MAXIMUM LIKELIHOOD ESTIMATION AND SIMULATION FOR ONE AND TWO STAGE SAMPLING OF NORMAL MIXTURES USING SAS* SOFTWARE

Wesley Kemmerlin
Kirby L. Jackson, University of South Carolina

ABSTRACT

In the investigation of a disease often a single continuous variable is measured for each individual and on the basis of the value of this variable an individual is classified as belonging to the diseased or nondiseased group. Assuming that the variable is distributed normally for each group with a possible different mean and variance then the observed data for a sample is distributed as a mixture of normal distributions. In addition if, in a second stage, a subset of the overall sample is measured on a variable that allows determination of the individuals true classification then more accurate estimation of the parameters for each normal distribution is possible. In this paper a procedure is developed using maximum likelihood estimation to estimate the parameters of the mixed normal populations under a multistage sampling procedure and perform simulations of possible sampling results. A SAS macro program is presented that can be used to perform the simulations and the required maximum likelihood estimation procedures. This general SAS program can be used for estimation of parameters in a simple mixed normal situation and can also be used when a two stage procedure is implemented.

I. INTRODUCTION TO PROBLEM

Suppose an investigator has a true and a fallible measuring device to classify subjects into one of two categories, diseased or nondiseased. The fallible device is a relatively inexpensive screening test which tends to misclassify subjects, whereas the true device is a much more expensive reference test which is subject to no misclassification. Using only the screening test on a sample of subjects from the population results in a biased or imprecise estimate of the prevalence of disease. A better estimate could be obtained if the reference test were used, but this is too expensive to do in most cases. A compromise between these two extremes can be obtained by using a two stage sampling scheme. The first stage consists of taking a sample of N people and applying a fairly cheap screening test to them. The second stage consists of taking a subsample of n people and performing the reference test on them. Under certain conditions this method is substantially more efficient than a one stage procedure.

II. STATISTICAL MODEL AND LIKELIHOOD THEORY

The statistical method for the two stage procedure may be described as follows:

Stage 1: Screening. Draw from frame a preliminary sample of N people. Give each person a test in which a continuous score will result with a higher score indicating a greater probability of the disease being present.

Stage 2: Reference test. Take a subsample of n people from the first stage and give them an exam in which the disease status (disease present or disease absent) is determined. The probability of selection into this sample may depend upon score on screening exam. Denote the diseased group as Group 1 and the nondiseased group as Group 2.

Assuming normal distributions for the first stage scores for the two groups, diseased and nondiseased, maximum likelihood estimation procedures are used to estimate the parameters of the mixed normal distribution. Five parameters must be estimated; the mean score for the diseased group (μ_1), the variance of the score for the diseased group (σ_1^2), the mean score for the nondiseased group (μ_2), the variance of the score for the nondiseased group (σ_2^2), and the prevalence of disease. Two different models were used for estimating the parameters. One model makes the assumption of a common variance (σ^2) for the scores, i.e. $\sigma_1^2 = \sigma_2^2$. In this model only four parameters are estimated. The second model does not make the equal variance assumption. Therefore, in this model, all five parameters must be estimated. Maximum likelihood estimates of these parameters were obtained by using formulas based on the EM (expectation-maximization) algorithm as described by Aitkin and Wilson (1980). These formulas are similar to those derived by Hosmer (1973) for a two stage sampling procedure with a simple random second stage sample. A SAS program for the basic estimation of a mixture of normals was also described by Berg(1986).

One disadvantage of the EM algorithm is that it does not automatically give an estimate of the asymptotic variance covariance matrix. The matrix of second derivatives were calculated for both the equal and unequal variance models. The inverse of the negative of this matrix evaluated at the parameter estimates was used as an estimate of the covariance matrix. The values across the diagonal of this matrix are

the asymptotic variances of the parameter estimates. Formulas for the first and second derivatives for the unequal variance model for a mixed normal distribution are given in Hasselblad(1966). These formulas were easily extended to the two stage procedure. In addition, formulas for the equal variance model were derived.

Let f_i, i = 1,2 denote the density function of a $N(u_i,\sigma_i 2)$ random variable where $-\infty < \mu_i < \infty$, $0 < \sigma_i 2 < \infty$.

Define:

$$f_3 = pf_1 + qf_2 \text{ where } 0<p<1 \text{ and } q=1-p.$$

Let $V=(p,u_1,\sigma_1,u_2,\sigma_2^2)$ denote the vector of unknown parameters. Also, let

$$f_0 = p^g(q)^{1-g}$$

where g=1 for the diseased group and 0 for the nondiseased group.

The likelihood function based on x, r, and g can be written as:

$$f(x,r,g) = f(x)\ f(r,g|x) = f(x)\ f(r|x)\ f(g|r,x).$$

Note that f(r|x) does not involve any of the parameters of the mixed distribution. Also, note that f(g|r,x)= f(g|x). Therefore, the likelihood is proportional to: f(x) f(g|x).

Let $x_1,x_2,\ldots x_N$ be a sample from a mixture of normals. After this sample is taken, a subsample of N is taken. This subsample is then broken down into 2 categories, diseases and nondiseased. Let n_1=number of observations classified as diseased in subsample, n_2=number of observations classified as nondiseased in subsample, and n_3=number of observations only in first sample. Hence, N is the total sample size, n_1 and n_2 are the number of observations taken where the component (disease present or disease absent) is known, and n_3 is the number taken from the mixed population. Thus the likelihood is proportional to

$$L(R) = \prod_{i=1}^{n_1} f_0 f_1 * \prod_{j=1}^{n_2} f_0 f_2 * \prod_{k=1}^{n_3} f_3$$

The important factor is that the likelihood is not affected by a sampling scheme that is based upon the first stage score.

III. MAXIMUM LIKELIHOOD ESTIMATION PROGRAM

A SAS macro, NORLIK, was written that performs maximum likelihood estimation on a mixture of two normals. The user sets initial parameter estimates and the macro calculates new parameter estimates. If the differences between the new estimates and the initial estimates are not small (as defined by user), another iteration of parameter calculation is performed. Iterations continue until all of the differences between the last estimate and the next to the last estimate are small enough to satisfy user or maximum number of iterations is exceeded. The asymptotic variance of the final estimate is then calculated.

The outline of NORLIK is as follows:

1. Input first stage scores
2. Sort first stage scores
3. Choose individuals for second stage sample
4. Input disease status variable for those taken in second stage
5. Merge data set containing first stage score with data set containing disease status variable
6. Calculate stratified data estimate of prevalence
7. Sum number in second stage who were classed as diseased
8. Put in initial parameter estimates
9. Sum first stage scores for individuals in second stage who were diseased and sum scores for those who were not diseased
10. Begin macro for maximum likelihood estimation
11. Calculate weights for those not in second stage
12. Use PROC MEANS to sum weights
13. Use Symput function to pass values from PROC MEANS to the next data step
14. Set initial estimates and calculate new parameter estimates
15. Calculate differences between last and next to last parameter estimates
16. Stop iterations if maximum number of iterations are exceeded
17. End iterations if all of the differences between the last estimate and the next to last estimate is neglible
18. Calculate second derivatives for those in second stage where disease was present
19. Calculate second derivatives for those in second state when disease was absent
20. Calculate second derivatives for those not in second stage
21. Calculate log likelihood
22. Sum three groups of second derivatives
23. Calculate asymptotic variance of parameter estimates using PROC MATRIX

NORLIK can be used to estimate parameters in a mixture given that a sampling scheme has been decided upon. However, in order to decide upon a sampling scheme, computer simulation is necessary. The program for that simulation is discussed in the next section.

IV. SIMULATION PROCEDURE

In order to decide upon the sampling scheme, computer simulation was performed using a shell program, SIMSHELL, with NORLIK embedded in the simulation procedure. The shell program generated data given user supplied parameters and accumulated summary statistics for each of the sampling schemes examined. An outline of the program is shown below.

1. Decide upon sample size and number of samples to be generated.
2. Initialization of parameters.
3. Generate data.
4. Enter macro, NORLIK
5. Store summary data.

V. APPLICATION OF PROGRAM

The investigator has an easy written screening exam for depression with a much more rigorous clinical exam possible. It is desired to estimate the prevalence of depression in a population and simultaneously to detect and perform the clinical exam on as many of the depressed as possible. The question to be answered is what is the most efficient sampling scheme consistant with the study objectives.

A total of 100 samples with 1,000 observations each were generated. The generation of the normal distribution of scores for the nondepressed group was created using a true simulation mean of 1 and a true simulation variance of 0.5. The generation of the normal distribution of scores for the depressed group was created using a true simulation mean of 2 and a true simulation variance of 0.5 for the equal variance model and 0.75 for the unequal variance model. A prevalence of 0.05 was used for the true simulation value. The simulation parameter values were used as initial estimates for the iterations.

Nineteen percent (n=190) of the observations from the first stage were selected for the second stage and from different types of selecting the second stage sample were examined. The four types are listed in Table I.

Table I

Sampling Types

Type	Description
1	Take the top 100 scores and a random 90 of the rest
2	Take a random 190 of the 1000
3	Take the top 190 scores
4	Take the top 50 scores, a random 50 of the next 100, and a random 90 of the rest

Two different models were examined through simulation, equal and unequal variance models. The models were not directly compared since they were based on different generated data. Analysis of four second stage sampling types was done for each model. Results for both models were similar in most cases. However, the unequal variance model gave more biased estimates of the parameters. This may have occurred because of different data in the two models.

The relative efficiency of the sampling types for the maximum likelihood estimates of the prevalence was calculated by first denoting the sampling type with the smallest variance of the prevalence, sampling type 4, as the reference variance. The reference variance was divided by the variance of each of the sampling types to obtain the relative efficiency. Table II shows the results of this calculation.

Table II

Relative efficiency (RE) of both the maximum likelihood (ML) and stratified data (S) estimates of the prevalence for the four sampling types for both the equal and unequal variance models

	Equal Variance	Unequal Variance
Type	RE of ML estimate	RE of ML estimate
4	1*	1*
1	0.8995	0.7996
3	0.8940	0.4584
2	0.7611	0.7633

*This is the reference variance

Sensitivity, percentage of depressed individuals classed as depressed, was an important issue in the choice of an overall screening procedure. The relative sensitivity was calculated by first denoting the sampling type with the highest sensitivity, sampling type 3, as the reference sensitivity. The sensitivities of the other three smapling types were divided by the reference sensitivity to obtain the relative sensitivity. Table III shows the results of this calculation.

Table III

Relative sensitivity (RS) of the four sampling types for both the equal and unequal variance models

Equal Variance		Unequal Variance	
Type	RS	Type	RS
3	1*	3	1*
1	0.8244	1	0.8574
4	0.7657	4	0.7983
2	0.2928	2	0.2983

*This is the reference sensitivity

Sampling type 3 could run into some difficulties when the assumption of normality is violated. If this assumption is violated, taking only the top scores from the first stage could lead to all subjects chosen for second stage being in the same depression group. If all subjects were in the same depression group, the ML estimate of the prevalence could not be obtained. Therefore, sampling type 3 should be used with caution.

In conclusion, sampling type 3 should not be used when estimation of the prevalence is the main objective of a study. This sampling type could be used when the investigator did not want a prevalence estimate but wanted to detect as many diseased individuals as possible. Sampling type 2 should not be used when detection of diseased individuals is the main objective of a study. Sampling type 2 also does not seem to be as efficient in estimating the prevalence as sampling types 1 and 4. However, this difference in efficiency was not statistically significant.

Based upon this study sampling types 1 and 4 are the sampling types of choice. To choose between these two types, the investigator's objectives must be considered. Sampling type 1 is better if the investigator needs to detect as many individuals as possible who are depressed while still getting a fairly efficient estimate of the prevalence. Sampling type 4 is better if the investigator attaches more importance to the efficiency of the prevalence estimate and less to the detection of depressed individuals.

VI. SUMMARY

This application shows the usefulness of SAS software in performing sophisticated maximum likelihood analyses necessary for designing a study. A SAS macro can be used to perform simulations to answer an important question based upon a relatively complicated maximum likelihood program embedded in a more general shell macro.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Carol Garrison for providing the motivation for investigating this problem through her studies of adolescent depression and Dr. Mark Schluchter for advice and help in investigating the problem.

For more information, you may contact:
Kirby L. Jackson
Department of Epidemiology and Biostatistics
University of South Carolina
Columbia, SC 29208

References

Aitkin, M. and Wilson, G.T. (1980). Mixture Models, Outliers, and the EM Algorithm. Technometrics 22, 325-331.

Berg, Richard L. (1986). Mixture of Larval Instars. Procedings SUGI ,829-833.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. Technometrics 8, 431-444.

Hosmer, D.W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics 29, 761-770.

Kemmerlin, W.R. (1986). Efficiency in Two Stage Sampling for a Normal Mixture. Master's thesis, University of South Carolina.