SAMPLE SIZE CALCULATION FOR COMPARING TIME-TO-EVENT DATA

HANSHENG WANG
Guanghua School of Management,
Peking University
Department of Business Statistics
& Econometrics
Beijing, P. R. China

SHEIN-CHUNG CHOW
Millennium Pharmaceuticals, Inc.
Cambridge, Massachusetts

1 INTRODUCTION

In clinical research, the occurrence of certain *events* (e.g., adverse events, disease progression, relapse, or death) is often of particular interest to the investigators, especially in the area of cancer trials. In most situations, these events are undesirable and unpreventable. In practice, it would be beneficial to patients if the test treatment could delay the occurrence of such events. As a result, the time-to-event has become an important study endpoint in clinical research. When the event is death, the time-to-event is defined as the patient's survival time. Consequently, the analysis of time-to-event data is referred to as *survival analysis*.

The statistical method for the analysis of time-to-event data is very different from those commonly used methods for other types of data, such as continuous and binary response, for two major reasons. First, time-to-event data are subject to censoring (e.g., right, left, or interval censoring). In other words, the exact value of the response is unknown; however, it is known that the value is larger or smaller than an observed censoring time or within an observed censoring interval. Second, the time-to-event data are usually highly skewed, which makes many standard statistical methods designed for normal data not applicable. In this article, the authors will focus on procedures for sample size calculation based on time-to-event data, which is subject to right censoring. Three commonly employed testing procedures, namely exponential model, Cox's proportional hazard model, and the log-rank test, are discussed.

The remaining of this article is organized as follows. In Section 2, sample size formulae based on exponential model are derived. In Section 3, Cox's proportional hazard model is discussed. In Section 4, sample size formulae based on log-rank test are presented. For each model, the sample size formulae for testing equality, superiority, and equivalence/non-inferiority are derived. Finally, this article concludes with a general discussion.

2 EXPONENTIAL MODEL

In this section, the sample size formulas for testing equality, superiority, and equivalence/noninferiority are derived based on exponentially distributed time-to-event data. In clinical research, the exponential model is probably one of the most commonly used simple parametric models for time-to-event data. It only involves one parameter (i.e., time-independent hazard rate λ). Many other important parameters (e.g., median survival time) can be computed based on the λ. Therefore, the effect of different treatments are usually directly compared in terms of the hazard rate λ.

Suppose that a two-arm parallel design is conducted to compare a new treatment with a standard therapy. The duration of the trial is expected to be T with T_0 accrual time. Each patient will enter the trial independently with entry time a_{ij}, where $i = 0, 1$ is the indicator for treatment group and j is the patient's identification number in the ith treatment group. It is assumed that a_{ij} follows a continuous distribution with the density function given by

$$g(z) = \frac{\gamma e^{-\gamma z}}{1 - e^{-\gamma T_0}}, \quad 0 \le z \le T_0$$

where γ is the parameter for the patient's accrual pattern. More specifically, $\gamma = 0$ corresponds to a uniform patient entry, $\gamma > 0$ indicates the fast patient entry, and $\gamma < 0$ implies a lagging patient entry. Denoted by t_{ij} to be the time-to-event of the jth patient

in the ith treatment group, it is assumed that t_{ij} is exponentially distributed with hazard rate λ_i. Note that t_{ij} are not always observable because of the duration of the trial. Therefore, the observable variable is $(x_{ij}, \delta_{ij}) = (\min(t_{ij}, T - a_{ij}), I\{t_{ij} \le T - a_{ij}\})$. In other words, only the smaller values of $T - a_{ij}$ and t_{ij} are observable. It can be shown that the likelihood function of x_{ij} is given by

$$L(\lambda_i) = \frac{\gamma^n e^{-\gamma \sum_{j=1}^{n_i} a_{ij}}}{(1 - e^{-\gamma T_0})^n} \lambda_i^{\sum_{j=1}^{n_i} \delta_{ij}} e^{-\lambda_i \sum_{j=1}^{n_i} x_{ij}}$$

Thus, the maximum likelihood estimator (MLE) for λ can be obtained as

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{n_i} \delta_{ij}}{\sum_{j=1}^{n_i} x_{ij}}$$

According to the Central Limit Theorem and Slusky's Theorem, it can be established that

$$\sqrt{n_i}(\hat{\lambda}_i - \lambda_i) \to_d N(0, \sigma^2(\lambda_i))$$

where

$$\begin{aligned} \sigma^2(\lambda_i) &= \frac{\lambda_i^2}{E(\delta_{ij})} \\ &= \lambda_i^2 \left[1 + \frac{\gamma e^{-\lambda_i T}(1 - e^{(\lambda_i - \gamma)T_0})}{(\lambda_i - \gamma)(1 - e^{-\gamma T_0})} \right]^{-1} \end{aligned} \tag{1}$$

For more technical details, the readers may refer to Lachin and Foulkes (1) and Chow et al. (2).

2.1 Test for Equality

As indicated earlier, the treatment effect is usually defined as the difference in the hazard rates between treatment groups. Thus, the parameter of the interest is given by $\epsilon = \lambda_1 - \lambda_2$, where λ_i is the hazard rate of the ith treatment. In order to test for equality, the following hypotheses are usually considered:

$$H_0 : \epsilon = 0 \quad \text{versus} \quad H_a : \epsilon \neq 0$$

Under the null hypothesis, the following test statistic:

$$T = (\hat{\lambda}_1 - \hat{\lambda}_2) \left[\frac{\sigma^2(\hat{\lambda}_1)}{n_1} + \frac{\sigma^2(\hat{\lambda}_2)}{n_2} \right]^{-1/2}$$

is asymptotically distributed as a standard normal random variable. Therefore, for a given significance level α, the null hypothesis would be rejected if

$$\left| (\hat{\lambda}_1 - \hat{\lambda}_2) \left[\frac{\sigma^2(\hat{\lambda}_1)}{n_1} + \frac{\sigma^2(\hat{\lambda}_2)}{n_2} \right]^{-1/2} \right| > z_{\alpha/2}$$

where $z_{\alpha/2}$ is the $(\alpha/2)$th upper quantile of a standard normal distribution. On the other hand, under the alternative hypothesis that $\epsilon \neq 0$, the power of the above test can be approximated by

$$\Phi \left(|\lambda_1 - \lambda_2| \left[\frac{\sigma^2(\lambda_1)}{n_1} + \frac{\sigma^2(\lambda_2)}{n_2} \right]^{-1/2} - z_{\alpha/2} \right)$$

Hence, the sample size needed for achieving the desired power of $1 - \beta$ can be obtained by solving the following equation:

$$|\lambda_1 - \lambda_2| \left[\frac{\sigma^2(\lambda_1)}{n_1} + \frac{\sigma^2(\lambda_2)}{n_2} \right]^{-1/2} - z_{\alpha/2} = z_\beta$$

Assume that $n_1/n_2 = \kappa$, it follows that

$$\begin{aligned} n_1 &= \kappa n_2 \\ n_2 &= \frac{(z_{\alpha/2} + z_\beta)^2}{(\lambda_1 - \lambda_2)^2} \left[\sigma^2(\lambda_1)/\kappa + \sigma^2(\lambda_2) \right] \end{aligned} \tag{2}$$

2.2 Test for Noninferiority/Superiority

From a statistical point of view, the problem of testing noninferiority and superiority can be unified by the following hypotheses:

$$H_0 : \epsilon \le \delta \quad \text{versus} \quad H_a : \epsilon > \delta$$

where δ is the superiority or non-inferiority margin. For a given significance level α, the null hypothesis would be rejected if

$$(\hat{\lambda}_1 - \hat{\lambda}_2 - \delta) \left[\frac{\sigma^2(\hat{\lambda}_1)}{n_1} + \frac{\sigma^2(\hat{\lambda}_2)}{n_2} \right]^{-1/2} > z_\alpha$$

On the other hand, under the alternative hypothesis that $\epsilon > \delta$ is true, the power of the above test can be approximated by

$$\Phi \left((\epsilon - \delta) \left[\frac{\sigma^2(\lambda_1)}{n_1} + \frac{\sigma^2(\lambda_2)}{n_2} \right]^{-1/2} - z_\alpha \right)$$

Thus, the sample size required for achieving the desired power of $1-\beta$ can be obtained by solving the following equation:

$$(\epsilon-\delta)\left[\frac{\sigma^2(\lambda_1)}{n_1}+\frac{\sigma^2(\lambda_2)}{n_2}\right]^{-1/2}-z_\alpha=z_\beta$$

Assume that $n_1/n_2=\kappa$, it follows that:

$$n_1=\kappa n_2$$
$$n_2=\frac{(z_\alpha+z_\beta)^2}{(\epsilon-\delta)^2}\left[\sigma^2(\lambda_1)/\kappa+\sigma^2(\lambda_2)\right]$$

2.3 Test for Equivalence

Therapeutic equivalence can be established by testing the following interval hypotheses:

$$H_0:|\epsilon|\geq\delta \quad \text{versus} \quad H_a:|\epsilon|<\delta$$

The purpose is to reject the null hypothesis of inequivalence and conclude the alternative hypothesis of equivalence. The above interval hypothesis can be partitioned into the following two one-sided hypotheses:

$$H_{01}:\epsilon\geq\delta \quad \text{versus} \quad H_{a1}:\epsilon<\delta$$
$$H_{02}:\epsilon\leq-\delta \quad \text{versus} \quad H_{a2}:\epsilon>-\delta$$

Note that tests for the above two one-sided hypotheses are operationally equivalent to the test for the interval hypotheses for equivalence. As a result, for a given significance level α, the null hypothesis would be rejected if

$$(\hat\lambda_1-\hat\lambda_2-\delta)\left[\frac{\sigma^2(\hat\lambda_1)}{n_1}+\frac{\sigma^2(\hat\lambda_2)}{n_2}\right]^{-1/2}<-z_\alpha$$

and

$$(\hat\lambda_1-\hat\lambda_2+\delta)\left[\frac{\sigma^2(\hat\lambda_1)}{n_1}+\frac{\sigma^2(\hat\lambda_2)}{n_2}\right]^{-1/2}>z_\alpha$$

On the other hand, under the alternative hypothesis that $|\epsilon|<\delta$ is true, the power of the above tests can be approximated by

$$\Phi\left((\delta-\epsilon)\left[\frac{\sigma^2(\lambda_1)}{n_1}+\frac{\sigma^2(\lambda_2)}{n_2}\right]^{-1/2}-z_\alpha\right)+$$
$$\Phi\left((\delta+\epsilon)\left[\frac{\sigma^2(\lambda_1)}{n_1}+\frac{\sigma^2(\lambda_2)}{n_2}\right]^{-1/2}-z_\alpha\right)-1$$

Therefore, assuming that $n_1/n_2=\kappa$, the sample size needed for achieving the power of $1-\beta$ is given by

$$n_1=\kappa n_2$$
$$n_2=\frac{(z_\alpha+z_{\beta/2})^2}{\delta^2}\left[\sigma^2(\lambda_1)/\kappa+\sigma^2(\lambda_2)\right] \quad \text{if } \epsilon=0$$
$$n_2=\frac{(z_\alpha+z_\beta)^2}{(\delta-|\epsilon|)^2}\left[\sigma^2(\lambda_1)/\kappa+\sigma^2(\lambda_2)\right] \quad \text{if } \epsilon\neq0 \tag{3}$$

2.4 An Example

Consider a cancer trial comparing a new treatment with a standard therapy in treating patients with a specific cancer. The trial is planned for a total of 4 years ($T=4$) with 2-year accrual ($T_0=2$). Uniform patient entry ($\gamma=0$) is assumed for both treatment groups. According to a pilot study, it is estimated that the hazard rate for treatment and the standard therapy are given by 1.50 ($\lambda_1=1.50$) and 2.00 ($\lambda_2=2.00$), respectively. Thus, according to Equation (1), the variance function can be obtained as

$$\sigma^2(\lambda_i)=\lambda_i^2\left(1+\frac{e^{-\lambda_i T}-e^{-\lambda(T-T_0)}}{\lambda_i T_0}\right)^{-1}$$

This result yields $\sigma^2(1.50)=2.29$ and $\sigma^2(2.00)=4.06$. Assume equal sample size allocation between treatment groups ($\kappa=1$). By Equation (2), the sample size required for achieving an 80% ($\beta=0.20$) power at 5% level of significance ($\alpha=0.05$) is given by

$$n=\frac{(z_{\alpha/2}+z_\beta)^2}{(\lambda_2-\lambda_1)^2}\left(\frac{\sigma^2(\lambda_1)}{k}+\sigma^2(\lambda_2)\right)$$
$$=\frac{(1.96+0.84)^2}{(2.00-1.50)^2}(2.29+4.06)\approx200$$

Hence, a total of 400 patients (200 patients per treatment group) are needed for achieving an 80% power.

3 COX'S PROPORTIONAL HAZARDS MODEL

In clinical research, Cox's proportional hazard model (3) is probably one of the most

commonly used regression models for time-to-event data. More specifically, let t_i be the time-to-event for the ith subject, C_i be the corresponding censoring time, and z_i be the associated d-dimensional covariates (e.g., treatment indication, demographical information, medical history). Also, let $h(t|z)$ be the hazard rate at time t given the covariates z. It is assumed that $h(t|z) = h(t|0)e^{b'z}$, where b is the regression coefficient. For the purpose of simplicity, assume that the treatment indication is the only covariate available here. For a more general situation, the readers may refer to Schoenfeld (4).

3.1 Test for Equality

In order to test for equality, the following statistical hypotheses are considered:

$$H_0 : b = 0 \quad \text{versus} \quad H_a : b \neq 0$$

Consider the following test statistic:

$$T = \left[\sum_{k=1}^{d}\left(I_k - \frac{Y_{1i}}{Y_{1i}+Y_{2i}}\right)\right] \times \left[\sum_{k=1}^{d}\left(\frac{Y_{1i}Y_{2i}}{(Y_{1i}+Y_{2i})^2}\right)\right]^{-1/2}$$

where Y_{ij} denotes the number of subjects at risk just prior the ith observed event. T is the test statistic derived based on partial likelihood developed by Cox (3). As can be seen, it is the same as the commonly used log-rank test statistic. As a result, the procedures introduced in this section can also be viewed as the sample size calculation procedures for log-rank test but under proportional hazard assumption.

Under the null hypothesis, it can be verified that T is asymptotically distributed as a standard normal random variable. Therefore, for a given significance level α, the null hypothesis would be rejected if $|T| > z_{\alpha/2}$. On the other hand, under the alternative hypothesis that $b \neq 0$ is true, T is approximately distributed as a normal random variable with unit variance and mean given by $\log(b)(np_1p_2d)^{1/2}$, where p_i is the proportion of patients in the ith treatment group and d is the proportion of the patients with observed events. Hence, the sample size needed for achieving the power of $1-\beta$ is given by

$$n = \frac{(z_{\alpha/2}+z_\beta)^2}{b^2p_1p_2d} \tag{4}$$

For more technical details, the readers may refer to Schoenfeld (4, 5) and Chow et al. (2).

3.2 Test for Noninferiority/Superiority

As discussed earlier, the hypotheses of testing noninferiority and superiority can be unified by the following statistical hypotheses

$$H_0 : b \leq \delta \quad \text{versus} \quad H_a : b > \delta$$

where δ is the superiority or non-inferiority margin. Similarly, consider the following test statistic

$$T = \left[\sum_{k=1}^{d}\left(I_k - \frac{Y_{1i}e^{\delta}}{Y_{1i}e^{\delta}+Y_{2i}}\right)\right] \times \left[\sum_{k=1}^{d}\left(\frac{Y_{1i}Y_{2i}e^{\delta}}{(Y_{1i}e^{\delta}+Y_{2i})^2}\right)\right]^{-1/2}$$

The null hypothesis would be rejected at the α level of significance if $T > z_\alpha$. On the other hand, under the alternative hypothesis that $b > \delta$ is true, T is approximately distributed as a normal random variable with unit variance and mean given by $(b-\delta)(np_1p_2d)^{1/2}$. Hence, the sample size needed for achieving the power of $1-\beta$ is given by

$$n = \frac{(z_{\alpha/2}+z_\beta)^2}{(b-\delta)^2p_1p_2d} \tag{5}$$

3.3 Test for Equivalence

To establish therapeutic equivalence between a test treatment and a control, similarly consider the following interval hypotheses:

$$H_0 : |b| \geq \delta \quad \text{versus} \quad H_a : |b| < \delta$$

Testing the above interval hypotheses is equivalent to testing the following two one-sided hypotheses:

$$H_{01} : b \geq \delta \quad \text{versus} \quad H_{a1} : b < \delta$$
$$H_{02} : b \leq -\delta \quad \text{versus} \quad H_{a2} : b > \delta$$

As a result, the null hypothesis should be rejected at the α level of significance if

$$\left[\sum_{k=1}^{d}\left(I_k - \frac{Y_{1i}e^{\delta}}{Y_{1i}e^{\delta} + Y_{2i}}\right)\right] \times \left[\sum_{k=1}^{d}\left(\frac{Y_{1i}Y_{2i}e^{\delta}}{(Y_{1i}e^{\delta} + Y_{2i})^2}\right)\right]^{-1/2} < -z_{\alpha}$$

and

$$\left[\sum_{k=1}^{d}\left(I_k - \frac{Y_{1i}e^{-\delta}}{Y_{1i}e^{-\delta} + Y_{2i}}\right)\right] \times \left[\sum_{k=1}^{d}\left(\frac{Y_{1i}Y_{2i}e^{-\delta}}{(Y_{1i}e^{-\delta} + Y_{2i})^2}\right)\right]^{-1/2} > z_{\alpha}$$

On the other hand, under the alternative hypothesis that $|b| < \delta$ is true, the sample size needed for achieving the power of $1 - \beta$ is given by

$$n = \frac{(z_{\alpha} + z_{\beta/2})^2}{\delta^2 p_1 p_2 d} \quad \text{if } b = 0$$
$$n = \frac{(z_{\alpha} + z_{\beta})^2}{(\delta - |b|)^2 p_1 p_2 d} \quad \text{if } b \neq 0$$

3.4 An Example

Consider the same example as described in Section 2.4. Suppose that a constant hazard ratio of $e^{1.5}(b = 1.5)$ is assumed between the standard therapy and the test compound. It is also assumed that sample size will be evenly distributed between the two groups ($p_1 = p_2 = 0.50$). On the other hand, based on a pilot study, it is observed that about 20% ($d = 0.20$) of patients will experience death before the end of the study. According to Equation (4) the sample size needed for achieving an 80% power ($\beta = 0.20$) for establishment of therapeutic equivalence at the 5% ($\alpha = 0.05$) level of significance is given by

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2}{b^2 p_1 p_2 d} = \frac{(1.96 + 0.84)^2}{1.5^2 \times 0.50 \times 0.50 \times 0.20} \approx 70$$

Therefore, a total of 70 patients (35 patients per treatment group) are needed for achieving the desired power for establishment of therapeutic equivalence between the test treatment and the standard therapy. On the other hand, if the investigator wishes to show a superiority with a superiority margin of 40% ($\delta = 0.40$), then, according to Equation (5), the sample size needed is given by

$$n = \frac{(z_{\alpha} + z_{\beta})^2}{(b - \delta)^2 p_1 p_2 d} = \frac{(1.64 + 0.84)^2}{(1.50 - 0.40) \times 0.50 \times 0.50 \times 0.20} \approx 102$$

Hence, a total of 102 patients (51 patients per treatment group) are needed for showing superiority of the test treatment over the standard therapy with an 80% power at the 5% level of significance.

4 LOG-RANK TEST

In practice, it is commonly encountered that the time-to-event data are neither exponentially distributed nor do they satisfy the proportional hazard assumption. In such a situation, a nonparametric test is often considered for evaluation of the treatment effect. The hypotheses of interest are then given by

$$H_0 : S_1(t) = S_2(t) \quad \text{versus} \quad H_a : S_1(t) \neq S_2(t)$$

For testing the above hypotheses, the following log-rank test statistic is usually considered:

$$T = \left[\sum_{i=1}^{d}\left(I_i - \frac{Y_{1i}}{Y_{1i} + Y_{2i}}\right)\right] \times \left[\sum_{i=1}^{d}\left(\frac{Y_{1i}Y_{2i}}{(Y_{1i} + Y_{2i})^2}\right)\right]^{1/2}$$

where the sum is over all observed events, I_i is the indicator of the first group, and Y_{ij} is number of patients at risk just before the jth death in the ith group. The null hypothesis would be rejected at the α level of significance if $|T| > z_{\alpha/2}$. Under the alternative hypothesis, however, the asymptotic distribution of T is complicated. As a result, procedure for sample size calculation is also complicated. In this section, the procedure suggested by Lakatos (6, 7) will be introduced.

To calculate the desired sample size, Lakatos (6, 7) suggested that the trial period should be first divided into N equally spaced intervals. The length of the interval should be sufficiently small so that the hazard rate and number of patients at risk can be roughly treated as a constant within each interval. Therefore, θ_i can be defined as the ratio of the hazard of the event in the ith interval, ϕ_i as the ratio of patients in the two treatment groups at risk in the ith interval, and d_i as the number of deaths within the ith interval and $\rho_i = d_i/d$, where $d = \sum d_i$. It is further defined that

$$\gamma_i = \frac{\phi_i\theta_i}{1+\phi_i\theta_i} - \frac{\phi_i}{1+\phi_i} \text{ and } \eta_i = \frac{\phi_i}{(1+\phi_i)^2}$$

Assuming equal sample size allocation between treatment groups, the sample size needed for achieving the power of $1-\beta$ is given by $n = 2d/(p_1+p_2)$, where p_i is the proportion of the observed events in the ith treatment group and

$$d = \frac{(z_{\alpha/2}+z_\beta)^2(\sum_{i=1}^N w_i^2\rho_i\eta_i)}{(\sum_{i=1}^N w_i\rho_i\gamma_i)^2}$$

More details can be found in Lakatos (6, 7) and Chow et al. (2).

4.1 An Example

To illustrate the procedure for sample size calculation based on the log-rank test statistic, consider a two-year cardiovascular trial with the hazard rates of 0.50 and 1.00 for a test treatment and a control, respectively. It is assumed that the time-to-event is exponentially distributed. However, a log-rank test is used for comparing the two treatments. Assuming equal sample size allocation between treatment groups, the time interval [0, 2] can be partitioned into 20 equally spaced intervals. Within each interval, the parameters needed (e.g., θ_i, ρ_i) can be computed according to the formula introduced in the previous section. As a result, sample size required for achieving a 90% power at the 5% level of significance is given by 139. For more detailed procedures regarding this example, the readers may refer to Lakatos (7) and Chow et al. (2).

5 DISCUSSION

Sample size calculation for time-to-event data plays an important role in assuring the success of the clinical trial. In this entry, three commonly used sample size calculation procedures for time-to-event endpoint are briefly introduced. The first one is mainly due to Lachin and Foulkes (1) and is based on exponentially distributed data. The second one is mainly due to Schoenfeld (4, 5), which is based on Cox's proportional hazard model. The last one is due to Lakatos (6), which requires no parametric or semiparametric assumption. Comparing those three procedures, they requires less and less parametric or model assumptions, which is a merit from a statistical modeling point of view. However, the complexity required for sample size calculation also increases as the parametric assumption becomes weaker. From a practical point of view, which procedure should be used not only depends on its statistical properties, but credit should also be given to the procedure, which is computationally simple and stable.

REFERENCES

1. J. M. Lachin and M. A. Foulkes, Evaluation of sample size and power for analysis of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and sttratification. *Biometrics* 1986; **42**: 507–519.
2. S. C. Chow, J. Shao, and H. Wang, *Sample Size Calculation in Clinical Research*. New York: Marcel Dekker, 2003.
3. D. R. Cox, Regression models and life tables. *J. Royal Stat. Soc. B* 1972; **34**: 187–220.

4. D. A. Schoenfeld, Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983; **39**: 499–503.
5. D. A. Schoenfeld, The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 1981; **68**: 316–319.
6. E. Lakatos, Sample size determination in clinical trials with time-dependent rates of losses and noncomplicance. *Controlled Clin. Trials* 1986; **7**: 189–199.
7. E. Lakatos, Sample size based on the log-rank statistic in complex clinical trials. *Biometrics* 1988; **44**: 229–241.