Adaptive Sample Size Calculation for Industrial Experiments

Submitted to
the Department of Statistics
of the University of Dortmund

In Fulfilment of
the Requirements for the Degree of
Doctor of Natural Sciences

by

Clovis Njontie Kouakep

Dortmund 2005

Supervisor:	Prof. Dr. Joachim Hartung, Dr. Guido Knapp
Co-Supervisor:	Prof. Dr. Götz Trenkler
Date of the oral examination:	December 20, 2005

CONTENTS

1 Introduction

In today's technological world, achievement and improvement of customer satisfaction, development of highly sophisticated products in record time, while improving productivity, product field reliability and quality have increased the need for more testing of materials, components and systems. Engineers, statisticians, scientists and others need to draw conclusion from scanty data. In order to select the best source of supply for use in new designs, computer manufacturers have to test electronic parts from each of several vendors. It is also the case for automobile manufacturers who have to inspect incoming shipments of bolts and nuts to compare lot quality with acceptance specifications. Chemical companies or Aircraft materials laboratories make fatigue-test for specimens of new metals with different combinations of chemical elements in order to compare the effects of these elements on tensile strength. An important step in the design of all these experiments is the determination of the number of specimens to be tested. Determining the appropriate sample size for an investigation in industrial experiments is an essential step in the statistical design of the project and it is usually a difficult one. The number of specimens in the investigation must be large enough to provide a reliable answer to the question addressed. Sample size is important for economic reasons; an undersized study can be a waste of resources for not having the capacity to produce useful results, while an oversized one uses more resources than are necessary. Historically, learning the techniques of sample size determination and power analysis have been difficult, because of relatively complex mathematical considerations and numerous different formulas.

In fixed sample size designs, one has to know in advance, the relevant alternative which should be detected with a given power and reliable guesses about nuisance parameters needed in the sample size formula. For example, for comparing means in a two-group parallel design the knowledge of the effect size and its variability are required to calculate the sample size needed for achieving a specified power when using a test at a given significance level. A misjudgement of the variance may lead to a seriously over- or under-powered study.

There has been considerable recent interest in achieving greater flexibility in sample size calculations, either because the sponsor is very uncertain about the underlying effect size or because of uncertainty about nuisance parameters like variance. Adaptive sample size calculations that preserve the type-I error and power are possible in both settings, especially in conjunction with group sequential trials where the data are routinely monitored. Adaptive sampling designs for statistical experiments are ones where the accruing data from experiments are used to adjust the experiment as it is being run.

Unfortunately, adaptive procedures are more complicated to design and to analyse and they tend to be more difficult to implement than fixed sample size procedures. Because of this, adaptive designs are usually overlooked in favour of simpler, though less efficient, fixed designs.

Adaptive designs can be divided into two categories:

- Designs with interim analysis which allow both sample size adaptation and early stopping (Bauer and Köhne: 1994, Proschan and Hunsberger: 1995, Lehmacher and Wassmer: 1999)
- Internal pilot study design that allows adaptation of the sample size during the ongoing experiment using the estimated variance obtained from an interim analysis. Wittes and Brittain (1990) provided a specific plan to design such pilot studies and for incorporating the data from the pilot phase into the final study results.

Shen and Fisher (1999) present a method of sequential analysis for industrial experiments, which uses all prior data to assign a weight to the next data, by using the **self-designing** method as introduced by Fisher (1998). This weight is used to guarantee the integrity of the variance of the final test statistic so that the overall type I error rate is preserved. The variance estimate and the effect size will be updated at each step and therefore the overall sample size. Extension was made by Hartung (2001) who presented a completely self-designing rule by taking the inverse normal transformation of the p-values within the classical Pocock (1977) design and Hartung and Knapp (2003) who proposed a flexible and effective adaptive method that allows for a completely self-designing of a group sequential experiment and a decision about stopping for significance of the sequential test results at each stage.

In this work, methods of adaptive sample size determination in industrial experiment for the comparison of two production process (or product, machine, systems) will be presented, further developed and the operating characteristics will be compared. The thesis is organised as follows:

Chapter 2 presents definitions of terms usually used in the context of hypothesis testing.

In chapter 3, a detailed discussion on sample size methodologies for some important situations in industrial experiments is provided: difference of means for two groups with normally distributed outcomes, equivalence of means for two groups with normally distributed outcome, comparison of two groups with reliability data. In the case of difference of means, we compared the exact computation of the sample size with an approximated computation to see that the difference is non relevant. By the equivalence test, methods of the computation of the sample size in case of the difference and the ratio of the mean of the two groups are presented and we conduct a small simulation study to access the power to simultaneously conclude that two means are both statistically different and equivalent. Beginning with a short review of the basic concepts of reliability studies, we present the methodologies of calculation of sample size using the log-rank test when the reliability function is exponentially distributed or follows a Weibull distribution.

Chapter 4 is an exposition of the methodology of internal pilot study and the self-designing procedure. By the internal pilot study, the proposal of Wittes and Brittain (1990) and the choice of the pilot sample will be of interest. The self-designing procedures as proposed by Shen and Fisher (1999) and Hartung (2001) are presented. Furthermore, following the idea of Yin and Shen (2005), we presented a way to combine the classical group sequential method to the self-designing procedure of Hartung. The resulting design will not only update the sample size and stop for futility at each stage, but also can stop at each stage for strong efficacy. The operating characteristics of the new design are evaluated and compared to those of the self-designing procedure using group sequential techniques of Pocock (1977) and O'Brien and Fleming (1979).

In chapter 5, we used for the first time procedure of internal pilot study to reestimate the sample size in the one-sided equivalence test with normally distributed outcomes. Using techniques similar to those of Kieser und Friede (2000), the exact type I error rate is computed using the pooled-variance to reestimate the variance. Characteristics of the reestimated sample size are also investigated through simulations. The work of Friede and Kieser (2003) is presented in the case where the one-sample variance has been used to reestimate the variance. The reestimated sample sizes of the two variance procedure are compared through simulations. Secondly, a simple way to use the self-designing procedure of Hartung (2001) in the one-sided equivalence test is proposed, to obtain an adequate sample size, to preserve the type I error and to gain power.

Chapter 6 deals with sample size adaptation for reliability study. After a detailed presentation of the theory of sequential analysis for reliability study as proposed by Tsiatis (1981, 1982), a flexible design method of updating sample size based on the idea of Hartung and Knapp (2003) is proposed and illustrated. The test statistic is a linear rank statistic similar to that proposed by Shen and Cai (2003) and the test has independent increments. Specimen entry is staggered. We illustrated the application of the proposed design using an engineering example. The performances of the procedure are investigated and compared to the usual log-rank test with fixed sample design under exponential failure time distribution (proportional hazard function) through a simulation study. Since the log-rank test behaves poorly in the case of non proportional hazard, a test statistic based on the integrated weighted difference in the Kaplan-Meier estimates of the reliabilities functions of the two groups is proposed. The performance of the proposed procedure is also evaluated under Exponential and Weibull failure time distribution through simulations. The proposed adaptive procedure is compared with adaptive procedure based on the log-rank test. Chapter 7 gives a summary of the work.

2 Hypothesis Testing

A statistical hypothesis is an assertion or conjecture concerning one or more populations. For instance, suppose we want to determine if a new product will meet a pre-defined design standard, or if process A will be better than process B, we perform a hypothesis test. The objective of hypothesis testing is to decide, based on the information derived from a sample, which of the following two claims is more likely:

1. The null **hypothesis,** H_0**,** is a statement that there is no difference between the groups being compared, with respect to the variable of interest.
2. The **alternative hypothesis,** H_a **or** H_1**,** is a statement that the null hypothesis is not true. It is the idea that there is a difference between the groups being compared, with respect to the variable of interest. In the context of power analysis and sample size determination, this difference is termed the "**effect size**".

Here are some examples of null hypotheses:

- The mean life of a new product at design stress level meets or is equal to a specified standard value.
- The average performance of product design A is the same as the average performance of product design B.
- There is no difference in the average quality of materials from supplier X and the average quality of materials from supplier Y.

In order to choose between the null and the alternative hypothesis, a **test statistic** or a **P-value** is calculated based on the available data. A P-value is the lowest level at which the observed value of the test statistic is significant. The P-value will be compared to a predetermined value in order to make the decision.

A null hypothesis can only be rejected. The acceptance of a hypothesis merely implies that the data does not give sufficient evidence to refuse it. On the other hand, rejection implies that the sample evidence refuses it.

In hypothesis testing there will always be the possibility of making a wrong decision. As shown in table 1, there are two kinds of errors:

- **Type I error** occurs when we reject the null hypothesis when it is true.
- **Type II error** occurs when we fail to reject the null hypothesis when it is false.

Reality **Decision**	H_0 **is true**	H_0 **is false**
Fail to reject H_0	Correct Decision	Type II error
Reject H_0	Type I error	Correct Decision

Table 2.1: *Possible situations in testing a statistical hypothesis*

The probability of type I error (sometimes called Producer's risk) is usually designated "alpha" or α, and statistical tests are designed to ensure that α is suitably small.

The probability of type II error (Consumer's risk) is designated β.

The **power of the test** $1-\beta$ is the probability of rejecting H_0 given that an alternative is true. **The more powerful the test, the better it is**.

In the following chapter, formulas for sample size determination of certain commonly used statistical distributions are presented, when the experimental objective has been formulated as a test of hypothesis. We will only consider the one-sided test of hypothesis, with the sample results used to detect deviations from the test in a single direction. The objective is the control of the power of an α-test for certain alternatives.

3 Fixed sample size

Sample size methodology has been well developed and standardised for some statistical methods, such as; paired and pooled t-tests, binomial proportion comparisons, equivalence test, regression models, correlation and simple survival analysis models. For some of these models, sample size calculations are exact in the sense of utilising a mathematical formula. In the absence of exact mathematical results, approximate formulas can sometimes be used, if not, simulation provides a viable alternative.

There is a small amount of published literature including Mace (1964), Kraemer and Thiemann (1987), Cohen (1988), Desu and Raghavarao (1990). There are numerous articles, especially in biostatistics journals, concerning sample size determination for specific tests and there is a growing amount of software for sample size determination, including nQuery Advisor (Elashoff, 2000), PASS (Hintze, 2000) and online calculators such as Lenth (2000).

One of the most popular approaches to sample size determination involves studying the power of a test of hypothesis. It is the approach emphasised here.

The sample size in the context of hypothesis testing is determined by controlling the power of an α-level test for certain alternatives. The power approach involves these elements:

- Specification of the underlying probability model for the data.
- Specification of a hypothesis test (Difference, Equivalence, ...) on a parameter θ.
- Specification of the significance level or type I error rate α of the test.
- Specification of an effect size $\tilde{\theta}$ that reflects an alternative of scientific interest.
- Specification of a goal value of the power at the test when $\theta = \tilde{\theta}$.
- Computations of the power function of the test. (Other parameters needed are estimated or given. In fact there are three parameters under our control which define the power of the study. These are the sample size, the effect size and α. Considering these three parameters and the power of the study together, specification of any three will allow the determination of the fourth.)

In the following sample size methodologies for the important situation in industrial studies will be presented.

3.1 Comparison of means of normally distributed data

Tests concerning two means represent a set of very important analytical tools for the engineer. We consider an experiment comparing two production groups with independent normally distributed outcomes with expectations μ_E for the experimental group and μ_S for the standard group, and with unknown but common variance σ^2. Two independent random samples of size n_E and n_S, respectively, are drawn from the two production groups.

Let X_E and X_S designate the normally distributed outcome of interest for the experimental and standard production groups respectively. We know that the random variable

$$Z = \sqrt{\frac{n_E n_S}{n_E + n_S}} \frac{(\overline{X}_E - \overline{X}_S) - (\mu_E - \mu_S)}{\sigma} \tag{3.1}$$

has a standard normal distribution. It serves as a basis for the development of the test procedures involving two means.

With $\theta = \mu_E - \mu_S$, the null hypothesis is usually two-sided as follow:

$$H_0 : \theta = 0 \qquad \text{vs.} \qquad H_1 : \theta \neq 0.$$

A point estimate of the unknown common variance σ^2 can be obtained by pooling the sample variances S_S^2 and S_E^2. Denoting the pooled estimator by S_p^2, we write

$$S_p^2 = \frac{(n_S - 1)S_S^2 + (n_E - 1)S_E^2}{n_S + n_E - 2}. \tag{3.2}$$

The test statistic is given by

$$T = \sqrt{\frac{n_E n_S}{n_E + n_S}} \frac{(\overline{X}_E - \overline{X}_S)}{S_p}. \tag{3.3}$$

The statistic T has under the null hypothesis a t-distribution with $n_S + n_E - 2$ degrees of freedom, so that the two-sided hypothesis is rejected when $|t| > t_{1-\alpha/2, n_S + n_E - 2}$. $t_{1-x,m}$ denotes the $(1-x)$ percentile of the central t-distribution with m degrees of freedom.

For simplicity and without loss of generality we consider in the following a balanced design, that is equal sample sizes in both production groups $n_S = n_E = n$ (In the general case, $\exists\ \xi$, such that $n_E = \xi \cdot n_S$).

Approximate computation of the sample size

We suppose the variance σ^2 is known.

For a specific alternative $\tilde{\theta} = \mu_S - \mu_E$, the power of the test is given by

$$1-\beta = P(|\overline{X}_E - \overline{X}_S| > c \text{ when } \theta = \tilde{\theta}). \tag{3.4}$$

Therefore,

$$\beta = P(-c < \overline{X}_E - \overline{X}_S < c \text{ when } \theta = \tilde{\theta})$$

$$= P\left[\sqrt{\frac{n}{2}}\,\frac{-c-\tilde{\theta}}{\sigma} < \sqrt{\frac{n}{2}}\,\frac{\overline{X}_E - \overline{X}_S - \tilde{\theta}}{\sigma} < \sqrt{\frac{n}{2}}\,\frac{c-\tilde{\theta}}{\sigma} \text{ when } \theta = \tilde{\theta}\right].$$

Under the alternative hypothesis $\theta = \tilde{\theta}$

$$Z = \sqrt{\frac{n}{2}}\,\frac{\overline{X}_E - \overline{X}_S - \tilde{\theta}}{\sigma}$$

is standard normal distributed. The critical value c is given by

$$c = \sigma\sqrt{\frac{2}{n}}\; z_{\alpha/2}\,.$$

z_x is the 100x percentile point of the standard normal distribution.

Thus

$$\beta = P\left[-z_{\alpha/2} - \sqrt{\frac{n}{2}}\,\frac{\tilde{\theta}}{\sigma} < Z < z_{\alpha/2} - \sqrt{\frac{n}{2}}\,\frac{\tilde{\theta}}{\sigma}\right].$$

This equation yields

$$-z_\beta \approx z_{\alpha/2} - \sqrt{\frac{n}{2}}\,\frac{\tilde{\theta}}{\sigma}$$

from which we conclude that

$$n \approx \frac{2(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\tilde{\theta}^2}.$$

The approximate sample sizes $n_S = n_E = n$ needed to give a power of $1-\beta$ when $\theta = \tilde{\theta}$ for a two-sided α-level test of $H_0 : \theta = 0$ is given by $[n]+1$ where $[\cdot]$ is the greatest integer function.

For the one-sided test $H_0 : \theta = 0$ against $H_1 : \theta > 0$, the expression for the required sample size when $n_S = n_E = n$ is given by

$$n = \frac{2(z_\alpha + z_\beta)^2 \sigma^2}{\tilde{\theta}^2}. \tag{3.5}$$

Exact computation of the sample size

Recall that the test statistic is given by

$$T = \sqrt{\frac{n}{2}} \frac{(\overline{X}_E - \overline{X}_S)}{S_p}. \tag{3.6}$$

Under the alternative hypothesis H_1 is T noncentrally t-distributed with $2n-2$ degrees of freedom and noncentrality parameter

$$NC = \frac{\tilde{\theta}}{S_p}\sqrt{\frac{n}{2}}. \tag{3.7}$$

The power of the test is given by

$$\begin{aligned} 1-\beta &= P(|T| > t_{1-\alpha/2,\,2(n-1)}) \\ &= P(T < t_{1-\alpha/2,\,2(n-1)}) + 1 - P(T < t_{\alpha/2,\,2(n-1)}). \end{aligned} \tag{3.8}$$

The relationship between the power and the effect size is shown on figure 3.1. The power is an increasing function of the effect size.

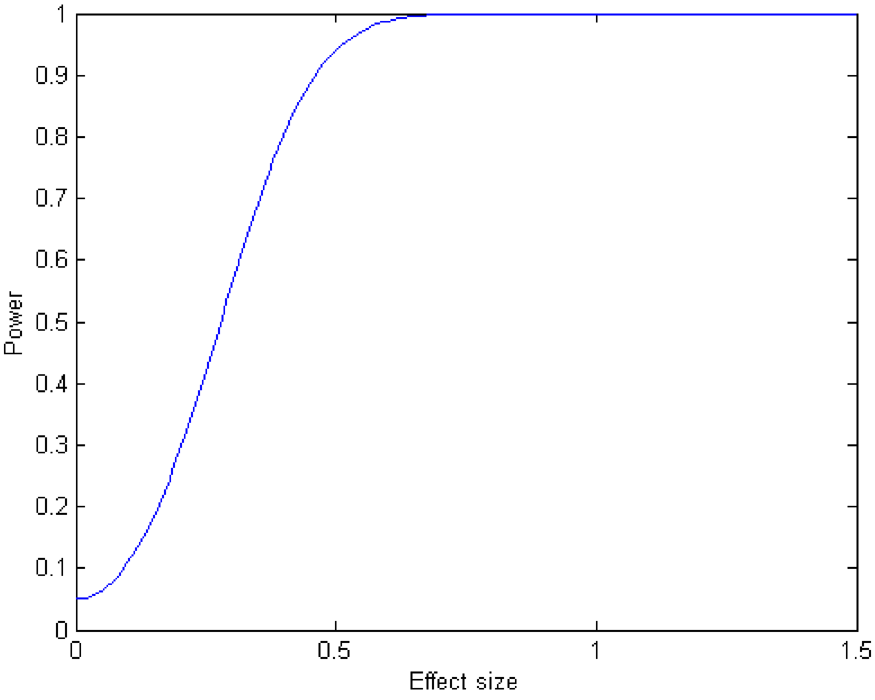


Figure 3.1: *Power of the test by the exact method for $\alpha = 0.05$, $n = 100$ and $S_p = 1$ and different values of the effect size.*

$$\beta = P(T < t_{\alpha/2,\,2(n-1)}) - P(T < t_{1-\alpha/2,\,2(n-1)})$$

$$= F_{t,NC,2(n-1)}(t_{\alpha/2,\,2(n-1)}) - F_{t,NC,2(n-1)}(t_{1-\alpha/2,\,2(n-1)}),$$

with $F_{t,NC,2(n-1)}$ denoting the distribution function of the noncentral t-distribution with $2n-2$ degrees of freedom and noncentrality parameter NC given by

$$F_{t,NC,2(n-1)}(t) = \frac{1}{2^{n-2}\Gamma(n-1)} \int_0^\infty x^{2n-1} \exp(-\frac{x^2}{2}) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{tx}{\sqrt{2(n-1)}}-NC} \exp(-\frac{u^2}{2})\, du\, dx. \qquad (3.9)$$

This equation has to be solved iteratively for n. The sample sizes $n_S = n_E = n$ needed to yield a power of $1-\beta$ when $\theta = \tilde{\theta}$ for a two-sided α-level test of $H_0 : \theta = 0$ is given by $[n]+1$, where n satisfies the equation

$$\beta = F_{t,NC,2(n-1)}(t_{\alpha/2,\,2(n-1)}) - F_{t,NC,2(n-1)}(t_{1-\alpha/2,\,2(n-1)}). \qquad (3.10)$$

Comparison of the two methods

To compare the two computation methods of the sample size, we computed the required sample size per group to give a power of 90% for a two-sided 5%-level test for different values of the standardized effect size $\tilde{\theta}/\sigma$. The computations have been done with the statistical software R. Results are given in table 2.1.

stand. effect size	**n (exact method)**	**n (approximate method)**	**Difference**
0.1	2102	2102	0
0.2	526	526	0
0.3	234	234	0
0.4	132	132	0
0.5	86	85	1
0.6	60	59	1
0.7	44	43	1
0.8	34	33	1
0.9	27	26	1
1	23	22	1
1.5	11	10	1

Table 3.1: *Comparison of the exact and approximate sample size*

The difference in sample sizes obtained using the normal approximation or the exact method is minimal as we can observe in table 2.1. For large sample size, the normal approximation will generally provide good estimates for sample size.

Summary

Considering the formula of the sample size given by the approximate method, recall that one step in the sample size problem requires eliciting an effect size of scientific interest. It is the smallest difference that is thought to provide meaningful improvement. The effect size is generally unknown and difficult to assess at the planning stage of the study and it is up to statisticians to elicit this information from the researchers involved in the study. To detect a small effect size, a large sample size is required and a small decrease in the effect size can result in a drastic change on the required sample size as shown by the figure 3.1 below. This is due to the fact that the sample size has a quadratic relationship with the effect to be detected.

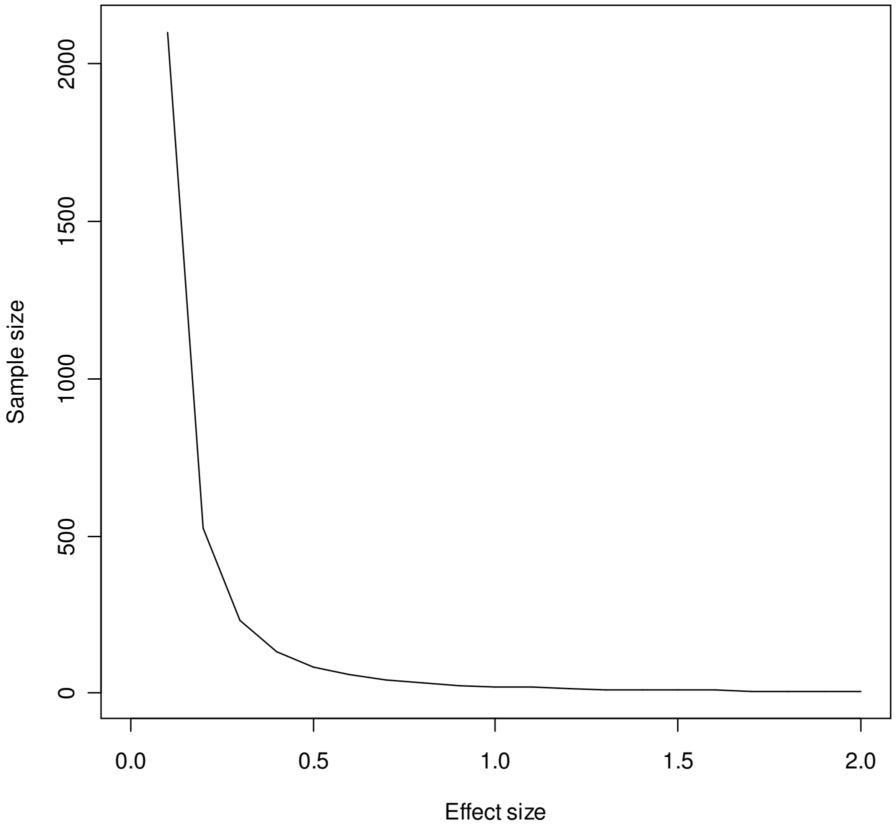


Figure 3.2*: Relationship between the sample size per group and the effect size by* $\alpha = 0.05$, $\beta = 0.1$ *and* $\sigma = 1$.

Recall that another element required for the computation of the sample size is the variability of the target variable. Since the variability of the target variable is typically not known, estimates based upon the past observed data are often used, especially when planning a study that utilizes a similar target or the same design. Another method to elicit the variability is to conduct a pilot study whose primary purpose is to provide a data-based estimate of the variance of the target variable. This will be the subject of chapter 4.

When the variance is large, a large sample size will be needed to obtain a precise estimate of the target. As the variance becomes larger, the sample size increases as shown in figure 3.2.

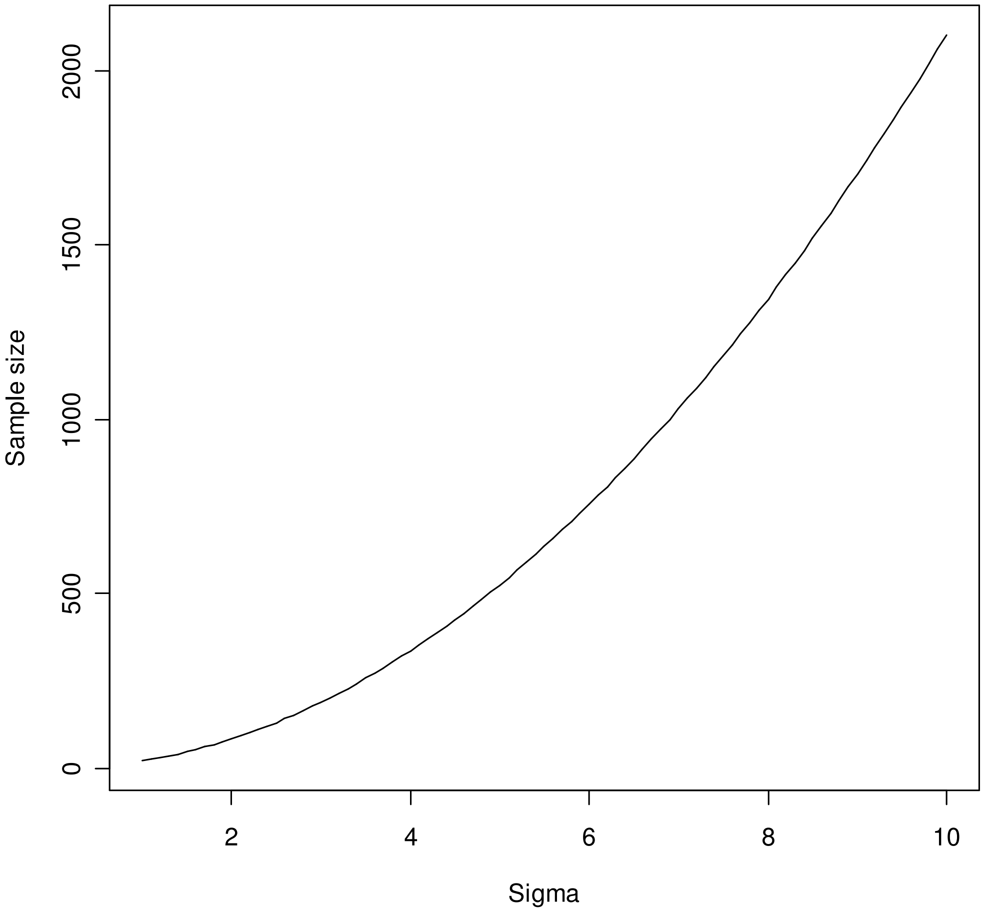


Figure 3.3: *Relationship between sample size per group and* σ *by* $\alpha = 0.05$, $\beta = 0.1$ *and* $\tilde{\theta} = 1$.

3.2 Testing equivalence with normally distributed outcomes

Many industrial experiments aim at showing equivalence between an experimental production group under development and an existing standard production group. The aim of such trials is usually to demonstrate that the groups differ by not more than a defined amount, which means the groups are equivalent. It is appropriate to use when the experimental group is hypothesised to be at least as good as the standard group in the primary variable and the experimental group has some advantages in secondary variables compared to the standard group. This fact necessitates a role reversal in the defining of the hypotheses; the specification of no difference in the alternative and a difference in the null.

Many of the currently employed methods of equivalence testing were developed in the 1970's and 1980's in the field of bio statistic and medicine (Metzler,1974; Westlake,1976,1979; Schuirmann,1981,1987; Anderson and Hauck, 1981,1983).

3.2.1 Formulation of the test and sample size formula

We consider experiments comparing two production groups with independent normally distributed outcomes with expectations μ_E for the experimental group and μ_S for the standard group, and with unknown but common variance σ^2. For simplicity and without loss of generality we consider a balanced design, which means an equal sample size in both production groups with a total of N observations.

Let X_E and X_S designate the normally distributed outcome of interest for the experimental and standard production group respectively. For equivalence testing it is reasonable to assume that the sign of the corresponding population means μ_E and μ_S are both positive.

The equivalence hypotheses are typically two-one-sided as follows:

$$H_0 : |\theta| \geq \delta \text{ against } H_1 : |\theta| < \delta , \quad \delta > 0 ,$$

where δ indicates the maximum difference allowed for an experimental group to be considered equivalent with a standard group, and θ is the difference (additive model) or ratio (multiplicative model) of means for the experimental group and the standard group. δ is sometimes defined as a percentage of the mean for the standard group.

The power is the probability of accepting equivalence when the groups are in fact equivalent, that is, the difference or the ratio of the success measure is within the prespecified boundaries. In the following we briefly describe statistical techniques for sample size involved in cases of additive model and multiplicative model.

Additive model

In this case, $\theta = \mu_E - \mu_S$, and the corresponding test problem is formulated as follow:

- The null hypothesis for the two one-sided test (TOST) is

  $$H_0^a : |\mu_E - \mu_S| \geq \delta_a$$

  or equivalently

  $$H_0^a : \mu_E - \mu_S \geq \delta_a \quad \text{or} \quad \mu_E - \mu_S \leq -\delta_a .$$

- The alternative hypothesis is

  $$H_1^a : |\mu_E - \mu_S| < \delta_a$$

  or equivalently

  $$H_1^a : -\delta_a < \mu_E - \mu_S < \delta_a .$$

Therefore the null hypothesis H_0^a can be tested by simultaneous testing of the following two one-sided hypotheses:

$$H_{01}^a : \mu_E - \mu_S \geq \delta_a \quad \text{versus} \quad H_{11}^a : \mu_E - \mu_S < \delta_a$$

and

$$H_{01}^a : \mu_E - \mu_S \leq -\delta_a \quad \text{versus} \quad H_{11}^a : \mu_E - \mu_S > -\delta_a .$$

The test statistic involved is the usual statistic for testing the difference between two population means with unknown variance. For the first one-sided test, the test statistic is given by

$$T_1^a = \sqrt{\frac{N}{4}} \frac{\overline{X}_E - \overline{X}_S - \delta_a}{S_p},$$

where S_p is the pooled standard deviation of the two samples, $\overline{X}_E$ and $\overline{X}_S$ the sample means of the experimental and standard production group.

The null hypothesis can be rejected at level α if $T_1^a \leq -t_{1-\alpha,N-2}$ where $t_{1-\alpha,N-2}$ is the $(1-\alpha)$ percentile of the central $t-$distribution with $N-2$ degrees of freedom.

The test statistic of the second one-sided test is similar to the first and is given by

$$T_2^a = \sqrt{\frac{N}{4}} \frac{\overline{X}_E - \overline{X}_S + \delta_a}{S_p} .$$

The null hypothesis can be rejected at level α if $T_2^a \geq t_{1-\alpha,N-2}$.

H_0^a is rejected at level α if $T_1^a \leq -t_{1-\alpha,N-2}$ and $T_2^a \geq t_{1-\alpha,N-2}$.

The power of the test is the probability of rejecting that the two means are different by at least δ_a when the means are in fact equivalent:

$$1-\beta = P(\ T_1^a \leq -t_{1-\alpha,N-2} \text{ and } T_2^a \geq t_{1-\alpha,N-2} \mid \theta = \tilde{\theta}_a, -\delta_a < \tilde{\theta}_a < \delta_a\ ) .$$

The method for calculation of the exact sample size can be found in Phillips (1990).

As shown in the case of difference of mean in section 2.1, an approximate sample size per group can be derived.

The sample size per group required for the rejection of H_{01}^a to give a power $1-\beta$ of an α-level test at a specified alternative $\tilde{\theta}_a = \mu_E - \mu_S$ is

$$n_1 = 2\frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{\left(\tilde{\theta}_a - \delta_a\right)^2} .$$

The sample size per group required for the rejection of H_{02}^a to give a power $1-\beta$ of an α-level test at a specified alternative $\tilde{\theta}_a = \mu_E - \mu_S$ is given by

$$n_2 = 2\frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{\left(\tilde{\theta}_a + \delta_a\right)^2} .$$

Thus the sample size per group n required for the rejection of H_0^a is

$$n = \begin{cases} \max(n_1, n_2) \text{ if } \tilde{\theta}_a \neq 0 \\ 2\dfrac{\left(z_\alpha + z_{\beta/2}\right)^2 \sigma^2}{\delta_a^2} \text{ if } \tilde{\theta}_a = 0 \end{cases} .$$

Multiplicative model

In this case, $\theta = \dfrac{\mu_E}{\mu_S}$ and the corresponding test problem is formulated as follow:

$$H_0^m : \left|\frac{\mu_E}{\mu_S}\right| \geq \delta_m \qquad \text{versus} \qquad H_1^m : \left|\frac{\mu_E}{\mu_S}\right| < \delta_m .$$

Another equivalent formulation of the test problem is given by

$$H_0^m : \frac{\mu_E}{\mu_S} \geq \delta_m^1 \qquad \text{or} \qquad \frac{\mu_E}{\mu_S} \leq \delta_m^2 \qquad \text{versus} \qquad H_1^m : \delta_m^2 < \frac{\mu_E}{\mu_S} < \delta_m^1 .$$

This formulation makes sense only if μ_E and μ_S have the same sign. We assumed that μ_E and μ_S are both positive and $\delta_m^1 = 1/\delta_m^2$ for the invariance of the test by permuting μ_E and μ_S.

The two one-sided test problem are

$$H_{01}^m : \frac{\mu_E}{\mu_S} \geq \delta_m^1 \qquad \text{versus} \qquad H_{11}^m : \frac{\mu_E}{\mu_S} < \delta_m^1$$

and

$$H_{02}^m : \frac{\mu_E}{\mu_S} \leq \delta_m^2 \qquad \text{versus} \qquad H_{12}^m : \frac{\mu_E}{\mu_S} > \delta_m^2 .$$

The test statistic involves the size-α likelihood ratio test proposed by Sasabuchi (1988) (Kieser and Hauschke, 1999). H_{01}^m is rejected if $T_m^1 \leq -t_{N-2,1-\alpha}$ and H_{02}^m is rejected if $T_m^2 \geq t_{N-2,1-\alpha}$, where

$$T_m^i = \sqrt{\frac{N}{2\left(1+\left(\delta_m^i\right)^2\right)}} \frac{\overline{X}_E - \delta_m^i \overline{X}_S}{S_p}, \qquad i = 1, 2 .$$

The power of the test is

$$P(T_m^1 \leq -t_{N-2,1-\alpha} \text{ and } T_m^2 \geq t_{N-2,1-\alpha} \mid \delta_m^1 < \frac{\mu_E}{\mu_S} < \delta_m^2).$$

The vector $\left(T_m^1, T_m^2\right)$ has a bivariate noncentral t-distribution with noncentrality parameters NC_1 and NC_2:

$$NC_i = \frac{\mu_E - \delta_m^i \mu_S}{S_p} \sqrt{\frac{n}{\left(1+\left(\delta_m^i\right)^2\right)}} = \sqrt{\frac{n}{1+\left(\delta_m^i\right)^2}} \frac{\theta_m - \delta_m^i}{CV_S} \qquad i = 1, 2$$

and correlation coefficient (Hauschke et al.,1999)

$$\rho = \frac{1 + \delta_m^1 \delta_m^2}{\sqrt{\delta_m^1 \delta_m^2 + \delta_m^1 + \delta_m^2 + 1}} .$$

CV_S denotes the coefficient of variation of the standard group:

$$CV_S = \frac{S_p}{\mu_S}.$$

Approximate sample size

We considered the first one-sided test

$$H_{01}^m : \frac{\mu_E}{\mu_S} \geq \delta_m^1 \qquad \text{versus} \qquad H_{11}^m : \frac{\mu_E}{\mu_S} < \delta_m^1.$$

For a specified alternative

$$\theta_m = \frac{\mu_E}{\mu_S} < \delta_m^1,$$

T_m^1 follows the noncentral t-distribution with $N-2$ degrees of freedom and noncentrality parameter

$$\sqrt{N/2}\,\vartheta = \sqrt{N/2}\sqrt{\frac{1}{1+\left(\delta_m^1\right)^2}}\,\frac{\theta_m - \delta_m^1}{CV_S}.$$

For the computation of the sample size, we need the expression of the power

$$1-\beta = P(T_m^1 < -t_{N-2,1-\alpha} \mid \theta_m, CV_S).$$

Hence

$$-t_{N-2,1-\alpha} = t_{N-2,\beta}(\sqrt{N/2}\,\vartheta),$$

with $t_{N-2,\beta}(\sqrt{N/2}\,\vartheta)$ denoting the $(1-\beta)$ percentile of the noncentral t-distribution with $N-2$ degrees of freedom and noncentrality parameter $\sqrt{N/2}\,\vartheta$.

We can express the percentiles of the noncentral t-distribution by those of the central t-distribution by using the approximation

$$-t_{N-2,\beta}(\sqrt{N/2}\,\vartheta) = -t_{N-2,\beta} + \sqrt{N/2}\,\vartheta$$

in other to facilitate the determination of the sample size. Hence the required sample size per group can be approximated by the smallest integer for which the following inequality is fulfilled

$$n_m^1 \geq \left(1+\left(\delta_m^1\right)^2\right)\frac{(t_{2n_m^1-2,\alpha} + t_{2n_m^1-2,\beta})^2\, CV_S^2}{(\theta_m - \delta_m^1)^2}.$$

This inequality can be solved iteratively for n_m^1.

An approximate formula to determine the sample size per group required to give a power $1-\beta$ of the α-level test H_{01}^{m} at the given alternative $\theta_m < \delta_m^1$ is

$$n_m^1 = \left(1+\left(\delta_m^1\right)^2\right)\frac{(z_{1-\alpha}+z_{1-\beta})^2 CV_S^2}{(\theta_m-\delta_m^1)^2}.$$

By analogy, the approximated formula to determine the sample size per group required to produce a power $1-\beta$ of the α-level test H_{02}^{m} at the given alternative $\theta_m > \delta_m^2$ is

$$n_m^2 = \left(1+\left(\delta_m^2\right)^2\right)\frac{(z_{1-\alpha}+z_{1-\beta})^2 CV_S^2}{(\theta_m-\delta_m^2)^2}.$$

Therefore, an approximate formula to determine the sample size per group required to give a power $1-\beta$ of the α-level test H_0^m at the given alternative $\delta_m^2 < \theta_m < \delta_m^1$ is given by (Kieser and Hauschke, 1999)

$$n_m = \begin{cases} n_m^1 \text{ when } 1 < \theta_m < \delta_m^1 \\ n_m^2 \text{ when } \delta_m^1 < \theta_m < 1 \\ \left(1+\left(\delta_m^2\right)^2\right)\dfrac{(z_{1-\alpha}+z_{1-\beta/2})^2 CV_S^2}{(1-\delta_m^2)^2} \text{ when } \theta_m = 1 \end{cases}.$$

3.2.2 Combinations of significance and equivalence test

If one performs a standard significance test and equivalence test on the same data set, then there are four possibilities by decision making about the rejection of the null hypothesis (Rogers et al., 1993). The four possibilities are given in table 3.2.

Equivalence Test	**Significance Test**	**Situation**
Reject	Reject	Equivalent and different
Reject	Fail to reject	Equivalent
Fail to reject	Reject	Different
Fail to reject	Fail to reject	Equivocal

Table 3.2: *Possible situations in simultaneous testing of significance and equivalence*

In the following we considered the additive formulation of the equivalence test and we designed a small simulation study to access the power of simultaneously concluding that two means are both statistically different and equivalent.

For the simulation we took $\sigma = \sqrt{2}$, five different sample sizes per group ($n = 10$, 50, 100, 200, 500), four different values for the effect size ($\theta = 0$, 0.1, 0.25, 0.4) and three values of the equivalence bound ($\delta_a = 0.1$, 0.25, 0.5).

Hence we have a fully randomised design with $5 \times 4 \times 3 = 60$ cases. For each case, 100000 simulations were run in the statistical software R and the proportion of rejections of the null hypothesis of the significance test, equivalence test and simultaneous rejection of both tests are given in the followings tables (tables 3.4 through 3.7).

Equivalence Bound (δ_a)	**Sample size per Group (n)**	**Equivalent**	**Different**	**Equivalent & Different**
0.1	10	0	0,04900	0
0.1	50	0	0,05096	0
0.1	100	0	0,04883	0
0.1	200	0	0,04894	0
0.1	500	0	0,05031	0
0.25	10	0	0,049	0
0.25	50	0	0,05096	0
0.25	100	0	0,04893	0
0.25	200	0,09585	0,04894	0
0.25	500	0,74760	0,05031	0
0.5	10	0,00001	0,04900	0
0.5	50	0,09719	0,05096	0
0.5	100	0,60568	0,04883	0
0.5	200	0,94172	0,04894	0,00298
0.5	500	0,99991	0,05031	0,05054

Table 3.3: *Proportion of rejections of the null hypothesis (100000 simulations) of the statistical significance and equivalence test, $\alpha = 0.05$, $\tilde{\theta} = 0$.*

Equivalence Bound (δ_a)	**Sample size per Group (n)**	**Equivalent**	**Different**	**Equivalent & Different**
0.1	10	0	0,05155	0
0.1	50	0	0,06419	0
0.1	100	0	0,07931	0
0.1	200	0	0,10782	0
0.1	500	0	0,20020	0
0.25	10	0	0,05155	0
0.25	50	0	0,06419	0
0.25	100	0	0,07931	0
0.25	200	0,07485	0,10782	0
0.25	500	0,50204	0,20020	0
0.5	10	0	0,05155	0
0.5	50	0,09196	0,06419	0
0.5	100	0,54865	0,07931	0
0.5	200	0,87714	0,10782	0,00431
0.5	500	0,99755	0,20020	0,19775

Table 3.4: *Proportion of rejections of the null hypothesis (100000 simulations) of the statistical significance and equivalence test, $\alpha = 0.05$, $\theta = 0.1$*

Equivalence Bound (δ_a)	**Sample size per Group (n)**	**Equivalent**	**Different**	**Equivalent & Different**
0.1	10	0	0,06527	0
0.1	50	0	0,14125	0
0.1	100	0	0,23694	0
0.1	200	0	0,41907	0
0.1	500	0	0,79806	0
0.25	10	0	0,06527	0
0.25	50	0	0,14125	0
0.25	100	0	0,23694	0
0.25	200	0,02032	0,41907	0

0.25	500	0,04956	0,79806	0
0.5	10	1	0,06527	0
0.5	50	0,06670	0,14125	0
0.5	100	0,32551	0,23694	0
0.5	200	0,55016	0,41907	0,00830
0.5	500	0,87342	0,79806	0,67148

Table 3.5: *Proportion of rejections of the null hypothesis (100000 simulations) of the statistical significance and equivalence test,* $\alpha = 0.05$, $\theta = 0.25$

Equivalence Bound (δ_a **)**	**Sample size per Group (** n **)**	**Equivalent**	**Different**	**Equivalent & Different**
0.1	10	0	0,09044	0
0.1	50	0	0,28804	0
0.1	100	0	0,51375	0
0.1	200	0	0,80473	0
0.1	500	0	0,99412	0
0.25	10	0	0,09044	0
0.25	50	0	0,28804	0
0.25	100	0	0,51375	0
0.25	200	0,00201	0,80473	0
0.25	500	0,00044	0,99412	0
0.5	10	0	0,09044	0
0.5	50	0,03532	0,28804	0
0.5	100	0,12153	0,51375	0
0.5	200	0,17528	0,80473	0,00618
0.5	500	0,30087	0,99412	0,29499

Table 3.6: *Proportion of rejections of the null hypothesis (100000 simulations) of the statistical significance and equivalence test,* $\alpha = 0.05$, $\theta = 0.4$

The results of the simulation study show that

- for $\theta < \delta_a$, the proportion of rejection of the null hypothesis of the equivalence test and the significance test approach unity as n increase. The convergence of the equivalence test is slow when $\tilde{\theta}$ is nearly equal to δ_a.
- for $\theta > \delta_a$, the proportion of the null hypothesis of the significance test approaches unity and the proportion of rejections of the equivalence test tends to zero as n increases.
- The proportion of simultaneous rejection of both tests was only for large values of n and δ_a different for zero. This is probably due to the fact that we assume homoscedasticity and normality of the data of the two groups being compared.

3.3 Sample size methods for reliability studies

In reliability studies the Normal distribution is not of major importance. The most common distributions in reliability analysis are the Exponential, Weibull, and Log-Normal distributions. In this section we present sample size methods to compare the failure time distribution between two independent groups (products, processes, machines). The Exponential and the Weibull distribution have been chosen to model the failure time.

3.3.1 Basic concepts

Due to rapid advances in technology, achievement and improvement of customer satisfaction, development of highly sophisticated products, the statistical analysis of reliability data has become a topic of considerable interest to statisticians and engineers. In today's technological world nearly everyone depends upon the continued functioning of a wide array of complex machinery and equipment for their everyday health, safety, mobility and economic welfare. We expect our cars, computers, electrical appliances, lights, televisions, etc. to function whenever we need them - day after day, year after year. When they fail the results can be a desaster: injury, loss of life and/or costly lawsuits can occur. More often, repeated failure leads to annoyance, inconvenience and a lasting customer dissatisfaction that can play havoc with the responsible company's marketplace position.

It takes a long time for a company to build up a reputation for reliability, and only a short time to be branded as "unreliable" after shipping a flawed product. Continual assessment of new product reliability and ongoing control of the reliability of everything shipped are critical necessities in today's competitive business arena.

The main purposes for collecting reliability data are as follows:

- Early detection of bad design, poor production processes or materials, defective parts, etc.;
- Comparison of two or more production processes or competing products;
- Prediction of product reliability, future claims, product warranty costs;
- Observation of the target of a new product development ;
- Providing needed inputs for system-failure risk assessment

Reliability data are typically censored, that means the exact failure times are not known. When reliability data are analysed, some units are unfailed, and their failure times are known only to be beyond their present running times. Such data are said to be **right censored** or **censored on the right.** Similarly a data is said to be **left censored** or **censored on the left** if the failure time is known only to be before a certain time. Removing unfailed units from an experiment at a prespecified time is known as **time censoring** or **type I censoring**. An experiment that is terminated after a specified number of failures is knows as **type II censoring**.

3.3.2 Reliability and related functions

The definitions in this section are taken from Meeker and Escobar (1998). In reliability analysis, each individual subject is followed up to some time, at which time either a failure is observed to occur or follow-up is curtailed without observation of a failure. When the time T of a failure is observed to an instant of time, the failure times have a right continuous distribution function $F(t) = P(T \leq t)$, t>0. The complement of the cumulative distribution function is the right continuous **survival distribution**

$$S(t) = 1 - F(t) = P(T > t)$$

and gives the probability of surviving beyond time t. For a continuous distribution the hazard function that describes the instantaneous probability of the failure among those still at risk, or those still free of the failure, is defined as

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{P(t < T \leq \Delta t + t \mid T > t)}{P(T > t)} = \frac{f(t)}{S(t)},$$

where $f(t)$ is the probability density function defined as

$$f(t) = \frac{dF(t)}{dt}.$$

The quantile t_p is the inverse of the cumulative distribution function and it is the time at which a specified proportion p of the population fails. For $0 < p < 1$, the p quantile of $F(t)$ is generally defined as the smallest time t such that

$$P(T \leq t) = F(t) \geq p .$$

Another important function for the analysis of reliability data is the **Likelihood Function.** The form of the likelihood function depends on the assumed probability model, the form of the available data and the question of the study. Assuming n independent observations, the likelihood can be written as the joint probability of the data and it is given by

$$L(p) = L(p;\mathrm{DATA}) = C\prod_{i=1}^{n} L_i(p;\mathrm{data}_i) = \prod_{i=1}^{n} [f(t_i;p)]^{\delta_i}[1 - F(t_i;p)]^{1-\delta_i} .$$

p is the vector of parameter to be estimated, $L_i(p;\mathrm{data}_i)$is the probability of the observation i data_i is the data for observation i, C a constant, $f(t_i;p)$ and $F(t_i;p)$ are the probability distribution function and the cumulative distribution function, respectively, of the specified distribution,

$$\delta_i = \begin{cases} 1 & \text{if } t_i \text{ is an exact failure} \\ 0 & \text{if } t_i \text{ is a right - censored observation} \end{cases} .$$

Another expression of the total likelihood is

$$L(p, \mathrm{DATA}) = C\prod_{i=1}^{m+1} [F(t_i)]^{l_i}[F(t_i) - F(t_{i-1})]^{d_i}[1 - F(t_i)]^{r_i} ,$$

where

$$n = \sum_{i=1}^{m+1} (l_i + d_i + r_i),$$

which d_i observations interval censored in t_{i-1} and t_i, l_i observations left-censored at t_i and r_i observations right-censored at t_i. The maximum likelihood estimate of $F(t)$ is obtained by providing a p that maximise $L(p)$.

For reliability applications, quantiles, failure probabilities and the hazard function are of higher interest than distribution moments. Many of the used statistics are either **location-scale distributions** or closely related. A random variable X follows a location-scale distribution if its cumulative distribution function can be expressed as (Meeker and Escobar, 1998)

$$F(x;\mu,\sigma) = \Phi\left(\frac{x-\mu}{\sigma}\right), \qquad \mu \in IR \text{ and } \qquad \sigma > 0 ,$$

where Φ does not depend on any unknown parameters. μ is a location parameter and σ a scale parameter. In the following some location-scale distributions are introduced.

Exponential Distribution

The exponential distribution occupies an important position in the analysis of reliability data. The two-parameter exponential distribution will be written with its cumulative distribution function in the form

$$F(t;\lambda,\gamma)=1-\exp\left(-\frac{t-\gamma}{\lambda}\right),\ t>\gamma\,,$$

in which case $\lambda>0$ is a scale parameter and γ is both a location and a threshold parameter. The density distribution function is given by

$$f(t;\lambda,\gamma)=\frac{1}{\lambda}\exp\left(-\frac{t-\gamma}{\lambda}\right)$$

and the hazard function is given by

$$h(t;\lambda,\gamma)=\frac{f(t;\lambda,\gamma)}{1-F(t;\lambda,\gamma)}=\frac{1}{\lambda}\,.$$

The p quantile is given by

$$t_p=\gamma-\lambda\log(1-p)\,.$$

The mean and variance of T which follows an exponential distribution are

$$E(T)=\gamma+\lambda \quad \text{and} \qquad Var(T)=\lambda^2\,.$$

If we set $\gamma=0$, we obtain the one-parameter exponential distribution, which is the simplest distribution commonly used in reliability analysis.

The hazard function is constant over the time; this means that for a unit still at risk, the probability of failing in the next small interval is independent of the age of the unit. This characteristic makes that the exponential distribution will not be appropriate for modelling population items subject to some combination of fatigue and corrosion like electronic components.

The likelihood function is given by

$$L(p)=L(\lambda,t)=\frac{1}{\lambda^n}\exp\left(-\sum_{i=1}^{n}\frac{t_i}{\lambda}\right)$$

and the maximum likelihood estimator of λ is

$$\hat{\lambda}=\bar{t}_n=\frac{1}{n}\sum_{i=1}^{n}t_i\,.$$

Weibull distribution

The two-parameter Weibull cumulative distribution function is given by

$$F(t;\eta,\beta)=1-\exp\left(-\left(\frac{t}{\eta}\right)^{\beta}\right), \qquad t>0,$$

where $\beta>0$ is a shape parameter and $\eta>0$ is a scale parameter.

The density distribution function is

$$f(t;\eta,\beta)=\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}\exp\left(-\left(\frac{t}{\eta}\right)^{\beta}\right)$$

and the hazard function is

$$h(t;\eta,\beta)=\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}, \qquad t>0.$$

The mean and the variance are given respectively by

$$\eta\Gamma\left(1+\frac{1}{\beta}\right) \quad \text{and} \quad \eta^2\left(\Gamma\left(1+\frac{2}{\beta}\right)-\Gamma^2\left(1+\frac{1}{\beta}\right)\right),$$

where

$$\Gamma(u)=\int_0^\infty x^{u-1}\exp(-x)\,dx$$

is the gamma function.

The p quantile of the Weibull distribution is given by

$$t_p=\eta[-\log(1-p)]^{1/\beta}.$$

The Weibull distribution can be used to model failure-time data with decreasing or increasing hazard function. That is why it can provide reliable models in many empirical studies

The likelihood function is

$$L(p)=L(\eta,\beta,t)=\frac{\beta^n}{\eta^{n\beta}}\left(\prod_{i=1}^{n}t_i\right)^{\beta-1}\exp\left(-\sum_{i=1}^{n}\left(\frac{t_i}{\eta}\right)^{\beta}\right),$$

the maximum likelihood estimates of the parameters are given by

$$\hat{\eta}=\left(\frac{1}{n}\sum_{i=1}^{n}t_i^{\hat{\beta}}\right)^{1/\hat{\beta}}$$

and

$$\hat{\beta} = \left[\frac{\sum_{i=1}^{n} t_i^{\hat{\beta}} \ln t_i}{\sum_{i=1}^{n} t_i^{\hat{\beta}}} - \frac{1}{n} \sum_{i=1}^{n} \ln t_i \right]^{-1} .$$

The second equation formula can be solved iteratively and one can show that this equation has a unique positive solution. (Meeker and Escobar, 1998)

3.3.3 Comparing survival distributions

Comparison of survival distribution is particularly important in industrial experiments where groups of items have been randomised to different processes and we are required to make inferences about the effect of the processes on reliability. In principle, because survival times are not normally distributed, nonparametric tests that are based on the rank ordering of survival times should be applied. A wide range of nonparametric tests can be used in order to compare survival times. We denote the number of survival distributions to be compared by *K* and the corresponding survivor functions by $S_1(t), S_2(t), \cdots, S_K(t)$. The null hypothesis is

$$H_0 : S_1(t) = S_2(t) = \cdots = S_K(t), \qquad \forall t .$$

The following five different (mostly nonparametric) tests for censored data are available: Gehan's generalised Wilcoxon test, the Cox-Mantel test, the Cox's *F* test , the log-rank test, and Peto and Peto's generalised Wilcoxon test. In the following, the log-rank test will be presented.

The log-rank test

In survival analysis, a log-rank test compares the equality of *K* survival functions by creating a sequence of *K*x2 contingency tables (*K* survival functions by failure observed/failure not observed at that time) one at each (uncensored) observed event time, and calculating a statistic based on the observed and expected values for these contingency tables. This test is also known as the Mantel-Cox (Mantel-Haenszel) test.

Suppose that *K*=2. Suppose $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$ are *m* distinct failure times across two groups and that at time $t_{(j)}$, d_{1j} units in Group I failed and d_{2j} units in Group II failed, for

$j=1,2,\cdots,m$. Also, suppose n_{1j} units in Group I are still alive just before time $t_{(j)}$ while n_{2j} units in Group II are still alive just before time $t_{(j)}$. Let $n_j = n_{1j} + n_{2j}$ and $d_j = d_{1j} + d_{2j}$. Then, we have the following table,

Group	Number of failures at $t_{(j)}$	Number surviving beyond $t_{(j)}$	Total at risk at $t_{(j)}$
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	$n_j = n_{1j} + n_{2j}$

Table 3.7: *Contingency table for two groups*

Under the null hypothesis of no difference between the groups, all subsets of size d_j of the n_j units at risk at $t_{(j)}$ are equally likely to comprise the set of failures at this time. Thus, conditionally on $d_j = d_{1j} + d_{2j}$, under the null hypothesis d_{1j} has a hypergeometric distribution. Under the null hypothesis, the mean is given by

$$E(d_{1j}) = \frac{n_{1j} d_j}{n_j}$$

and the variance by

$$Var(d_{1j}) = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

The log-rank test is

$$L = \frac{\left(\sum_{j=1}^{m} \left(d_{1j} - \frac{n_{1j} d_j}{n_j} \right) \right)^2}{\sum_{j=1}^{m} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}}.$$

Denoting

$$E = \sum_{j=1}^{m} E(d_{1j})$$

the expected number - under the null hypothesis - if no difference between the groups and

$$O = \sum_{j=1}^{m} d_{1j}$$

the observed number of failures in group I,

L can be expressed as

$$L = \frac{(O-E)^2}{Var(O-E)}.$$

When the number of failure is not too small,

$$\frac{\sum_{j=1}^{m}\left(d_{1j} - E(d_{1j})\right)}{\sqrt{\sum_{j=1}^{m} Var(d_{1,j})}} = \frac{\sum_{j=1}^{m}\left(d_{1j} - \frac{n_{1j}d_j}{n_j}\right)}{\sqrt{\sum_{j=1}^{m}\frac{n_{1j}n_{2j}d_j(n_j-d_j)}{n_j{}^2(n_j-1)}}}$$

is under the null hypothesis approximately standard normal distributed. Therefore, L has approximately a χ_1^2 distribution (chi-square distribution with one degree of freedom).

If the number of groups being compared is bigger than 2 (*K*>2), the log-rank test is based on the corresponding *K*x2 contingency table at each $t_{(j)}$.

Group	Number of failures at $t_{(j)}$	Number surviving beyond $t_{(j)}$	Total at risk at $t_{(j)}$
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
...	...	...	...
K	d_{Kj}	$n_{Kj} - d_{Kj}$	n_{Kj}
Total	d_j	$n_j - d_j$	$n_j = n_{1j} + n_{2j} + \cdots + n_{Kj}$

Table 3.8: *Contingency table for K groups*

For each $t_{(j)}$, the distribution of $D_j = (d_{1j}, d_{2j}, \cdots, d_{K-1j})^T$ under the null hypothesis is multivariate hypergeometric.

The log-rank statistic is expressed as

$$L = (\vec{O} - E)^T V^{-1} (\vec{O} - E)$$

where $\vec{O} = \sum_{j=1}^{m} D_j$, $E = \sum_{j=1}^{m} E_j$, $V = \sum_{j=1}^{m} V_j$.

E_j and V_j denote the mean and variance-covariance matrix of the multivariate hypergeometric distribution at each $t_{(j)}$.

If we denote for $i = 1,2,\cdots,K$ and $j = 1,2,\cdots,m$,

n_{ij} the total of units at risk in *i*th group at *j*th ordered failure time,

d_{ij} the observed total of failures in *i*th group at *j*th ordered failure time,

E_{ij} the expected total of failure in *i*th group at *j*th ordered failure time,

$$E_{ij} = \frac{n_{ij}}{n_j} d_j .$$

$$O_i - E_i = \sum_{j=1}^{m} \left(d_{ij} - E_{ij}\right), \qquad Var\left(O_i - E_i\right) = \sum_{j=1}^{m} \frac{n_{ij}\left(n_j - n_{ij}\right)d_j\left(n_j - d_j\right)}{n_j^{\,2}\left(n_j - 1\right)},$$

$$Cov\left(O_i - E_i, O_l - E_l\right) = \sum_{j=1}^{m} \frac{-n_{ij}n_{lj}d_j\left(n_j - d_j\right)}{n_j^{\,2}\left(n_j - 1\right)} .$$

$$\left(\vec{O} - E\right) = \left(O_1 - E_1, \cdots, O_{K-1} - E_{K-1}\right)^T .$$

$$V = \left(V_{il}\right) \qquad i = 1,2,\cdots,K-1 \text{ and } l = 1,2,\cdots,K-1,$$

with

$$V_{il} = \begin{cases} Var\left(O_i - E_i\right) & \text{if } \quad i = l \\ Cov\left(O_i - E_i, O_l - E_l\right) & \text{if } \ i \neq l \end{cases} .$$

The log-rank statistic is given by

$$L = \left(\vec{O} - E\right)^T V^{-1} \left(\vec{O} - E\right).$$

Under the null hypothesis L is chi-square distributed with $K-1$ degrees of freedom ($L \sim \chi^2_{K-1}$).

3.3.4 Sample size determination procedures

Sample size determination for reliability data is particularly complex. The problem of finding the sample size in trials with reliability data reduces to that of determining the length of the experiment required to obtain a desired power for an α -level test, where the survival distributions, accrual rate, follow-up period are specified.

In many instances we wish to compare the failure-time distributions between two independent groups of subjects or machines or processes, which is done here by comparing the hazard rates of the two groups (h_S for the standard group and h_E for the experimental group).

The null hypothesis to be tested is

$$H_0 : \theta = 1,$$

against the alternative hypothesis

$$H_1 : \theta \neq 1,$$

with

$$\theta = \frac{h_E}{h_S}.$$

The computation of sample sizes itself consists of two parts. First, the number of events (failures, d) required to detect a certain effect is determined. In the second step the necessary sample size is calculated depending on the probability of a failure during the study.

We assume uniform accrual during the accrual period of A years and the follow-up period is of τ additional years. The probability of a failure is given by

$$P(Failure) = \int_0^A P(failure\ and\ entry\ at\ t)\,dt$$

$$= \int_0^A P(failure \mid entry\ at\ t)P(entry\ at\ t)\,dt$$

$$= 1 - \frac{1}{A}\int_0^A P(no\ failure \mid entry\ at\ t)\,dt$$

$$= 1 - \frac{1}{A}\int_0^A S(A + \tau - t)\,dt$$

$$= 1 - \frac{1}{A}\int_{\tau}^{\tau+A} S(u)du\,.$$

One-parameter exponential survival

The exponential distribution is one of the most common life distribution used in reliability testing. It is the only distribution with a constant failure rate function. The exponential distribution represents a good model for testing units/products that have no early failures (units have passed the burn-in period) and no significant wear-out mechanism. The following discussion is based on the work of George and Desu (1974). The design problem is solved by making the following assumptions:

- The survival time distributions for the standard and experimental group are exponential with hazard rates h_S and h_E.
- Unit accrual follows a Poisson process with rate a (>0). Therefore the number of units entering the experiment will be distributed as a Poisson variable with mean $a \cdot A$.
- The number of failures in both groups is equal; $d_S = d_E = d$.

With the assumption of exponential survival, $S(t) = \exp(-ht)$ the probability of a failure is given by

$$P(failure) = 1 - \frac{1}{A}\int_{\tau}^{A+\tau} \exp(-hu)du = 1 - \frac{1}{hA}\exp(-h\tau)(1-\exp(-hA)).$$

The test statistic for the test problem $H_0 : \theta = 1$ against the alternative hypothesis $H_1 : \theta > 1$ is the ratio of the maximum likelihood estimators of the rate parameters

$$\hat{\theta} = \frac{\hat{h}_E}{\hat{h}_S},$$

with $\hat{h}_i = 1/\bar{t}_i$, where $\bar{t}_i$ is the mean survival time in group $i = E, S$.

The test consists of rejecting H_0 if $\hat{\theta} > c$. One require c such that for the type I error α,

$$P(\hat{\theta} > c \mid \theta = 1) = \alpha$$

and for the power of the test $1-\beta$ at the alternative $\tilde{\theta} > 1$,

$$P(\hat{\theta} > c \mid \theta = \tilde{\theta}) = 1 - \beta.$$

Exact method

It is known that $\hat{\theta}$ has an F distribution with $2d$ and $2d$ degrees of freedom (Mace, 1964). Therefore, the critical value c of the test statistic is given by

$$c = F_{1-\alpha}(2d, 2d)$$

and the power requirement imposes the condition

$$P(\hat{\theta} > F_{1-\alpha}(2d, 2d) \mid \theta = \tilde{\theta}) = 1 - \beta.$$

Thus, the required number of failures d in each group needed to give a power of $1-\beta$ at the alternative $\tilde{\theta} > 1$ for an α-level test of $H_0 : \theta = 1$ must satisfy the relation

$$F_{1-\alpha}(2d, 2d) = \tilde{\theta} F_{\beta}(2d, 2d)$$

which needs to be solved iteratively.

Approximate method

$W = \log\hat{\theta}$ is approximately normally distributed with mean $\log\theta$ and variance $2/d$.

For the type I error, we have

$$P(W > c \mid \theta = 1) \le \alpha \;\Rightarrow\; c = z_{1-\alpha}\sqrt{2/d}\;.$$

The power at the alternative $\tilde{\theta} > 1$ is

$$P(W > c \mid \theta = \tilde{\theta}) = 1 - \beta \;\Rightarrow\; c = \log\tilde{\theta} - z_{1-\beta}\sqrt{2/d}\;.$$

z_x is the 100*x* percentile point of the standard normal distribution.

Equality of the two c yield the required number of failures d in each group

$$d = \frac{2(z_{1-\alpha} + z_{1-\beta})^2}{(\log\tilde{\theta})^2}.$$

The following table shows the difference between the exact and the approximate method.

$\tilde{\theta}$	d **(exact method)**	d **(approximate method)**	**Difference**
1.1	1886	1886	0
1.2	516	516	0
1.3	250	249	1
1.4	152	152	0
1.5	105	105	0
1.6	78	78	0
1.7	62	61	1
1.8	51	50	1
1.9	43	42	1
2	37	36	1
2.5	21	21	0

Table 3.9: *Difference between the number of failures required to detect a significance difference using the exact and the approximate method,* $\alpha = 0.05$ *and* $\beta = 0.10$

The difference between the exact and the approximate method (as shown in table 3.8) is minimal. The approximate method has the advantage that it is easy to compute, although with modern computing techniques this advantage is minimal.

The required duration of the study is computed on the basis of the expected total number of failure, given a total sample size of $n = a \cdot A$. It is the smallest length of accrual and follow-up such that

$$a \cdot A \cdot \left(\frac{P_S}{h_S} + \frac{P_E}{h_E} \right) = 2d \left(\frac{1}{h_S} + \frac{1}{h_E} \right),$$

with

$$P_i = 1 - \frac{1}{h_i A} \exp(-h_i \tau)(1 - \exp(-h_i A)), \qquad i = E, S.$$

These minima are obtained for the follow-up $\tau = 0$ and the accrual length A satisfying

$$\left(A - \frac{2d}{a} \right)\left(\frac{1}{h_S} + \frac{1}{h_E} \right) = \frac{1 - \exp(-h_S A)}{h_S^2} + \frac{1 - \exp(-h_E A)}{h_E^2}.$$

The necessary total sample size can be calculated by

$$n = 2 \cdot d \cdot \frac{\left(\frac{1}{h_S} + \frac{1}{h_E} \right)}{\left(\frac{P_S}{h_S} + \frac{P_E}{h_E} \right)} \quad = \quad 4 \cdot \frac{(z_{1-\alpha} + z_{1-\beta})^2}{(\log \tilde{\theta})^2} \cdot \frac{\left(\frac{1}{h_S} + \frac{1}{h_E} \right)}{\left(\frac{P_S}{h_S} + \frac{P_E}{h_E} \right)}.$$

Schoenfeld and Richter (1982) developed an approximate formula for power and sample size of their parametric test statistic. Considering the median survival of the specimens of the two groups m_E and m_S, $R = m_E / m_S$ can express the difference between the two groups. The hypotheses in their development are

$$H_0 : \frac{m_E}{m_S} = 1 \quad \text{and} \qquad H_1 : \frac{m_E}{m_S} = R.$$

Assuming an exponential survival distribution for each group with parameters λ_E und λ_S respectively, let t_i be the total time on study and d_i be the number of failures on group i, $i = E, S$. The asymptotic distribution of the test statistic

$$T = \frac{\ln\left(\frac{t_S d_E}{t_E d_S} \right)}{\sqrt{\frac{1}{d_E} + \frac{1}{d_S}}}$$

under the alternative hypothesis is used to obtain, with n specimens on each group, the following expression for power

$$1-\beta=\Phi\left(\frac{\sqrt{n}\ln(R)}{\sqrt{\frac{1}{p_E}+\frac{1}{p_S}}}-z_{1-\alpha}\right),$$

where Φ is the standard normal distribution function, α the type I error rate.
The sample size in each group is computed according to

$$n=\frac{\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{(\ln(R))^2}\left(\frac{1}{p_E}+\frac{1}{p_S}\right).$$

p_i is the probability that a specimen in group i fail in the course of the study and is given by

$$p_i=1-P_i(A)S_i(\tau),\quad i=E,S\ ,$$

where the survival function of the exponential in group i, S_i and is given by

$$S_i(\tau)=\exp\left(-\frac{\tau\ln(2)}{m_i}\right).$$

The probability that a specimen not failed to the end of the accrual period after entry, P_i, is given by

$$P_i=\frac{1}{A}\int_0^A S_i(A-v)dv=\frac{1-\exp\left(\frac{-A\ln 2}{m_i}\right)}{\frac{A\ln 2}{m_i}}.$$

For alternative formulations of this problem, we refer to Schoenfeld (1983), Lachin & Foulkes (1986).

Weibull survival

Although an exponential distribution may provide an acceptable approximation to the distribution of survival times over relatively short intervals, the adequacy of the characterization of the distribution on more substantial proportion is no guarantee because the hazard function is constant over time. Therefore, it may be more flexible to assume the Weibull than the exponential distribution. Thus an extension of the sample size calculations assuming Weibull survival can give a better appreciation about the study at the design stage. Based on the development of Schoenfeld and Richter (1982) for the exponential case, Moonseong et al. (1998) suggested an extension of the sample size formula for the case of the Weibull distribution. Assuming the same shape parameter κ belongs to the two groups, the corresponding survival function can be expressed depending on the corresponding median m_i as

$$S_i(t) = \exp\left(-\ln 2 \left(\frac{t}{m_i} \right)^{\kappa} \right),$$

and the corresponding hazard function is given by

$$h_i(t) = \frac{\kappa \ln 2}{m_i} \left(\frac{t}{m_i} \right)^{\kappa - 1}.$$

The shape parameter κ indicates the degree of acceleration of hazard over time.

- If $\kappa > 1$, the hazard accelerates.
- If $\kappa < 1$, the hazard decelerates
- If $\kappa = 1$, the hazard is constant, which is the case of the exponential distribution. Thus the exponential distribution is a special case of the Weibull distribution.

Another important and well known relationship between the two distributions is the following: If the survival time T is Weibull distributed, T^{κ} is exponentially distributed. Therefore, because of the test problem

$$H_0 : \frac{m_E}{m_S} = 1 \quad \text{and} \quad H_1 : \frac{m_E}{m_S} = R$$

under the Weibull distribution, the alternative hypothesis becomes equivalent to

$$H_1 : \left(\frac{m_E}{m_S} \right)^{\kappa} = R^{\kappa}.$$

The sample size formula developed by Schoenfeld and Richter (1982) can be found by replacing R by R^{κ}.

$$n = \frac{\left(z_{1-\alpha} + z_{1-\beta} \right)^2}{\kappa^2 (\ln(R))^2} \left(\frac{1}{p_E} + \frac{1}{p_S} \right).$$

p_i is the probability that a specimen in group i will fail in the course of the study and is given by

$$p_i = 1 - P_i(A) S_i(\tau),$$

where S_i is the survival function of the Weibull in group i, and the probability that a specimen not failed up to the end of the accrual period after entry, P_i, is given by

$$P_i = \frac{1}{A}\int_0^A S_i(A-v)dv \;=\; \frac{1}{A}\int_0^A \exp\left(-\ln 2\left(\frac{A-v}{m_i}\right)^{\kappa}\right) dv\,,$$

which can be solved by numerical integration.

The power can also be derived and is given by

$$1-\beta = \Phi\left(\frac{\kappa\sqrt{n}\,\ln(R)}{\sqrt{\frac{1}{p_E}+\frac{1}{p_S}}} - z_{1-\alpha}\right).$$

In the following, we investigate the effect of the shape parameter on the power calculation. The parameters are set as follows:

- The ratio R is between 1 and 1.5.
- The accrual period is 2 months and the follow-up period is for 3 additional months.
- The shape parameter in both groups is $\kappa = 1$, 2, 3.
- The median for the standard group is $m_S = 1$.
- $n = 100$ and $\alpha = 0.05$.

The computation is made in Mathematica 4. Figure shows the power curve of the test for the tree values of the shape parameter $\kappa = 1$, 2, 3. (blue colour for $\kappa = 3$, red colour for $\kappa = 2$ and green colour for $\kappa = 1$ or exponential survival). The power is increasing in κ for any value of R. It clearly indicates that the power depends on the shape parameter κ. This parameter must be identified correctly when the underlying distribution of the survival is Weibull. A reason for this dependency is the fact that when the hazard rate increases, a small difference of the median induces a greater power or a small sample size when the distribution is Weibull.

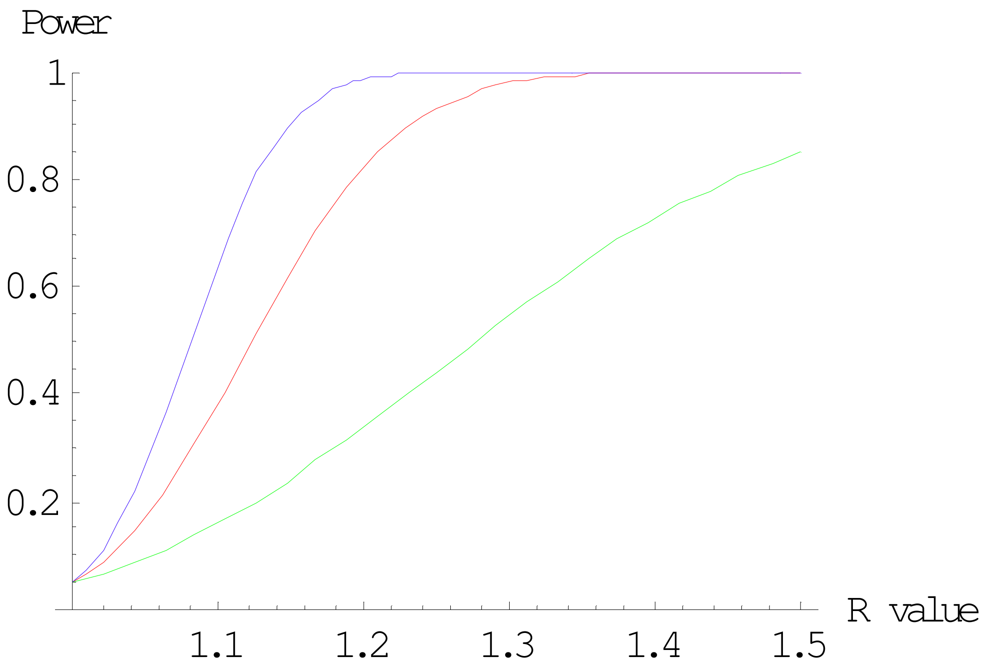


Figure 3.4: *Power curves for* $\kappa = 1, 2, 3$.

With the same specifications for the parameter, we computed the sample size required to obtain a power of 0.9 for the test problem. The computation is done with Mathematica 4 and the results are summarized in Table 3.10. We can observe very important changes on the sample size for varying R and κ. The required sample size is decreasing in κ. This implies that the difference of the expected number of failures between the ratios during the study is increasing in κ. These indicate that the required sample size depends on the shape parameter as deduced from the power comparison above.

R	κ	n
1.05	1	7755
	2	1802
	3	800
1.1	1	2043
	2	473
	3	210
1.15	1	955
	2	220
	3	98
1.2	1	565
	2	130
	3	58
1.25	1	379
	2	87
	3	39
1.3	1	276
	2	63
	3	28
1.35	1	212
	2	48
	3	22
1.4	1	170
	2	39
	3	17
1.45	1	144
	2	32
	3	14
1.5	1	119
	2	27
	3	12

Table 3.10: *Comparison of the sample size per group for different values of* R *and* κ.

Sample size based on the log-rank statistic

Schoenfeld (1981) and Freedman (1982) presented a sample size formula for comparing two survival distributions using the log-rank test. Their methods are based on the asymptotic expectation and variance of the log-rank statistic. But the conditions under which they derived their formulae are very restrictive. Lakatos (1988) extends Freedman’s approach by modelling the survival curves that one could expect under very general conditions using a stochastic process. To calculate the sample size, he then used the asymptotic expectation and variance of the log-rank test applied to those curves. In the following, the sample size provided by Lakatos (1988) will be presented, beginning with the basic nonstationary Markov process.

The basic Markov process von Lakatos

We consider an industrial experiment that compares two production processes or groups (experimental and standard processes) in which each specimen is randomised into one of the two groups. In this nonstationary Markov process the experimental and the standard groups are modelled separately. Each specimen randomised to the experimental group is considered to be a complier initially and is in the state A_E. The probability of having a failure in a giving period of time is P_E. Thus as the experiment progresses, a transition to a different state occurs. If a specimen has a failure, it is transferred to a state *E*. Those specimens who become lost to follow-up and can not be followed for the failure of interest or competing risk are transfered to state *L*. If the specimen no longer complies with the experimental processes, it is transferred to the state A_C. Assuming that there is no time lag in the effectiveness of the processes at a given time, t, a specimen is in one of the four states (L, E, A_E, A_C) with the corresponding vector of occupancy probabilities D_t. The basic initial distribution is

$$D_{t_0} = \begin{bmatrix} \text{Loss} \\ \text{Event} \\ \text{Active Complier} \\ \text{Active Noncomplier} \end{bmatrix} = \begin{bmatrix} L \\ E \\ A_E \\ A_C \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

In the discrete formulation, the transition matrices $T_{k-1,k}$ are constructed such that $t_{k-1,k}(s_1, s_2)$ is the probability of transferring from state s_1 to state s_2 during the time interval $[t_{k-1}, t_k]$.

$$D_{t_k} = T_{k-1,k} D_{t_{k-1}}, \qquad k = 1,2,\cdots,m,$$

where t_m is the end of the experiment. This model provides a sequence of m distributions $\{D_{t_k}, \ k = 1,2,\cdots,m\}$.

Transitions can take place at any time. The probability of failing in the interval $[t_{k-1}, t_k]$ is

$$\frac{1 - S(t_k)}{S(t_{k-1})}$$

with $S(.)$ denoting the cumulative survival distribution function. This cumulative survival distribution can take any form. A continuous process can be approximated by replacing each transition matrix T by $\prod_{l=1}^{I} T_l$, where each time specimen has been divided into I equal intervals, and each off-diagonal element of T_l is given by an appropriate term of the form

$$\frac{1 - S(t_l)}{S(t_{l-1})}.$$

For constant hazard rate within each time specimen, this amounts to replacing each off-diagonal entry y in T by

$$1 - (1 - y)^{\frac{1}{I}}.$$

Derivation of sample size formula

We begin by introducing the formula of the total number of failures d derived by Freedman (1982). This formula is derived by considering the expectation and the variance of the log-rank statistic. We know that the log-rank statistic

$$T = \frac{\sum_{j=1}^{d} \left(d_{1j} - \frac{n_{1j} d_j}{n_j} \right)}{\sqrt{\sum_{j=1}^{d} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^{\ 2} (n_j - 1)}}}$$

under the null hypothesis is approximately standard normally distributed. If we consider a constant hazard ratio θ (the ratio of the hazards in the two groups does not change with time), then the expectation and the variance of the statistic T are

$$E(T)=\frac{\sum_{j=1}^{d}\left(\frac{\phi_j\theta}{1+\phi_j\theta}-\frac{\phi_j}{1+\phi_j}\right)}{\sqrt{\sum_{j=1}^{d}\frac{\phi_j}{(1+\phi_j)^2}}}$$

and

$$V(T)=\frac{\sum_{j=1}^{d}\frac{\phi_j\theta}{(1+\phi_j\theta)^2}}{\sum_{j=1}^{d}\frac{\phi_j}{(1+\phi_j)^2}},$$

where ϕ_j denotes the ratio of the units at risk in the two groups before the *j*th failure. Under the assumption that the ratio of the number of unit in each group at risk just before each failure is equal to 1 ($\phi_j=1$), this is

$$E(T)=\sqrt{d}\,\frac{\theta-1}{\theta+1}$$

and

$$V(T)=\frac{4\theta}{(\theta+1)^2}.$$

Therefore, using the fact that the log-rank statistic is asymptotically normally distributed, one may show that the total number of failures required to give a power of $1-\beta$ at the alternative $\theta>1$ for an α-level test is given by

$$d=\left(\frac{\theta+1}{\theta-1}\right)^2\left(z_\alpha+2\frac{\sqrt{\theta}}{\theta+1}z_\beta\right)^2.$$

In fact if

$$T\sim N(E(T),V(T)),$$

then

$$P(T>z_\alpha)=P\left(Z>\frac{z_\alpha-E(T)}{\sqrt{V(T)}}\right)=P\left(Z>\frac{(\theta+1)z_\alpha-\sqrt{d}(\theta-1)}{2\sqrt{\theta}}\right),$$

with

$$Z\sim N(0,1).$$

Therefore

$$P(T>z_\alpha)=1-\beta \qquad \Leftrightarrow \qquad \frac{(\theta+1)z_\alpha-\sqrt{d}(\theta-1)}{2\sqrt{\theta}}=-z_\beta.$$

This equation yields

$$d = \left(\frac{\theta+1}{\theta-1}\right)^2 \left(z_\alpha + 2\frac{\sqrt{\theta}}{\theta+1} z_\beta\right)^2 .$$

For

$$\frac{2\sqrt{\theta}}{\theta+1}$$

near 1,

$$d \approx \left(\frac{\theta+1}{\theta-1}\right)^2 \left(z_\alpha + z_\beta\right)^2,$$

which is the formula usually used for practical applications.

If we considered the Tarone-Ware statistic which is expressed as

$$T_W = \frac{\sum_{j=1}^{d} w_j \left(d_{1j} - \frac{n_{1j} d_j}{n_j}\right)}{\sqrt{\sum_{j=1}^{d} w_j^{\,2} \left(\frac{n_{1j} n_{2j} d_j \left(n_j - d_j\right)}{n_j^{\,2}\left(n_j - 1\right)}\right)}} ,$$

where w_j is the *j*th Tarone-Ware weight. If we divide the experiment in I independent time intervals, the expectation can be written as

$$E\left(T_W\right) = \frac{\sum_{l=1}^{I} \sum_{j=1}^{d_l} w_{lj} \left[\frac{\phi_{jl}\theta_l}{1+\phi_{jl}\theta_l} - \frac{\phi_{jl}}{1+\phi_{jl}}\right]}{\sqrt{\sum_{l=1}^{I} \sum_{j=1}^{d_l} w_{lj}^{\,2} \frac{\phi_{jl}}{\left(1+\phi_{jl}\right)^2}}},$$

where

d_l is the number of failure in the lth interval,

w_{lj} is the jth Tarone-Ware weight in the lth interval,

ϕ_{jl} is the ratio of items in the two groups at risk just prior to the jth failure in the lth interval,

θ_{jl} is the ratio of the hazard in the two groups of failing just before the jth failure in the lth interval.

When $w_{lj}=1$ for all l and j, the log-rank is obtained. Assuming constant ratio of units under risk before failure j in the lth interval, $\phi_{jl} = \phi_l$, the expectation of the log-rank statistic can be written as

$$E(T_W) = \frac{\sum_{l=1}^{I} d_l \frac{\phi_l\theta_l}{1+\phi_l\theta_l} - \frac{\phi_l}{1+\phi_l}}{\sqrt{\sum_{l=1}^{I} d_l \frac{\phi_l}{(1+\phi_l)^2}}} .$$

Letting

$$\rho_l = \frac{d_l}{d} ,$$

where

$$d = \sum_{l=1}^{I} d_l ,$$

$$\gamma_l = \frac{\phi_l\theta_l}{1+\phi_l\theta_l} - \frac{\phi_l}{1+\phi_l} , \qquad \eta_l = \frac{\phi_l}{(1+\phi_l)^2} ,$$

then

$$E(T_W) = \frac{\sum_{l=1}^{I} d_l\gamma_l}{\sqrt{\sum_{l=1}^{I} d_l\eta_l}} = \frac{d \sum_{l=1}^{I} \rho_l\gamma_l}{\sqrt{d}\sqrt{\sum_{l=1}^{I} \rho_l\eta_l}} .$$

Denoting

$$E(D) = \frac{\sum_{l=1}^{I} \rho_l\gamma_l}{\sqrt{\sum_{l=1}^{I} \rho_l\eta_l}} ,$$

the expectation of the log-rank statistic can be written as

$$E(T_W) = E(D)\sqrt{d} .$$

Using the asymptotic normal distribution of the log-rank with expectation $E(T_W)$ and variance one, we have for an α -level one-sided test to achieve a power of $1-\beta$

$$E(T_W) = z_\alpha + z_\beta .$$

Therefore ($E(D)$ is independent of d_l and d)

$$E(T_W) = z_\alpha + z_\beta = E(D)\sqrt{d} \qquad \Rightarrow \qquad \sqrt{d} = \frac{z_\alpha + z_\beta}{E(D)} = \frac{z_\alpha + z_\beta}{\dfrac{\sum_{l=1}^{I} \rho_l \gamma_l}{\sqrt{\sum_{l=1}^{I} \rho_l \eta_l}}}$$

and

$$d = \frac{(z_\alpha + z_\beta)^2 \sum_{l=1}^{I} \rho_l \eta_l}{\left(\sum_{l=1}^{I} \rho_l \gamma_l \right)^2} .$$

The quantities ρ_l , γ_l, η_l can be easily determined from the sequences of vectors of occupancy probabilities under the Markov model.

The required total sample size is determined using the cumulative failure rates P_S and P_E from the final distributions of the Markov model:

$$n = \frac{2d}{(P_S + P_E)} \qquad = \qquad \frac{2(z_\alpha + z_\beta)^2 \sum_{l=1}^{I} \rho_l \eta_l}{(P_S + P_E)\left(\sum_{l=1}^{I} \rho_l \gamma_l \right)^2} .$$

4 Sample size adaptation methods

We saw in chapter 3 that in case of normally distributed outcomes, the sample size is determined by the type I and type II error rates, the effect size and the variance of the outcomes. The variance is generally unknown at the design stage of the study and it is often the case that the variance will be estimated based on the past studies. However, due to many factors influencing the study conditions like equipment and population, there is still uncertainty about whether the assumed value of the variance is appropriate for the current study. An approach to solving this problem consists of conducting a pilot study whose primary purpose is to provide a data-based estimate of the variance of the outcomes. Wittes and Brittain (1990) provided a specific plan to design such pilot studies and for incorporating the data from the pilot phase into the final study results. In this work, we deal with the internal pilot study approach proposed by Wittes and Brittain (1990). Another approach will be the self-designing method proposed by Fisher (1998). This method consists of using all information available prior to a stage to estimate the sample size and the weight for the next step of the experiment. Shen and Fisher (1999) gave a method to construct the final test statistic based on the weighted average of the sequentially collected information for the case of normal variables with known variances. Hartung (2001) presented a completely self-designing rule by taking the inverse normal transformation of the p-values within the classical Pocock (1977) design. Hartung and Knapp (2003) proposed a flexible effective and adaptive method that allows for a completely self-designing of a group sequential experiment and a decision about stopping for significance of the sequential test results at each stage. All those approaches will be used in this work for further investigations. The self-designing method is a little bit different for the classical group sequential design where we have to test the null hypothesis for rejection after each stage. Pocock provided clear lines for group sequential tests with given type I error and power. O’Brien and Fleming (1979) proposed an alternative to Pocock’s repeated significance tests. By the classical group sequential design, there is no possibility to change the maximum sample size to obtain the desired power. Therefore, a flexible design combining the advantages of group sequential and self-designing methods is of interest. In the following, methods of internal pilot study and self-designing are presented. We also introduce a method of combining the classical group sequential method and the adaptive self-designing method of Hartung (2001) and compared the performance of the proposed method with the adaptive self-designing method of Hartung.

4.1 Internal pilot study

The general procedure for internal pilot study can be described as follows:

- A reasonable guess σ_0 of the variability is used to calculate a preliminary sample size per group n_0.
- After observation of $n_1 \leq n_0$ units per group, we use these observations to obtain an update estimate σ_1 of the variability. The new estimate, is employed in the sample size formula to obtain an updated sample size per group $\hat{n}$.
- The final sample size per group n_f is chosen.
- After observation - if necessary - of $n_2 = n_f - n_1$ more units per group, the hypothesis test is performed using all the n_f observations per group.

This procedure is also called the two-stage design. It has been introduced by Stein (1945). Stein's procedure uses only the pilot variance estimate in the final statistic. Accordingly, Wittes and Brittain (1990) proposed a two-stage design like Stein's, but using the t-test at the end of the study (all the data are used for the final variance estimate).

4.1.1 The proposal of Wittes and Brittain

We used the same notation as in section 3.1. The internal pilot study approach proposed by Wittes and Brittain (1990) proceeds as follows.

- A preliminary sample size n_0 per group is computed by using a reasonable guess of the variance σ_0.
- After observation of $n_1 \leq n_0$ units per group, we use these observations to obtain an update estimate σ_1 of the variability. This estimate of the variability is obtained by computing the observed variance within each group and pooling them:

$$S_1^2 = \frac{S_{E,1}^2 + S_{S,1}^2}{2}.$$

- The new estimate of the variance is employed in the sample size formula to obtain an update sample size per group $\hat{n}$:

$$\hat{n} = n_0 \left(\frac{\sigma_1}{\sigma_0} \right)^2$$

- The final sample size per group n_f is chosen equal to $\max(n_0, \hat{n})$
- After observation if necessary of $n_2 = n_f - n_1$ more units per group, the hypothesis test is performed using all the n_f observations per group. The test statistic is expressed as

  $$T = \sqrt{\frac{n_f}{2}} \frac{\overline{X}_E - \overline{X}_S}{S_f},$$

  where S_f is the pooled variance estimate using all the n_f units per group and $\overline{X}_E$ and $\overline{X}_S$ are computed using all the n_f units per group.
- H_0 is rejected for $T > t_{1-\alpha/2, 2(n_f - 1)}$.

For Wittes and Brittain procedure, the final sample size per group n_f should not be smaller than the originally planned sample size. Birkett and Day (1994) pointed out that this could result in an unnecessarily large sample size if the prior estimate of the variance σ_0 is too large. They proposed the rule $n_f = \max(n_1, \hat{n})$ so that the final sample size cannot be smaller than the size of the internal pilot study.

Simulations done by Wittes and Brittain (1990) and later by Birkett and Day (1994) show that the type I error rate may exceed the nominal level. Wittes et al. (1999) conducted the type I error rate using numerical integration and their proposed adjustments. Denne and Jennison (1999) proposed a test based on Stein's two-stage test by using an internal pilot study to estimate variance and thus the final sample size. Kieser and Friede (2000) quantified the maximum excess of the type I error rate for normally distributed outcomes.

4.1.2 Choice of the pilot sample

The choice of the sample size of the pilot study is very important for the experiment. Low pilot sample, which represents adaptation early in the course of the experiment, has the advantage of allowing timely reallocation of resources to the study, but it may lead to an imprecise estimate of the variance. Large pilot sample on the other hand has the benefit that one can estimate the variance quite precisely, but it may be impractical for economic or logistical reasons.

For the choice of the pilot sample, many proposals have been done by the researchers.

For the Stein's two-stage design, Seelbinder (1953) applied the minimax principle. He suggested that n_1 should be chosen to minimise the maximum of $E(\hat{n}) - n$ over some range of values of σ. Following the same target, Moshman (1958) proposed a rule that not only keeps $E(\hat{n})$ small, but also the probability of an extremely large total sample size, over some range of values of σ. In order to control the relative weight given to each of the desired features, he included an additional parameter and suggested that this parameter should be chosen independently from statistical considerations. Wittes and Brittain (1990) for the simulation studies used $n_1 = 0.5n_0$. Sandvik et al. (1996) proposed a method which aims to make the pilot as large as possible whilst controlling the probability of the pilot being larger than the appropriate fixed sample size. These methods require a pre-estimate of the true variance gain from the experiment. Denne and Jennison (1999) proposed a rule for choosing the size of the internal pilot which also requires some pre-study knowledge about likely values for the true variance. They took a value that minimises the ratio $E(\hat{n})/n$ for the true values of the variance and they proposed a strategy of finding a value for this ratio that is near to the minimum.

4.2 Self-designing method

We consider a balanced experiment comparing two production groups with independent normally distributed outcomes with expectations μ_E for the experimental group and μ_S for the standard group, and with common variance σ^2. X_E and X_S denote the normally distributed outcome of interest for the experimental and standard production groups respectively. $\theta = \mu_E - \mu_S$ is the parameter of interest and we desired to conduct an α-level test for the null hypothesis $H_0 : \theta = 0$ against the one-sided alternative $\theta > 0$. As developed in section 3.1, the expression for the required sample size per group $n_S = n_E = n$ to give a power of $1-\beta$ when $\theta = \tilde{\theta}$ is

$$n = \frac{2(z_\alpha + z_\beta)^2 \sigma^2}{\tilde{\theta}^2}.$$

Suppose that after observation of a fraction r of the planned sample size n, the difference in means $\theta = \tilde{\theta}_r < \tilde{\theta}$, then after observing all the n units, the conditional power is low if the true effect size is $\tilde{\theta}_r$ rather than the hypothesized $\tilde{\theta}$ and it is unlikely that H_0 will be rejected. The sample size per group we would have required to give a power of $1-\beta$ at the alternative $\theta = \tilde{\theta}_r$ is $(\tilde{\theta} / \tilde{\theta}_r)^2 n$, so that one may wish to increase the sample size to meet this condition. Therefore under H_0 the final test statistic Z will not follow a normal distribution because it is a function of the first stage data and the test that reject H_0 when $Z > z_\alpha$ does not have the level α. Fisher (1998) proposed a method which allows changes to the sample size at an interim analysis while still preserving the type I error rate. In the following, the proposal of building the test statistic of Shen and Fisher (1999) is presented. This proposal follows the general setting of self-designing trials introduced by Fisher (1998). This will be followed by the self-designing method of Hartung (2001).

4.2.1 The proposal of Shen and Fisher

The experiment is divided into an infinite number of stages. The maximal size sample pro group of each stage $\{K_l,\ l = 1, 2, \cdots\}$ of the experiment is fixed for the beginning. As mentioned above the aim is to test the null hypothesis $H_0 : \theta = 0$ against the one-sided alternative $\theta > 0$. We denote $\overline{X}_{E,l}$ and $\overline{X}_{S,l}$ the means of the experimental or standard group in the *l*-the block with size K_l respectively and the difference of the means in the *l*-the block $\overline{X}_l = \overline{X}_{E,l} - \overline{X}_{S,l}$. The test statistic of the *l*-the block of data is given by

$$Z_l = \sqrt{\frac{K_l}{2}} \frac{\overline{X}_{E,l} - \overline{X}_{S,l}}{\sigma}.$$

The mean of Z_l is

$$\sqrt{\frac{K_l}{2}} \frac{\theta}{\sigma}$$

and the variance is one. The final test statistic Z has the form

$$Z = \sum_{l=1}^{L} w_l Z_l = \sum_{l=1}^{\infty} w_l Z_l \,,$$

where the nonnegative weights w_l are functions of $Z_1, \cdots, Z_{l-1}$. With probability one, there exists a positive finite random number L such that

$$\sum_{l=1}^{\infty} w_l^2 = \sum_{l=1}^{L} w_l^2 = 1 \ (w_l = 0 \quad \forall\, l > L).$$

Theorem 1 of Fisher (1998) states that under the null hypothesis Z is standard normally distributed. Therefore the weights affect only the variance of the test statistic and not the mean. At a given level α of the test, the null hypothesis H_0 is rejected, if $Z > z_\alpha$. The weight at the *l*-the stage is computed based on the accumulated data prior to the *l*-the stage and w_1 is estimate from the initial design parameters. That implies the weight at each stage will be determined iteratively. The decision of going to the next stage or stopping the experiment will be taken using a lower stopping boundary $z(l)$. The experiment will be stopped at the *l*-the stage if

$$\sum_{j=1}^{l} \sqrt{K_j} \frac{Z_j}{\sqrt{N_l}} < z(l),$$

where $z(l)$ is a lower stopping boundary and

$$N_l = \sum_{j=1}^{l} K_j ,$$

otherwise the procedure will be continued as long as

$$\sum_{j=1}^{l} w_j^2 < 1 .$$

There is not a rule for the selection of $z(l)$, that is why this selection is based on subjective opinion. Because the weight functions are constructed iteratively, we need a weight for the first stage. This can be obtained using pre-specified design parameters. The weight for the first block can be chosen equal to

$$w_1 = \left(\frac{n_1}{K_1} \right)^{1/2} ,$$

where the minimum fixed sample size per group n_1 to achieve a power of β at the alternative $\tilde{\theta}$ is given by

$$n_1 = \frac{2(z_\alpha + z_\beta)^2 \sigma^2}{\tilde{\theta}^2} .$$

The additional sample size at the *l*-the stage to ensure a power of $1-\beta$ given the observations up to stage $l-1$ can be obtained by solving the following equation for n_l

$$P\left(\sum_{j=1}^{l} w_j Z_j + \sqrt{1 - \sum_{J=1}^{l-1} w_j^2} \left(\sqrt{\frac{n_l}{2}} \right)^{1/2} \frac{\overline{X}_l^*}{\sigma} > z_\alpha \mid \theta = \tilde{\theta}_{l-1}, Z_1, \cdots, Z_{l-1} \right) = 1 - \beta ,$$

where $\overline{X}_l^*$ is the mean difference of two groups in n_l samples per group, $\tilde{\theta}_{l-1}$ is the sample mean up to the *l*-1-the stage.

$$P\left(\sum_{j=1}^{l} w_j Z_j + \sqrt{1 - \sum_{J=1}^{l-1} w_j^2} \left(\sqrt{\frac{n_l}{2}} \right)^{1/2} \frac{\overline{X}_l^*}{\sigma} > z_\alpha \mid \theta = \tilde{\theta}_{l-1}, Z_1, \cdots, Z_{l-1} \right) = 1 - \beta \qquad \Leftrightarrow$$

$$P\left(\left(\sqrt{\frac{n_l}{2}}\right)^{1/2}\frac{\overline{X}_l^*}{\sigma}\le\frac{z_\alpha-\sum_{j=1}^{l-1}w_jZ_j}{\sqrt{1-\sum_{J=1}^{l-1}w_j^2}}\mid\theta=\tilde{\theta}_{l-1},Z_1,\cdots,Z_{l-1}\right)=\beta\,.$$

But conditionally on $Z_1,\cdots,Z_{i-1}$, the variable

$$\left(\sqrt{\frac{n_l}{2}}\right)^{1/2}\frac{\overline{X}_l^*}{\sigma}$$

is normally distributed. Its mean is

$$\sqrt{\frac{n_l}{2}}\frac{\theta}{\sigma}$$

and variance is one.

Therefore

$$\Phi\left(\frac{z_\alpha-\sum_{j=1}^{l-1}w_jZ_j}{\sqrt{1-\sum_{J=1}^{l-1}w_j^2}}-\sqrt{\frac{n_l}{2}}\frac{\tilde{\theta}_{l-1}}{\sigma}\right)=\beta\,,$$

where Φ denotes the cumulative distribution function of the standard normal distribution. The additional sample size per group in the l-the stage to ensure power of $1-\beta$ is then given by

$$n_l=2\frac{\sigma^2}{\tilde{\theta}_{l-1}^2}\left\{\frac{z_\alpha-\sum_{j=1}^{l-1}w_jZ_j}{\sqrt{1-\sum_{j=1}^{l-1}w_j^2}}+z_\beta\right\}^2\,,$$

and the weight at the i-the stage is defined to be

$$w_l=\begin{cases}\left(\dfrac{K_l\sqrt{1-\sum_{j=1}^{l-1}w_j^2}}{n_l}\right)^{1/2} & \text{for}\quad l=2,\cdots,m-1\\[2ex] \sqrt{1-\sum_{j=1}^{m-1}w_j^2} & \text{for}\ l=m\end{cases}\quad.$$

Another proposal is the following class of weight function with the particularity that the early stages have more weight. The weight at the first stage is defined above and the later one are given by

$$w_l = \begin{cases} \left(\dfrac{K_l}{2n_i}\right)^{1/2} & \text{for } l = 2,\cdots,m-1 \\ \sqrt{1-\sum\limits_{j=1}^{m-1} w_j^2} & \text{for } l = m \end{cases}.$$

For logistical and economical reasons, it may be necessary to allow early termination of the experiment and "failed to reject" H_0. Explicitly, we will continue the experiment as long as

$$\sum_{j=1}^{l} \sqrt{K_j} \frac{Z_j}{\sqrt{N_l}} > z(l) \text{ and } l < m.$$

If

$$\sum_{j=1}^{l} \sqrt{K_j} \frac{Z_j}{\sqrt{N_l}} \le z(l),$$

then stopped the experiment without rejecting H_0 and set $l = m$.

The lower boundary $z(l)$ at the l-the stage is the Wald-type constant likelihood boundary and is given by

$$z(l) = \frac{NC_l}{2} + \frac{\log\left(\dfrac{\beta}{1-\alpha}\right)}{NC_l},$$

where NC_l denotes the noncentrality parameter which is

$$NC_l = \sqrt{\frac{N_l}{2}} \frac{\tilde{\theta}}{\sigma}.$$

4.2.2 The proposal of Hartung

The aim is to test the null hypothesis $H_0 : \theta = 0$ against the one-sided alternative $\theta > 0$. The experiment is formally divided into an infinite number of disjoint stage $1, 2, \cdots, l, \cdots$. Only a random finite number of stages will be observed according to a self-designing procedure. n_l specimens are enrolled in the experiment and randomised across the two production groups in the stage l. Based on their responses the test statistic T_l of the l the stage is computed.

Under the null hypothesis, T_l has a continuous distribution function $F_{l,0}$. Therefore, the p-values

$$p_l = 1 - F_{k,0}(T_l)$$

are uniformly distributed on the interval $(0,1)$, such that

$$u_l = \Phi^{-1}(1 - p_l)$$

is normally distributed with mean 0 and variance 1, where Φ^{-1} denotes the inverse of the standard normal distribution function Φ. The final test statistic U has the form

$$U = \sum_{l=1}^{L} w_l u_l = \sum_{l=1}^{\infty} w_l u_l \;\; (w_l = 0 \;\; \forall\, l > L),$$

where the nonnegative weights w_l are functions of $u_1, \cdots, u_{l-1}$. With probability one, there exists a positive finite random number L such that

$$\sum_{l=1}^{\infty} w_l^2 = \sum_{l=1}^{L} w_l^2 = 1 .$$

Theorem 1 of Fisher (1998) states that under the null hypothesis U is standard normally distributed. Therefore the weights affect only the variance of the test statistic and not the mean. At a given level α_G of the test, the null hypothesis H_0 is rejected, if $U > \Phi^{-1}(1 - \alpha_G)$.

If the number of specimens in the lth stage n_l are determined upon knowledge of the previous study stages, the distribution of the u_l and the independence of u_l, u_k, $l \neq k$ still holds. Given the global type I and II error rates α_G, β_G, the number of specimens in the first stage n_1, $w_1 \leq 1$, the type II error rate β_g for generating the sequential additional number of specimens $n_l = \hat{n}_{l-1}$, $\beta_g \geq 0.2$ and can also be determined as a function of l, the self-designing rule is characterized by the formula

$$R = R(\alpha_G, \beta_G; n_1; w_1; \beta_g, \varepsilon; \alpha_L),$$

where ε is a lower bound for the weights w_l, that is $\varepsilon \leq w_l$ and α_L as defined by $\Phi^{-1}(\alpha_L)$, is a lower bound for

$$\frac{1}{\sqrt{l}} \sum_{j=1}^{l} u_j .$$

$\alpha_L = 0.6$ or $\alpha_L = \alpha_L(l)$, increasing with l starting even at zero. The notation $a_l = \hat{a}_{l-1}$ indicates that a_l is estimated upon all the knowledge obtained in the previous study stages before the beginning of stage l. If

$$\frac{1}{\sqrt{l}}\sum_{j=1}^{l} u_j \le \Phi^{-1}(\alpha_L),$$

H_0 is not rejected and the experiment stops at that stage.

Now, let w_j, p_j, u_j be given for $j = 1,\cdots,l-1$, with

$$U_{l-1} = \sum_{j=1}^{l-1} w_j u_j \,,$$

then if for stage l in the equation

$$P\left(U_{l-1} + \sqrt{1-\sum_{j=1}^{l-1} w_j^{\;2}} \cdot \hat{u}_l > \Phi^{-1}(\alpha_G) \mid \theta = \hat{\theta}_{l-1} > 0 \right) = 1-\beta$$

we put $\beta = \beta_G$ and

$$w_l = w_{l,G} = \sqrt{1-\sum_{j=1}^{l-1} w_j^{\;2}} \,,$$

we would have

$$\sum_{j=1}^{l-1} w_j^{\;2} + w_{l,G}^{\;\;2} = 1$$

and the final test statistic

$$U_{l,G} = U_{l-1} + w_{l,G} \cdot u_l(\beta_G)$$

would hold level α_G and power $1-\beta_G$ conditional on $\hat{\theta}_{l-1} > 0$. $u_l(\beta_G)$ results from $n_l(\beta_G)$ specimens in stage l. If we put $\beta = \beta_g > \beta_G$, we take $w_l = w_{l,g} < w_{l,G}$ so that the experiment does not stop after stage l, when the test results obtainable with $n_l(\beta_g)$ specimens of $n_l(\beta_G) > n_l(\beta_g)$ specimens in stage l is reduced. Therefore, letting

$$\beta_g = \beta_g(l) \underset{l \to L}{\rightarrow} \beta_G \,,$$

the termination of the experiment can be accelerated.

The potential additional number of specimens m_l and M_l for stage l are defined by Hartung (2001) as

$$m_l = S_l(\hat{p}_l, \beta_g) \qquad \text{and} \qquad M_l = S_l(\hat{p}_l, \beta_G),$$

where

$$\hat{p}_l = 1 - \Phi(\hat{u}_l) = 1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha_G) - U_{l-1}}{\sqrt{1-\sum_{j=1}^{l-1} w_j^{\;2}}} \right), \tag{4.1}$$

and $q = S_l(\alpha, \beta)$ means that q is the smallest, finite number such that in a sample of size q the test of H_0 by the statistic U_l has level α and power $1-\beta$, $\hat{\theta}_{l-1}$ given.

For power $1-\beta_G$ these potential additional number of specimens in stage l would lead to the levels $\hat{\alpha}_l(m_l)$ and $\hat{\alpha}_l(M_l)$ respectively, given by the following implicit equations

$$m_l = S_l(\hat{\alpha}_l(m_l), \beta_G) \quad \text{and} \qquad M_l = S_l(\hat{\alpha}_l(M_l), \beta_G),$$

m_l and M_l given.

Defining the weight function W_l by

$$W_l = \frac{\Phi^{-1}\left(1 - \frac{\hat{\alpha}_l(m_l)}{2}\right)}{\Phi^{-1}\left(1 - \frac{\hat{\alpha}_l(M_l)}{2}\right)} \sqrt{1 - \sum_{j=1}^{l-1} W_j^{\,2}}\,, \qquad W_1 = w_1 \le 1 \text{ given,} \tag{4.2}$$

the weight w_l and the additional number of specimens n_l in stage l are then given respectively by

$$w_l = \begin{cases} W_l & \text{if} \quad W_l \ge \varepsilon \\ \sqrt{1 - \sum_{j=1}^{l-1} w_j^{\,2}} & \text{if} \quad W_l < \varepsilon \end{cases} \tag{4.3}$$

and

$$f_l = \begin{cases} m_l & \text{if} \quad W_l \ge \varepsilon \\ M_l & \text{if} \quad W_l < \varepsilon \end{cases}. \tag{4.4}$$

If $W_l < \varepsilon$, then put $l = L$ and the experiment stops at the Lth stage. With n_l specimens in the lth stage, the test statistic T_l and p-value p_l, we obtain the intermediate result

$$U_l = U_{l-1} + w_l \Phi^{-1}(1 - p_l)$$

or for $l = L$ the final result U_L.

The adaptation procedure can be summarized as follows:

1. Define the global type I and II error rates α_G and β_G, the type II error rate β_g for generating the sequential number of specimens n_l, the lower bound ε for the weights w_l and finally α_L.
2. Choose the starting configuration n_1 and w_1 for the first stage.
3. After study part "stage l" $l \ge 1$, calculate T_l, w_l, p_l, u_l,

$$U_l = \sum_{j=1}^{l} w_j u_j \, .$$

If

$$\frac{1}{\sqrt{l}} \sum_{j=1}^{l} u_j \le \Phi^{-1}(\alpha_L),$$

H_0 is not rejected and the experiment stops at that stage.

If

$$\frac{1}{\sqrt{l}} \sum_{j=1}^{l} u_j > \Phi^{-1}(\alpha_L),$$

then go to the next step.

4. Compute the weight function W_{l+1} and finally the weight w_{l+1} and the number of specimens n_{l+1} for stage $l+1$.
5. If $W_{l+1} < \varepsilon$, the study stops and H_0 is rejected, if the final test statistic $U > \Phi^{-1}(1-\alpha_G)$. If $W_{l+1} \ge \varepsilon$, then go to step 3 and replace l by $l+1$.

4.3 Combining the adaptive self-designing of Hartung with the classical group sequential design

The following discussion is based on the idea of Yin and Shen (2005). We consider a balanced experiments comparing two production groups with independent normally distributed outcomes with expectations μ_E for the experimental group and μ_S for the standard group, and with common variance σ^2. X_E and X_S denote the normally distributed outcome of interest for the experimental and standard production groups respectively. $\theta = \mu_E - \mu_S$ is the parameter of interest and we wish to conduct an α-level test for the null hypothesis $H_0 : \theta = 0$ against the one-sided alternative $\theta > 0$.

By the self-designing method, we only have the possibility to stop the experiment for futility and not for efficacy like in group sequential method. The null hypothesis is tested for rejection only at the final stage. We investigated the fact of introducing one more interim analysis into the self-designing procedure of Hartung (2001), using techniques of group sequential method for early stopping for efficacy. Let us make a short review of classical group sequential design.

4.3.1 Classical group sequential design

We denote $\overline{X}_{E,l}$ and $\overline{X}_{S,l}$ the means of the experimental or standard group in the *l*-the stage respectively and the difference of the means in the *l*-the stage $\overline{X}_l = \overline{X}_{E,l} - \overline{X}_{S,l}$. The maximum number of stages to be performed L is prespecified at the beginning of the experiment. The test statistic of the *l*-the stage of data is given by

$$T_l = \sum_{i=1}^{l} \sqrt{\frac{i \cdot n}{2}} \frac{\overline{X}_{E,i} - \overline{X}_{S,i}}{\sigma}.$$

At the *l*-the stage, $l = 1, 2 \cdots, L-1$, the null hypothesis H_0 is rejected and the experiment is stopped if $T_l \geq b_l$ for a properly chosen boundary b_l. Otherwise the null hypothesis is not rejected and the experiment continues to stage $l+1$. At the last stage L, the null hypothesis is rejected if $T_L \geq b_L$, otherwise, the null hypothesis is not rejected.

Now, how do we choose the boundaries b_l, $l = 1, 2 \cdots, L$ so that the overall significance level is α? Of interest is the probability of rejecting the null hypothesis when it is true to be α. The testing procedure will have level α if the boundaries are chosen such that under the null hypothesis,

$$P\{T_1 < b_1, T_2 < b_2, \cdots, T_L < b_L\} = 1 - \alpha.$$

Two boundaries have been discussed extensively in the literature. Those are the Pocock boundary; Pocock (1977) and the O'Brien and Fleming boundary; O'Brien and Fleming (1979). The Pocock boundary used the same critical value at each stage (Reject H_0 at stage l if $T_l \geq b_P$) and the O'Brien and Fleming boundary used smaller critical value at each stage so that it is hard to reject the null hypothesis early in the study and the final test as similar as possible to a fixed sample test (Reject H_0 at stage l if $T_l \geq b_{OF} / \sqrt{l}$).

Values of b_P and b_{OF} are obtained for given α and L.

4.3.2 Adaptation procedure

In the self-designing procedure, the experiment can be terminated at the second stage due to futility or spending all the weight. Therefore we have to perform only one interim analysis after observing the data of the first stage and the second and last efficacy test at the end of the experiment. After observing the data of the first stage, we compute the test statistic T_1.

- If $T_1 < \Phi^{-1}(\alpha_L)$, H_0 is not rejected and the experiment stops.
- If $T_1 \geq b_1$, H_0 is rejected and the experiment stops

  If $T_1 \geq \Phi^{-1}(\alpha_L)$ and $T_1 < b_1$, then go to the next step.
- Compute the weight function W_2 and finally the weight w_2 and the number of specimens n_2 for stage 2. The procedure continues normally as described in the last section where only a stop for futility is applied at the end of the observation of the data of each stage.

4.3.3 A simulation study

The operating characteristic of the new design is compared by simulation to those of the self-designing of Hartung (2001) in terms of empirical size, empirical power, average sample number and the number of stages performed. We want to detect an improvement of 0.5 in the difference of the means of the experimental and the standard group. In the adaptive self-designing of Hartung, the parameters are set as follows: $\alpha_G = 0.05$, $\alpha_L = 0.6$, $\beta_G = 0.1$, $\varepsilon = 0.1$, $\sigma = \sqrt{2}$, $n_1 = N/2$ = 138, where N is the sample size for the fixed-sample design when θ is correctly specified, $n_{\min} = n_1/8$, $w_1 = \sqrt{2}/2$. The upper bounds for the sample size spending functions m_l and M_l are set to 275 in a first time and 200 in the second time, corresponding to two different strategies. For the new design, the Pocock boundary (Design 1) and the O'Brien and Fleming boundary (Design 2) are used for stop for efficacy at the end of the first stage. Specifically for $\alpha = 0.05$ and $L = 2$, the Pocock boundary is $b_1 = 2.178$ and the O'Brien and Fleming boundary is $b_1 = 2.797$. θ vary from 0.1 to 0.7 under the alternative hypothesis. For each case, 100000 independents replications were performed. The results of the simulations are summarized in table 4.1 and table 4.2 for the two strategies, respectively.

Hartung (2001)				Design 1			Design 2		
$\overline{\theta}$	Size	ASN	Stage	Size	ASN	Stage	Size	ASN	Stage
0	0.044	257.7	1.6	0.050	254.7	1.5	0.043	257.5	1.6
$\overline{\theta}$	Power	ASN	Stage	Power	ASN	Stage	Power	ASN	Stage
0.1	0.16	307.2	2.0	0.17	299.6	1.9	0.16	305.6	2.0
0.2	0.38	341.9	2.6	0.40	329.8	2.2	0.38	340.8	2.5
0.3	0.63	348.6	3.1	0.65	329.0	2.5	0.63	344.9	3.0
0.4	0.81	327.4	3.4	0.84	300.6	2.5	0.82	320.5	3.2
0.5	0.90	290.3	3.5	0.93	256.8	2.3	0.92	279.0	3.0
0.6	0.94	250.9	3.3	0.97	214.9	1.9	0.96	235.2	2.7
0.7	0.96	216.7	3.0	0.99	180.4	1.6	0.98	197.8	2.2

Table 4.1: *Empirical Size (Size), Empirical power (Power), Average sample number (ASN), Number of performed stages (Stage) for different design methods with the strategy* $M_l \le 275$, $m_l \le 275$.

Hartung (2001)				Design 1			Design 2		
$\overline{\theta}$	Size	ASN	Stage	Size	ASN	Stage	Size	ASN	Stage
0	0.045	224.5	1.6	0.051	221.5	1.5	0.046	224.2	1.6
$\overline{\theta}$	Power	ASN	Stage	Power	ASN	Stage	Power	ASN	Stage
0.1	0.14	261.2	1.9	0.16	255.7	1.8	0.14	260.7	1.9
0.2	0.33	290.3	2.4	0.35	279.0	2.1	0.34	288.5	2.3
0.3	0.57	300.9	2.9	0.60	282.3	2.3	0.58	297.8	2.8
0.4	0.77	291.93	3.2	0.80	266.0	2.3	0.78	284.8	3.0
0.5	0.88	267.6	3.3	0.92	235.5	2.1	0.90	257.0	2.9
0.6	0.93	238.2	3.2	0.97	203.1	1.8	0.95	223.3	2.6
0.7	0.95	211.1	3.0	0.99	174.8	1.5	0.98	192.4	2.2

Table 4.2: *Empirical Size (Size), Empirical power (Power), Average sample number (ASN), Number of performed stages (Stage) for different design methods with the strategy* $M_l \le 200$, $m_l \le 200$.

The simulations suggest that under the null hypothesis, the new designs yield a test with the specified size and use quite the same sample size as the self-designing of Hartung. Evidently the empirical size by Design 2 (O'Brien and Fleming boundary) is larger than by Design 1 (Pocock boundary) because of the more stringent stopping conditions of the O'Brien and Fleming procedure at early stages. The fact of using different upper bound for the sample size function m_l and M_l does not have important impact on the size property of all the designs. The largest difference between the size is observed by Design 2 with values of 0.043 (strategy $M_l \leq 275,\ m_l \leq 275$) and 0.046 (Strategy $M_l \leq 200,\ m_l \leq 200$). Under the alternative hypothesis, the new designs for both strategies lead to a gain of power near the adaptive self-designing of Hartung with smaller sample size. The difference between the power is larger when the true difference between the means is underestimated as when the true underlying difference between the means is overestimated. A larger upper bound for the sample size functions m_l and M_l results in a larger power. It is not surprising that the number of performed stages in the new designs is smaller than in the design Hartung due to the fact that the first one results by the insertion of another stopping rule in the later one. The choice of the strategy has no relevant influence in the average number of stages to be performed.

5 Sample size adaptation for the one-sided equivalence test

We consider experiments comparing two production groups with independent normally distributed outcomes with expectations μ_E for the experimental group and μ_S for the standard group, and with unknown but common variance σ^2. For simplicity and without loss of generality we consider a balanced design, which is an equal sample size in both production groups with a total of N observations.

Let X_E and X_S designate the normally distributed outcome of interest for the experimental and standard production groups respectively. For equivalence testing, it is reasonable to assume that the sign of the corresponding population means μ_E and μ_S are both positive.

Recall that for the additive model, $\theta = \mu_E - \mu_S$, the one-sided-equivalence hypotheses are typically as follows:

$$H_0 : \mu_E - \mu_S \geq \delta_a \qquad \text{versus} \qquad H_1 : \mu_E - \mu_S < \delta_a .$$

The test statistic involved is the usual statistic for testing the difference between two population means with unknown variance. For the first one-sided test, the test statistic is

$$T_a = \sqrt{\frac{N}{4}}\,\frac{\overline{X}_E - \overline{X}_S - \delta_a}{S_p}, \tag{5.1}$$

where S is the pooled standard deviation of the two samples, $\overline{X}_E$ and $\overline{X}_S$ the sample means of the experimental and standard production group.

The null hypothesis can be rejected at level α if $T_a \leq -t_{1-\alpha,N-2}$ where $t_{1-\alpha,N-2}$ is the $(1-\alpha)$ percentile of the central $t-$distribution with $N-2$ degrees of freedom.

The sample size per group required for the rejection of H_0 to give a power $1-\beta$ of an α-level test at a specified alternative $\tilde{\theta}_a = \mu_E - \mu_S$ is expressed as

$$n = 2\frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{\left(\tilde{\theta}_a - \delta_a\right)^2}. \tag{5.2}$$

In the following, two methods of sample size adaptation are presented. The first using the internal pilot study procedure and the second the self-designing procedure

5.1 Sample size adaptation using internal pilot study

The sample size adjustment procedure considered for our investigation can be described as follows:

- A reasonable guess of the variability is used to calculate a preliminary sample size per group n_0. To do this, we replace the population standard deviation σ in the formula (5.2) by an estimate σ_0.
- After observation of $n_1 \le n_0$ units per group, we use these observations to obtain an update estimate σ_1 of the variability. The new estimate, respectively, is employed in the sample size formula (5.2) to obtain an update sample size per group $\hat{n}$.
- The final sample size per group n_f is chosen equal to $\max(n_0, \hat{n})$ if the final sample size can not be lower than initially planned (Wittes and Brittain, 1990) or equal to $\max(n_1, \hat{n})$ as proposed by Birkett and Day (1994). After observation if necessary of $n_2 = n_f - n_1$ more units per group, the hypothesis test is performed using all the n_f observations per group.

5.1.1 Variance estimator, distribution of the test statistic and actual type I error rate

The sample size of the one-sided test problem $H_0 : \mu_E - \mu_S \ge \delta_a$ versus $H_1 : \mu_E - \mu_S < \delta_a$ is recalculated according to formula (5.2) by replacing the unknown true variance by an estimated gain from the data of the pilot study. In the following, some methods for variance estimation and the derived distribution of the test statistic, sample size and the actual type I error rate are given.

Pooled sample variance

The usual pooled variance estimate is expressed as

$$S_p^2 = \frac{S_E^2 + S_S^2}{2}.$$

Therefore the test statistic obtained for all the n_f observation per group is given by

$$T_a = \sqrt{\frac{n_f}{2}}\frac{\overline{X}_E - \overline{X}_S - \delta_a}{S_p}.$$

Because n_1 is chosen arbitrary, the total simple size per group n_f is no more a constant but depends on the data of the pilot study. Therefore T_a no longer follows under $H_0 : \mu_E - \mu_S \geq \delta_a$ a central t-distribution. To avoid this problem, we consider the test statistic

$$T_a^* = \sqrt{\frac{n_f}{2}}\frac{\overline{X}_E - \overline{X}_S - \delta_a}{S_p^*}, \tag{5.3}$$

with (Kieser und Friede: 2000)

$$S_p^{*2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_f - 2},$$

where S_i^2, $i = 1, 2$ are the pooled variance estimate of the observations before $(i = 1)$ and after $(i = 2)$ the interim analysis:

$$S_i^2 = \frac{S_{E,i}^2 + S_{S,i}^2}{2}, \quad i = 1, 2.$$

$S_{E,i}^2$ and $S_{S,i}^2$ denote the variance of the experimental or standard group before ($i = 1$) or after ($i = 2$) sample size adaptation.

In the fixed sample situation with n_f observations per group, T_a^* followed under the null hypothesis the central t-distribution with $2n_f - 4$ degrees of freedom.

Distribution of the test statistic T_a^*

The key idea is to decompose the test statistic into components for which the joint density can be derived.

Let

$$D_i = \sqrt{\frac{n_i}{2}}\frac{\overline{X}_{E,i} - \overline{X}_{S,i} - \delta_a}{\sigma}, \quad i = 1, 2,$$

then the test statistic T_a^* can be written as

$$T_a^* = \sigma \frac{\sqrt{\frac{n_1}{n_f}}D_1 + \sqrt{\frac{n_2}{n_f}}D_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_f - 2}}}.$$

$\overline{X}_{E,i}$ and $\overline{X}_{S,i}$ denote the means of the experimental or standard group before ($i=1$) or after ($i=2$) sample size adaptation.

Furthermore if we use the following notation

$$V_i = \frac{2(n_i-1)S_i^2}{\sigma^2},\ i=1,2,$$

then the test statistic can be written as

$$T_a^* = \frac{\sqrt{\frac{n_1}{n_f}}D_1 + \sqrt{\frac{n_2}{n_f}}D_2}{\sqrt{\frac{V_1+V_2}{2n_f-4}}}.$$

V_1 is χ^2-distributed with $2(n_1-1)$ degrees of freedom. Conditional on V_1, V_2 is χ^2-distributed with $2(n_2-1)$ degrees of freedom. Under the null hypothesis, D_1 is normally distributed with expectation zero and variance one. Conditional on V_1, D_2 is normally distributed with expectation zero and variance one. D_1 and V_1 are independent, D_2 and V_2 are independent given V_1. Therefore, under the null hypothesis the joint density of D_1, V_1, D_2, V_2 can be written as

$$f(d_1,v_1,d_2,v_2) = \phi(d_1)\, g_{\chi^2_{2(n_1-1)}}(v_1)\, \phi(d_2)\, g_{\chi^2_{2(n_2-1)}}(v_2),$$

where ϕ denotes the density of the standard normal distribution and $g_{\chi^2_n}$ the density of the chi-square distribution with n degrees of freedom.

Furthermore if we use the following notation

$$D = \sqrt{\frac{n_1}{n_f}}D_1 + \sqrt{\frac{n_2}{n_f}}D_2 = \frac{1}{\sigma}\sqrt{\frac{n_f}{2}}\left(\overline{X}_E - \overline{X}_S - \delta_a\right),$$

D will be normally distributed with mean $\frac{1}{\sigma}\sqrt{\frac{n_f}{2}}\left(\mu_E - \mu_S - \delta_a\right)$ and variance 1. The test statistic will be written as

$$T_a^* = \frac{D}{\sqrt{\frac{V_1+V_2}{2n_f-4}}}.$$

Under the null hypothesis the joint density of D, V_1, V_2 will be written as

$$f(d,v_1,v_2) = \phi(d)\ g_{\chi^2_{2(n_1-1)}}(v_1)\ g_{\chi^2_{2(n_2-1)}}(v_2).$$

The pooled variance estimator used for sample size recalculation can be written as

$$S_1^2 = \frac{\sigma^2 V_1}{2(n_1 - 1)}.$$

Therefore, the sample size adaptation formula reads

$$\hat{n} = \frac{2(z_{1-\alpha} + z_{1-\beta})^2}{(\theta_a - \delta_a)^2} S_1^2 = \frac{2(z_{1-\alpha} + z_{1-\beta})^2}{(\theta_a - \delta_a)^2} \frac{\sigma^2 V_1}{2(n_1 - 1)} \underset{(5.2)}{=} \frac{n V_1}{2(n_1 - 1)}, \tag{5.4}$$

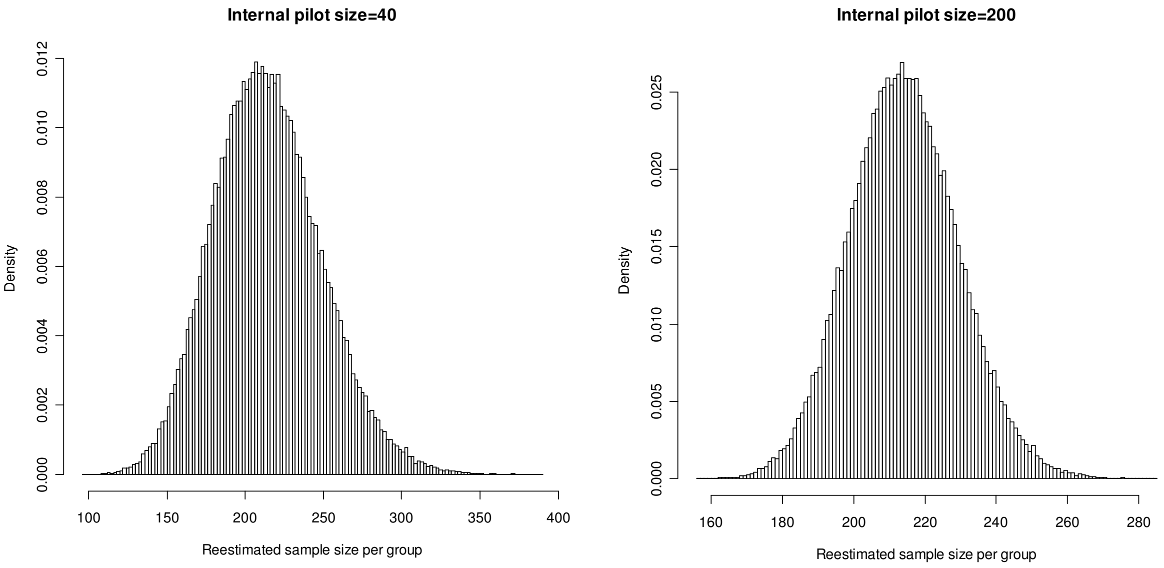
where n denotes the true required sample size per group.

The actual type I error α_a can be calculated by integration of the joint density f over the rejection region of the test.

$$\alpha_a = \int_0^\infty \int_0^\infty \int_{-\infty}^{-\left(t_{2n_f - 4, 1-\alpha}\right)\sqrt{\frac{v_1 + v_2}{2n_f - 4}} + \frac{\delta_a}{\sigma}\sqrt{\frac{n_f}{2}}} f\left(d, v_1, v_2 \mid H_0\right) dd\, dv_2\, dv_1. \tag{5.5}$$

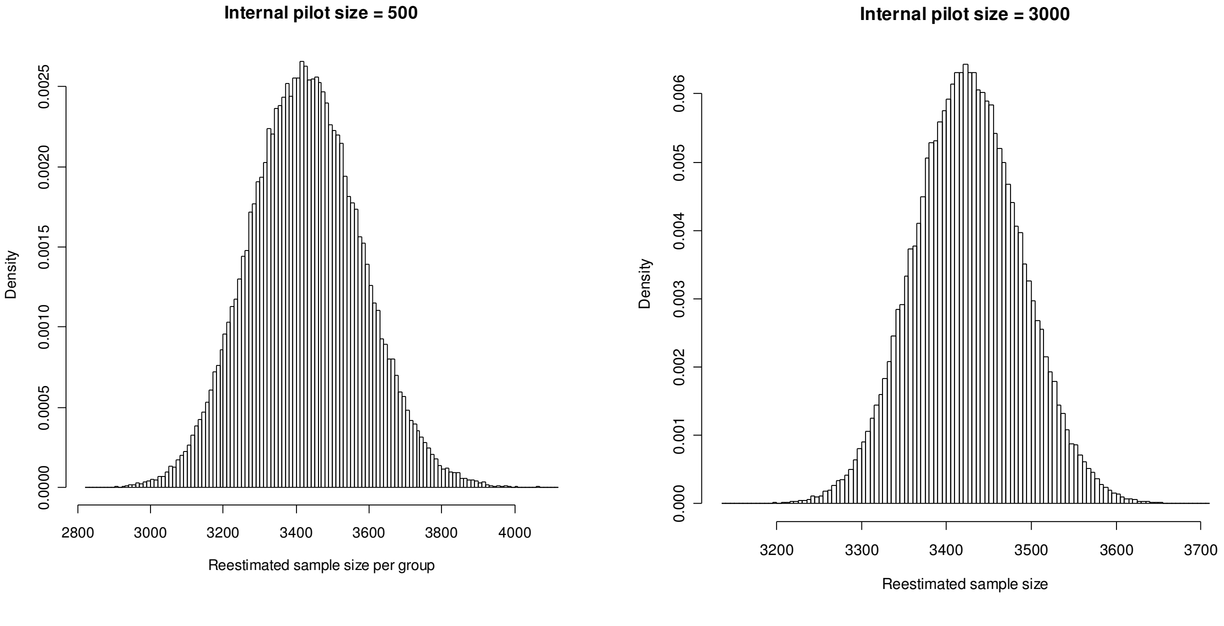
It is a function of α, n_1, n and δ_a / σ. Because α_a is a function of the true required sample size n, it is possible to obtain an upper bound $\alpha_a^{\max}$ by maximising α_a over n(Kieser and Friede, 2000).

Simulation studies have been conducted to assess the distribution of the reestimated sample size per group $\hat{n}$ for different design parameters. The simulation has been run 100000 times in the statistical package R.



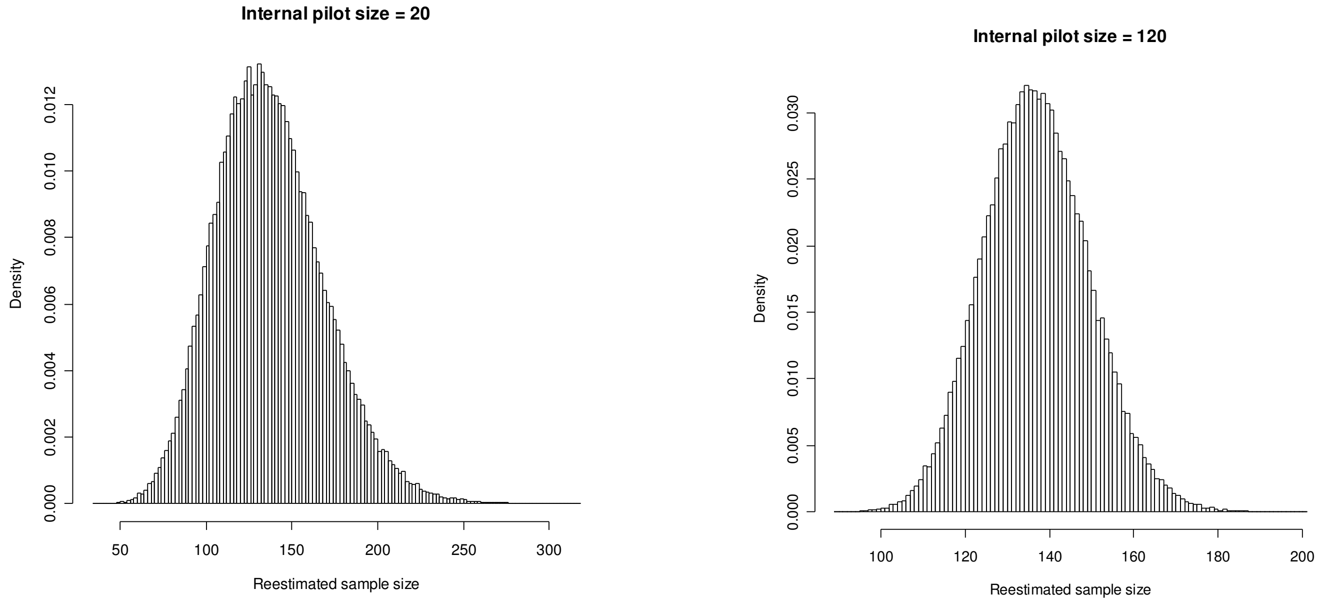
Mean:	214.20	214.20
Std:	34.3419	15.1565
Median:	212.50	213.8

Figure 5.1: *Distribution of* $\hat{n}$ *for* $\alpha = 0.05$, $1-\beta = 0.90$, $\sigma = \sqrt{2}$, $\delta_a = 0.1$, $\tilde{\theta}_a = 0.5$, *that is* $n = 215$



Mean:	3426	3425
Std:	153.1103	62.7463
Median:	3424	3425

Figure 5.2: *Distribution of* $\hat{n}$ *for* $\alpha = 0.05$, $1-\beta = 0.90$, $\sigma = \sqrt{2}$, $\delta_a = 0.4$, $\tilde{\theta}_a = 0.5$, *that is* $n = 3426$



Mean:	137.10	137
Std:	31.3711	12.57982
Median:	134.80	136.60

Figure 5.3: *Distribution of* $\hat{n}$ *for* $\alpha = 0.05$, $1-\beta = 0.90$, $\sigma = \sqrt{2}$, $\delta_a = 1$, $\tilde{\theta}_a = 0.5$, *that is* $n = 137$

Figure 5.1, 5.2, 5.3 shows the results of the simulation of the distribution of the reestimated sample size for different value of the equivalence bound. If we compare the standard deviations for all the cases considered (equivalence bound smaller, bigger or near to the true alternative $\tilde{\theta}_a$), it appears that it is bigger for small internal pilot size than for large ones. The adaptation rule leads to a final sample size near to the true sample size.

One-sample variance procedure

The following discussion is based on the work of Friede and Kieser (2003). The idea is to estimate the variance, ignoring the group assignments.

The one-sample variance calculated from the data merged over the two production groups is given by

$$S_{OS}^2 = \frac{1}{2n_1 - 1}\sum_{j=1}^{2n_1}(X_j - \overline{X})^2,$$

with

$$X_j = \begin{cases} X_{Ej} \text{ if } X_j \in \text{ group E} \\ X_{Sj} \text{ if } X_j \in \text{ group S} \end{cases}.$$

$\overline{X}$ denotes the overall mean of the merged sample.

Considering the fact that the equal sample size in both production groups, S_{OS}^2 can be written as (Friede and Kieser, 2003)

$$S_{OS}^2 = \frac{1}{2n_1 - 1}\left[\sum_{j=1}^{n_1}\left(X_{Ej} - \overline{X}\right)^2 + \sum_{j=1}^{n_1}\left(X_{Sj} - \overline{X}\right)^2\right]$$

$$= \frac{1}{2n_1 - 1}\left(2(n_1 - 1)S_1^2 + \frac{n_1}{2}\left(\overline{X}_{E,1} - \overline{X}_{S,1}\right)^2\right).$$

Using the fact that $\overline{X}_{E,1} - \overline{X}_{S,1}$ is normally distributed with mean θ and variance $2\sigma^2 / n_1$,

$$E(S_{OS}^2) = \sigma^2 + \frac{n_1}{2(2n_1 - 1)}\theta^2 .$$

But in the case of equivalence study, the alternative hypothesis states an irrelevant distance between the groups. Therefore θ tends to zero and S_{OS}^2 tends to be an unbiased estimate of the variance σ^2.

Introducing the variance σ^2, S_{OS}^2 can be written as

$$S_{OS}^2 = \frac{\sigma^2}{2n_1 - 1}\left(\frac{2(n_1 - 1)S_1^2}{\sigma^2} + \frac{n_1}{2\sigma^2}\left(\overline{X}_{E,1} - \overline{X}_{S,1}\right)^2\right).$$

With the same notation as in the case of “pooled sample variance” and noting that

$$D_1 + \sqrt{\frac{n_1}{2}}\frac{\delta_a}{\sigma} = \frac{\left(\overline{X}_{E,1} - \overline{X}_{S,1}\right)}{\sigma}\sqrt{\frac{n_1}{2}},$$

S_{OS}^2 can be expressed by

$$S_{OS}^2 = \frac{\sigma^2}{2n_1 - 1}\left(V_1 + \left(D_1 + \sqrt{\frac{n_1}{2}}\frac{\delta_a}{\sigma}\right)^2\right).$$

Therefore, the sample size adaptation formula reads

$$\hat{n} = \frac{2(z_{1-\alpha} + z_{1-\beta})^2}{(\theta_a - \delta_a)^2} S_{OS}^2 = \frac{2(z_{1-\alpha} + z_{1-\beta})^2}{(\theta_a - \delta_a)^2}\frac{\sigma^2}{2n_1 - 1}\left(V_1 + \left(D_1 + \sqrt{\frac{n_1}{2}}\frac{\delta_a}{\sigma}\right)^2\right)$$

$$= n\ \frac{1}{2n_1 - 1}\left(V_1 + \left(D_1 + \sqrt{\frac{n_1}{2}}\frac{\delta_a}{\sigma}\right)^2\right).$$

The test statistic obtained for all the n_f observation per group is given by

$$T_a^{OS} = \sqrt{\frac{n_f}{2}} \frac{\overline{X}_E - \overline{X}_S - \delta_a}{\tilde{S}_{OS}},$$

with

$$\tilde{S}_{OS}^2 = \frac{1}{2n_f - 1} \sum_{j=1}^{2n_f} (X_j - \overline{X})^2 = \frac{1}{2n_f - 1} \left[\sum_{j=1}^{2n_1} (X_j - \overline{X})^2 + \sum_{j=1}^{2n_2} (X_j - \overline{X})^2 \right].$$

Set

$$V_2^* = V_2 + \frac{n_1 n_2}{\sigma^2 n_f} \left(\left(\overline{X}_{E,1} - \overline{X}_{E,2} \right)^2 + \left(\overline{X}_{S,1} - \overline{X}_{S,2} \right)^2 \right).$$

Then the one-sample variance calculated from the data merged over the production groups can be written as

$$\tilde{S}_{OS}^2 = \sigma^2 \frac{V_1 + V_2^*}{2(n_f - 1)},$$

and the test statistic becomes

$$T_a^* = \frac{\sqrt{\frac{n_1}{n_f}} D_1 + \sqrt{\frac{n_2}{n_f}} D_2}{\sqrt{\frac{V_1 + V_2^*}{2(n_f - 1)}}}.$$

Under the null hypothesis D_1 is normally distributed with expectation zero and variance one. V_1 is χ^2-distributed with $2(n_1 - 1)$ degrees of freedom. Conditional on V_1 and D_1 is V_2^* χ^2-distributed with $2n_2$ degrees of freedom. Conditional on V_1 and D_1 is D_2 normally distributed with expectation zero and variance one. D_1 and V_1 are independent, D_2 and V_2^* are independent given V_1 and D_1. Therefore, under the null hypothesis the joint density of D_1, V_1, D_2, V_2^* can be written as

$$f(d_1, v_1, d_2, v_2^*) = \phi(d_1)\, g_{\chi^2_{2(n_1-1)}}(v_1)\, \phi(d_2)\, g_{\chi^2_{2(n_2-1)}}(v_2^*),$$

where ϕ denotes the density of the standard normal distribution and $g_{\chi^2_n}$ the density of the chi-square distribution with n degrees of freedom.

The actual type I error rate can be computed by integration of the density f over the rejection region of the one-sided test. According to the adjustment procedure described in the section 3, the sample size can or cannot increase after the internal.

If the sample size is not increased, then $n_2 = 0$ and therefore $D_2 = 0,\ V_2^* = 0$.

The test statistic becomes

$$T_a^* = \frac{D_1}{\sqrt{\frac{V_1}{2(n_1-1)}}},$$

and the value of D_1 for with H_0 is rejected after the first stage is

$$a_1 = -t_{2(n_1-1),1-\alpha}\sqrt{\frac{v_1}{2(n_1-1)}} + \frac{\delta_a}{\sigma}\sqrt{\frac{n_1}{2}}.$$

The maximum value of V_1 for which the adjusted sample size per group is n_1, is

$$b_1 = n_1\frac{2n_1-1}{n}.$$

If the sample size increases after the first stage, the value of D_2 for which H_0 is rejected after the second stage is expressed as

$$a_2 = -\sqrt{\frac{n_f}{n_2}}\left(\sqrt{\left(\frac{v_1+v_2^*}{2\left(n_f-1\right)}\right)}\,t_{2\left(n_f-1\right),1-\alpha} - d_1\sqrt{\frac{n_1}{n_f}} + \frac{\delta_a}{\sigma}\sqrt{\frac{n_f}{2}}\right).$$

The remaining part b_2 of the variance not already spent by V_1 is

$$b_2 = \sqrt{b_1 - v_1} - \frac{\delta_a}{\sigma}\sqrt{\frac{n_1}{2}}.$$

The actual type I error rate is therefore given by (Friede and Kieser: 2003)

$$\alpha_a = \int_0^{b_1}\int_{a_1}^{\max(b_2,a_2)}\phi(d_1)g_{\chi^2_{2(n_2-1)}}(v_1)\,dd_1\,dv_1$$

$$+\int_{b_1}^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\int_{-\infty}^{-a_2} f(d_1,v_1,d_2,v_2^*)\,dd_2\,dv_2^*\,dd_1\,dv_1$$

$$+\int_0^{b_1}\int_{-\infty}^{-b_2}\int_0^{\infty}\int_{-\infty}^{-a_2} f(d_1,v_1,d_2,v_2^*)\,dd_2\,dv_2^*\,dd_1\,dv_1$$

$$+\int_0^{b_1}\int_{b_2}^{\infty}\int_0^{\infty}\int_{-\infty}^{-a_2} f(d_1,v_1,d_2,v_2^*)\,dd_2\,dv_2^*\,dd_1\,dv_1.$$

A simulation study has been conducted to assess the distribution of the reestimated sample size per group $\hat{n}$ for different design parameters. The simulation has been run 100000 times using the statistical package R.

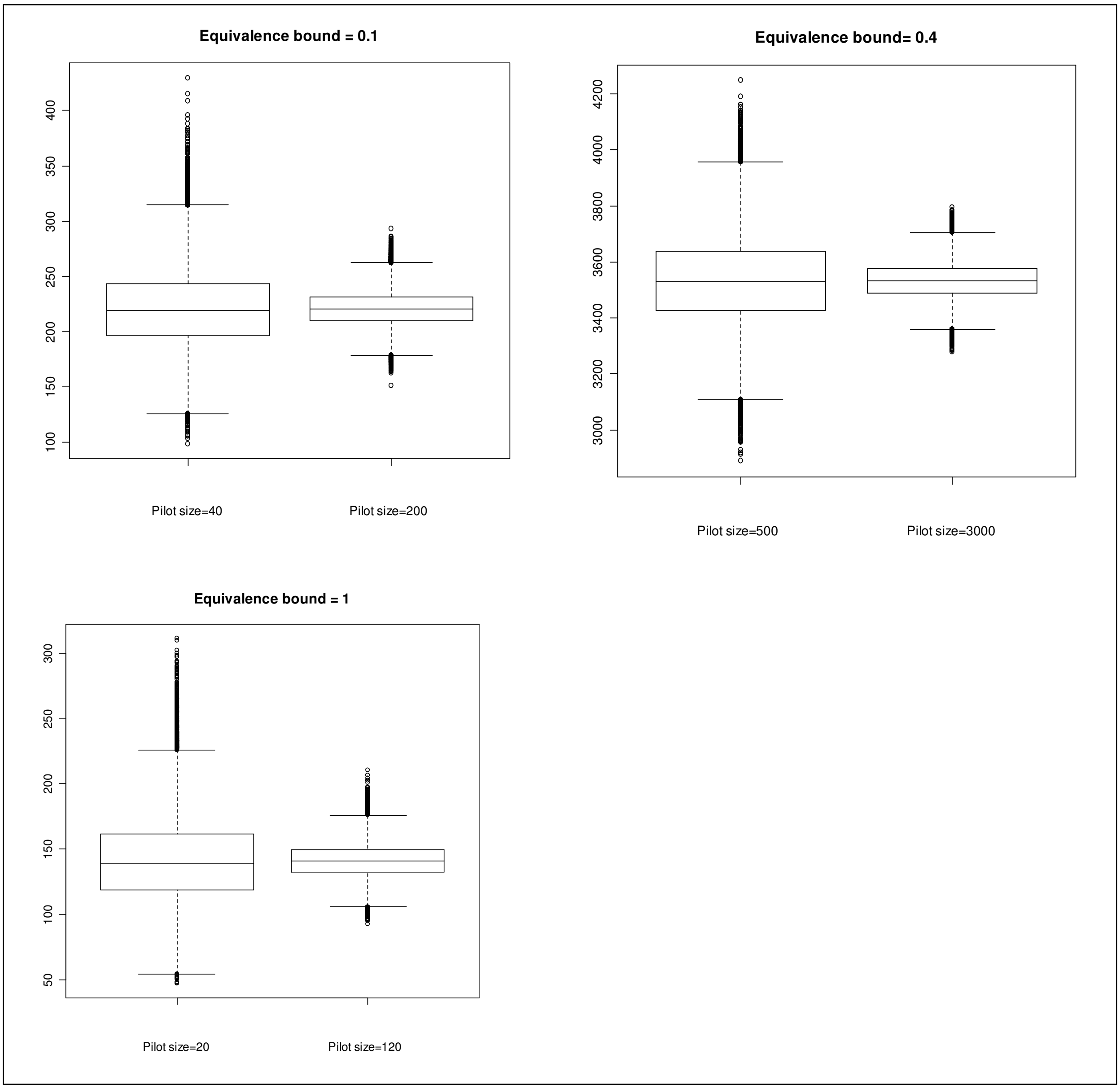


Figure 5.3 : *Distribution of* $\hat{n}$ *for* $\alpha = 0.05$, $1-\beta = 0.90$, $\sigma = \sqrt{2}$ *: Boxplots*

Figure 5.3 shows us the same phenomena as in the case of pooled variance estimate. The standard deviation is bigger for small internal pilot sample size than for large one.

5.1.2 Comparison of the two variance procedures

In other to compare the characteristics of the two variance estimators, we conduct a simulation study with the following parameters: $\alpha = 0.05$, $\beta = 0.1$, $\sigma = \sqrt{2}$, $\delta_a = 0.1, 0.4, 1$,

$\tilde{\theta}_a = 0.5$, and various values of the pilot sample size per group n_1. The simulations were done with the statistical package R and 100000 replications were run for each situation. The following tables (Table 5.1, Table 5.2, Table5.3) show mean standard deviation (SD), mean square error (MSE) of the adjusted sample size for various pilot sample n_1 and various equivalence bound δ_a. The MSE should be kept as small as possible. It is given by the sum of variance and squared bias,

$$MSE = Var + Bias^2 .$$

n_1	**Pooled variance procedure**			**One-sample variance procedure**		
	$\hat{n}$	SD	MSE	$\hat{n}$	SD	MSE
40	214.10	34.27351	1174.676	220.90	35.09456	1278.289
80	214	24.11959	581.761	220.7	24.81638	659.9911
120	214	19.67539	387.1233	220.8	20.24449	454.1485
160	214	16.92681	286.5216	220.7	17.43219	347.9556
200	214	15.20718	231.2657	220.7	15.66028	289.2108

Table 5.1 : *Simulated expected sample size per group* $\hat{n}$*, SD, MSE for pilot size* n_1 *(*$\alpha = 0.05,\ \beta = 0.1,\ \sigma = \sqrt{2},\ \tilde{\theta}_a = 0.5,\ \delta_a = 0.1 \quad \Rightarrow n = 214.0962$*)*

n_1	**Pooled variance procedure**			**One-sample variance procedure**		
	$\hat{n}$	SD	MSE	$\hat{n}$	SD	MSE
500	3426	153.5648	23582.16	3533	158.0700	36491.92
1000	3425	108.4648	11764.68	3532	111.6419	23880.48
1500	3426	88.17925	7775.585	3533	90.88672	19779.84
2000	3426	77.0680	59339.541	3533	79.3399	17819.63
2500	3426	68.4965	4691.803	3533	70.5663	16488.64
3000	3426	62.3672	3889.771	3533	64.2653	15664.56

Table 5.2 : *Simulated expected sample size per group* $\hat{n}$*, SD, MSE for pilot size* n_1 *(*$\alpha = 0.05,\ \beta = 0.1,\ \sigma = \sqrt{2},\ \tilde{\theta}_a = 0.5,\ \delta_a = 0.4 \quad \Rightarrow n = 3425.539$*)*

n_1	**Pooled variance procedure**			**One-sample variance procedure**		
	$\hat{n}$	SD	MSE	$\hat{n}$	SD	MSE
20	137.20	31.5411	994.888	141.60	32.1025	1051.583
40	137	21.9528	481.9282	141.40	22.5074	525.421
60	137.10	17.8133	317.3176	141.40	18.2840	353.2460
80	137.10	15.4949	240.0942	141.40	15.9233	272.4941
100	137	13.7444	188.9108	141.30	14.1386	218.4283
120	137	12.6025	158.8261	141.30	12.9619	186.3363
150	137	11.2306	126.1285	141	11.5739	152.2625

Table 5.3: *Simulated expected sample size per group $\hat{n}$, SD, MSE for pilot size n_1 ($\alpha = 0.05$, $\beta = 0.1$, $\sigma = \sqrt{2}$, $\tilde{\theta}_a = 0.5$, $\delta_a = 1 \Rightarrow n = 137.0216$)*

- The pooled variance procedure leads to mean sample sizes very close to the true sample size even for small pilot sample sizes. It is not the case for the one-sample variance where the mean sample size is larger than the true sample size. This phenomenon is more powerful in the case where the equivalence bound is close to the effect size because evidently we needed more observations to detect equivalence.
- The standard deviation of the pooled variance procedure is the smallest for every choice of the pilot sample size in comparison to the one-sample variance procedure, however the difference becomes smaller and smaller with increasing pilot sample size.
- The pooled variance procedure has a high precision in terms of mean square error.

5.2 Sample size adaptation using the self-designing of Hartung

We considered experiments comparing two production groups with independent normally distributed outcomes with expectations μ_E for the experimental group and μ_S for the standard group, and with unknown but common variance σ^2. We are interested in the one-sided equivalence test problem

$$H_0 : \mu_E - \mu_S \le -\delta_a \quad \text{versus} \quad H_1 : \mu_E - \mu_S > -\delta_a,$$

which is the test problem

$$H_0 : \theta' \leq 0 \qquad \text{versus} \qquad H_1 : \theta' > 0,$$

with $\theta' = \theta + \delta_a$, $\theta = \mu_E - \mu_S$.

The experiment is formally divided into an infinite number of disjoint stage $1, 2, \cdots, l, \cdots$. Only a random finite number of stages will be observed according to a self-designing procedure. n_l specimens are enrolled in the experiment and randomised across the two production groups in the stage l. We denote $\overline{X}_{E,l}$ and $\overline{X}_{S,l}$ the means of the experimental or standard group in the *l*-the stage respectively and the difference of the means in the *l*-the stage $\overline{X}_l = \overline{X}_{E,l} - \overline{X}_{S,l}$. The test statistic T_l of the l the stage is given by

$$T_l = \sqrt{\frac{n_l}{4}} \frac{\overline{X}_l + \delta_a}{S_p^l},$$

where S_p^l denotes the pooled variance estimate in the l the stage.

Under the null hypothesis, T_l is t-distributed with $n_l - 2$ degrees of freedom ($F_{n_l - 2}$), which is a continuous probability distribution. Therefore, the p-values

$$p_l = 1 - F_{n_l - 2}(T_l),$$

where are uniformly distributed on the interval $(0,1)$, such that

$$u_l = \Phi^{-1}(1 - p_l)$$

is normally distributed with mean 0 and variance 1 (Hartung: 2001).

The adaptation procedure follows as in section 4.2.2 with the same notations. The sample size spending function $S_l(\alpha, \beta)$ for stage l is given by

$$S_l(\alpha, \beta) = 4 \frac{(z_\alpha + z_\beta)^2 \hat{S}_p^{l\,2}}{\left(\hat{\theta}_{l-1} + \delta_a\right)^2},$$

where the estimator of the effect size at the end of the *l-1*the stage is

$$\hat{\theta}_{l-1} = \frac{\sum_{j=1}^{l-1} n_j \hat{\hat{\theta}}_j}{\sum_{j=1}^{l-1} n_j},$$

with

$$\hat{\hat{\theta}}_l = 2 \frac{T_l}{\sqrt{n_l}} - \delta_a$$

the estimated effect size in the *l*-the group of data, and the update-estimate of the variance in the *l* the stage is

$$\hat{S}_p^l = \frac{\sum_{j=1}^{l} (n_j - 1) S_p^j}{\sum_{j=1}^{l} n_j - l}.$$

In order to assess the operating characteristics of the proposed adapted-sample design and for comparison with the fixed-sample design, a small simulation study has been conducted for different design parameters. As type I and II error rates we take $\alpha = 0.05$ and $\beta = 0.1$. Then $\delta_a = 0.1$ and we need a sample size of 191 to obtain the desired power at $\theta = 0.5$. The parameters in the proposed adaptation procedure are set as follows: $\alpha_G = 0.05$, $\alpha_L = 0.6$, $\beta_G = 0.1$, $\varepsilon = \sqrt{0.1}$, $\sigma = \sqrt{2}$, $n_1 = N/2$ = 96, where N is the sample size for the fixed-sample design when θ is correctly specified, that is 192, $n_{\min} = n_1/8$, $w_1 = \sqrt{2}/2$. The upper bounds for the sample size spending functions m_l and M_l are set to 190. We simulated the two designs mentioned above 100000 times. Under the null hypothesis $\theta = -\delta_a = -0.1$ the fixed-sample design and the adapted-sample design results in an empirical size of 0.049 and 0.047 respectively. In table 5.4 we put together various power values of the fixed-sample design and the adapted-sample design. The average sample size number (ASN) of the adapted-sample design is also reported.

	Fixed-sample design	Proposed adapted-sample design	
θ	Empirical Power	Empirical Power	ASN
0.2	0.430	0.497	227.066
0.3	0.621	0.688	223.043
0.4	0.788	0.822	208.366
0.5	0.899	0.899	188.511
0.6	0.961	0.936	168.056
0.7	0.988	0.954	149.883
0.8	0.997	0.964	135.802

Table 5.4: *Simulated empirical power and average sample size number for the fixed-sample design and the adapted-sample design*

The adapted-sample design achieves the desired power at $\theta = 0.5$ with a slightly small average sample size number. When the true underlying difference θ is overestimated, the adapted-sample design results in a larger power. The largest difference between the two empirical powers is given at $\theta = 0.3$ with values of 0.621 and 0.688 for the fixed-sample design and the adapted-sample design respectively. These phenomena change when the true underlying difference θ is underestimated and the adapted-sample design results in a smaller power.

6 Sample size adaptation for reliability studies

In many industrial experiments, it is of interest to compare the failure time distribution of different production machines or processes. For censored survival data, linear rank statistics are commonly used for this purpose. These statistics pertain to the survival distribution over the entire follow-up period. With staggered entry and long-term follow-up, partial but increasing information for the experiment becomes available at successive monitoring times. The information is contained not only in the sample size, but also in the number of failures observed during the experiment. Sequential designs for monitoring reliability data are difficult because the martingale structure that underlies most techniques for reliability analysis relates to the internal time scale for units and does not apply to the monitoring time scale. (Oakes, 2001). Recent years have seen a growing interest in combining sequential designs and tests for censored reliability data. The current literature include Tsiatis (1982), Slud and Wei (1982), Tsiatis, Rosner and Tritchler (1985), Lan und Lachin (1990), Lin (1991), Lin, Shen, Ying and Breslow (1996), Lin, Yao and Ying (1999). It is very useful to know the joint distribution of the test statistic at different points of time in order to adjust for the effect of repeated significance testing. In the following, asymptotic joint distribution of the more general efficient scores test (Log-rank test is a special case) for the proportional hazards model calculated at different points in time is presented. This methodology was established by Tsiatis (1981, 1982). With little additional effort, the results of the properties of the test statistics are used for sample size adaptation using a self-designing rule similar to those of Hartung and Knapp (2003).

In subsection 1, we presented the general theory of sequential reliability as introduced by Tsiatis (1981, 1982). In subsection 2, we developed and illustrated a method of updating sample size using a nonparametric linear rank test, based on the idea of Hartung (2001) and Hartung and Knapp (2003).The test statistic is similar to that proposed by Shen and Cai (2003). The accrual duration and the follow-up duration are the parameters on which the strategies for adjustment are based. An example is given to illustrate the adaptation procedure. The performances of the method and the strategies are evaluated and compared to the usual log-rank test with fixed sample design under exponential failure time distribution (proportional hazard). The case of non proportional hazard is discussed in subsection 3 and illustrated with an example under the Weibull failure time distribution.

6.1 General theory of sequential reliability analysis

6.1.1 Notation and Formulas

Let the nonnegative random variable Y denote the real time of entry into the experiment and let the random variable V denote failure time. Let C denote the time from entry to censoring. We assumed that the hazard rate for failure is related to a covariate Z in a log linear fashion,

$$h(x \mid z) = h(x)\exp(\beta z),$$

where $h(x \mid z)$ denotes the hazard rate at time x given that the covariate Z is equal to z. The covariate Z is assumed to be a random variable with finite mean and variance. For the comparison of two machines or processes, the variable Z is equal to either zero or one, representing one machine or another. We wish to test the null hypothesis

$$H_0 : \beta = 0 \qquad \text{or} \qquad H_0 : h(x \mid z) = h(x)$$

for all $x \geq 0$. Under the null hypothesis the survival distribution of V at any time x can be expressed as

$$\exp(-\Lambda(x))$$

and the density as

$$h(x)\exp(-\Lambda(x)),$$

where

$$\Lambda(x) = \int_0^x h(u)\, du$$

denotes the cumulative hazard function.

The time of entry into experiment, Y, is assumed to be a bounded positive random variable with distribution function

$$H(y \mid z) = P(Y \leq y \mid Z = z),$$

which may depend on Z.

The distribution of the time to censoring also depends on Z and is given by

$$G(c \mid z) = P(C < c \mid Z = z).$$

$\overline{G}(c \mid z) = 1 - G(c \mid z))$ denotes the survival distribution. It is assumed that given the covariate Z, the random variables V, Y, C are conditionally independent.

Suppose that the experiment involves n units, which enter serially and are assigned to machines according to some random mechanism. Under the null hypothesis, the data can be expressed as n identically and independently distributed random vectors (V_i, Y_i, C_i, Z_i) for $i = 1, \cdots, n$.

If the data were to be examined at time t, the following variables could be observed:

- Time to failure or censoring

  $X(t) = \max\{\min(V, t-Y, C), 0\}$
- Indicator variable for failure

  $$\Delta(t) = \begin{cases} 1 \text{ if } V < \min(t-Y, C) \\ 0 \qquad \text{otherwise} \end{cases}$$

At time t and under the null hypothesis, the data can be expressed as n identically and independently distributed random vectors $(X_i(t), \Delta_i(t), Z_i)$ for $i = 1, \cdots, n$.

It is clear that at time t, some of the units may yet not have entered the experiment, which is why it will be seen later that the statistics computed at time t will depend only on the data observed up to that time.

The class of tests Θ of testing H_0, which are similar to that in Tarone and Ware (1977) is characterized by statistics of the form (Tsiatis: 1982)

$$S_n(t) = \sum_{i=1}^{n} \hat{Q}(t, X_i(t)) \Delta_i(t) \left\{ Z_i - \frac{\sum\limits_{j \in R(t, X_i(t))} Z_j}{n(t, X_i(t))} \right\},$$

where

$$R(t,x) = \{j \in \{1, \cdots n\} \mid X_j(t) \geq x\}$$

denotes the risk set at time x if the data where observed at real time t, $x \leq t$, and

$$n(t,x) = \sum_{j=1}^{n} I(X_j(t) \geq x),$$

$I(\cdot)$ denoting the indicator function.

The random function $\hat{Q}(t,x)$, $x \leq t$, corresponds to the weighting functions described by Tarone and Ware (1977) and is assumed to converge in probability in sup norm to a function $Q(t,x)$ such that

$$\int_0^t Q^2(t,x)\, h(x) \exp\{-\Lambda(x)\} dx < \infty .$$

For the log-rank test,

$$\hat{Q}(t,x) = Q(t,x) = 1 \text{ for all } t > 0, \quad 0 \leq x \leq t .$$

For the modified Wilcoxon test,

$$\hat{Q}(t,x) = \frac{n(t,x)}{n},$$

which converge in probability to the function

$$\begin{aligned} Q(t,x) = P(X(t) \geq x) &= E\{P(X(t) \geq x) \mid Z\} \\ &= E\{P(V \geq x,\ t - Y \geq x,\ C \geq x) \mid Z\} \\ &= \exp\{-\Lambda(x)\} E\{H(t - x \mid Z)\overline{G}(x \mid Z)\}. \end{aligned}$$

In the following section, the asymptotic distribution of the test statistic will be derived. All calculations are made assuming the null hypothesis is true and therefore the variables V, Y, C, Z are mutually independent.

6.1.2 Asymptotic distribution of the statistic

The key to deriving the joint distribution of the statistic $S_n(t)$ over time is to approximate it by a sum of identically and independently distributed random variables. Because at any time t a positive probability of the failure must be, it is assumed that

$$P\{Y < t,\ V < t - Y,\ V < C\} > 0$$

for any time t that $S_n(t)$ is to calculate.

Denoting

$$N_i(t,x) = I(X_i(t) \leq x,\ \Delta_i(t) = 1),$$

we get

$$\int_0^t dN_i(t,x) = \Delta_i(t)$$

and the statistic $S_n(t)$ can be written as

$$S_n(t) = \sum_{i=1}^{n} \int_0^t dN_i(t,x)\hat{Q}(t,x)\left\{Z_i - \sum_{j \in R(t,x)} \frac{Z_j}{n(t,x)}\right\}.$$

Noting that

$$R(t,x) = \{j \in \{1, \cdots n\} \mid X_j(t) \geq x\}$$

and therefore

$$\sum_{j \in R(t,x)} Z_j = \sum_{i=1}^{n} Z_i\ I(X_i(t) \geq x),$$

The expression of $S_n(t)$ can be rewrite as

$$S_n(t) = \sum_{i=1}^{n} \int_0^t \{dN_i(t,x) - h(x)I(X_i(t) \geq x)dx\}\hat{Q}(t,x)\left\{Z_i - \sum_{j \in R(t,x)} \frac{Z_j}{n(t,x)}\right\}.$$

It can be easily shown by using the law of large numbers that

$$\frac{1}{n}\sum_{j \in R(t,x)} Z_j$$

converges in probability to

$$E\{ZI(X(t) \geq x)\} = E[ZP\{(X(t) \geq x) \mid Z\}]$$

and

$$\frac{n(t,x)}{n}$$

converges in probability to

$$P(X(t) \geq x).$$

Therefore

$$\frac{\sum_{j \in R(t,x)} Z_j}{n(t,x)}$$

converges in probability to

$$\mu(t,x) = \frac{E[ZP\{(X(t) \geq t) \mid Z\}]}{P(X(t) \geq x)}$$

$$= \frac{\exp\{-\Lambda(x)\}E\{ZH(t-x \mid Z)\overline{G}(x \mid Z)\}}{\exp\{-\Lambda(x)\}E\{H(t-x \mid Z)\overline{G}(x \mid Z)\}} = \frac{E\{ZH(t-x \mid Z)\overline{G}(x \mid Z)\}}{E\{H(t-x \mid Z)\overline{G}(x \mid Z)\}}.$$

For fixed t, $N_i(t,x)$ is a counting process with intensity process $h(x)I(X_i(t) \geq x)$.

Therefore $J_i(t,y) = N_i(t,y) - \int_0^y h(u)I(X_i(t) \geq u)du$ is a martingale.

By adding and subtracting similar terms,

$$S_n(t) = \sum_{i=1}^{n} \int_0^t \{dN_i(t,x) - h(x)I(X_i(t) \geq x)dx\}\hat{Q}(t,x)\left\{Z_i - \sum_{j \in R(t,x)} \frac{Z_j}{n(t,x)}\right\}$$

$$= \overline{S}_n(t) + E_n(t),$$

where

$$\overline{S}_n(t) = \sum_{i=1}^{n} \int_0^t dJ_i(t,x)Q(t,x)\{Z_i - \mu(t,x)\}$$

and

$$E_n(t) = -\sum_{i=1}^{n}\int_0^t Q(t,x)dJ_i(t,x)\left\{\sum_{j\in R(t,x)}\frac{Z_j}{n(t,x)}-\mu(t,x)\right\}$$

$$+\sum_{i=1}^{n}\int_0^t \left\{\hat{Q}(t,x)-Q(t,x)\right\}dJ_i(t,x)\{Z_i-\mu(t,x)\}$$

$$-\sum_{i=1}^{n}\int_0^t \left\{\hat{Q}(t,x)-Q(t,x)\right\}dJ_i(t,x)\left\{\sum_{j\in R(t,x)}\frac{Z_j}{n(t,x)}-\mu(t,x)\right\}.$$

It can be shown that

$$\frac{1}{\sqrt{n}}E_n(t)$$

is a second-order term that is asymptotically negligible (Tsiatis: 1981, lemma 3.1 and Breslow and Crowley: 1974, Theorem 4). Hence the asymptotic distribution of

$\frac{S_n(t)}{\sqrt{n}}$ is the same as that of $\frac{\bar{S}_n(t)}{\sqrt{n}}$.

The statistic $\bar{S}_n(t)$ can be written as

$$\bar{S}_n(t)=\sum_{i=1}^{n}\int_0^t dJ_i(t,x)Q(t,x)\{Z_i-\mu(t,x)\}$$

$$=\sum_{i=1}^{n}\int_0^t d\left\{N_i(t,x)-\int_0^x h(u)I(X_i(t)\geq u)\,du\right\}Q(t,x)\{Z_i-\mu(t,x)\}$$

$$=\sum_{i=1}^{n}\left[\int_0^t dN_i(t,x)Q(t,x)\{Z_i-\mu(t,x)\}\right]$$

$$-\sum_{i=1}^{n}\left[-\int_0^t Q(t,x)d\left\{\int_0^x h(u)I(X_i(t)\geq u)\,du\right\}\{Z_i-\mu(t,x)\}\right]$$

$$=\sum_{i=1}^{n}\left[\int_0^t dN_i(t,x)Q(t,x)\{Z_i-\mu(t,x)\}\right]$$

$$-\sum_{i=1}^{n}\left[\int_0^t Q(t,x)h(x)I(X_i(t)\geq x)\,dx\{Z_i-\mu(t,x)\}\right]$$

$$=\sum_{i=1}^{n}\left[\Delta_i(t)Q(t,X_i(t))\{Z_i-\mu(t,X_i(t))\}-\int_0^{X_i(t)}Q(t,x)\{Z_i-\mu(t,x)\}h(x)dx\right].$$

$\overline{S}_n(t)$ is a sum of identically and independently distributed random variables and the asymptotic distribution can be derived by application of the central limit theorem. Therefore the following fundamental theorem can be proved.

<u>Theorem 1:</u> (Tsiatis: 1982, Theorem 3.1): Defining the statistics in the class of tests Θ

$$S_n^{(1)}(t_1) = \sum_{i=1}^{n} \hat{Q}_1(t_1, X_i(t_1))\Delta_i(t_1)\left\{ Z_i - \frac{\sum\limits_{j \in R(t_1, X_i(t_1))} Z_j}{n(t_1, X_i(t_1))} \right\}$$

and

$$S_n^{(2)}(t_2) = \sum_{i=1}^{n} \hat{Q}_2(t_2, X_i(t_2))\Delta_i(t_2)\left\{ Z_i - \frac{\sum\limits_{j \in R(t_2, X_i(t_2))} Z_j}{n(t_2, X_i(t_2))} \right\},$$

where $t_2 \geq t_1$, then the random vector

$$\frac{\{S_n^{(1)}(t_1), S_n^{(2)}(t_2)\}}{\sqrt{n}}$$

converges in distribution to a bivariate normal distribution with mean zero and covariance matrix

$$\Omega = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix},$$

where

$$\sigma_{11} = \int_0^{t_1} Q_1^{\,2}(t_1, x)\Phi(t_1, x)h(x)dx$$

$$\sigma_{22} = \int_0^{t_{21}} Q_2^{\,2}(t_2, x)\Phi(t_2, x)h(x)dx$$

$$\sigma_{12} = \int_0^{t_1} Q_1(t_1, x)Q_2(t_2, x)\Phi(t_1, x)h(x)dx$$

and

$$\Phi(t, x) = \exp\{-\Lambda(x)\}E\left\{\{Z - \mu(t, x)\}^2 H(t - x \mid Z)\overline{G}(x \mid Z)\right\}.$$

Proof

Recalling that the asymptotic distribution of the statistic

$$\frac{S_n(t)}{\sqrt{n}}$$

is the same as

$$\frac{\bar{S}_n(t)}{\sqrt{n}}.$$

The differences

$$\frac{S_n^{(1)}(t_1) - \bar{S}_n^{(1)}(t_1)}{\sqrt{n}} \quad \text{and} \quad \frac{S_n^{(2)}(t_2) - \bar{S}_n^{(2)}(t_2)}{\sqrt{n}}$$

converge in probability to zero. Therefore, the difference between any linear combination of

$$\frac{S_n^{(1)}(t_1)}{\sqrt{n}} \quad \text{and} \quad \frac{S_n^{(2)}(t_2)}{\sqrt{n}}$$

and the same linear combination of

$$\frac{\bar{S}_n^{(1)}(t_1)}{\sqrt{n}} \quad \text{and} \quad \frac{\bar{S}_n^{(2)}(t_2)}{\sqrt{n}}$$

will also converge in probability to zero. By an application of the Cramer-Wold device, the asymptotic joint distribution of

$$\frac{\left\{S_n^{(1)}(t_1), S_n^{(2)}(t_2)\right\}}{\sqrt{n}}$$

will be the same as that of

$$\frac{\left\{\bar{S}_n^{(1)}(t_1), \bar{S}_n^{(2)}(t_2)\right\}}{\sqrt{n}},$$

which is equal to

$$\frac{1}{\sqrt{n}}\left\{\sum_{i=1}^{n}\left[\Delta_i(t_1)Q_1\left(t_1, X_i(t_1)\right)\left\{Z_i - \mu\left(t_1, X_i(t_1)\right)\right\} - \int_0^{X_i(t_1)} Q_1\left(t_1, x\right)\left\{Z_i - \mu\left(t_1, x\right)\right\}h(x)dx\right],\right.$$

$$\left.\sum_{i=1}^{n}\left[\Delta_i(t_2)Q_2\left(t_2, X_i(t_2)\right)\left\{Z_i - \mu\left(t_2, X_i(t_2)\right)\right\} - \int_0^{X_i(t_2)} Q_2\left(t_2, x\right)\left\{Z_i - \mu\left(t_2, x\right)\right\}h(x)dx\right]\right\}.$$

This quantity is a normalized sum of identically and independently distributed random variables; which by application of the multivariate central limit theorem converge in distribution to a bivariate normal distribution. The next step of the proof is to calculate the first and second moment of the random vector

$$\left\{\left[\Delta(t_1)Q_1(t_1,X(t_1))\{Z-\mu(t_1,X(t_1))\}-\int_0^{X(t_1)}Q_1(t_1,x)\{Z-\mu(t_1,x)\}h(x)dx\right],\right.$$

$$\left.\left[\Delta(t_2)Q_2(t_2,X(t_2))\{Z-\mu(t_2,X(t_2))\}-\int_0^{X(t_2)}Q_2(t_2,x)\{Z-\mu(t_2,x)\}h(x)dx\right]\right\}.$$

The appropriate expectations shall be calculated by finding the mean of the conditional expectations with respect to Z because the entry rate and censoring distribution may depend on Z. Defining

$$\overline{Q}_i(x)=Q_i(t_i,x)\{z-\mu(t_i,x)\}. \qquad i=1,2,$$

the conditional expectation, given $Z=z$

- $E\left[\left\{\Delta(t)\overline{Q}_1(X(t))-\int_0^{X(t)}\overline{Q}_1(x)h(x)dx\right\}\mid Z=z\right]=0$

- $E\left[\left\{\Delta(t_1)\overline{Q}_1(X(t_1))-\int_0^{X(t_1)}\overline{Q}_1(x)h(x)dx\right\}\left\{\Delta(t_2)\overline{Q}_2(X(t_2))-\int_0^{X(t_2)}\overline{Q}_2(x)h(x)dx\right\}\mid Z=z\right]$

$$=\int_0^{t_1}\overline{Q}_1(x)\overline{Q}_2(x)h(x)\exp\{-\Lambda(x)\}H(t_1-x\mid Z)\overline{G}(x\mid z)dx.$$

In fact,

$$E[\{\Delta(t)\overline{Q}_1(X(t))\}\mid Z=z]=\int_0^{t}\overline{Q}_1(x)h(x)\exp\{-\Lambda(x)\}H(t-x\mid Z)\overline{G}(x\mid z)dx.$$

$$E\left[\int_0^{X(t)}\overline{Q}_1(x)h(x)dx\mid Z=z\right]=-\int_0^t\left\{\int_0^x\overline{Q}_1(u)h(u)du\right\}d\{P(X(t)\geq x)\}$$

$$=-\int_0^t\left\{\int_0^x\overline{Q}_1(u)h(u)du\right\}d[\{\exp(-\Lambda(x))\}H(t-x\mid z)\overline{G}(x\mid z)].$$

Integrating this quantity by parts we obtain

$$E\left[\int_0^{X(t)} \overline{Q}_1(x)h(x)dx \mid Z=z\right] = -\left[\int_0^{x} \overline{Q}_1(u)h(u)du \cdot \{\exp(-\Lambda(x))\} H(t-x \mid z)\overline{G}(x \mid z)\right]_0^t$$

$$+\int_0^t \overline{Q}_1(x)h(x)\exp\{-\Lambda(x)\}H(t-x \mid Z)\overline{G}(x \mid z)dx\,.$$

But

$$\left[\int_0^{x} \overline{Q}_1(u)h(u)du \cdot \{\exp(-\Lambda(x))\} H(t-x \mid z)\overline{G}(x \mid z)\right]_0^t = 0$$

because

$$H(0 \mid z) = P(Y \le 0 \mid Z=z) = 0 \qquad \text{and} \qquad \int_0^0 \overline{Q}_1(u)h(u)du = 0\,.$$

Therefore

$$E\left[\int_0^{X(t)} \overline{Q}_1(x)h(x)dx \mid Z=z\right] = \int_0^t \overline{Q}_1(x)h(x)\exp\{-\Lambda(x)\}H(t-x \mid Z)\overline{G}(x \mid z)dx$$

$$= E\left[\{\Delta(t)\overline{Q}_1(X(t))\} \mid Z=z\right]$$

and the first point is established.

For the second point,

$$E\left(\left\{\Delta(t_1)\overline{Q}_1(X(t_1)) - \int_0^{X(t_1)} \overline{Q}_1(x)h(x)dx\right\}\left\{\Delta(t_2)\overline{Q}_2(X(t_2)) - \int_0^{X(t_2)} \overline{Q}_2(x)h(x)dx\right\} \mid Z=z\right)$$

$$= E\{\Delta(t_1)\overline{Q}_1(X(t_1))\Delta(t_2)\overline{Q}_2(X(t_2)) \mid Z=z\}$$

$$-E\left\{\Delta(t_1)\overline{Q}_1(X(t_1)) \int_0^{X(t_2)} \overline{Q}_2(x)h(x)dx \mid Z=z\right\}$$

$$-E\left\{\Delta(t_2)\overline{Q}_2(X(t_2)) \int_0^{X(t_1)} \overline{Q}_1(x)h(x)dx \mid Z=z\right\}$$

$$+E\left(\left\{\int_0^{X(t_1)} \overline{Q}_1(x)h(x)dx\right\}\left\{\int_0^{X(t_2)} \overline{Q}_2(x)h(x)dx\right\} \mid Z=z\right).$$

$$E\{\Delta(t_1)\overline{Q}_1(X(t_1))\Delta(t_2)\overline{Q}_2(X(t_2)) \mid Z=z\} = E\{\Delta(t_1)\overline{Q}_1(X(t_1))\overline{Q}_2(X(t_2)) \mid Z=z\}$$

$$= \int_0^{t_1} \overline{Q}_2(x)\overline{Q}_1(x)h(x)\exp\{-\Lambda(x)\}H(t_1-x \mid Z)\overline{G}(x \mid z)dx\,.$$

Denoting for simplification of notation

$$\int_0^x \overline{Q}_i(u)h(u)du = \psi_i(x),$$

$$E\left\{\Delta(t_1)\overline{Q}_1(X(t_1))\int_0^{X(t_2)}\overline{Q}_2(x)h(x)dx \mid Z = z\right\} = \int_0^{t_1}\psi_2(x)\overline{Q}_1(x)h(x)\exp\{-\Lambda(x)\}H(t_1 - x \mid Z)\overline{G}(x \mid z)dx$$

$$E\left\{\Delta(t_2)\overline{Q}_2(X(t_2))\int_0^{X(t_1)}\overline{Q}_1(x)h(x)dx \mid Z = z\right\}$$

is evaluated by computing the integral in two regions, namely when

$$\{\Delta(t_1) = 1\} \qquad \text{and} \qquad \{\Delta(t_1 = 0),\ \Delta(t_2) = 1\},$$

therefore is equal to

$$\int_0^{t_1}\psi_1(x)\overline{Q}_2(x)h(x)\exp\{-\Lambda(x)\}H(t_1 - x \mid Z)\overline{G}(x \mid z)dx$$

$$+\int_0^{t_1}\int_{t_1-y}^{t_{2-y}}\psi_1(t_1 - y)\overline{Q}_2(x)h(x)\exp\{-\Lambda(x)\}\overline{G}(x \mid z)dxdH(y \mid z).$$

$$E\left(\left\{\int_0^{X(t_1)}\overline{Q}_1(x)h(x)dx\right\}\left\{\int_0^{X(t_2)}\overline{Q}_2(x)h(x)dx\right\} \mid Z = z\right)$$

is computed in three parts, namely when

$$\{\Delta(t_1) = 1\}, \quad \{\Delta(t_1 = 0),\ \Delta(t_2) = 1\}, \qquad \{\Delta(t_2) = 0\}.$$

The region $\{\Delta(t_2) = 0\}$ can be further divided into three sub regions:

1. $\{C < t_1 - Y,\ V > C,\ \text{for}\ 0 \le C \le t_1\}$
2. $\{t_1 - Y < C < t_2 - Y,\ V > C\}$
3. $\{C > t_2 - Y,\ V > t_2 - Y\}$.

Therefore,

$$E\left(\left\{\int_0^{X(t_1)}\overline{Q}_1(x)h(x)dx\right\}\left\{\int_0^{X(t_2)}\overline{Q}_2(x)h(x)dx\right\} \mid Z = z\right)$$

$$= \int_0^{t_1}\psi_2(x)\psi_1(x)h(x)\exp\{-\Lambda(x)\}H(t_1 - x \mid z)\overline{G}(x \mid z)dx$$

$$+\int_0^{t_1}\int_{t_1-y}^{t_{2-y}}\psi_1(t_1 - y)\psi_2(x)h(x)\exp\{-\Lambda(x)\}\overline{G}(x \mid z)dxdH(y \mid z)$$

$$-\int_0^{t_1} \psi_2(x)\psi_1(x)\exp\{-\Lambda(x)\}H(t_1 - x \mid z)\, d\overline{G}(x \mid z)$$

$$-\int_0^{t_1}\int_{t_1-y}^{t_2-y} \psi_1(t_1 - y)\psi_2(x)\exp\{-\Lambda(x)\}\, d\overline{G}(x \mid z)\, dH(y \mid z)$$

$$+\int_0^{t_1} \psi_1(t_1 - y)\psi_2(t_2 - y)\exp\{-\Lambda(t_2 - y)\}\overline{G}(t_2 - y \mid z)\, dH(y \mid z).$$

After some long calculations including integration by parts,

$$-E\left\{\Delta(t_1)\overline{Q}_1(X(t_1))\int_0^{X(t_2)}\overline{Q}_2(x)h(x)dx \mid Z = z\right\} - E\left\{\Delta(t_2)\overline{Q}_2(X(t_2))\int_0^{X(t_1)}\overline{Q}_1(x)h(x)dx \mid Z = z\right\}$$

$$+E\left(\left\{\int_0^{X(t_1)}\overline{Q}_1(x)h(x)dx\right\}\left\{\int_0^{X(t_2)}\overline{Q}_2(x)h(x)dx\right\} \mid Z = z\right)$$

is equal to zero and the second point is established.

The covariance σ_{12} is obtained by finding the expectation of

$$\int_0^{t_1}\overline{Q}_1(x)\overline{Q}_2(x)h(x)\exp\{-\Lambda(x)\}H(t_1 - x \mid Z)\overline{G}(x \mid z)dx.$$

This yields

$$\sigma_{12} = \int_0^{t_1} Q_1(t_1, x)Q_2(t_2, x)E\left[\{Z - \mu(t_1, x)\}\{Z - \mu(t_2, x)\}H(t_1 - x \mid z)\overline{G}(x \mid z)\right]h(x)\exp\{-\Lambda(x)\}dx$$

$$= \int_0^{t_1} Q_1(t_1, x)Q_2(t_2, x)E\left[\{Z - \mu(t_1, x)\}^2 H(t_1 - x \mid z)\overline{G}(x \mid z)\right]h(x)\exp\{-\Lambda(x)\}dx$$

$$+\int_0^{t_1} Q_1(t_1, x)Q_2(t_2, x)\{\mu(t_1, x) - \mu(t_2, x)\}$$

$$\times E\left[\{Z - \mu(t_1, x)\}H(t_1 - x \mid z)\overline{G}(x \mid z)\right]h(x)\exp\{-\Lambda(x)\}dx.$$

Using the fact that

$$\mu(t, x) = \frac{E\{ZH(t - x \mid Z)\overline{G}(x \mid Z)\}}{E\{H(t - x \mid Z)\overline{G}(x \mid Z)\}},$$

$$\int_0^{t_1} Q_1(t_1,x)Q_2(t_2,x)\{\mu(t_1,x)-\mu(t_2,x)\}$$

$$\times E\left[\{Z-\mu(t_1,x)\}H(t_1-x\mid z)\overline{G}(x\mid z)\right]h(x)\exp\{-\Lambda(x)\}dx \quad = \quad 0 .$$

Therefore

$$\sigma_{12} = \int_0^{t_1} Q_1(t_1,x)Q_2(t_2,x)E\left[\{Z-\mu(t_1,x)\}^2 H(t_1-x\mid z)\overline{G}(x\mid z)\right]h(x)\exp\{-\Lambda(x)\}dx .$$

σ_{11} and σ_{22} are special cases of σ_{12}. This completes the proof of the theorem. □

σ_{12} can be estimated by (Tsiatis: 1982)

$$\frac{1}{n}\left[\sum_{i=1}^{n}\left(\Delta_i(t_1)\hat{Q}_1(t_1,X_i(t_1))\hat{Q}_2(t_2,X_i(t_1))\sum_{j\in R(t_1,X_i(t_1))}\frac{\{Z_j-\hat{\mu}(t_1,X_i(t_1))\}^2}{n(t_1,X_i(t_1))}\right)\right],$$

where

$$\hat{\mu}(t_1,x) = \sum_{j\in R(t_1,x)}\frac{Z_j}{n(t_1,x)} .$$

In fact, σ_{12} can be written as

$$\int_0^{t_1} Q_1(t_1,x)Q_2(t_2,x)E\left[\{Z-\mu(t_1,x)\}^2 I(X(t_1)\geq x)\right]d\Lambda(x).$$

A consistent estimate of σ_{12} can be obtained by replacing the quantities in this expression by their appropriate estimates yielding to

$$\int_0^{t_1} \hat{Q}_1(t_1,x)\hat{Q}_2(t_2,x)\hat{E}\left[\{Z-\hat{\mu}(t_1,x)\}^2 I(X(t_1)\geq x)\right]d\hat{\Lambda}(x),$$

where

$$\hat{E}\left[\{Z-\hat{\mu}(t_1,x)\}^2 I(X(t_1)\geq x)\right] = \sum_{j\in R(t_1,x)}\frac{\{Z_j-\hat{\mu}(t_1,x)\}^2}{n}$$

and

$$\hat{\Lambda}(x) = \int_0^{x}\frac{dN_i(t_1,u)}{n(t_1,u)}$$

is the estimate of the cumulative hazard function given by Nelson (1969).

The random vector

$$\frac{1}{\sqrt{n}}\left\{S_n^{\ i}(t_i);\ i=1,\cdots,k\right\}$$

will converge to a multivariate normal with mean zero and covariance matrix with elements

$$\sigma_{ij};\ i,j=1,\cdots,k,$$

that can be estimated by $\hat{\sigma}_{ij}$. It can be shown using a multivariate version of Slutsky's theorem (Shorack, 2000), that the normalized score statistics

$$\left\{Z_i=\frac{1}{\sqrt{n}}\frac{S_n^{(i)}(t_i)}{\sqrt{\hat{\sigma}_{ii}}};\ i=1,\cdots,k\right\}$$

will converge in distribution to a multivariate normal with mean zero and covariance matrix with element that can be estimated by (Tsiatis, 1982)

$$\frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}};\ i,j=1,\cdots,k.$$

An important remark given in Tsiatis (1982) is the fact that if the weighting function $\hat{Q}(t,x)$ converge to a function $Q(x)$ independent of t, then $\sigma_{12}=\sigma_{11}$, which indicates that the process $S_n(t)$ has asymptotically independent increments.

Theorem 1 can help us to construct adaptive test by using test statistic within the class of Θ at time determinate by observing predetermined number of failures.

6.2 Adaptation under proportional hazard

6.2.1 Test statistic

Let $h_E(t)$ be the hazard function for the experimental group and $h_S(t)$ be the hazard function for the standard group. Assume that we have a proportional hazard model with constant log ratio

$$\theta=\ln\left(\frac{h_E(t)}{h_S(t)}\right).$$

A one-sided test will be performed with a significance level of α to detect a log-hazard ratio of $\tilde{\theta}>0$. The null hypothesis to be tested is

$$H_0:\theta=0 \qquad \text{against} \qquad H_1:\theta>0.$$

The study is characterized by an accrual period and a follow-up period. Subjects enter the experiment sequentially and are allocated symmetrically to the two groups and are followed until either they fail or are administratively censored at the end of the follow-up period. In these studies we distinguish theoretically between two time scales:

- Calendar time is the usual measure of time relative to some fixed event such as the start of the experiment.
- Follow-up time is the time of each unit, measured individually from entry until it either fails or is administratively censored.

We intend to monitor the experiment at calendar time t_l, $l = 1, 2, \cdots, L$. The study is formally divided into L **disjoint** study stages. After observing every $f_l \leq K_l$ failures during the lth stage or the *l*th observation period $(t_{l-1}, t_l]$, where $t_0 = 0$, $l = 1, 2, \cdots, L$, a test statistic $S_n(t_l)$ is computed, based on all cumulated data up to time t_l. Using the same notation as in the last section, the score statistic at calendar time t_l, $S_n(t_l)$ is given by

$$S_n(t_l) = \sum_{i=1}^{n} \Delta_i(t_l) \left\{ Z_i - \frac{\sum\limits_{j \in R(t_l, X_i(t_l))} Z_j}{n(t_l, X_i(t_l))} \right\}, \tag{6.1}$$

which is a special case of the general class of test statistic Θ given in the last section

$$S_n(t_l) = \sum_{i=1}^{n} \hat{Q}(t_l, X_i(t_l)) \Delta_i(t_l) \left\{ Z_i - \frac{\sum\limits_{j \in R(t_l, X_i(t_l))} Z_j}{n(t_l, X_i(t_l))} \right\},$$

with

$$\hat{Q}(t, x) = Q(t, x) = 1 \text{ for all } t > 0, \quad 0 \leq x \leq t.$$

From the preceding section, we know that the random vector

$$\frac{\{S_n(t_l), S_n(t_k)\}}{\sqrt{n}} \qquad \text{for} \qquad l \neq k$$

converges in distribution to a bivariate normal distribution with mean zero and covariance matrix with elements σ_{lk} and that the process $S_n(t)$ has asymptotically independent increments.

Gail, DeMets and Slud (1982) verified by simulations that under the null hypothesis, $S_n(t_l)$ given by the formula (6.2.1) (score log-rank statistic) has asymptotically independent increments. The extension to $\theta \neq 0$ is based on arguing that under local alternatives (θ close to zero), the variance and covariance structure of the $S_n(t_l)$ process will be preserved. Therefore for any $l \neq k$,

$$\operatorname{cov}(S_n(t_l) - S_n(t_{l-1}),\ S_n(t_k) - S_n(t_{k-1}))$$

converges to zero .

This assumption might deteriorate if θ is far from zero.

Thus, under the null hypothesis, the distribution of

$$\frac{(S_n(t_l) - S_n(t_{l-1}))}{\sqrt{n}}$$

converges to a normal distribution with mean zero and variance $\sigma_l^{\ 2} - \sigma_{l-1}^{\ 2}$, where

$$\sigma_l^{\ 2} = \sigma_{ll}.$$

We considered the statistic (Shen and Cai, 2003)

$$T_l = T(t_l) = \frac{1}{\sqrt{n}} \frac{\{S_n(t_l) - S_n(t_{l-1})\}}{\sqrt{\hat{\sigma}_l^{\ 2} - \hat{\sigma}_{l-1}^{\ 2}}}, \quad l = 1, 2, \cdots,$$

where $\hat{\sigma}_l^{\ 2}$ is a consistent estimate of $\sigma_l^{\ 2}$ as given in the last section.

T_l can express the standardized information cumulated during (t_{l-1}, t_l).

Theorem 2

Under the null hypothesis T_l and T_k are asymptotical independent for any $l \neq k$.

Note that a direct application of the Slutsky's theorem shows that T_l is asymptotically normally distributed with mean zero and variance 1. Now we have to prove that the random vector $\{T_l, T_k\}$ converges in distribution to a standard bivariate normal distribution.

The proof follows by application of theorem 1 with the fact that, $\text{cov}(T_l, T_k) \approx 0$ for any $l \neq k$.

Using the fact that $\hat{\sigma}_l^{\ 2}$ can be approximated by (Tsiatis et al., 1985)

$$\hat{\sigma}_Z^{\ 2}(t_l)\frac{d_l}{n},$$

where d_l denotes the expected number of failures at time t_l and $\hat{\sigma}_Z^{\ 2}(t_l)$ the empirical variance of the indicator variable Z as observed up to time t_l, the standardized test statistic can be rewritten as

$$T_l = \frac{S_n(t_l) - S_n(t_{l-1})}{\sqrt{d_l \hat{\sigma}_Z(t_l) - d_{l-1}\hat{\sigma}_Z(t_{l-1})}}, \qquad l = 1, 2, \cdots, L, \tag{6.2}$$

Because in reliability analysis, the number of failures carry primary information, the adaptation procedure is based on the number of failures, rather than on the actual simple size. Using ideas similar to those of Hartung and Knapp (2003), the adaptation procedure is as follows:

Under the null hypothesis, theT_l have a continuous distribution function $F_{l,0}$. Therefore, the p-values

$$p_l = 1 - F_{l,0}(T_l), \qquad l = 1, 2, \cdots, L$$

are uniformly distributed on the interval $(0,1)$, such that

$$u_l(\nu_l) = F^{-1}_{\chi^2(\nu_l)}(1 - p_l) \tag{6.3}$$

is chi-square distributed with ν_l degrees of freedom ($\chi^2(\nu_l)$) (Hedges and Olkin, 1985), where $F^{-1}_{\chi^2(\nu_l)}$ denotes the inverse of the standard chi-square distribution function with ν_l degrees of freedom and $u_l(\nu_l)$ is the $(1 - p_l)$-quantile of the $\chi^2(\nu_l)$-distribution.

The final test statistic up to stage l has the form

$$U_l = \sum_{j=1}^{l} u_j(\nu_j), \tag{6.4}$$

which under the null hypothesis H_0 follows a $\chi^2(\nu_\Sigma(l))$-distribution,

$$\nu_\Sigma(l) = \sum_{j=1}^{l} \nu_j \quad l = 1, 2, \cdots, L. \tag{6.5}$$

It is clear that$u_j(\nu_j) \geq 0$, $\forall\ j$ and therefore$\forall\ l^* \in \{1, 2, \cdots, L\}\quad U_{l^*} \leq U_L$. If

$$U_{l^*} \geq \chi^2(\nu_\Sigma(L))_{1-\alpha},$$

the null hypothesis is rejected at level α and the study can be stopped after stage l^* $\chi^2(\nu_\Sigma(L))_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the $\chi^2(\nu_\Sigma(L))$-distribution.

6.2.2 Adaptive designing

We begin the study by defining the total degrees of freedom to be available in the whole sequential trial

$$\nu_\Sigma(L) = \sum_{j=1}^{L} \nu_j .$$

The global critical value is defined as $\chi^2(\nu_\Sigma(L))_{1-\alpha}$ and the minimum number of degrees of freedom,$\nu_{\min}$ assigned to each stage, which will be realized, is defined.

For easy of representation, $\nu_{\min} = 1$ and$\nu_\Sigma(L) = L$. Therefore, the global critical value is

$$cv_\alpha = \chi^2(L)_{1-\alpha} .$$

The available degrees of freedom is divided into ν_1 and ν_2^* degrees of freedom in the first step with

$$1 \le \nu_1 < L \qquad \text{and} \qquad \nu_2^* = L - \nu_1 ,$$

so that under H_0

$$u_1(\nu_1) + u_2(\nu_2^*) \sim \chi^2(L).$$

ν_1 is the prespecified non-random degrees of freedom of the first stage and the null hypothesis H_0 is rejected at level α and the experiment stops if

$$U_1 = u_1(\nu_1) \ge cv_\alpha .$$

Otherwise, the experiment will continue.

If $\nu_2^* = 1$, we set $\nu_2 = \nu_2^* = 1$ (because of $\nu_{\min} = 1$) and the experiment definitely stops after the second stage. If $\nu_2^* \ge 2$, it can be divided into two parts ν_2 and ν_3^* with

$$\nu_2 \ge 1 \qquad \text{and} \qquad \nu_3^* = \nu_2^* - \nu_2$$

so that under H_0

$$u_1(\nu_1) + \{u_2(\nu_2) + u_3(\nu_2^*)\} \sim \chi^2(L).$$

ν_2 is the degrees of freedom assigned the second stage and the null hypothesis H_0 is rejected at level α and the experiment stops if

$$U_2 = u_1(\nu_1) + u_2(\nu_2) \ge cv_\alpha .$$

Otherwise, the experiment will continue.

If $\nu_3^* = 1$, we set $\nu_3 = \nu_3^* = 1$ and the experiment definitely stops after the third stage. If $\nu_3^* \ge 2$, it can be divided again into two parts ν_3 (degrees of freedom assigned to the third stage) and ν_4^* with

$$\nu_3 \ge 1 \qquad \text{and} \qquad \nu_4^* = \nu_3^* - \nu_3 ,$$

and so on.

After the $(l-1)$, the following scheme is obtained:

$$u_1(\nu_1) + u_2(L - \nu_1) \sim \chi^2(L) \quad \text{under } H_0$$

$$u_1(\nu_1) + [u_2(\nu_2) + u_3(L - \nu_1 - \nu_2)] \sim \chi^2(L) \quad \text{under } H_0$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$u_1(\nu_1) + \left[u_2(\nu_2) + \left\{ u_3(\nu_3) + \left(\cdots + \left[u_{l-1}(\nu_{l-1}) + u_l\left(L - \sum_{j=1}^{l-1} \nu_j \right) \right] \right) \right\} \right] \sim \chi^2(L) \text{ under } H_0$$

with

$$L-\sum_{j=1}^{l-1}\nu_j \geq 1 \text{ and } \quad \nu_j \geq 1.$$

Let us introduce the notation $a_l = \hat{a}_{l-1}$, which indicates that a_l is determinate or estimated upon knowledge of all the information obtained in the previous study stages before the beginning of stage l. Then $\nu_l = \hat{\nu}_{l-1}$.

The adaptation procedure is based on the number of failures, not the number of units because it is the first one which determines the power of the study. Given the a priori fixed number of failures of the first stage f_1, the additional number of failures in the lth stage $d_l - d_{l-1} = f_l$ is determined upon knowledge of the previous study stages, $f_l = \hat{f}_{l-1}$. Under the null hypothesis the distribution of the p-values p_l and the independence of p_l and p_k, $l \neq k$, still hold provided the continuity of the distribution of the test statistics. If the distribution of the test statistic under H_0 is not continuous, in order to ensure that the combination test does not exceed the pre-chosen type I error rate, the distribution of p_1 and the conditional distributions of p_l given $(p_1, \cdots, p_{l-1})$ have to be stochastically larger than the uniform distribution. (Brannath, Posch and Bauer, 2002)

In the following, an algorithm for determining the degrees of freedom and the additional number of failures in each stage based on the available knowledge prior to the stage that will be performed is given.

Denoting S_l the sample size spending function used after stage $(l-1)$, the minimum additional number of failures M_l for stage l (will be the last stage), holding the given power $1-\beta$ and type I error rate α conditionally on the results of the previous study parts is defined by Hartung and Knapp (2003)

$$M_l = S_{l-1}\left(1-F_{\chi^2\{L-\nu_\Sigma(l-1)\}}\left(cv_\alpha - U_{l-1}\right), \beta\right). \tag{6.6}$$

M_l is the additional number of failures to achieve a conditional power of $1-\beta$ because, in the conditional error function

$$1-F_{\chi^2\{L-\nu_\Sigma(l-1)\}}\left(cv_\alpha - U_{l-1}\right),$$

the whole remaining degrees of freedom is used.

At each stage, the parameter of interest θ will be estimated based on the knowledge of all the previous study parts as well as update estimates of other parameters like variances, $\theta_l = \hat{\theta}_{l-1}$.

These parameters are directly or indirectly involved in S_{l-1} and may not yet have stabilized. Therefore, only a part of M_l should be used as additional number of failures, that is

$$f_l = \varepsilon_l \cdot M_l, \quad \text{with} \quad 0 < \varepsilon_l < 1. \tag{6.7}$$

Hartung and Knapp (2003) proposed as degrees of freedom associated to stage l

$$\nu_l = \{L - \nu_\Sigma(l-1)\} \cdot \frac{f_l}{M_l} = \{L - \nu_\Sigma(l-1)\} \cdot \varepsilon_l . \tag{6.8}$$

The sequence $\{\varepsilon_l\}$ may be defined before the beginning of the study and is defined as

$$\{\varepsilon_l\} = \left\{ \frac{M_l}{m_l} \right\}, \tag{6.9}$$

where

$$m_l = S_{l-1}\left(1 - F_{\chi^2\{L-\nu_\Sigma(l-1)\}}(cv_\alpha - U_{l-1}), \beta_g\right) \tag{6.10}$$

for a fixed type II error rate β_g, which is larger than β. In a similar way is the power spending approach discussed in Bauer (1992). In other to adjust the effect of a too large chosen β_g, one may choose a minimum number of failure $f_{\min}$ to be observed in each stage, because a too large β_g increases the number of stages to be performed.

The adaptation procedure can be summarized as follows:

6. Define the type I and II error rates α and β, the type II error rate β_g for generating the sequential number of observed failures f_l, the minimum number of observed failure in each stage $f_{\min}$, the maximal number of observed failures in each stage K_l, the minimum number of degrees of freedom $\nu_{\min}$ and finally the total number of degrees of freedom $\nu_\Sigma(L)$.
7. Choose the starting configuration f_1 and ν_1 for the first stage.
8. After study part "stage l" $l \geq 1$, calculate

$$U_l = \sum_{j=1}^{l} u_j(\nu_j).$$

If $U_l \geq cv_\alpha$, H_0 is rejected and the experiment stops.

If $\nu_\Sigma(l) = \nu_\Sigma(L)$, the experiment stops and H_0 is rejected if $U_l \geq cv_\alpha$.

If $U_l < cv_\alpha$ and $\nu_\Sigma(l) < \nu_\Sigma(L)$ then go to the next step.

9. Compute the weight function W_{l+1}, the degrees of freedom ν_{l+1} and the number of observed failures f_{l+1} for stage $l+1$. The weight function is given by

$$W_{l+1} = \max\left[1,\ \left(L - \nu_{\Sigma}(l)\right) \cdot \max\left(\varepsilon_{l+1}, \frac{f_{\min}}{M_{l+1}}\right)\right] \tag{6.11}$$

and the degrees of freedom and the additional number of observed failures are set as follows (Hartung and Knapp, 2003)

$$\nu_{l+1} = \begin{cases} W_{l+1} & \text{if} \quad L - \nu_{\Sigma}(l) \geq W_{l+1} + 1 \\ L - \nu_{\Sigma}(l) & \text{otherwise} \end{cases} \tag{6.12}$$

and

$$f_{l+1} = \max\left[\frac{\nu_{l+1} \cdot M_{l+1}}{L - \nu_{\Sigma}(l)},\ f_{\min}\right]. \tag{6.13}$$

10. Then go to step 3 and replace l by $l+1$.

In the following, some strategies for adjusting the design under our self-designing method are presented.

6.2.3 Strategies for adjustment

Recall that the units enter the experiment sequentially and symmetrically and are followed until they fail or the study is terminated. There are three main factors that determine the number of failures for censored survival data. These are:

- The hazard rates for both groups (standard and experimental)
- The accrual rate
- The accrual time

Therefore, when the experimenter decides to continue the experiment, he can extend the follow-up duration after accrual, the accrual duration or increase the accrual rate. Evidently there are many other possibilities which are out of the scope of this work.

With fixed accrual rate and duration and the follow-up period varying, it is clear that by extending the follow-up time, we may observe extra failures and therefore more failures should be observed in a later stage of the study.

The second possibility is when we assume that the accrual rate and the follow-up duration are fixed. Therefore we can increase the number of failures observed by extending the accrual duration.

The last possibility is to fix the accrual and follow-up duration. By increasing the accrual rate, we may increase the number of failures observed. If we accrue more units to the experiment, we obtain the required number of failures sooner and the total study duration is reduced.

The design problem is solved by making the following assumptions:

- Uniform accrual during the accrual period of A years and the follow-up period is of τ additional years
- The survival time distributions for the standard and experimental group are exponential with hazard rates h_S and h_E.
- Unit accrual is according to a Poisson process with rate a (>0). Therefore the number of units entering the experiment will be distributed as a Poisson variable with mean $a \cdot A$.

The empirical type I error rates and power estimates are based on simulating 10000 independent experiments for various combinations of accrual rate, accrual duration, follow-up duration and hazard rates. For equal randomisation with fixed sample design, George & Desu (1974) formula for the total number of failures needed to achieve a power of $1-\beta$ at the alternative $\tilde{\theta} > 0$ is given by

$$d = \frac{4\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\tilde{\theta}^2}. \tag{6.14}$$

Let $h(t)$ be the hazard function. Recall that the probability of surviving beyond time t is

$$P(V > t) = \exp(-\Lambda(t)),$$

where

$$\Lambda(t) = \int_0^t h(u)\, du$$

denotes the cumulative hazard function.

The expected number of failures at calendar time t_l is

$$d_l = \int_0^A a\{1 - \exp[-\Lambda(t_l - u)]\} du.$$

For exponential survival with constant hazard h, the expected number of failures in each group at calendar time t_l is given by

$$d_l^{(h)} = \begin{cases} a\left(t_l - \dfrac{1-\exp(-ht_l)}{h}\right) & \text{if} \quad t_l \le A \\ & . \\ a\left(A - \dfrac{\exp(-ht_l)}{h}(\exp(hA)-1)\right) & \text{if} \quad t_l > A \end{cases} \tag{6.15}$$

For equal randomisation, the expected number of failures at time t_l, under the alternative hypothesis is

$$d_l = d(t_l) = \frac{1}{2}\left(d_l^{(h_E)} + d_l^{(h_S)}\right) \tag{6.16}$$

and the accrual rate a can be easily computed. For an accrual duration A, the required sample size is $n = A \cdot a$.

Since we required d failures to achieve the desired power, the minimum accrual duration must be

$$A_{\min} = \frac{d}{a}.$$

It is not necessary to keep the accrual open beyond time $A_{\max}$, where $A_{\max}$ satisfies

$$d_{A_{\max}} = d.$$

Therefore, the range of duration of the accrual period is

$$\frac{d}{a} \le A \le A_{\max}.$$

Given the accrual period A, we can estimate the follow-up period τ by solving the equation

$$d_{A+\tau} = d. \tag{6.17}$$

For equal randomisation with fixed sample design, George & Desu (1974) formula for the total sample size needed to achieve a power of $1-\beta$ at the alternative $\tilde{\theta} > 0$ is given by

$$n = d \cdot \frac{\left(\dfrac{1}{h_S} + \dfrac{1}{h_E}\right)}{\left(\dfrac{P_S}{h_S} + \dfrac{P_E}{h_E}\right)} = \frac{4\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\tilde{\theta}^2} \frac{\left(\dfrac{1}{h_S} + \dfrac{1}{h_E}\right)}{\left(\dfrac{P_S}{h_S} + \dfrac{P_E}{h_E}\right)}, \tag{6.18}$$

with

$$P_i = 1 - \frac{1}{h_i A}\exp(-h_i\tau)(1-\exp(-h_i A)), \qquad i = E, S \tag{6.19}$$

the probability of observing a failure in the *i*th group.

In the following sections, we used the following sample size spending function to compute formula (6.6) and (6.2.10);

$$S_{l-1}(\alpha,\beta)=\frac{4\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{\hat{\theta}_{l-1}^{\ 2}}, \tag{6.20}$$

with

$$\hat{\theta}_{l-1}=\frac{\sum_{j=1}^{l-1} n_j\hat{\theta}_j}{\sum_{j=1}^{l-1} n_j}, \tag{6.21}$$

where $\hat{\theta}_j$ is the Maximum Likelihood Estimate for θ and n_j is the number of units enrolled in the study up to the calendar time t_j.

6.2.4 Example

An engineer wishes to compare the time until stress corrosion crack initiation for a standard material manufactured by a company *A* with the time until crack initiation of a new formulation of this material manufactured by the company *B*. Let $h_A(t)$ be the hazard function for the standard material of company A and $h_B(t)$ be the hazard function of the material manufactured by company *B*. The engineer is interested in the test problem

$H_0:\theta=0$ vs. $H_1:\theta>0$

with

$$\theta=\ln\left(\frac{h_B(t)}{h_A(t)}\right).$$

More precisely, the engineer wants 90% power to detect an improvement of 0.597 on the log hazard ratio scaled by conducting a 5%-level one-sided test. Based on data collected from one historical study using the standard material, the value $h_A=0.1$ has been observed. The total degrees of freedom to be available in the whole sequential test is $L=10$. The type II error rate for generating the sequential number of observed failures is $\beta_g=0.8$. The minimum number of observed failures in each stage is $f_{\min}=12$. The global critical value is given by $cv_\alpha=\chi^2(10)_{0.95}=18.307$.

Data from one historical stress corrosion crack initiation study suggest the exponential distribution as a model for this type of data. That means, the exponential distribution is chosen to model the time to crack initiation. The test conditions need to be accelerated to ensure crack initiation failures without introducing a failure mode not usually experienced under normal operating conditions. The materials are then tested under stressful conditions, so that each hour of testing is equivalent to 10000 hours of actual use in the field. Specimens from both the standard and new material will be tested simultaneously using a machine loaded stress test at a single high stress condition and these are followed until either they fail or are administratively censored at the end of the follow-up period. The maximum number of specimen the engineer can test at one time is 200. Specimens are allocated uniformly during the accrual period of 4 hours and the follow-up period is of 6 additional hours. Allocation is according to a Poisson process with rate 50 per hour. The starting configuration for the first stage are chosen as $f_1 = 24$ and $\nu_1 = 2$.

After observing 24 failures in both groups in the first stage, we compute the statistic $S_{n_1} =$ 3.0414 according to formula (6.1), with n_1 representing the number of specimens enrolled in the study up to that time ($t_1 = 3.22$ hours $n_1 = 162$). We observe then a test value of T as $T_1 = 1.2378$ according to formula (6.2) with $S_n(t_0) = 0$ and $\sigma_Z(t_1) = 0.5015$. Thus the corresponding p-value is $p_1 = 0.1078$ and $u_1(\nu_1) = u_1(2) = 4.4532$. $U_1 = 4.4532 < cv_\alpha = 18.3070$. From the first stage we obtain the maximum likelihood estimate of θ , $\hat{\theta}_1 = 0.7386$. According to formula (6.6)

$$M_2 = S_1\left(1 - F_{\chi^2\{8\}}(cv_\alpha - U_1), \beta\right) = \frac{4\left(\Phi^{-1}\left(F_{\chi^2\{8\}}(cv_\alpha - U_1)\right) + \Phi^{-1}(1-\beta)\right)^2}{\hat{\theta}_1^{\,2}} \cong 53,$$

and formula (6.3.10)

$$m_2 = S_1\left(1 - F_{\chi^2\{8\}}(cv_\alpha - U_1), \beta_g\right) = \frac{4\left(\Phi^{-1}\left(F_{\chi^2\{8\}}(cv_\alpha - U_1)\right) + \Phi^{-1}(1-\beta_g)\right)^2}{\hat{\theta}_1^{\,2}} \cong 3.$$

Because the maximum number of failures to observe in each stage is set to 50, we decided to take $M_2 = 50$. Therefore the sequence for the second stage $\varepsilon_2 = m_2 / M_2 = 0.06$ and we compute the weight function $W_2 = 1.92$ (formula (6.11)) and the degrees of freedom of the second stage $\nu_2 = 2$ (formula (6.12)). The additional number of failures to observe in the second stage is then $f_2 = 13$. After observing 13 more failures, we compute the statistic S_{n_2} = 4.5613 with n_2 representing the number of specimens enrolled in the study up to that time ($t_2 = 3.86$ hours, $n_2 = 194$). We observe then a test value $T_2 = 0.8409$ with $\sigma_Z(t_2) = 0.5012$ and the corresponding p-value is $p_2 = 0.2001$, which implies $u_2(\nu_2) = u_2(2) = 3.2169$. Therefore

$U_2 = u_1(2) + u_2(2) =$ 7.6701. $U_2 < cv_\alpha$=18.3070 and the update estimate of θ is $\hat{\theta}_2 =$0.7290 (6.21). Hence, by (6.6), we obtain $M_3 = 50$ and by (6.10) $m_3 = 2$. The sequence for the third stage $\varepsilon_3 = 0.04$, the weight function is $W_3 = 1.44$ and by (6.12) $\nu_3 = 2$. By (6.13), the additional number of failures to observe in the third stage is $f_3 = 17$. After observing 17 more failures, we compute the statistic $S_{n_3} =$ 7.1246 by $t_3 = 4.77$ hours, $n_2 = 200$ and the test value is $T_3 = 1.2402$ which correspond to a p-value of $p_3 = 0.1074$. $u_3(\nu_3) = u_3(2) = 4.4616$ and $U_3 = u_1(2) + u_2(2) + u_3(2) = 12.1318$. $U_3 < cv_\alpha$=18.3070 and $v_\Sigma(3) = 6 < 10$, Therefore, the fourth stage must be performed. The update estimate of θ is $\hat{\theta}_3 =$0.7347 . Hence, we obtain $M_4 = 35$, $m_4 = 1$. The sequence for the fourth stage $\varepsilon_4 = 0.0285$, the weight function is $W_4 = 1.3714$ and $\nu_4 = 2$. The additional number of failures to observe in the fourth stage is then $f_4 = 18$. After observing 18 more failures, we compute the statistic $S_{n_4} =$ 6.7858 by $t_4 = 5.74$ hours, $n_4 = 200$ and we obtain a test value of $T_4 = -0.1593$ which corresponds to a p-value of $p_4 = 0.5632$. $u_4(\nu_4) = u_4(2) = 1.1479$ and $U_4 = u_1(2) + u_2(2) + u_3(2) + u_4(2)$ = 13.2797. $U_4 < cv_\alpha$=18.3070 and $v_\Sigma(4) = 8 < 10$, Therefore, the next stage must be performed. The update estimate of θ is $\hat{\theta}_4 =$0.7046. Hence, we obtain $M_5 = 58$ and we decided to take $M_5 = 50$ $m_5 = 3$. The sequence for the fourth stage $\varepsilon_5 = 0.06$, the weight function is $W_5 = 1$ and $\nu_5 = 1$. The additional number of failures to observe in the five stages is therefore $f_5 = 25$. After observing 25 more failures, we computed the statistic $S_{n_5} =$ 12.1119 and the test value is $T_5 = 2.1251$ which corresponds to a p-value of $p_5 = 0.0167$. $u_5(\nu_5) = u_5(1) = 5.7183$ and $U_5 = u_1(2) + u_2(2) + u_3(2) + u_4(2) + u_5(1)$ = 18.9980. $U_5 < cv_\alpha$ = 18.3070. Thus H_0 is rejected at level $\alpha = 0.05$ and we stopped the test. We need about 24 + 13 + 17 + 18 + 25 = 97 failures to observe in the study to detect an improvement of 0.597 with power 0.9. In a fixed-sample design, we would have needed to observe about 96 failures if the estimate of the log hazard ratio θ would have been known in advance.

6.2.5 Simulation

In this section, simulations are carried out to give an impression of the operating characteristics of the adaptation procedure for different cases and a comparison with the usual log-rank test with fixed sample size. In each simulation, a log-rank statistic is computed based on the calculated sample size from the fixed sample design at the determinate termination time. The performance of the log-rank from the fixed sample design is compared with the proposed procedure in terms of error rate, power, average number of failures and average study duration by simulation. The simulated values are based on 10000 independent replications. For the proposed procedure, for the computation of $S_n(t_l)$, n is replaced by the number of specimens enrolled in the study up to the end of the lth stage. Specimens, which enter the study at that time but have not failed, are censored.

Specimens are equally randomised, in each stage, to the two groups. The type I error rate $\alpha = 0.05$ and a power of $1-\beta = 0.9$ is specified. We use one-quarter of the number of failures required in the fixed sample design as f_1. The total number of available degrees of freedom are set equal to $L = 10$. The minimum number of failures to be observed in each stage is set equal to $f_{\min} = f_1 / 2$. The required number of failures for the l-th stage is computed using (6.7) where the sequence $\{\varepsilon_l\}$ is determined as in (6.9) with $\beta_g = 0.8$.

First Strategy: Fixed accrual rate and accrual period and varying follow-up duration

We want to detect an improvement of $\tilde{\theta} = 0.5$ on the log hazard ratio scale with $h_E = 0.0607$ and $h_S = 0.1$. The number of failures required in the fixed sample design is equal to 138 (cf (6.13)) when the true log hazard ratio is correctly specified. With a 4 months accrual duration and a 3 months follow-up after accrual, the fixed accrual rate is computed according to formula (6.16) and is equal to $a = 106$ units per month. The follow-up period τ vary from 4 to 7 months. The usual log-rank test is computed at the end of 7 months for 424 accrued units. A log-rank score statistic $S_n(t_l)$ given by (6.1) is calculated after observing

$$\sum_{i=1}^{l} f_i$$

failures and the procedure works as described before.

Preliminary simulations showed that the global critical value given by

$$cv_\alpha = \chi^2(L)_{1-\alpha}$$

was very conservative; therefore we adjust as

$$cv_\alpha = \chi^2(L)_{1-\alpha} - 6.5.$$

We also introduce a lower bound for an early acceptance of H_0, that is

$$U_l \leq \chi^2(\nu_\Sigma(l))_{1-\alpha_L},$$

where $\alpha_L = 0.6$.

We considered here with two different cases corresponding to the upper bounds in the sample size functions M_l and m_l. The first case is given with

$M_l \leq 120$, $m_l \leq 120$ and the second with $M_l \leq 100$, $m_l \leq 100$. The use of upper bounds for M_l and m_l influences directly not only the degrees of freedom, but also the number of failures to be observed in the next stage which are functions of the sequence ε_l (Hartung and Knapp, 2003).

The results of the simulations are given in table 6.2.1 with the following abbreviation:

ES: Empirical Size

EP: Empirical Power

ANF: Average Number of Failures

ASD: Average Study of Duration

Here for the proposed method, the use of a larger bound for M_l and m_l results in a larger power, a greater average number of failures and greater average study duration. The greater difference between the power is observed at $\theta = 0.45$ and is more than 3%. The average number of failures with $M_l \leq 120$, $m_l \leq 120$ is considerably larger when the true underlying alternatives, θ, are overestimated by $\tilde{\theta} = 0.5$ than with $M_l \leq 100$, $m_l \leq 100$. Comparing now the proposed procedure and the fixed-sample design, considering the bounds $M_l \leq 100$, $m_l \leq 100$, we observed that when the alternative hypothesis is correctly specified, the proposed procedure achieves power similar to that from the fixed-sample design with a considerably smaller average number of failures and a smaller average study duration. The difference between the average number of failures is greater when the true underlying alternatives, θ, are considerably overestimated or underestimated. In general, the proposed procedure results in a smaller average study duration. For small values for θ, the proposed procedure results in a slightly loss of power and a slightly gain of power for greater values of θ. Considering the bounds $M_l \leq 120$, $m_l \leq 120$, the proposed strategy which allows an increase in the number of failures by extending the follow-time results in a larger power.

θ	**Characteristics**	$M_l \le 120$ $m_l \le 120$	$M_l \le 100$ $m_l \le 100$	**Fixed-sample design**
Under H_0 ($\theta = 0$)	ES	0.0501	0.0539	0.0476
	ANF	142.7387	132.8587	165.1637
	ASD	6.1847	5.8376	7
0.35	EP	0.572	0.5462	0.5603
	ANF	151.6892	134.216	145.0005
	ASD	7.3079	6.5669	7
0.4	EP	0.6883	0.6588	0.666
	ANF	151.2503	133.3302	142.4735
	ASD	7.4049	6.6290	7
0.45	EP	0.7859	0.7553	0.7607
	ANF	150.2099	132.2322	140.033
	ASD	7.4829	6.6821	7
0.5	EP	0.8638	0.8358	0.8351
	ANF	148.3828	130.399	137.6446
	ASD	7.5240	6.7071	7
0.55	EP	0.9188	0.8977	0.8911
	ANF	145.6095	127.5866	135.333
	ASD	7.5270	6.6897	7
0.6	EP	0.9518	0.9375	0.9309
	ANF	142.3617	124.8595	133.1214
	ASD	7.5046	6.6711	7
0.65	EP	0.972	0.9646	0.959
	ANF	138.7742	121.6478	130.9758
	ASD	7.4638	6.6309	7
0.7	EP	0.9855	0.9829	0.976
	ANF	134.4241	117.9606	128.90
	ASD	7.3862	6.5650	7

Table 6.2.1: *Empirical size and Power, Average Number of failure for the design with the first strategy*

Second Strategy: Fixed accrual rate and follow-up duration, but varying accrual duration

We considered the same hazard rates for the standard and the experimental group as in the first strategy. With an accrual rate varying from 4 to 7 months and a fixed follow-up of 3 month, the fixed accrual rate is computed according to formula (6.2.16) and is equal to $a = 84$ per months. The results are summarized in table 6.2.2 and will help us to compare the two strategies. We can observe a slightly gain of power compared to the first strategy. The average study duration is also a bit lower in this case. For the rest, the observed phenomena is similar to that of the first strategy.

θ	**Characteristics**	$M_l \leq 120$ $m_l \leq 120$	$M_l \leq 100$ $m_l \leq 100$	**Fixed-sample design**
Under H_0 ($\theta = 0$)	ES ANF ASD	0.0475 142.7717 6.3725	0.0505 132.9402 6.1516	0.0422 165.0002 7
0.3	EP ANF ASD	0.4518 152.0618 7.1053	0.4312 134.7154 6.6416	0.444 147.0837 7
0.4	EP ANF ASD	0.6966 151.3941 7.2289	0.6637 133.4563 6.7354	0.6702 141.7929 7
0.45	EP ANF ASD	0.7988 150.2331 7.2645	0.7615 131.9816 6.7535	0.7656 139.2836 7
0.5	EP ANF ASD	0.873 148.2141 7.2712	0.843 130.1936 6.7571	0.8391 136.8714 7
0.55	EP ANF ASD	0.9253 146.0274 7.2695	0.9011 128.0235 6.7450	0.8944 134.5749 7

0.6	EP ANF ASD	0.9603 143.0786 7.2397	0.9448 125.431 6.7148	0.933 132.3658 7
0.65	EP ANF ASD	0.9774 139.2037 7.1764	0.9695 122.0988 6.6562	0.9608 130.2026 7
0.7	EP ANF ASD	0.9889 135.0838 7.0979	0.9842 118.5523 6.5844	0.9766 128.1318 7

Table 6.2.2: *Empirical size and Power, Average Number of failure for the design with the second strategy*

The second strategy may work more effectively when hazard rates are high because the number of failures increases. The performance of the second strategy are now evaluated with relatively high hazard rates $h_E = 0.25$ and $h_S = 0.5$ corresponding to $\tilde{\theta} = 0.693$. The number of failures required in the fixed sample design is equal to 72 (cf (6.2.13)) when the true log hazard ratio is correctly specified. With accrual duration varying from 4 to 7 months and a 3 months follow-up after accrual, the fixed accrual rate per month is computed according to formula (6.2.16) and is equal to $a = 18$ per months. The usual log-rank test is computed at the end of 7 months for 126 accrued units. We adjust the global critical value as

$$cv_\alpha = \chi^2(L)_{1-\alpha} - 5$$

and we also introduce a lower bound for an early acceptance of H_0, that is

$$U_l \leq \chi^2(v_\Sigma(l))_{1-\alpha_L},$$

where $\alpha_L = 0.6$.

The upper bounds in the sample size functions M_l and m_l are given with

$M_l \leq 72$, $m_l \leq 72$ and the second with $M_l \leq 60$, $m_l \leq 60$. The results are summarized in table 6.2.3. We observed that extending the accrual period is effective for high hazard rates because a slight gain in power is achieved compared to low hazard rates given in table 6.2.2.

θ	**Characteristics**	$M_l \leq 72$ $m_l \leq 72$	$M_l \leq 60$ $m_l \leq 60$	**Fixed-sample design**
Under H_0	ES	0.0501	0.0523	0.0419
$\theta = 0$	ANF	90.8462	84.2982	91.1127
	ASD	6.82.89	6.5379	7
	EP	0.533	0.5174	0.537
0.45	ANF	87.4881	77.08	83.3331
	ASD	7.2486	6.5947	7
	EP	0.6274	0.603	0.6292
0.5	ANF	86.7372	76.1244	82.4085
	ASD	7.2615	6.5806	7
	EP	0.7118	0.6861	0.7071
0.55	ANF	85.8742	75.0941	81.4685
	ASD	7.27024	6.5647	7
	EP	0.7875	0.7597	0.7789
0.6	ANF	84.3906	73.9449	80.5266
	ASD	7.2427	6.5383	7
	EP	0.8459	0.8251	0.8371
0.65	ANF	82.8645	72.3906	79.57
	ASD	7.2092	6.4849	7
	EP	0.8915	0.8692	0.877
0.693	ANF	81.3651	71.220	78.7732
	ASD	7.1735	6.4481	7
	EP	0.931	0.9141	0.9213
0.75	ANF	79.2352	69.339	77.7067
	ASD	7.1130	6.3820	7
	EP	0.9558	0.9448	0.9464
0.8	ANF	77.2976	67.6589	76.7568
	ASD	7.0512	6.3168	7

Table 6.2.3: *Empirical size and Power, Average Number of failure for the design with the second strategy with relative high hazard rates*

6.3 Sample size adaptation under non proportional hazard

The log-rank statistic is the commonly used two-sample nonparametric test statistic and is adequate in situation of proportional hazard (the ratio of the hazards is constant over time). But in non proportional case, it does not have a good interpretation with respect to any particular alternative of interest. More precisely, it is not sensitive to the duration and magnitude of the difference in survival over time; that is why it behaves poorly. To overcome this problem, Pepe and Fleming (1989) proposed a class of test based on the integrated weighted difference in the Kaplan-Meier estimates. In this section we develop and illustrate a method of updating sample size using a test statistic which belongs to the class of test based on the integrated weighted difference in the Kaplan-Meier estimates proposed by Pepe and Fleming (1989, 1991). The adaptation method is based on the work of Hartung and Knapp (2003). Accumulating data will be reviewed sequentially to adjust the sample size and the study can be terminated early if a large group difference is observed. The proposed test statistic is a linear weighted Kaplan-Meier which can update the sample size based on the observed data.

In subsection 1, notation required in the sequential framework and the test statistic is introduced. The asymptotic joint distribution of the test statistic at different calendar time or different interim analysis time is also derived. Particular attention will be given on the conditions under which the test statistic has a joint independent increments structure. In subsection 2, we present methods for adapting the test statistic to the adaptation procedures. A simulation study to obtain the operating characteristics of the proposed method and for a comparison with the sequential log-rank test will close this section.

6.3.1 Notation and test statistic

Let the nonnegative random variable Y denote the real time of entry into the experiment and let the random variable V denote failure time. Let C denote the time from entry to censoring. The survival distribution of V at any time x can be expressed as

$$\exp(-\Lambda(x))$$

and the density as

$$h(x)\exp(-\Lambda(x)),$$

where $h(x)$ denotes the hazard rate at time x and

$$\Lambda(x) = \int_0^x h(u)\,du$$

denotes the cumulative hazard function.

The time of entry into experiment, Y, is assumed to be a bounded positive random variable with distribution function

$$H(y) = P(Y \leq y).$$

The distribution of the time to censoring is given by

$$G(c) = P(C < c).$$

$S(c) = 1 - G(c))$ denotes the survival distribution. It is assumed that the random variables V, Y, C are independent within each group.

Denoting $S_E(t)$ and $S_S(t)$ the survival function of the experimental and standard groups respectively, we want to test if the two survival functions are equal. Considering the median survival of the specimens of the two groups m_E and m_S, $\theta = m_E / m_S$ can express the difference between the two groups. A one-sided test will be performed with a significance level of α to detect a median ratio of $\tilde{\theta} > 1$. The null hypothesis to be tested is

$$H_0 : \theta = 1 \qquad \text{against} \qquad H_1 : \theta > 1.$$

The study is characterized by an accrual period A and a follow-up period τ. Subjects enter the experiment sequentially and are allocated symmetrically to the two groups and are followed until either they fail or are administratively censored at the end of the follow-up period.

Suppose that the experiment involved for each group i, n_i units, $i = E, S$, which enter the experiment serially and are assigned to machines according to some random mechanism. The data can be expressed as identically and independently distributed random vectors (V_{ij}, Y_{ij}, C_{ij}) for $i = E, S$, $j = 1, \cdots, n_i$.

If the data were to be examined at time t, the following variables could be observed:

- Time to failure or censoring

  $$X_{ij}(t) = \max\{\min(V_{ij}, t - Y_{ij}, C_{ij}), 0\}$$

- Indicator variable for failure

  $$\Delta_i(t) = \begin{cases} 1 \text{ if } V_{ij} < \min(t - Y_{ij}, C_{ij}) \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

At time t, the data can be expressed as n identically and independently distributed random vectors $\left(X_{ij}(t), \Delta_{ij}(t)\right)$ for $i = E, S$, $j = 1, \cdots, n_i$.

Defining the counting process

$$N_{ij}(t,x) = I\left(X_{ij}(t) \le x,\ \Delta_{ij}(t) = 1\right),$$

which is the indicator function that unit j in group i was observed to fail at calendar time t before duration x, $x \le t$, $I(\cdot)$ denoting the indicator function and define

$$\overline{N}_i = \sum_{j=1}^{n_i} N_{ij}(t,x), \qquad i = E, S .$$

and

$$\overline{N}(t,x) = \overline{N}_E(t,x) + \overline{N}_S(t,x).$$

Similarly, define the number of units at risk at calendar time t and at failure time x, $x \le t$,

$$n_i(t,x) = \sum_{j=1}^{n_i} I\left(X_{ij}(t) \ge x\right)$$

and let

$$n(t,x) = n_E(t,x) + n_S(t,x).$$

We define the Kaplan Meier estimate of the survival function $S_i(t)$, if we used the data available at calendar time t as

$$\hat{S}_i(t,x) = \prod_{w \le x} \left\{ 1 - \frac{d\overline{N}_i(t,w)}{n_i(t,w)} \right\}, \ i = E, S .$$

Define the σ-algebra $\Im(t,x)$ generated by failures happening before calendar time t and failure time x, $x \le t$; formally:

$$\Im(t,x) = \sigma\left\{ I\left(Y_{ij} \le t\right), Y_{ij} I\left(Y_{ij} \le t\right), I\left(V_{ij} \le \min\left(w, t - Y_{ij}\right)\right), I\left(C_{ij} \le \min\left(w, t - Y_{ij}, V_{ij}\right)\right)\right\}$$

for $i = E, S$, $j = 1, \cdots, n_i$, $0 \le w \le x$.

Under some independence conditions and for fixed t

$$\overline{N}_i(t,x) - \int_0^x h_i(w) n_i(t,w) dw, \quad i = E, S$$

is a martingale with respect to the filtration $\left(\Im(t,x)\right)_{x \ge 0}$. The Nelson-Aalen estimators are given by

$$\hat{\Lambda}_i(t,x) = \int_0^x \frac{I\left(n_i(t,w) > 0\right)}{n_i(t,w)} \overline{N}_i(t,w) dw, \qquad i = E, S .$$

Note that

$$\sqrt{n_i}\left[\hat{S}_i(t,x) - \exp\left(-\hat{\Lambda}_i(t,x)\right)\right]$$

converges in probability to zero as shown by Breslow and Crowley (1974).

For $x \le t$ and under the null hypothesis, we define the statistic

$$Z_n(t,x) = \sqrt{n}\int_0^x \frac{M(t,w)}{n(t,w)}dw\,,$$

where

$$M(t,x) = \overline{N}(t,x) - \int_0^x h(w)n(t,w)dw\,.$$

According to Murray and Tsiatis (1999),

$$\operatorname{cov}(Z_n(t_1,x_1), Z_n(t_2,x_2)) = \int_0^{\min(x_1,x_2)} \frac{h(w)}{S(w)C(t_2,w)}dw$$

asymptotically, where

$$C(t,x) = P(Y_{ij} < t - x, C_{ij} > x)$$

and

$$\operatorname{cov}\left[\left(\hat{S}(t_1,x_1) - S(x_1)\right), \left(\hat{S}(t_2,x_2) - S(x_2)\right)\right] \approx \frac{S(x_1)S(x_2)}{n(t_2)} \int_0^{\min(x_1,x_2)} \frac{h(w)}{S(w)H(t_2,w)}dw\,,$$

where

$$n(t) = \sum_{i=1}^{2}\sum_{j=1}^{n_i} I(Y_{ij} \le t)$$

is the total sample size enrolled at calendar time t and

$$H(t,x) = \frac{P(Y_{ij} < t - x, C_{ij} > x)}{P(Y_{ij}) \le t}$$

is the censoring reliability distribution amount specimens entered by calendar time t.

If we considered the case where only one analysis is performed at the end of the study, we can focus our attention only on the internal time of the specimens. Let $t_{\max}$ be the last point where a consistent reliability estimate $\hat{S}(t_{\max})$ may be defined, the statistic

$$Q = \sqrt{\frac{n_E \cdot n_S}{n}} \int_0^{t_{\max}} \left\{\hat{S}_E(w) - \hat{S}_S(w)\right\} dw$$

compared the reliability functions of the two groups during the first $t_{\max}$ periods of study. Pepe and Fleming (1989) show that under the null hypothesis, Q converge in distribution to a normal distribution with mean zero and variance given by

$$\sigma^2 = \sum_{i=1}^{2} p_{3-i} \int_0^{t_{\max}} \frac{\left\{A^{t_{\max}}(w)\right\}^2 h(w)}{S(w)H_i(w)} dw .$$

p_i is the probability of failing in group i, $h(x)$ and $S(x)$ are the hazard and reliability functions, respectively, common to the two groups under the null hypothesis, and $A^{t_{\max}}$ is given by

$$A^{t_{\max}}(x) = \int_x^{t_{\max}} S(w) dw .$$

A natural estimator of σ^2 can be obtained by replacing the unknown quantities by their consistent estimates,

$$\hat{\sigma}^2 = \sum_{i=1}^{2} \hat{p}_{3-i} \int_0^{t_{\max}} \frac{\left\{\hat{A}^{t_{\max}}(w)\right\}^2}{\tilde{S}(w)\hat{H}_i(w)\overline{n}(w)} d\overline{N}(w).$$

$\hat{p}_i = \frac{n_i}{n}$, $\tilde{S}(x)$ is the pooled Kaplan-Meier estimate of groups E and S ,$\hat{H}_i$ is the estimate of the censoring time reliability probability, $\overline{N}(x)$ is the observed number of failures at time x, $\overline{n}(x)$ is the observed number of specimens still at risk at time x and

$$\hat{A}^{t_{\max}}(x) = \int_x^{t_{\max}} \tilde{S}(w) dw .$$

We consider the test statistic Q evaluated at calendar time t_l (Murray and Tsiatis, 1999)

$$Q_l = Q(t_l) = \sqrt{\frac{n_E(t_l) \cdot n_S(t_l)}{n(t_l)}} \int_0^{\tau} \left\{\hat{S}_E(t_l, w) - \hat{S}_S(t_l, w)\right\} dw .$$

Without loss of the generality, we consider a balanced design at each calendar time, that is

$$n_E(t_l) = n_S(t_l) \qquad \forall\ t_l ,$$

then

$$Q_l = \sqrt{\frac{n(t_l)}{4}} \int_0^{\tau} \left\{\hat{S}_E(t_l, w) - \hat{S}_S(t_l, w)\right\} dw . \tag{6.22}$$

Under the null hypothesis

$$\operatorname{cov}(Q_{l-1},Q_l)=\frac{\sqrt{n(t_{l-1})\cdot n(t_l)}}{4}\operatorname{cov}\left[\int_0^{\tau}\left\{\hat{S}_E(t_{l-1},w_{l-1})-\hat{S}_S(t_{l-1},w_{l-1})\right\}dw_{l-1},\int_0^{\tau}\left\{\hat{S}_E(t_l,w_l)-\hat{S}_S(t_l,w_l)\right\}dw_l\right]$$

$$=\frac{\sqrt{n(t_{l-1})\cdot n(t_l)}}{4}\left[\int_0^{\tau}\int_0^{\tau}\operatorname{cov}\left(\hat{S}_E(t_{l-1},w_{l-1}),\hat{S}_E(t_l,w_l)\right)dw_{l-1}dw_l+\int_0^{\tau}\int_0^{\tau}\operatorname{cov}\left(\hat{S}_S(t_{l-1},w_{l-1}),\hat{S}_S(t_l,w_l)\right)dw_{l-1}dw_l\right]$$

$$=\frac{\sqrt{n(t_{l-1})\cdot n(t_l)}}{4}\left[\int_0^{\tau}\int_0^{\tau}\operatorname{cov}\left(\hat{S}_E(t_l,w_{l-1}),\hat{S}_E(t_l,w_l)\right)dw_{l-1}dw_l+\int_0^{\tau}\int_0^{\tau}\operatorname{cov}\left(\hat{S}_S(t_l,w_{l-1}),\hat{S}_S(t_l,w_l)\right)dw_{l-1}dw_l\right]$$

$$=\frac{\sqrt{n(t_{l-1})\cdot n(t_l)}}{4}\left\{\left[\int_0^{\tau}\int_0^{\tau}\frac{S(w_{l-1})S(w_l)}{n_E(t_l)}\int_0^{\min(w_{l-1},w_l)}\frac{h(w)dw}{S(w)H_E(t_l,w)}\,dw_{l-1}\,dw_l\right]\right.$$

$$\left.+\left[\int_0^{\tau}\int_0^{\tau}\frac{S(w_{l-1})S(w_l)}{n_S(t_l)}\int_0^{\min(w_{l-1},w_l)}\frac{h(w)dw}{S(w)H_S(t_l,w)}\,dw_{l-1}\,dw_l\right]\right\}$$

$$=\frac{1}{2}\sqrt{\frac{n(t_{l-1})}{n(t_l)}}\left\{\left[\int_0^{\tau}\frac{[A^{\tau}(w)]^2h(w)}{S(w)H_E(t_l,w)}\,dw\right]+\left[\int_0^{\tau}\frac{[A^{\tau}(w)]^2h(w)}{S(w)H_S(t_l,w)}\,dw\right]\right\}.$$

Therefore

$$\operatorname{var}(Q_l)=\frac{1}{2}\left\{\left[\int_0^{\tau}\frac{[A^{\tau}(w)]^2h(w)}{S(w)H_E(t_l,w)}\,dw\right]+\left[\int_0^{\tau}\frac{[A^{\tau}(w)]^2h(w)}{S(w)H_S(t_l,w)}\,dw\right]\right\}\tag{6.23}$$

and

$$\operatorname{cov}(Q_{l-1},Q_l)=\sqrt{\frac{n(t_{l-1})}{n(t_l)}}\operatorname{var}(Q_l),$$

so that it relates directly to the variance of Q_l.

If we considered the standardized statistic

$$Q^*(t)=\frac{\sqrt{n(t)}}{2}\frac{Q(t)}{\operatorname{var}(Q(t))},$$

then

$$\frac{Q^*(t)}{\sqrt{\operatorname{var}(Q^*(t))}}=\frac{Q(t)}{\sqrt{\operatorname{var}(Q(t))}},$$

so that the two statistics can be used interchangeably.

$$\operatorname{cov}\left(Q^*(t_{l-1}),Q^*(t_l)\right)=\operatorname{cov}\left(\frac{\sqrt{n(t_{l-1})}}{2}\frac{Q(t_{l-1})}{\operatorname{var}(Q(t_{l-1}))},\frac{\sqrt{n(t_l)}}{2}\frac{Q(t_l)}{\operatorname{var}(Q(t_l))}\right)$$

$$=\frac{\sqrt{n(t_{l-1})}}{2}\cdot\frac{\sqrt{n(t_l)}}{2}\cdot\frac{\operatorname{cov}(Q_{l-1},Q_l)}{\operatorname{var}(Q_{l-1})\operatorname{var}(Q_l)}$$

$$=\frac{\sqrt{n(t_{l-1})}}{2}\cdot\frac{\sqrt{n(t_l)}}{2}\cdot\frac{\sqrt{\frac{n(t_{l-1})}{n(t_l)}}\operatorname{var}(Q_l)}{\operatorname{var}(Q_{l-1})\operatorname{var}(Q_l)}$$

$$=\frac{n(t_{l-1})}{4}\cdot\frac{1}{\operatorname{var}(Q_{l-1})}$$

$$=\operatorname{var}\left(Q^*(t_{l-1})\right).$$

Therefore, the test statistic Q has an independent increment structure and for any $l \neq k$, $\operatorname{cov}(Q_l - Q_{l-1}, Q_k - Q_{k-1})$ converge in probability to zero.

$\sigma_x{}^2 = \operatorname{var}(Q_x)$ can be estimated consistently by $\hat{\sigma}_x{}^2$ which is obtained by replacing the corresponding quantities with their empirical estimators. Thus, under the null hypothesis, $Q_l - Q_{l-1}$ can express the information cumulated during the period (t_{l-1}, t_l). $Q_l - Q_{l-1}$ converge in distribution to a normal distribution with mean zero and variance $\sigma_l{}^2 + \sigma_{l-1}{}^2$. By a direct application of the Slutsky's theorem, the test statistic

$$T_l = \frac{Q_l - Q_{l-1}}{\sqrt{\hat{\sigma}_l{}^2 + \hat{\sigma}_{l-1}{}^2}} \tag{6.24}$$

is asymptotically normally distributed with mean zero and variance one. Furthermore, for any $l \neq k$ the random vector $\{T_l, T_k\}$ converge in distribution to a bivariate standard normal distribution. Therefore T_l and T_k are asymptotically independent for any $l \neq k$.

6.3.2 Adaptation procedure

The procedure for sample size adaptation used is the self-designing procedure of Hartung and Knapp (2003) as described in section 6.2.2. Depending on the underlying reliability distribution, we will use two strategies to adjust the design: The strategy of fixing the accrual period and varying the follow-up period and the strategy of fixing the follow-up period and varying the accrual period. The design problem is solved by making the following assumptions:

- Uniform accrual during the accrual period of A years and the follow-up period is of τ additional years
- Unit accrual is according to a Poisson process with rate a (>0). Therefore the number of units entering the experiment will be distributed as a Poisson variable with mean $a \cdot A$.

The expected number of failures at calendar time t_l is given by

$$d_l = \int_0^{\min(A, t_l)} a\{1 - \exp[-\Lambda(t_l - u)]\} du ,$$

where

$$\Lambda(t) = \int_0^t h(u)\, du$$

denotes the cumulative hazard function. For equal randomisation, the expected number of failures at time t_l, under the alternative hypothesis is given by

$$d_l = d(t_l) = \frac{1}{2}\left(d_l^{(h_E)} + d_l^{(h_S)}\right) .$$

Then we will consider three configurations of the design:

- The reliability time distributions for the standard and experimental group are Exponential with medians respectively m_S and m_E (**Configuration I**).
- The reliability time distributions for the standard and experimental group are Weibull with medians respectively m_S and m_E (**Configuration II**)
- The reliability time distributions for the standard group is Exponential with median respectively m_S and that for the experimental group is Weibull with median m_E (**Configuration III**).

For equal randomisation with fixed sample design, formula for the total number of failures needed to achieve a power of $1-\beta$ at the alternative $\tilde{\theta} > 0$ is given by

$$d = \frac{4\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\tilde{\theta}^2}.$$

Under **configuration I** and **II** $\tilde{\theta}$ is expressed as

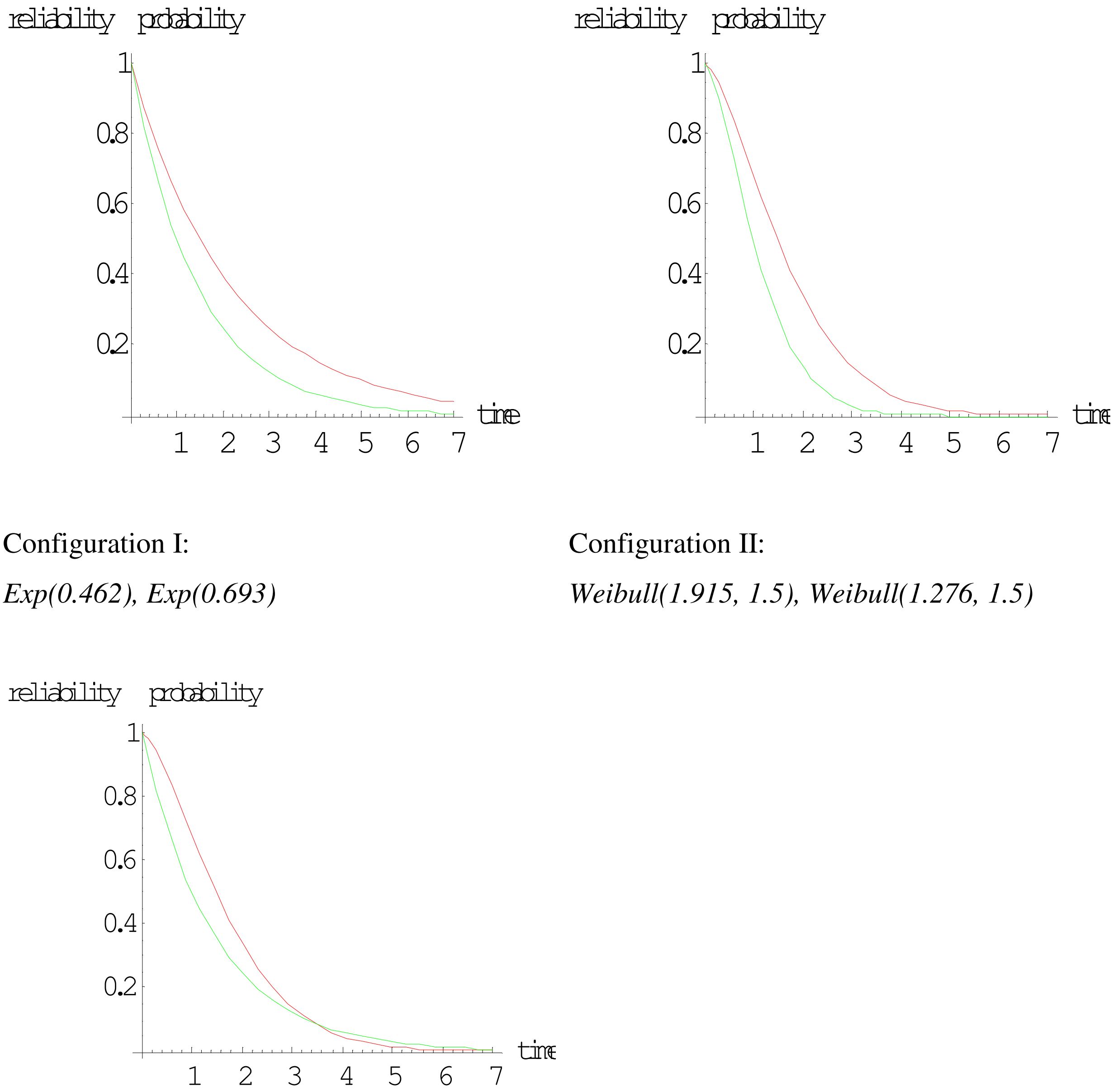
$$\tilde{\theta} = \log\left(\frac{m_E}{m_S}\right) \text{ and } \tilde{\theta} = \kappa \log\left(\frac{m_E}{m_S}\right)$$

respectively. κ is the shape parameter common to the two groups to be compared.

6.3.3 Simulation

This methodology will be illustrated via a simulation study. We want 90% power to detect an improvement of 0.5 on the median ratio scaled ($m_S = 1$, $m_E = 1.5$) by conducting a 5%-level one-sided test. We use one-quarter of the number of failures required in the fixed sample design as f_1. The total number of available degrees of freedom are set equal to $L = 10$. The minimum number of failures to be observed in each stage is set equal to $f_{\min} = f_1 / 2$. The required number of failures for the l-th stage is computed using (6.3.7) where the sequence $\{\varepsilon_l\}$ is determined as in (6.3.9) with $\beta_g = 0.8$. The reliability curves used in simulation are displayed in figure 6.1. The curve of the experimental group has a red colour and that for the standard group a green colour. The parameters of the underlying reliability distributions correspond to the medians $m_S = 1$ and $m_E = 1.5$ for the standard and experimental group respectively.

For the Weibull distribution, the shape parameter κ is set to 1.5 corresponding to an increasing hazard function.



Configuration I:

Exp(0.462), Exp(0.693)

Configuration II:

Weibull(1.915, 1.5), Weibull(1.276, 1.5)

Configuration III:

Exp(0.693), Weib(1.915, 1.5)

Figure 6.1: *Reliability curves used in simulation under the tree configurations*

The simulated values are based on 1000 independent replications run in the statistical software R.

Configuration I

In the case, we considered that the reliability of the experimental group and the standard group are exponentially distributed with medians $m_E = 1.5$ and $m_S = 1$ respectively. Therefore the total number of failures requires achieving a power of 90% for a 5% level test is 209. We considered the strategy of varying the accrual period for 4 to 7 months. Therefore, according to formula (6.16), we need to enrol 40 specimens per group and per month in the study so that at the end of the 7 th month, the total sample size will be equal to 560. After observing f_l failure during the *l*th stage, we compute the test statistic T_l for the new procedure and the procedure works as described in section 6.2.2. For purposes of comparison, we also compute the test statistic for the procedure based on the log-rank test (**Log-rank**) developed in section 6.2 under the same design. Preliminary simulations showed that the global critical value given by

$$cv_\alpha = \chi^2(L)_{1-\alpha}$$

was very conservative; therefore we adjust in the new procedure as

$$cv_\alpha = \chi^2(L)_{1-\alpha} - 0.5$$

and in the Log-rank as

$$cv_\alpha = \chi^2(L)_{1-\alpha} - 7.7\,.$$

We also introduce a lower bound for an early acceptance of H_0, that is

$$U_l \leq \chi^2(v_\Sigma(l))_{1-\alpha_L},$$

where $\alpha_L = 0.6$.

We considered here the upper bounds in the sample size functions M_l and m_l as $M_l \leq 180$, $m_l \leq 180$. The empirical size, the empirical power, the average number of failures observed, the average study duration are presented in table 6.4 for both the new procedure and the Log-rank. The new procedure gives a type I error rate of 0.049 and the Log-rank a type I error rate of 0.046. This indicates that the new procedure maintains the type I error rate to the pre-specified degree. The new procedure provides a power of 0.893 which is very good although it is slightly lower than the power for the Log-rank. It is not surprising that the Log-rank has a very high power because it is known that the log-rank test is locally most powerful against the proportional hazards alternatives. The average study duration and the average number of observed failures are not significantly different for the two procedures. This indicates that the new procedure is comparable to the procedure based on the log-rank test under the proportional hazard alternatives.

Hypotheses	Characteristics	New procedure	Log-rank
Null Hypothesis ($\theta = 1$)	Empirical Size	0.049	0.046
	Average Number of Failures	201.659	218.04
	Average Study Duration	3.803	4.054
Alternative Hypothesis ($\theta = 1.5$)	Empirical Power	0.893	0.917
	Average Number of Failures	232.26	224.07
	Average Study Duration	4.526	4.400

Table 6.4: *Empirical size, empirical power, average number of failures and average study duration for the new procedure and the Log-rank under configuration I*

Configuration II

We considere that the reliability function of the experimental group and the standard group are Weibull distributed with medians $m_E = 1.5$ and $m_S = 1$ respectively. More precisely, the reliability of the experimental group follows a Weibull(1.915, 1.5) and that for the standard group a Weibull(1.276, 1.5). Therefore the total number of failures requires achieving a power of 90% for a 5% level test is 93. We considered the strategy of fixed accrual period for 1.5 months and follow-up duration varying for 1.5 to 2.3 months. Therefore, according to formula (6.16), we need to enrol 50 specimens per group and per month in the study so that at the end of the 2.3th month, the total sample size will be equal to 230. We adjusted the critical value in the Log-rank procedure as

$$cv_\alpha = \chi^2(L)_{1-\alpha} - 1.5 .$$

We also introduce a lower bound for an early acceptance of H_0, that is

$$U_l \le \chi^2(\nu_\Sigma(l))_{1-\alpha_L} ,$$

where $\alpha_L = 0.6$.

We considered here the upper bounds in the sample size functions M_l and m_l as

$M_l \le 500$, $m_l \le 500$. The empirical size, the empirical power, the average number of failures observed, the average study duration are presented in table 6.5 for both the new procedure and the Log-rank. The new procedure gives a type I error rate of 0.054 and the Log-rank a type I error rate of 0.05. This indicates that the new procedure maintains the type I error rate to a satisfactory degree. The new procedure does not attained high efficiency in this case although its power is slightly bigger than the Log-rank. The average number of failures in the new procedure is smaller than in the Log-rank and the new procedure stops the study earlier than the Log-rank.

Hypotheses	Characteristics	New procedure	Log-rank
	Empirical Size	0.054	0.05
Null Hypothesis	Average Number of Failures	266.315	449.691
($\theta = 1$)	Average Study Duration	2.90	3.991
	Empirical Power	0.758	0.749
Alternative Hypothesis	Average Number of Failures	95.124	126.982
($\theta = 1.5$)	Average Study Duration	2.275	2.633

Table 6.5: *Empirical size, empirical power, average number of failures and average study duration for the new procedure and the Log-rank under configuration II*

Configuration III

We considered that the reliability of the standard group is exponentially distributed with medians $m_S = 1$ and that of the experimental group follows a Weibull(1.915, 1.5) with median $m_E = 1.5$. We considered the strategy of varying the accrual period for 2 to 4 months.

In this case, there is no formula for determining the total number of failures requires achieving a power of 90% for a 5% level test. Therefore we used the usual log-rank test in the fixed-design to determine the number of failures needed. A power of 90% and 5% type I error is reached with the usual log-rank with fixed-sample in the same study design when 270 failures occurred. Therefore, according to formula (6.16), we need to enrol 54 specimens per group and per month in the study so that at the end of the 4 *th* month, the total sample size will be equal to 432. We adjusted the critical value in the Log-rank procedure as

$$cv_\alpha = \chi^2(L)_{1-\alpha} - 4 .$$

We considered here the upper bounds in the sample size functions M_l and m_l as

$M_l \leq 180$, $m_l \leq 180$. The empirical size, the empirical power, the average number of failures observed, the average study duration are presented in table 6.6 for both the new procedure and the Log-rank. The new procedure gives a type I error rate of 0.051 and the Log-rank a type I error rate of 0.051. The Log- rank results in a substantial loss of power. In contrast, the proposed procedure is effective to achieve a larger power. The new procedure gains 5.3% more power than the Log-rank. The Log-rank tends to end the experiment earlier than the new procedure.

Hypotheses	**Characteristics**	**New procedure**	**Log-rank**
	Empirical Size	0.051	0.051
Null Hypothesis	Average Number of Failures	361.728	246.781
($\theta = 1$)	Average Study Duration	6.821	4.292
	Empirical Power	0.928	0.874
Alternative Hypothesis	Average Number of Failures	199.002	159.482
($\theta = 1.5$)	Average Study Duration	3.398	3.075

Table 6.6: *Empirical size, empirical power, average number of failures and average study duration for the new procedure and the Log-rank under configuration III*

These results clearly indicate the advantages of the new procedure over the Log-rank under nonproportional alternatives.

7 Conclusion

Sample size determination is a vital tool for experiment planning and almost always difficult. It requires care in eliciting scientific objectives and in obtaining suitable quantitative information prior to the experiment. We presented in chapter 3 formulas for sample size for some currently used statistical methodologies and study designs in industry for the purpose of testing of hypothesis. Performing a valid sample size determination requires estimates of the variability in the data, as well as defining the effect size sought. Misspecification of these two parameters results in either an over-sized study, that is neither economical nor time-consuming to the experimenter, or an under-powered study, that may give inconclusive results. Two methodologies to avoid this problem have been presented in chapter 4: The internal pilot study as proposed by Wittes and Brittain that allows recalculation of the sample size during an ongoing experiment using the estimated variance obtained from an interim analysis. The self-designing procedure is the second methodology. The proposal of Shen and Fisher (1999) and the proposal of Hartung (2001) have been described. Furthermore, following the idea of Yin and Shen (2005), a design has been proposed, that combined the advantages of self-designing of Hartung and classical group sequential design by introducing a stop for efficacy in the self-designing procedure of Hartung (2001). The resulting design can update the sample size at each stage, the maximum number of stage to be performed is not fixed in advance and the experiment can be stopped with rejection of the null hypothesis after the first stage. Compared to the self-designing procedure of Hartung (2001), the proposed design presents better characteristics in terms of average sample number and power.

In chapter 5 the methods of internal pilot study and self-designing have been used for sample size adaptation in equivalence study. Two methods of estimating the variance in the internal pilot study procedure have been compared through simulations. The comparison was for the situation where the data follow a normal distribution and the equivalence is formulated in terms of difference between the means of the outcome variable of interest. Here are some important conclusions:

- Since the effect size is expected to be quite small in non-inferiority or equivalence experiments, the within-group variance is under the alternative hypothesis close to the total variance estimated by the one-sample variance. Therefore if the initial variance guess is right, it results in too large sample size as needed.
- In both cases, the type I error rate can be quantified and can be controlled.

- In general, we suggest the use of the pooled variance estimated because it presents better characteristics in terms of power and average sample number.

By the adaptation using the self designing procedure, the illustrated method in comparison to the fixed-sample design has small average sample number when achieving the same power.

In chapter 6, a self-designing rule similar to that of Hartung and Knapp (2003) has been used for sample size reestimation for censored reliability data with a staggered entry. To do that, we needed for the case of proportional hazard functions the results of the work of Tsiatis (1981, 1982) for the construction of the test statistics that we used in each stage, and those of Gail, DeMets and Slud (1981) and Tsiatis (1982) to proof the asymptotical independency of the test statistic at each stage and therefore the use of the self-designing procedure. The test statistic is linear rank statistic based on the log-rank test. The number of failures is determined sequentially based on observed data or cumulated information. For the two strategies of adjusting the design investigated by simulations, the strategy of continuing the experiment by extending the accrual period and fixing the follow-up period works well than the strategy of extending the follow-up period and fixing the accrual period. Because the log-rank test behaves poorly for nonproportional hazards, a method of updating sample size using a test statistic which belong to the class of test based on the integrated weighted difference in the Kaplan-Meier estimates proposed by Pepe and Fleming (1989, 1991) has been developed and illustrated. The asymptotically independency of the new statistic was proved using results of the work of Murray and Tsiatis (1999). Simulation results show the advantages of this new statistic over the later one based on the log-rank statistic under some alternatives. The new adaptive procedure to compare two reliability curves based on the Kaplan-Meier estimates provides a useful alternative to adaptive procedure based on rank statistics for censored reliability data.

In chapter 5, we investigated the sample size reestimation for equivalence experiments using the internal pilot study procedure. More attention needs to be given to situations where the variances in the two groups are not equal or the outcomes of interest are not normally distributed. Of interest will be also the case where more than two groups are to be compared.

The adaptation procedure for censored reliability data proposed in chapter 6 may not be completely efficient. It would be desirable to investigate more efficient weight functions and strategies. Moreover, it would be interesting to generalize the approach in the following way: Divide the calendar time into a finite number of intervals and consider the situation where the accrual rate is piecewise constant over each interval or the hazard rates for failure are piecewise constant over each interval but the hazard ratio is constant. These are situations which may happen in the production process.

References

Anderson S. and Hauck W. W. (1981): *Testing Equivalence in Comparative Bioavailability Trials.* Presented at August, 1981 Joint Statistical Meetings in Detroit, Michigan.

Anderson S. and Hauck W. W. (1983): *A new procedure for testing equivalence in comparative bioavailability and other clinical trials*, Communication in Statistics: Theory and Methods 2, 2663-2692.

Bauer P. (1992): *The choice of sequential boundaries based on the concept of power spending,* Biometrie und Informatik in Medizin und Biologie 23, 3-15.

Bauer P. and Köhne K. (1994): *Evaluation of experiments with adaptive interim Analyses*, Biometrics **50**, 1029-1041.

Benner A. (2000): *A Comparison of Sample Size Methods for Survival Studies*, DKFZ, Heidelberg.

Birkett M. A., Day S. J. (1994). *Internal pilot studies for estimating the sample size,* Statistics in Medicine 13, 2455-2463.

Brannath W., Posch M. and Bauer P. (2002): *Recursive combination tests*, Journal of the American Statistical Association 97, 236-244.

Breslow N. and Crowley J. (1974): *A large sample study on the life table and product limit estimates under random censorship*, Annals of Statistics 2, 437-453.

Castelloe J. (2000): *Sample Size Computations and Power Analysis with the SAS System*, Cary, NC, SAS Institute, Inc., Paper 265-25.

Cohen J. (1988): *Statistical Power Analysis for the Behavioural Sciences*, Academic Press, New York, 2nd edn.

Denne J. S., Jennison C. (1999). *Estimating the sample size for a t-test using an internal pilot*. Statistics in Medicine 18, 1575-1585.

Desu M. M. and Raghavarao D. (1990): *Sample Size Methodology*, Academic Press, Boston.

Elashoff J. (2002): *nQuery Advisor Release 5.0*, Statistical Solutions, Cork, Ireland.

Fisher L. (1998). *Self-Designing Clinical Trials*. Statistics in Medicine 17, 1551-1562.

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

Friede T., Kieser M. (2003). *Blinded sample size reassessment in non-inferiority and equivalence trials,* Statistics in Medicine 22, 995-1007.

Friede T., Kieser M. (2002). *On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation,* Statistics in Medicine 21, 165-176.

Friede T., Kieser M. (2001). *A comparison of methods for adaptive sample size adjustment,* Statistics in Medicine 20, 3861-3873.

Friede T., Kieser M. (2001). *Sample size adjustment in clinical trials for proving equivalence*, Drug Information Journal 35, 1401-1408.

Gail M., DeMets D. L., Slud E. V. (1982). *Simulation studies on increments of the two-sample logrank score test for survival time data, with application to group sequential boundaries.* In Survival Analysis, eds, J. Crowley and R. A. Johnson, Hayward CA: Institute of mathematical statistics, pp.287-301.

George S. and Desu M. (1974): *Planning the size and duration of a clinical trial studying the time to some critical event*, Journal of Chronic Diseases, 27: 15-24.

Hartung J. (2001). *A Self-Designing Rule for Clinical Trials with Arbitrary Response Variables*, Controlled Clinical Trials 22, 111-116.

Hartung J., Knapp G. (2003): *A New Class of Completely Self-Designing Clinical Trials*, Biometrical Journal 45, 3-19.

Hauschke D., KieserM., Diletti E. and Burke M. (1999): *Sample size determination for proving equivalence based on the ratio of two means for normally distributed data*, Statistics in Medicine 18, 93-105.

Hedges L. V. and Olkin I. (1985*): Statistical methods for meta-analysis*, Academic Press, Orlando.

Hintze J. (2000): *PASS 200*, Number Cruncher Statistical Systems, Kaysville, UT.

Kieser M., Friede T. (2000). *Re-calculating the sample size in internal pilot study designs with control of the type I error rate*, Statistics in Medicine 19, 901-911.

Kieser M. and Hauschke D. (1999): *Approximate sample sizes for testing hypotheses about the ratio and difference of two means.* Journal of Biopharmaceutical Statistics 9, 641-650.

Kraemer H. C. and Thieman S. (1987): *How many subjects? Statistical Power Analysis in Research*, Sage Publications, Newbury Park, CA.

Lachin J. M. and Foulkes M. A. (1986): *Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, non-compliance and stratification*, Biometrics, 42, 507-519.

Lakatos E. (1988): *Sample sizes based on the log-rank statistic in complex clinical trials*, Biometrics 42, 229-241.

Lan K. K. G. and Lachin J. M. (1990): *Implementation of group sequential Logrank tests in a maximum duration trial*, Biometrics 46, 759-770.

Lehmacher W., Wassmer G. (1999): *Adaptive sample size calculations in group sequential trials,* Biometrics 55, 1286–1290.

Lenth R. V. (2000): *Java applets for power and sample size*, http://www.stat.uiowa.edu/~rlenth/power/.

Lin D. Y. (1991): *Nonparametric sequential testing in clinical trials with incomplete multivariate observations*, Biometrika 78, 123-131.

Lin D. Y., Shen L., Ying Z. and Breslow N. E. (1996): *Group sequential designs for monitoring survival probabilities*, Biometrics 52, 1033-1041.

Lin D. Y., Yao Q. and Ying Z. (1999): *A general theory on stochastic curtailment for censored survival data*, Journal of the American Statistical Association 94, 510-521.

Mace A. E. (1964): *Sample size determination*, Rheinhold, New York.

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

Metzler C. M. (1974): *Bioavailability – A Problem in Equivalence*, Biometrics 30, 309-317

Moonseong H., Myles S. F. and David B. A. (1998): *Power and sample size for survival analysis under the Weibull distribution when the whole lifespan is of interest*, Mechanisms of ageing and development 102, 45-53.

Moshman J. (1958): *A method for selecting the size of the initial sample in Stein's two sample procedure.* Annals of Mathematical Statistics 29, 1271-1275.

Murray S. and Tsiatis A. (1999): *Sequential Methods for Comparing Years of Life Saved in the Two-Sample Censored Data Problem*, Biometrics 55, 1085-1092.

Nelson W. (1969): *Hazard plotting for incomplete failure data*, Journal of Quality Technology 1, 27-52.

Oakes D. (2001): *Biometrika Centenary: Survival Analysis*, Biometrika 88, 99-142.

O'Brien P. C. and Fleming T. R. (1979): *A multiple testing procedure for clinical trials.* Biometrics 35, 549-556.

Pepe M. S. and Fleming T. R. (1989): *Weighted Kaplan-Meier Statistics – A class of distance tests for censored survival data*, Biometrics 45, 497-507.

Pepe M. S. and Fleming T. R. (1991): *Weighted Kaplan-Meier Statistics: Large sample and optimality considerations,* Journal of the Royal Statistical Society, Series B 53, 341-352.

Phillips K. (1990): *Power of the two one-sided tests procedure in bioequivalence*, Journal of Pharmokinetics and Biopharmaceutics 18, 137-144.

Pocock, S. J. (1977). *Group sequential methods in the design and analysis of clinical trials.* Biometrika 64, 191-199.

Proschan M. A. and Hunsberger S. A. (1995): *Designed extension of studies based on conditional power.* Biometrics 51**,** 1315-1324.

Rogers J., Howard K. and Vessey J. (1993): *Using significance tests to evaluate equivalence between two experimental groups*, Psychological Bulletin 113 (3), 553-565.

Sandvik L, Erikssen J, Mowinckel P., Rodland EA (1996). *A method for determining the size of internal pilot studies*, Statistics in Medicine 15, 1587-1590.

Sasabuchi S. (1988): *A multivariate one-sided test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor.* Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics 42, 9-19.

Schoenfeld D. A. (1981): *The asymptotic properties of nonparametric tests for comparing survival distributions,* Biometrika 68, 316-319.

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

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

Schuirmann D. (1981): *On hypothesis testing to determine if the mean of the normal distribution is contained in a known interval*, Biometrics 37, 617-

Schuirmann D. (1987): *A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability*. Journal of Pharmokinetics and Biopharmaceutics 15, 657-680.

Seelbinder B. M. (1953): *On Stein's two-stage sampling scheme*, Annals of Mathematical Statistics 24, 640-649.

Sheiner L.B. (1992): *Bioequivalence revisited*, Statistics in Medicine, 11, 1777-1788.

Shen Y. and Cai J. (2003): *Sample size reestimation for clinical trials with censored survival data*, Journal of the American Statistical Association 98, 418-426.

Shen Y., Fisher L. (1999). *Statistical inference for self-designing clinical trials with one-sided hypothesis*. Biometrics 55, 190-197.

Shorack G. R. (2000*): Probability for Statisticians*, Springer-Verlag, New York, Inc.

Stein C. (1945). *A two-sample test for a linear hypothesis whose power is independent of the variance.* Annals of Mathematical Statistics 16, 243-258.

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

Tsiatis A. A. (1981). *The asymptotic joint distribution of the efficient score test for the proportional hazards model calculated over time.* Biometrika 68, 311-315.

Tsiatis A. A. (1982). *Repeated significance testing for a general class of statistics used in censored survival data*. Journal of American Statistical Association 77, 855-861.

Tsiatis A., Rosner G. L., Tritchler D. L. (1985): *Group sequential tests with censored survival data adjusting for covariates*, Biometrika 72, 365-373.

Westlake W. J. (1976): *Symmetric confidence Intervals for Bioequivalence Trials*, Biometrics 32, 741-744

Westlake W. J. (1979): *Design and Statistical Evaluation of Bioequivalence Studies in Man*. Principles and Perspective in Drug Bioavailability, Karger, Basel, 192-210

Wittes J., Brittain E. (1990). *The role of internal pilot studies in increasing the efficiency of clinical trials,* Statistics in Medicine 9, 65-72.

Wittes J., Schabenberger O., Zucker D., Brittain E., Proschan M. (1999): *Internal pilot studies: Type I error rate of the naive t-test*, Statistics in Medicine 18, 3481-3491.

Yin G., Shen Y. (2005): *Self-Designing Trial Combined with Classical Group Sequential Monitoring,* Journal of Biopharmaceutical Statistics 15, 667-675.