CHAPTER 9

Survival Analysis and Reliability

9.1 SURVIVAL ANALYSIS

The term "survival analysis" is obviously self-explanatory. The objectives are to determine the proportion of a population that will survive past a particular point in time, as well as determining the rate at which death or "failure," in general, is occurring at some point in time, and what the factors are that can cause a change in the survival rate. Two-group survival studies are often used, with survival examined for each group over a specified length of time and sample size determined to detect a specified difference in the survival proportions at the conclusion of a study for a selected power value.

Industrial reliability analysis, on the other hand, is concerned with the functionality of manufactured items rather than people. More specifically, in that context, reliability is the probability that the unit will function adequately for the intended length of time. That type of reliability should be distinguished from other types of reliability. For example, interobserver reliability studies are conducted to investigate the reproducibility and level of agreement among people, as in rating products such as food items. Sample size can be determined for these other types of reliability studies, and this has been discussed in the literature, but in this chapter we will consider only industrial-type reliability studies since this parallels survival analysis since a unit of production will eventually die (or be discarded), just as a person will die.

The emphasis in this chapter, however, is on determining sample size for survival studies, as comparatively little attention has been given to determining sample size for reliability analyses. In general, sample size computations for survival studies are based on either (1) an estimate of surviving proportions for one or two groups at a fixed point in time, or (2) a model for an entire survival curve.

Sample Size Determination and Power, First Edition. Thomas P. Ryan.

Determining sample size for a survival analysis can be very challenging. Quoting from Simon (2004), "sample size is a tricky problem, and it is especially tricky for survival data models." The author pointed out that it is the number of deaths (or events, in general, not the number of patients) that is most important. This same point has been emphasized by other authors, including Collett (2003, p. 300). Simon (2004) also stated: "You can make some simplifying assumptions, but generally, power calculations for a survival data model are a mess." This is partly due to the nonconstant nature of the number of people who are part of the study at a given point in time, with some people entering, some dropping out, and some dying along the way. Consequently, some software that has sample size capability for survival analysis make some simplifying assumptions that serve to make sample size determination less complicated, although obviously also less precise. There is one well-known software package that allows the user to provide information regarding accrual periods, dropout rate, hazard rate, and sample size for each period and computes power using simulations. It cannot be used to directly determine sample size for survival studies, however. Another well-known software package permits the same type of information to be entered but determines sample size rather than computing power. Freeware for survival analysis falls into both categories.

Noncompliance is one of the complications as participants can discontinue use of study medication. Almost all existing sample size methods assume that when patients do discontinue, they do so independently of their risk of an endpoint, so that noncompliance is noninformative. Jiang, Snapinn, and Iglewicz (2004) pointed out that this is not always the case, however, and introduced a modified version of the sample size method proposed by Lakatos (1988) in the presence of informative noncompliance.

A detailed comparison of software and freeware for survival analysis is provided in Section 9.4. Since there is so much to chose from, a user needs to be careful in selecting software or freeware that will correspond to input that the user can provide and will give output in the form(s) that is needed.

9.1.1 Logrank Test

The logrank test is the most commonly used method in clinical trial work for comparing the total survival of two or more groups of individuals (Lakatos and Lan, 1992). The test can be used in clinical trial work to compare two types of treatments, such as two types of drugs. This might consist of a standard drug recommended for a certain physical condition, and a new drug that shows promise as possibly being superior to the standard drug.

There are several versions of a logrank test, depending on what is assumed. One version is the simplest logrank test because it is actually a nonparametric type of test as no assumption is made about the shape of the survival curve for any of the groups, nor is there an assumption about the distribution of survival

times. This is the method of analysis advocated by Bland and Altman (2004) and can be illustrated as follows. We will assume an example with two groups. The null hypothesis is that there is no difference between the two populations in the probability that an event (such as death) occurs at any point in time. Computations are performed each time there is an event and are based on the times of those events. For example, assume that there are 50 people in group #1 and 40 people in group #2, and that the first death occurs in week #4, and the person was in group #1. The probability of a person dying in week #4 is 1/90 since there were 90 people alive at the start of the week. Under the null hypothesis, the expected number of deaths in group #1 is $50/90 = 5/9$ and $40/90 = 4/9$ for group #2. Assume that the second death occurs in week #10 and the person is in group #2. There were 89 people alive at the start of that week, so the risk of dying was 1/89. The expected number of deaths were $49/89 = .55$ for group #1 and $40/89 = .44$ for group #2, respectively.

These calculations continue each time a death occurs and the deaths and expected number are summed for each group. The analysis that is then performed is just a simple chi-square test using the observed and expected figures for each group. For example, assume that the number of deaths in group #1 is 16 and the expected number is 13.82, with the number of deaths in group #2 being 18 and the expected number is 20.18. The usual χ^2 goodness-of-fit statistic is then computed as $\sum_g \left(O_g - E_g\right)^2 / E_g$, with g denoting the group, O representing the observed value, and E denoting the expected value. For this example, $\chi^2 = (16 - 13.82)^2/13.82 + (18 - 20.18)^2/20.18 = 0.58$. The number of degrees of freedom is the number of groups minus 1. The calculated value of χ^2 is quite small, for any number of degrees of freedom and for any significance level, and since $\chi^2_{.05,1} = 3.84$, the null hypothesis would not be rejected.

This is a simple test that has both advantages and disadvantages. The primary advantage is that it avoids the specification of a survival distribution for each population, which is likely to be skewed since there can be many events that occur early and much fewer that occur later. Specifying a specific skewed distribution could be difficult, however, and the assumption would not be possible to test with any reasonable amount of power in the absence of a considerable amount of data. The disadvantages of the test are that it is strictly a significance test and as such cannot provide an estimate of the size of the difference between the groups or a confidence interval on the mean difference.

Some readers may be interested in the short tutorial on the logrank test given by Bland and Altman (2004). Yateman and Skene (1992) may also be of interest because they approximated survival and loss distributions with piecewise exponential distributions and patient entry with a piecewise linear distribution, which they claimed significantly reduces the required computation and enables the expected number of deaths to be routinely evaluated.

As discussed in Section 9.1, power is based on the number of events (deaths), not the sample size, so it is desirable to determine the required number of deaths,

although obviously this cannot be fixed as one fixes a sample size. Nevertheless, it is important to know how may deaths are needed. Collett (2003, p. 300) gave the formula for the required number of deaths for the logrank test, while pointing out that the formula also applies to the Cox proportional hazards model (see Section 9.1.5). The formula is

$$d = \frac{4(Z_{\alpha/2} + Z_{\beta})^2}{\theta_R^2} \tag{9.1}$$

with θ_{R} denoting the true log-hazard ratio and $Z_{\alpha/2}$ and Z_{β} as defined and used in earlier chapters. The log-hazard ratio is defined as $\log(S_2)/\log(S_1)$, with S_1 and S_2 denoting the proportion surviving in the first and second groups, respectively. The designation of first and second group doesn't matter for Eq. (9.1) because the log of the hazard ratio if the latter had been defined as $\log(S_1)/\log(S_2)$ would just be the negative of the log of the hazard ratio as originally defined here, and since that number is squared in Eq. (9.1), the manner in which the hazard ratio is defined relative to the subscripts is immaterial.

Equation (9.1) is based on the assumption that there are equal proportions of individuals assigned to each treatment group. If the proportions differ, so that the proportions are π and $(1 - \pi)$ for the two groups, the appropriate formula is then

$$d = \frac{(Z_{\alpha/2} + Z_{\beta})^2}{\pi(1 - \pi)\theta_R^2}$$

Since $\pi(1 - \pi)$ has its maximum value when $\pi = 0.5$, the required number of deaths is inflated by the disparity in the proportions, although that inflation will be slight unless the respective proportions differ greatly.

The derivation of this formula is given by Collett (2003) and is based on the assumption that the logrank statistic has approximately a normal distribution. Thus, the formula is an approximation since the Z-variates in the formulas for equal and unequal proportions are based on the assumption of a normal distribution.

Hsieh (1987) gave a simple method for determining sample size for the logrank test (or t-test) for designs with unequal sample sizes, with simulation used to compare the performance of that method with other methods. Hsieh (1992) compared that method with several others and found that no method was uniformly best.

Zhang and Quan (2009) pointed out that the proportional hazards model won't always hold when the logrank test is used and derived formulas for calculating the asymptotic power of the logrank test under a nonproportional hazards alternative. In general, a modified approach will often be necessary when the logrank test is used since the assumptions for its use won't necessarily be met. This was also addressed by Song, Kosorok, and Cai (2008), who proposed modified logrank

tests that are robust with respect to different data-generating processes and gave the accompanying sample size formulas. Jung, Kang, McCall, and Blumenstein (2005) proposed modification of the logrank test for noninferiority trials with a survival endpoint and developed an accompanying sample size formula. Lu and Pajak (2000) considered sample size and power for a logrank test when change point(s) in treatment effect are given. Jung and Hui (2002) considered rank tests in general for comparing more than two survival distributions and gave both an asymptotic and approximate sample size formula. Jung (2008) introduced a sample size formula for weighted rank test statistics used with paired two-sample survival data. Gail (1985) compared a logrank test with a test of two proportions relative to sample size and power and found that each test has comparative advantages and disadvantages under certain conditions.

9.1.1.1 Freedman Method

Lakatos and Lan (1992) compared several methods of determining sample size for the logrank test. One of these is due to Freedman (1982), whose sample size formula is based on the expected value and variance of the logrank test. With that method, sample size is determined as follows. Let θ denote the hazard ratio, which can be expressed in several ways, one of which is $\theta = \log(S_1)/\log(S_2)$, with S_1 denoting the proportion of subjects in group #1 who are surviving at the end of the clinical trial, and S_2 defined the same way for group #2. The hazard ratio is assumed to remain constant throughout the trial and time periods are not stated, nor is accrual considered.

Assume that in a particular clinical trial those proportions are $S_1 = .3$ and $S_2 = .4$ and that there are an equal number of subjects who begin. Then if θ is defined as $\theta = \log(S_1)/\log(S_2)$, we obtain $\theta = \log(.3)/\log(.4) = 1.314$. Assume further that the power is to be .80 and $\alpha = .05$ and that a two-sided test is to be used with equal group sizes. The sample size, n, is then determined as

$$\begin{aligned} n &= \left[\frac{\theta+1}{\theta-1}\right]^2 (z_{\alpha/2} + z_\beta)^2 / (2 - S_1 - S_2) \\ &= \left[\frac{1.314+1}{1.314-1}\right]^2 (1.96 + 0.8416)^2 / (2 - .3 - .4) \\ &= 327.96 \end{aligned}$$

so $n = 328$ would be used and this would be the sample size used in each of the two groups. [Note that if θ had been defined as $\theta = \log(S_2)/\log(S_1)$, the bracketed expression would have been $(1 + 1.314)/(1 - 1.314)$ and that would have produced the same sample size since $(1.314 - 1)^2 = (1 - 1.314)^2$.]

This solution can also be produced by several software. Specifically, PASS would be used for this example by first selecting "Survival" from the menu, and

then selecting "Logrank Tests (Freedman)," then entering the values for α, power, S_1, and S_2, and specifying that a two-sided test is to be used. PASS shows the power to be .8001—virtually identical to the target value—with $n_1 = n_2 = 328$ given as the solution, in accordance with the hand computation. The output also shows the total number of events (i.e., deaths) that must be observed for the two groups combined (427, in this case) in order for the stated power to be attained.

The user of nQuery would first select "Survival (Time to Event)" and then select "Logrank test for equality of survival curves." That output also shows $n_1 = n_2 = 328$ but shows 421 for the total number of events, in disagreement with the result given by PASS. [Although not listed in the menu, the reference for the procedure is Freedman (1982), suggesting that sample size is being determined using the Freedman approach.]

Similarly, the Freedman method is the default for the logrank test in Stata, which gives the same sample size for each, $n = 328$, but gives 428 as the "estimated number of events," thus differing slightly from the 427 given by PASS and more than slightly from the 421 events given by nQuery. This discrepancy may be due to different assumptions or different methods of estimation, but that isn't clear. The expected number of events, E, is given by $4(z_{\alpha/2} + z_\beta)^2/\log^2(1/\theta)$. This computation yields $E = 420.8$, which supports the solution given by nQuery.

If a one-sided test is desired, $z_{\alpha/2}$ in the formula for n would be replaced by z_α, so 1.645 would be used in place of 1.96. For this example, the sample size would then be computed as $n = 258.36$, so $n = 259$ would be used. This is the sample size given by PASS and nQuery, with the former indicating that the power is .8002 and nQuery showing that 332 events for the two groups combined are required to reach that power. Stata also gives $n = 259$ but gives 336 as the necessary number of events, again disagreeing with nQuery. Hand computation gives $E = 331.45$, which again supports the solution of nQuery.

9.1.1.2 Other Methods

A parametric test can be justified if there is faith in the distributional assumption that must be made. For example, survival might be considered to be exponential over time. The software nQuery has the capability for sample size determination for a test of equal exponential survival for two groups. For example, if $\alpha = .05$, a two-sided test is used, power = .90, the exponential parameter (mean) is 2 for one group and 3 for the other group, the total number of required events is 256.

Cantor (1992), however, explained that the assumption of an exponential distribution is inappropriate when a nonzero proportion of the population is expected to have indefinite survival. A Gompertz model was stated as being a reasonable alternative for such a scenario. A method was given for calculating the required accrual time for a clinical trial in which the treatments have Gompertz survival distributions that satisfy the proportional hazards assumption. A computer program that performs the necessary computations was also provided.

For the case when stratification is present, Ahnn and Anderson (1995) derived the sample size formula for the stratified logrank test, extending the result of Palta and Amini (1985). See also Ahnn and Anderson (1998).

9.1.2 Wilcoxon–Breslow–Gehan Test

A variation of the logrank test is used if a person considers events at different times to be of different importance, such as deaths early in a survival curve being of more importance than later deaths if a new treatment that is purported to be beneficial in avoiding early mortality is part of a comparison study. Thus, early deaths would have a high weight (i.e., penalty). In general, different times would be weighted unequally and the chi-square statistic would be

$$\chi^2 = \sum_{gt} \frac{R_t(O_{gt} - E_{gt})^2}{R_t^2 E_{gt}}$$

with R_t denoting the weight that is assigned at time t. This test has various names, one of which is the Wilcoxon–Breslow–Gehan test in recognition of the fact that it is a modification of the Wilcoxon test and has been proposed by both Breslow and Gehan.

This test can be performed in Stata, in addition to the logrank test and other variations of the logrank test. More specifically, the `stpower logrank` command in Stata estimates required sample size, power, and effect size for survival analysis comparing survivor functions in two groups using the logrank test. It provides options for unequal allocation of subjects to the two groups, possible withdrawal of subjects from the study, plus uniform accrual of subjects into the study.

It does not, however, provide the capability of determining sample size for the Wilcoxon–Breslow–Gehan test, nor does other sample size determination software.

9.1.3 Tarone–Ware Test

This is another weighted logrank test and it was proposed by Tarone and Ware (1977), who claimed that it may be more powerful against a range of alternatives than the logrank test and Wilcoxon–Breslow–Gehan test. Ahnn and Anderson (1998) gave a sample size formula for testing the equality of at least two survival distributions using the Tarone–Ware class of test statistics in the case of nonproportional hazards, time-dependent losses, noncompliance, and drop-in. Chow, Shao, and Wang (2008) described the Tarone–Ware test in their Section 7.4.1, which included the formula for sample size, and they gave an example of its use in Section 7.4.2.

9.1.4 Other Tests

Schumacher (1981) proposed distribution-free tests for use with censored data, with one test based on McNemar's (1947) test.

9.1.5 Cox Proportional Hazards Model

The Cox proportional hazards model (Cox, 1972) is also known as Cox regression, which was the topic of Section 5.3, where it was presented in some detail and will be presented in less detail here, with an emphasis on applications in survival analysis. It is covered in this chapter because the model is used in the comparison of survival curves for treatment groups and is often used to adjust for covariates such as patient age and disease stage.

As such, an understanding of the term "proportional hazards" is needed. As defined in Section 5.3, a hazard function is the probability of failure (i.e., death in survival analysis) at time t divided by the probability of survival until time t. As such, a hazard function is a function of time and a set of covariates, which might vary over time. The proportional hazards model separates these two effects, so that the model is written

$$h(t; \mathbf{x}) = \lambda(t) \exp[G(\mathbf{x}; \beta)]$$

using the notation of McCullagh and Nelder (1989), with $\lambda(t)$ denoting the baseline hazard and $G(\mathbf{x}; \beta)$ the model that contains the parameters and the set of covariates.

As explained by McCullagh and Nelder (1989), it is conventional, but not necessary, to assume that the covariates have a multiplicative effect on the hazard. If so, the model could then be written

$$h(t; \mathbf{x}) = \lambda(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m) \tag{9.2}$$

where $X_1, X_2, \ldots, X_m$ denote the covariates, and $\lambda(t)$ is the baseline hazard, with $X_1, X_2, \ldots, X_m$ the covariates. The word "proportional" comes from the $\lambda(t)$ term and the fact that there is no other term on the right side of the model that is a function of the time, t. The term $\lambda(t)$ might seemingly be a continuous, smooth function of time, but for the proportional hazards model it is defined only at times that events occur and, as stated by McCullagh and Nelder (1989), plays the role of a blocking factor in a blocked experimental design (see Section 6.2.5).

The model can be written in the form of a linear model by dividing each side of Eq. (9.2) by $\lambda(t)$ and then taking the logarithm of each side, producing

$$\log\left(\frac{h\{(t), (X_1, X_2, \ldots, X_m)\}}{\lambda(t)}\right) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \tag{9.3}$$

Notice that this is similar to the way that the alternative form of the logistic regression model was produced at the beginning of Section 5.2.

The development of sample size formulas for comparing survival curves for the proportional hazards model dates at least from Schoenfeld (1983) and includes Schoenfeld and Borenstein (2005). Schoenfeld's sample size formula for the logrank test for two-sample censored data is well known and widely used according to Eng and Kosorok (2005), who derived a sample size formula based on the limiting distribution of the two-sided supremum-weighted logrank statistic. Gangnon and Kosorok (2004) gave a simple formula for weighted logrank statistics applied to clustered survival data with variable cluster sizes and arbitrary treatment assignment within clusters that reduces to the Schoenfeld's (1983) formula when there is either no clusters or else independence within clusters. Xie and Waksman (2003) also derived a sample size formula for clustered survival data.

Schoenfeld (1983) illustrated how adjusting for covariates permits a reduction in sample size for the same power, giving an example in which the necessary sample size for 80% power was reduced from 272 to 212 when there was an adjustment for covariates.

Despite the popularity of the proportional hazards model in survival analysis work, adequacy of model fit should be assessed whenever it is applied, as with any other model. In particular, if the proportional hazards assumption is not met, the relative risk for variables in the model can be either overestimated or underestimated, with the result that the power for testing the corresponding parameters is reduced. Stablein, Carter, and Novak (1981) discussed how the appropriateness of the proportional hazards model in a particular setting can be tested.

What if the proportional hazards assumption is not met and there is a time dependency? There are two schools of thought on this: (1) use weighted estimation in Cox regression and (2) model the time-dependent effects.

As noted in Section 5.3, maximum likelihood won't always be the best estimation method for the Cox proportional hazards model (see also Heinze and Schemper, 2001), which presents a problem relative to sample size determination because the sample size determination methods that have been proposed, such as Hsieh and Lavori (2000), have been presented assuming that maximum likelihood is the method of estimation. Heinze and Dunkler (2008) presented a solution to the monotone likelihood problem that can occur with Cox regression and extended Firth's (1993) procedure to Cox regression with time-dependent effects.

9.1.6 Joint Modeling of Longitudinal and Survival Data

Although joint modeling of longitudinal and survival data has become popular in recent years, not much attention has been devoted to design aspects. Chen,

Ibrahim, and Chu (2011) derived a sample size formula for estimating the effect of the longitudinal process in joint modeling and extended Schoenfeld's (1983) sample size formula for estimating the overall treatment effect in the joint modeling setting.

9.1.7 Multistage Designs

Desseaux and Porcher (2007) presented a flexible design with sample size reevaluation for survival trials in the context of a two-stage design.

9.1.8 Comparison of Software and Freeware

Since survival analysis is such a commonly used statistical tool, it is not surprising that there is adequate software and freeware from which to choose. The capabilities and input requirements vary considerably, however, and this is potentially confusing to users who might use multiple software packages.

nQuery has two options for its logrank test with the difference between the two being the components of the total input information. Specifically, the user specifies the significance level and indicates whether the test is one-sided or two-sided.

PASS has several options, one of which determines sample size using the method of Freedman (1982). The way that this routine differs from the other logrank routines in PASS is indicated by the following statement from the PASS online help system: "Time periods are not stated. Rather, it is assumed that enough time elapses to allow for a reasonable proportion of responses to occur. If you want to study the impact of accrual and follow-up time, you should use one of the other logrank modules also contained in *PASS*." Thus, it is possible to account for accrual and follow-up time with PASS, but not with the Freedman logrank routine.

Power and Precision also has the capability for sample size determination for comparing two survival curves and provides several options, not in terms of test selection but relative to how a study is performed. Specifically, the user can choose between a hazard rate that is either constant or varies; either no attrition, constant attrition, or attrition that varies; and either subjects entering prior to the first study interval, entering during that study at a constant rate, or entering at a rate that varies.

Stata will perform sample size calculations for survival studies, including the logrank test, through its `stpower` command. Specifically, `stpower logrank` is used for determining sample size and power when the logrank test is used. Options that are provided include being able to account for unequal allocation of subjects between the two groups, withdrawal of subjects from the study, and uniform accrual of subjects into the study.

The `stpower cox` command in Stata estimates the required sample size, power, and effect size using Cox proportional hazard models with one or more covariates. Its options include accounting for possible correlation between the covariate of interest and other predictors, and for withdrawal of subjects from the study. The other Stata command for sample size determination and power in survival analysis is `stpower exponential`, which can be used for estimating sample size and power for comparing two exponential survivor functions.

In addition to the built-in capabilities, some menu driven programs have been contributed that add to Stata's sample size determination capability for survival analysis. This includes the programs of Royston and Babiker (2002) and Barthel, Royston, and Babiker (2005), with the latter updating the former. Barthel, Babiker, Royston, and Parmar (2006) contributed a Stata program, ART, that utilized a general framework for sample size calculation in survival studies for comparing two or more survival distributions using any one of a class of tests that includes the logrank test. The program has considerable flexibility and allows for the possible presence of nonuniform patient entry, nonproportional hazards, loss to follow-up, and treatment changes including crossover between treatments. Of course, power can be determined using simulation and Feiverson (2002) showed how to write programs that estimate the power of virtually any test that Stata can perform.

The overall survival analysis capabilities in Stata, including sample size determination, are described in Cleves, Gould, Gutierrez, and Marchenko (2008).

There is other software that have been available during the past few decades. For example, Dupont and Plummer (1990) described freeware that they contributed that can be used for studies with a survival response measure as well as dichotomous and continuous outcomes, with the software downloadable from `http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize`.

An applet for determining sample size when a survival curve is to be compared against a historical control is available at `http://www.cct.cuhk.edu.hk/stat/survival/Dixon1988.htm`.

Although software can render tables and nomograms unnecessary, they are still undoubtedly in use by many people. Chapter 8 of Machin, Campbell, Tan, and Tan (2009), which was cited in Chapter 7, contains three sample size tables for comparing survival rates and Schoenfeld and Richter (1982) provided nomograms for calculating the number of patients needed for a clinical trial that has survival as an endpoint.

9.2 RELIABILITY ANALYSIS

It should be noted that there are different types of reliability, as the reliability of a test instrument used in education is another type of reliability, and the term "reliability" is used often in that literature. Although sample sizes can, of course,

also be determined for nonindustrial reliability, as in Shoukri, Asyali, and Donner (2004), that will not be the focus of this section.

Meeker, Hahn, and Doganaksoy (2004) discussed the planning, including sample size determination, of life tests to demonstrate reliability. They gave an example for which the requirement was for a newly designed bearing for a washing machine to run flawlessly on 99% of all units for 4000 cycles. That is, the reliability was to be $R = .99$ and the objective was to demonstrate that with $100(1 - \alpha)\% = 90\%$ confidence. Thus, $\alpha = .10$. Meeker et al. (2004) gave the required sample size as $n = \log(\alpha)/\log(R)$, with this formula having been given previously by Hahn (1974). Thus, $n = \log(.10)/\log(.99) = 229.105$, so $n = 230$ bearings would be used.

They stated that the requisite number of test units can be reduced by running each unit beyond the specified lifetime, provided that some assumptions can be made about the distribution of time to failure based on experience and knowledge of the failure mechanism. Meeker and Escobar (1998) showed that under the assumption of a Weibull distribution as the life distribution, a zero failure demonstration test run for k multiples of the specified lifetime requires that

$$n = \frac{1}{k^{\beta}} \left[\frac{\log(\alpha)}{\log(R)} \right]$$

units be tested, with β the shape parameter of the Weibull distribution, α the significance level for the test, and R the reliability. As the title of their paper indicates, McKane, Escobar, and Meeker (2005) addressed sample size and number of failure requirements for demonstration tests when location-scale and log-location-scale distributions are involved and there is failure censoring. That is, a test is run until a specified number of failures occurs. They addressed sample size determination using graphs, estimating the necessary sample size from a graph. Unfortunately, whereas software for sample size determination in survival analysis is plentiful, software for determining sample size in various types of (industrial) reliability tests seems to be virtually nonexistent.

9.3 SUMMARY

Much has been written about sample size determination for tests used in survival analysis. In particular, several different methods of sample size determination have been proposed for the logrank test. Software developers have responded to this level of activity, with the result that PASS, in particular, has 15 procedures for determining sample size. There has not been a similar interest in software for reliability testing, however. Consequently, users need to rely on sample size formulas given in research papers and in books such as Mathews (2010), which covers more reliability tests than were given in this chapter.

REFERENCES

Ahnn, S. and S. J. Anderson (1995). Sample size determination for comparing more than two survival distributions. *Statistics in Medicine*, **14**, 2273–2282.

Ahnn, S. and S. J. Anderson (1998). Sample size determination in complex clinical trials using the log-rank test. *Statistics in Medicine*, **17**, 2525–2534.

Barthel, F. M.-S., P. Royston, and A. Babiker (2005). A menu-driven facility for complex sample size calculations in randomized controlled trials with a survival or a binary outcome. *Stata Journal*, **5**, 123–129.

Barthel, F. M.-S., A. Babiker, P. Royston, and M. K. B. Parmar (2006). Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, and loss to follow-up and cross-over. *Statistics in Medicine*, **25**, 2521–2542.

Bland, J. M. and D. G. Altman (2004). The logrank test. *British Medical Journal*, **328** (Number 7447, May 1), 1073. (Available online with free subscription at `http://www.bmj.com/cgi/content/full/328/7447/1073`.)

Cantor, A. B. (1992). Sample size calculations for the log rank test: A Gompertz model approach. *Journal of Clinical Epidemiology*, **45**(10), 1131–1136.

Chen, L. M., J. G. Ibrahim, and H. Chu (2011). Sample size and power determination in joint modeling of longitudinal and survival data. *Statistics in Medicine*, **30**(18), 2295–2309.

Chow, S.-C., J. Shao, and H. Wang (2008). *Sample Size Calculations in Clinical Research,* 2nd edition. Boca Raton, FL: Chapman and Hall/CRC.

Cleves, M. A., W. W. Gould, R. G. Gutierrez, and Y. Marchenko (2008). *An Introduction to Survival Analysis Using Stata*, 2nd edition. College Station, TX: Stata Press.

Collett, D. (2003). *Modeling Survival Data in Medical Research*, 2nd edition. London: Chapman and Hall.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Desseaux, K. and R. Porcher (2007). Flexible two-stage design with sample size reassessment for survival trials. *Statistics in Medicine*, **26**, 5002–5013.

Dupont, W. D. and W. D. Plummer (1990). Power and sample size calculations: A review and computer program. *Controlled Clinical Trials*, **11**, 116–128.

Eng, K. H. and M. R. Kosorok (2005). A sample size formula for the supremum log-rank statistic. *Biometrics*, **61**, 86–91.

Feiverson, A. H. (2002). Power by simulation. *Stata Journal*, **2**, 107–124.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine*, **1**, 121–129.

Gail, M. H. (1985). Applicability of sample size calculations based on a comparison of proportions for use with the log rank test. *Controlled Clinical Trials*, **6**, 112–119.

Gangnon, R. E. and M. R. Kosorok (2004). Sample-size formula for clustered survival data using weighted log rank statistics. *Biometrika*, **91**, 263–275.

Hahn, G. J. (1974). Minimum size sampling plans. *Journal of Quality Technology*, **6**(3), 121–127.

Heinze, G. and D. Dunkler (2008). Avoiding infinite estimates of time-dependent effects in small-sample survival studies. *Statistics in Medicine*, **27**(30), 6455–6469.

Heinze, G. and M. Schemper (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, **57**(1), 114–119.

Hsieh, F. Y. (1987). A simple method of sample size calculation for unequal-sample-size designs that use the logrank or *t*-test. *Statistics in Medicine*, **6**(5), 577–581.

Hsieh, F. Y. (1992). Comparing sample size formulas for trials with unbalanced allocation using the logrank test. *Statistics in Medicine*, **11**, 1091–1098.

Hsieh, F. Y. and P. W. Lavori (2000). Sample size calculations for the Cox proportional hazards model with nonbinary covariates. *Controlled Clinical Trials*, **21**(6), 552–560.

Jiang, Q., S. Snapinn, and B. Iglewicz (2004). Calculation of sample size in survival trials: The impact of informative noncompliance. *Biometrics*, **60**, 800–806.

Jung, S.-H. (2008). Sample size calculation for the weighted rank statistics with paired survival data. *Statistics in Medicine*, **27**, 3350–3365.

Jung, S.-H. and S. Hui (2002). Sample size calculation for rank tests comparing *k* survival populations. *Lifetime Data Analysis*, **8**, 361–373.

Jung, S.-H., S. Kang, L. McCall, and B. Blumenstein (2005). Sample size computation for two-sample noninferiority log-rank test. *Journal of Biopharmaceutical Statistics*, **15**, 969–979.

Lakatos, E. (1988). Sample sizes based on the logrank test in complex clinical trials. *Biometrics*, **44**, 229–241.

Lakatos, E. and K. K. G. Lan (1992). A comparison of sample size methods for the logrank test. *Statistics in Medicine*, **11**(2), 179–191.

Lu, J. and T. F. Pajak (2000). Statistical power for a long-term survival trial with a time-dependent treatment effect. *Controlled Clinical Trials*, **21**(6), 561–573.

Machin, D., M. J. Campbell, S.-B. Tan, and S.-H. Tan (2009). *Sample Size Tables for Clinical Studies*. London: BMJ Books.

Mathews, P. (2010). *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Fairport Harbor, OH: Mathews Malnar and Bailey, Inc.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

McKane, S. W., L. A. Escobar, and W. Q. Meeker (2005). Sample size and number of failure requirements for demonstration tests with log-location-scale distributions and failure censoring. *Technometrics*, **47**, 182–190.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.

Meeker, W. Q. and L. A. Escobar (1998). *Statistical Methods for Reliability Data*. New York: Wiley.

Meeker, W. Q., G. J. Hahn, and N. Doganaksoy (2004). Planning life tests for reliability demonstration. *Quality Progress*, August, 80–82.

Palta, M. and S. B. Amini (1985). Consideration of covariates and stratification in sample size determination for survival studies. *Journal of Chronic Diseases*, **38**, 801–809.

Royston, P. and A. Babiker (2002). A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or binary outcome. *Stata Journal*, **2**, 151–163.

Schoenfeld, D. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics*, **39**, 499–503.

Schoenfeld, D. and M. Borenstein (2005). Calculating the power or sample size for the logistic and proportional hazards models. *Journal of Statistical Computation and Simulation*, **75**, 771–785.

Schoenfeld, D. and J. Richter (1982). Nomograms for calculating the number of patients needed for a clinical trial with survival as an endoint. *Biometrics*, **38**, 163–170.

Schumacher, M. (1981). Power and sample size determination for survival time studies with special regard to the censoring mechanism. *Methods of Information in Medicine*, **20**, 110–115.

Shoukri, M. M., M. H. Asyali, and A. Donner (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research*, **13**, 251–271.

Simon, S. (2004). Sample size for a survival data model. (Electronic resource at `www.pmean.com/04/survival.html`.)

Song, R., M. R. Kosorok, and J. Cai (2008). Robust covariate-adjusted log-rank statistics and corresponding sample size formula for recurrent events data. *Biometrics*, **64**, 741–750.

Stablein, D. M., W. H. Carter, Jr., and J. W. Novak (1981). Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials*, **2**, 149–159.

Tarone, R. E. and J. H. Ware (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, **64**, 156–160.

Williamson, J. M., H.-M Lin, and H.-Y. Kim (2009). Power and sample size calculations for current status survival analysis. *Statistics in Medicine*, **28**, 1999–2011.

Xie, T. and J. Waksman (2003). Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Statistics in Medicine*, **22**(18), 2835–2846.

Yateman, N. A. and A. M. Skene (1992). Sample sizes for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions. *Statistics in Medicine*, **11**, 1103–1113.

Zhang, D. and H. Quan (2009). Power and sample size calculation for logrank test with a time lag in treatment effect. *Statistics in Medicine*, **28**, 2617–2638.

EXERCISE

9.1. Use PASS or other software to verify the sample size of 328 that was given in Section 9.1.1.1. If you use PASS, you will notice that the output includes the number of events for each group.