Sample Size Calculation Based on Finite Mixture Model

Zejiang Yang*

Abstract

In a clinical study, the primary endpoint may have different distributions in different cases. Thus, the primary analysis could be based on the finite mixture model. In order to consider these different distributions, this paper will present a method of sample size calculation based on the finite mixture model. We assume that the components in the finite mixture model are normally distributed and the corresponding mixing proportions have a multinomial distribution. According to the Central Limit Theorem, the sample size can be determined using sample mean and variance of this finite mixture model, which are calculated through the conditional expectation and conditional variance. This method is applied to an example for showing the impact of different assumptions on the sample size calculation.

Key Words: Sample Size, Mixture Model, Conditional Variance, Multinomial Distribution

1. Introduction

In clinical studies, subjects may respond to treatment quite differently. This may depend on genetic polymorphism or other baseline characteristics. Thus a finite mixture model is an appropriate choice in modeling and analyzing these clinical data. Kessler and McDowell (2012) introduced SAS FMM Procedure for analyzing data with finite mixture models. Schlattmann P. (2009) discussed finite mixture models in much details.

However, available statistical literature, and many current statistical sample size software such as PASS, nQuery and SAS do not provide a direct solution on the sample size calculation based on the finite mixture model. While the primary endpoint in a clinical trial follows a finite mixture model, how to calculate the sample size for this study is a major interest. This motivates my research on sample size calculation based on a finite mixture model. In this paper, we assume that the components in the finite mixture model are normally distributed and the corresponding mixing proportions have a multinomial distribution.

2. Model Setting

Assume that the primary endpoint in a clinical study follows a finite mixture model. Let Y_T and Y_C be the primary endpoints for active treatment group and control group respectively, which can be simply expressed in the following notation:

$$Y_T \sim \sum_{i=1}^{K} p_{ti} X_{ti} \quad Y_C \sim \sum_{i=1}^{K} p_{ci} X_{ci} \tag{1}$$

We may further assume that

- For $i = 1, 2, ...K$,
  $E(X_{ti}) = \mu_{ti} \quad Var(X_{ti}) = \sigma_{ti}^2$
  $E(X_{ci}) = \mu_{ci} \quad Var(X_{ci}) = \sigma_{ci}^2$

*Biostatistics Consultancy Group, INC Research/inVentiv Health, 3201 Beechleaf Court, Suite 600, Raleigh, NC 27604

- p_{ti} and p_{ci} are the probability of belonging to category i for active treatment group and control group respectively, and $E(p_{ti}) = \theta_{ti}$ and $E(p_{ci}) = \theta_{ci}$, where $\sum_{i=1}^{K} \theta_{ti} = 1$ and $\sum_{i=1}^{K} \theta_{ci} = 1$.

Note p_{ti} (p_{ci}) for $i = 1, 2, ...K$ are dependent.

Let us denote $F_T(x)$ and $F_C(x)$ are the distribution functions of Y_T and Y_C respectively, $F_{ti}(x)$ and $F_{ci}(x)$ are the distribution function of X_{ti} and X_{ci} respectively, then based on Theorem of Total Probability,

$$F_T(x) = \sum_{i=1}^{K} p_{ti} F_{ti}(x) \quad F_C(x) = \sum_{i=1}^{K} p_{ci} F_{ti}(x)$$

Thus their corresponding density functions $f_T(x)$, $f_C(x)$, $f_{ti}(x)$ and $f_{ci}(x)$ have the following relationship:

$$f_T(x) = \sum_{i=1}^{K} p_{ti} f_{ti}(x) \quad f_C(x) = \sum_{i=1}^{K} p_{ci} f_{ci}(x) \tag{2}$$

which are the common definition of the finite mixture model.

3. Statistical Properties

Assume that in a clinical study, the primary endpoint is observed for n subjects in active treatment group and n subjects in control groups respectively. For active treatment group, $i = 1, 2, ...K$ and $j = 1, 2, ...n$, let

$$Z_{tij} = \begin{cases} 1 & \text{if jth subject in category } i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

then $n\hat{p}_{ti} = \sum_{j=1}^{n} Z_{tij}$ ($i = 1, 2, ...K$) has a multinomial distribution with parameters n and θ_{t1}, θ_{t2}, ... θ_{tK}. Similarly for control group, Z_{cij} has similar definition as in (3), thus $n\hat{p}_{ci} = \sum_{j=1}^{n} Z_{cij}$ has a multinomial distribution with parameters n and θ_{c1}, θ_{c2}, ... θ_{cK}. Note that

$$E(\hat{p}_{ti}) = \theta_{ti} \qquad E(\hat{p}_{ci}) = \theta_{ci} \tag{4}$$

$$E(\hat{p}_{ti}\hat{p}_{tj}) = \frac{(n-1)\theta_{ti}\theta_{tj}}{n} \qquad E(\hat{p}_{ci}\hat{p}_{cj}) = \frac{(n-1)\theta_{ci}\theta_{cj}}{n}$$

$$E(\hat{p}_{ti}^2) = \frac{n(n-1)\theta_{ti}^2 + n\theta_{ti}}{n^2} \qquad E(\hat{p}_{ci}^2) = \frac{n(n-1)\theta_{ci}^2 + n\theta_{ci}}{n^2}$$

Thus for $i = 1, 2...K$, and $j = 1, 2, ...K$ ($i \neq j$)

$$Var(\hat{p}_{ti}) = E(\hat{p}_{ti}^2) - (E(\hat{p}_{ti}))^2 = \frac{\theta_{ti}(1-\theta_{ti})}{n} \tag{5}$$

$$Cov(\hat{p}_{ti}, \hat{p}_{tj}) = -\frac{\theta_{ti}\theta_{tj}}{n} \tag{6}$$

Similarly,

$$Var(\hat{p}_{ci}) = \frac{\theta_{ci}(1-\theta_{ci})}{n} \tag{7}$$

$$Cov(\hat{p}_{ci}, \hat{p}_{cj}) = -\frac{\theta_{ci}\theta_{cj}}{n} \tag{8}$$

Assume $X_{ti1}, X_{ti2}, ...X_{tin_{ti}}$ are the n_{ti} samples from active treatment group and category i ($i = 1, 2..., K$), where $\sum_{i=1}^{K} n_{ti} = n$, and $X_{ci1}, X_{ci2}, ...X_{cin_{ci}}$ are the n_{ci} samples from

control group and category i ($i = 1, 2..., K$), where $\sum_{i=1}^{K} n_{ci} = n$. Then the sample mean for active treatment group is

$$\bar{X}_t = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_{ti}} X_{tij} = \sum_{i=1}^{K}\frac{n_{ti}}{n}\bar{X}_{ti} = \sum_{i=1}^{K}\hat{p}_{ti}\bar{X}_{ti} \tag{9}$$

where $\hat{p}_{ti} = n_{ti}/n$, and

$$\bar{X}_{ti} = \frac{1}{n_{ti}}\sum_{j=1}^{n_{ti}} X_{tij} \tag{10}$$

which is the sample mean for active treatment group in category i, thus $E(\bar{X}_{ti}|\hat{p}_{ti}) = \mu_{ti}$ and $Var(\bar{X}_{ti}|\hat{p}_{ti}) = \sigma_{ti}^2/n_{ti} = \sigma_{ti}^2/(n\hat{p}_{ti})$. From (4) to (6),

$$\begin{aligned} \mu_t &= E(\bar{X}_t) = \textstyle\sum_{i=1}^{K} E(\hat{p}_{ti}\bar{X}_{ti}) = \sum_{i=1}^{K} E[\hat{p}_{ti}E(\bar{X}_{ti}|\hat{p}_{ti})] \\ &= \textstyle\sum_{i=1}^{K} \mu_{ti}E(\hat{p}_{ti}) = \sum_{i=1}^{K} \mu_{ti}\theta_{ti} \end{aligned} \tag{11}$$

$$\begin{aligned} Var(\hat{p}_{ti}\bar{X}_{ti}) &= Var[E(\hat{p}_{ti}\bar{X}_{ti}|\hat{p}_{ti})] + E[Var(\hat{p}_{ti}\bar{X}_{ti}|\hat{p}_{ti})] \\ &= Var[\hat{p}_{ti}E(\bar{X}_{ti}|\hat{p}_{ti})] + E[\hat{p}_{ti}^2 Var(\bar{X}_{ti}|\hat{p}_{ti})] \\ &= Var[\hat{p}_{ti}\mu_{ti}] + E[\hat{p}_{ti}^2\frac{\sigma_{ti}^2}{n\hat{p}_{ti}}] \\ &= \mu_{ti}^2 Var[\hat{p}_{ti}] + \frac{\sigma_{ti}^2}{n}E[\hat{p}_{ti}] \\ &= \frac{\mu_{ti}^2\theta_{ti}(1-\theta_{ti}) + \sigma_{ti}^2\theta_{ti}}{n} \end{aligned} \tag{12}$$

and for $i \neq j$,

$$\begin{aligned} Cov(\hat{p}_{ti}\bar{X}_{ti}, \hat{p}_{tj}\bar{X}_{tj}) &= E(\hat{p}_{ti}\bar{X}_{ti}\hat{p}_{tj}\bar{X}_{tj}) - E(\hat{p}_{ti}\bar{X}_{ti})E(\hat{p}_{tj}\bar{X}_{tj}) \\ &= E[\hat{p}_{ti}\hat{p}_{tj}E(\bar{X}_{ti}|\hat{p}_{ti})E(\bar{X}_{tj}|\hat{p}_{tj})] \\ &-E[\hat{p}_{ti}E(\bar{X}_{ti}|\hat{p}_{ti})]E[\hat{p}_{tj}E(\bar{X}_{tj}|\hat{p}_{tj})] \\ &= \mu_{ti}\mu_{tj}[E(\hat{p}_{ti}\hat{p}_{tj}) - E(\hat{p}_{ti})E(\hat{p}_{tj})] \\ &= \mu_{ti}\mu_{tj}Cov(\hat{p}_{ti}, \hat{p}_{tj}) \\ &= -\frac{\mu_{ti}\mu_{tj}\theta_{ti}\theta_{tj}}{n} \end{aligned} \tag{13}$$

Therefore

$$\begin{aligned} Var(\bar{X}_t) = \quad \frac{\sigma_t^2}{n} &= \frac{1}{n}[\sum_{i=1}^{K}(\mu_{ti}^2\theta_{ti}(1-\theta_{ti}) + \sigma_{ti}^2\theta_{ti}) \\ &-2\sum_{1<=i<j<=K}\mu_{ti}\mu_{tj}\theta_{ti}\theta_{tj}] \end{aligned} \tag{14}$$

Similarly for control group,

$$\mu_c = E(\bar{X}_c) = \sum_{i=1}^{K}\mu_{ci}\theta_{ci} \tag{15}$$

$$\begin{aligned} Var(\bar{X}_c) = \quad \frac{\sigma_c^2}{n} &= \frac{1}{n}[\sum_{i=1}^{K}(\mu_{ci}^2\theta_{ci}(1-\theta_{ci}) + \sigma_{ci}^2\theta_{ci}) \\ &-2\sum_{1<=i<j<=K}\mu_{ci}\mu_{cj}\theta_{ci}\theta_{cj}] \end{aligned} \tag{16}$$

4. Sample Size Calculation

Let the sample mean difference between active treatment group and control group be

$$D = \bar{X}_t - \bar{X}_c = \sum_{i=1}^{K} \hat{p}_{ti}\bar{X}_{ti} - \sum_{i=1}^{K} \hat{p}_{ci}\bar{X}_{ci} \tag{17}$$

Then

$$\mu_D = E(\bar{X}_t) - E(\bar{X}_c) = \sum_{i=1}^{K} (\mu_{ti}\theta_{ti} - \mu_{ci}\theta_{ci}) \tag{18}$$

$$Var(D) = \sigma_D^2 = Var(\bar{X}_t) + Var(\bar{X}_c) = \frac{\sigma_t^2 + \sigma_c^2}{n} \tag{19}$$

For a large sample size n,

$$\frac{\bar{X}_t - \bar{X}_c - \mu_D}{\sigma_D} \tag{20}$$

approximately has a standard normal distribution. Thus

$$\frac{\bar{X}_t - \bar{X}_c}{\sigma_D} \tag{21}$$

has a normal distribution with mean of $\mu^* = \mu_D/\sigma_D$ and standard deviation of 1, which is denoted as $N(\mu^*, 1)$. Under Hypothesis

$$H_0 : \mu_t = \mu_c \quad H_A : \mu_t > \mu_c$$

the statistical power

$$P(\frac{\bar{X}_t - \bar{X}_c}{\sigma_D} > z_\alpha) = P(N(\mu^*, 1) > z_\alpha) = P(N(0, 1) > z_\alpha - \mu^*) = 1 - \beta \tag{22}$$

Thus $z_\alpha - \mu^* = -z_\beta$, it can easily derived that the sample size per group is

$$n = (\sigma_t^2 + \sigma_c^2) \left(\frac{z_\alpha + z_\beta}{\mu_D} \right)^2 \tag{23}$$

The above sample size calculation is based on one-sided test. For two-sided test, we only replace α with $\alpha/2$ in (23).

5. Example

Assume that the primary variable in active treatment group and control group follows the mixture models with two components.

$$Y_T \sim p_t N(\mu_{t1}, \sigma_{t1}^2) + (1 - p_t)N(\mu_{t2}, \sigma_{t2}^2)$$

$$Y_C \sim p_c N(\mu_{c1}, \sigma_{c1}^2) + (1 - p_c)N(\mu_{c2}, \sigma_{c2}^2)$$

where p_t follows a binomial distribution $B(n, \theta_t)$ and p_c follows a binomial distribution $B(n, \theta_c)$. Table 1 shows the sample size results under different scenarios.

Table 1: **Sample Size Based on a Finite Mixture Model**

θ_t	μ_{t1}	μ_{t2}	σ_{t1}	σ_{t2}	θ_c	μ_{c1}	μ_{c2}	σ_{c1}	σ_{c2}	α	Power	n
0.3	5	8	6	7	0.3	4	5	6	6	0.05	0.8	90
0.4	5	8	6	7	0.4	4	5	6	6	0.05	0.8	106
0.5	5	8	6	7	0.5	4	5	6	6	0.05	0.8	126
0.6	5	8	6	7	0.6	4	5	6	6	0.05	0.8	152
0.7	5	8	6	7	0.7	4	5	6	6	0.05	0.8	189
0.3	5	8	6	7	0.3	4	5	6	6	0.05	0.9	124
0.4	5	8	6	7	0.4	4	5	6	6	0.05	0.9	146
0.5	5	8	6	7	0.5	4	5	6	6	0.05	0.9	174
0.6	5	8	6	7	0.6	4	5	6	6	0.05	0.9	211
0.7	5	8	6	7	0.7	4	5	6	6	0.05	0.9	261

6. Summary

This paper developed a sample size calculation method for a finite mixture model under the assumptions that the components in the finite mixture model are normally distributed and the corresponding mixing proportions have a multinomial distribution. Actually, it can be seen from Section 3 that this method is also applied to the situation when the components in a finite mixture model are not normally distributed. This method can incorporate information from all components of a finite mixture model into the sample size calculation. The sample size can be determined using sample mean and variance of the finite mixture model, which are calculated through the conditional expectation and conditional variance. It is recommended to use this method in those cases when the primary endpoint in a clinical study follows a finite mixture model.

REFERENCES

Schlattmann P. (2009), *Medical Applications of Finite Mixture Models*, Springer-Verlag, Berlin Heidelberg

Kessler D. and McDowell A. (2012) "Introducing the FMM Procedure for Finite Mixture Models," *SAS Global Forum, Statistics and Data Analysis*