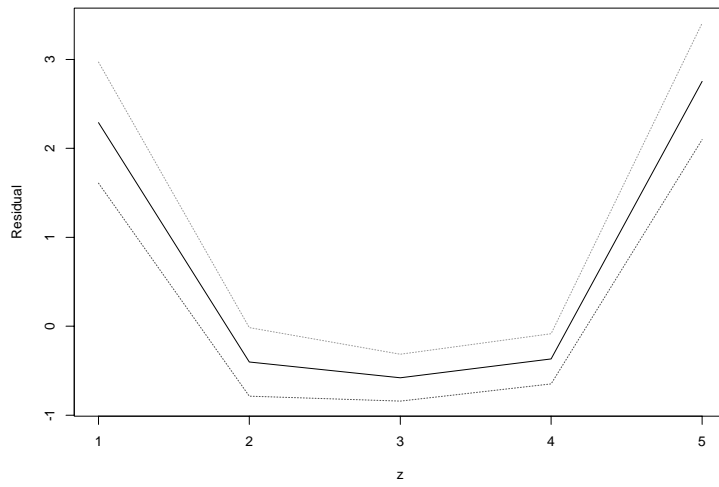
**Figure 5.** Residual plot: Hepatitis A virus infection data



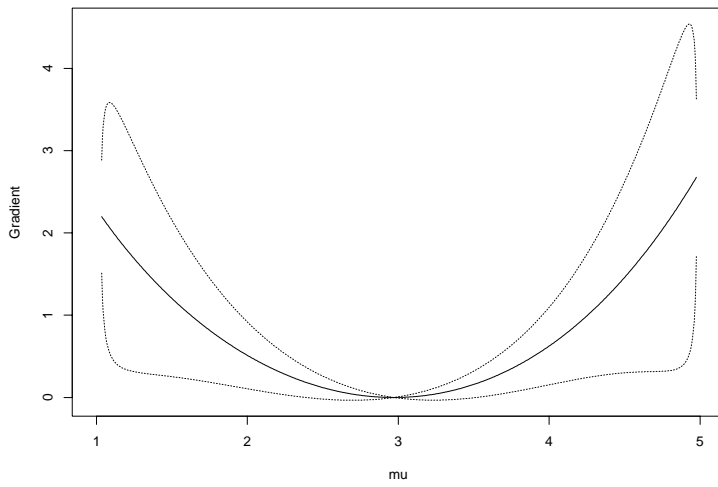
**Figure 6.** Gradient plot: Hepatitis A virus infection data



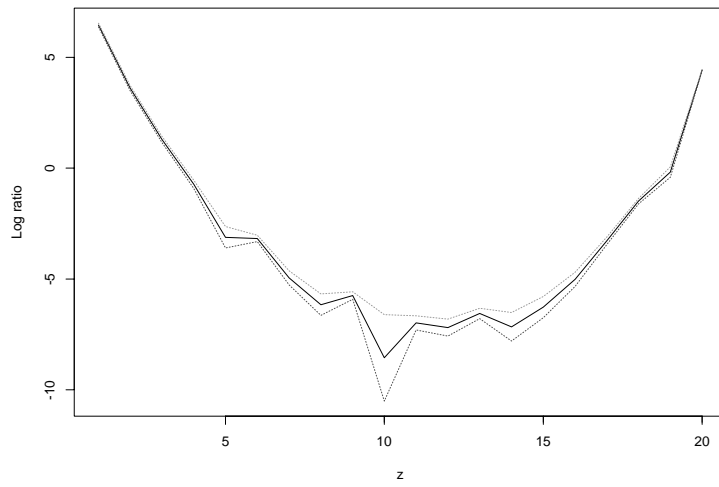
**Figure 7.** Log ratio plot: Meadow vole data



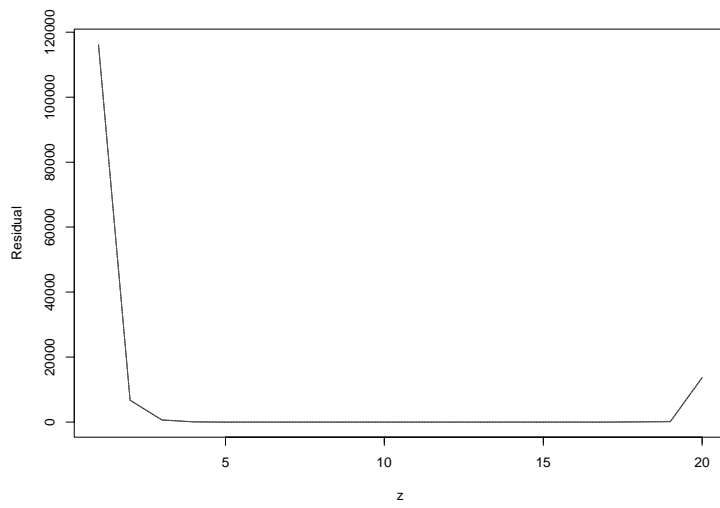
**Figure 8.** Residual plot: Meadow vole data



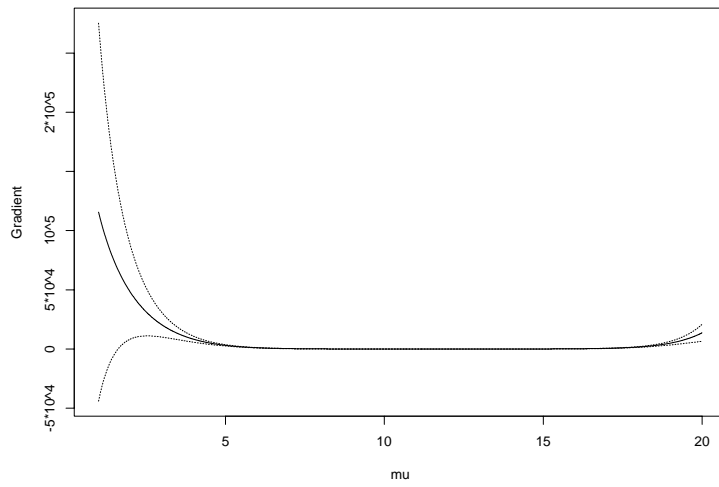
**Figure 9.** Gradient plot: Meadow vole data



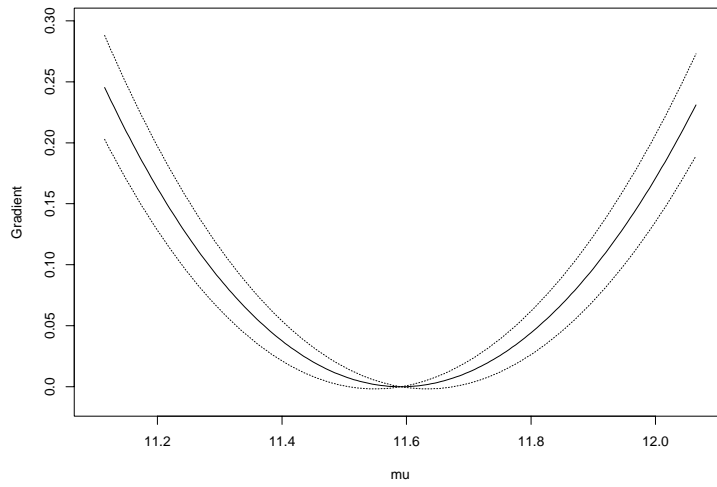
**Figure 10.** Log ratio plot: Hong Kong bird data



**Figure 11.** Residual plot: Hong Kong bird data



**Figure 12.** Gradient plot: Hong Kong bird data



**Figure 13.** Localized gradient plot : Hong Kong bird data