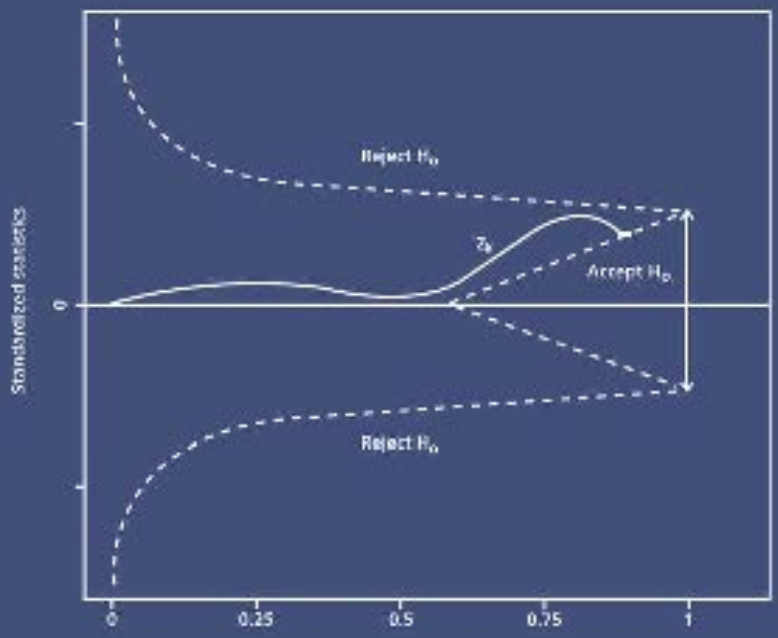
Sample Size Calculations In Clinical Research



Shein-Chung Chow
Jun Shao
Hansheng Wang

the same region or from region to region, while generalizability is referred to as whether the clinical results can be generalized to other similar patient populations within the same region or from region to region. When the sponsor of a newly developed or approved drug product is interested in getting the drug product into the marketplace from one region (e.g., where the drug product is developed and approved) to another region, it is a concern that differences in ethnic factors could alter the efficacy and safety of the drug product in the new region. As a result, it is recommended that a bridging study be conducted to generate a limited amount of clinical data in the new region in order to extrapolate the clinical data between the two regions (ICH, 1998a).

In practice, it is often of interest to determine whether a clinical trial that produced positive clinical results provides substantial evidence to assure reproducibility and generalizability of the clinical results. In this chapter, the reproducibility of a positive clinical result is studied by evaluating the probability of observing a positive result in a future clinical study with the same study protocol, given that a positive clinical result has been observed. The generalizability of clinical results observed from a clinical trial will be evaluated by means of a sensitivity analysis with respect to changes in mean and standard deviation of the primary clinical endpoints of the study.

1.1.4 Substantial Evidence with a Single Trial

Although the FDA requires that at least two adequate and well-controlled clinical trials be conducted for providing substantial evidence regarding the effectiveness of the drug product under investigation, a single trial may be accepted for regulatory approval under certain circumstances. In 1997, FDA published the *Modernization Act* (FDAMA), which includes a provision (Section 115 of FDAMA) to allow data from one adequate and well-controlled clinical trial investigation and confirmatory evidence to establish effectiveness for risk/benefit assessment of drug and biological candidates for approval under certain circumstances. This provision essentially codified an FDA policy that had existed for several years but whose application had been limited to some biological products approved by the Center for Biologic Evaluation and Research (CBER) of the FDA and a few pharmaceuticals, especially orphan drugs such as zidovudine and lamotrigine. As it can be seen from Table 1.1.2, a relatively strong significant result observed from a single clinical trial (say, p-value is less than 0.001) would have about 90% chance of reproducing the result in future clinical trials.

Consequently, a single clinical trial is sufficient to provide substantial evidence for demonstration of efficacy and safety of the medication under study. However, in 1998, FDA published a guidance which shed the light

Test for Superiority

To show superiority of a test drug over an active control agent or a standard therapy, we may consider the following hypotheses:

$$H_0 : \mu_T - \mu_S \leq \delta \quad \text{versus} \quad H_a : \mu_T - \mu_S > \delta. \tag{1.2.3}$$

The rejection of the above null hypothesis suggests that the difference between the test drug and the standard therapy is greater than a clinically meaningful difference. Therefore, we may conclude that the test drug is superior to the standard therapy by rejecting the null hypothesis of (1.2.3). Note that the above hypotheses are also known as hypotheses for testing *clinical* superiority. When $\delta = 0$, the above hypotheses are usually referred to as hypotheses for testing *statistical* superiority.

Test for Equivalence

In practice, unless there is some prior knowledge regarding the test drug, usually we do not know the performance of the test drug as compared to the standard therapy. Therefore, hypotheses (1.2.2) and (1.2.3) are not preferred because they have pre-determined the performance of the test drug as compared the standard therapy. As an alternative, the following hypotheses for therapeutic equivalence are usually considered:

$$H_0 : |\mu_T - \mu_S| \geq \delta \quad \text{versus} \quad H_a : |\mu_T - \mu_S| < \delta. \tag{1.2.4}$$

We then conclude that the difference between the test drug and the standard therapy is of no clinical importance if the null hypothesis of (1.2.4) is rejected.

It should be noted that a valid sample size calculation can only be done based on appropriate statistical tests for the hypotheses which can reflect the study objectives under a valid study design. It is then suggested that the hypotheses be clearly stated when performing a sample size calculation. Each of the above hypotheses has different requirement for sample size in order to achieve a desired power or precision of the corresponding tests.

1.2.4 Primary Study Endpoint

A study objective (hypotheses) will define what study variable is to be considered as the primary clinical endpoint and what comparisons or investigations are deemed most clinically relevant. The primary clinical endpoints depend upon therapeutic areas and the indications that the test drugs sought for. For example, for coronary artery disease/angina in cardiovascular system, patient mortality, is the most important clinical endpoint

perform a sample size re-estimation with unblinding of the treatment codes as even the significance level will be adjusted for potential statistical penalty for the unblinding. Thus, sample size re-estimation without unblinding the treatment codes has become very attractive. Shih (1993) and Shih and Zhao (1997) proposed some procedures without unblinding for sample size re-estimation within interim data for double-blinded clinical trials with binary outcomes. Detailed procedure for sample size re-estimation with unblinding will be given in Chapter 8.

In practice, it is suggested that procedure for sample size re-estimation be specified in the study protocol and should be performed by an external statistician who is independent of the project team. It is also recommended that a data monitoring committee (DMC) be considered to maintain the scientific validity and integrity of the clinical trial when performing sample size re-estimation at the interim stage of the trial. More details regarding sample size re-estimation are provided in Chapter 8.

1.4 Aims and Structure of the Book

1.4.1 Aim of the Book

As indicated earlier, sample size calculation plays an important role in clinical research. Sample size calculation is usually performed using an appropriate statistical test for the hypotheses of interest to achieve a desired power for detection of a clinically meaningful difference. The hypotheses should be established to reflect the study objectives for clinical investigation under the study design. In practice, however, it is not uncommon to observe discrepancies among study objectives (or hypotheses), study design, statistical analysis (or test statistic), and sample size calculation. These inconsistencies often result in (i) wrong test for right hypotheses, (ii) right test for wrong hypotheses, (iii) wrong test for wrong hypotheses, or (iv) right test for right hypotheses with insufficient power. Therefore, the aim of this book is to provide a comprehensive and unified presentation of statistical concepts and methods for sample size calculation in various situations in clinical research. Moreover, the book will focus on the interactions between clinicians and biostatisticians that often occur during various phases of clinical research and development. This book is also intended to give a well-balanced summarization of current and emerging clinical issues and recent developed statistical methodologies in the area of sample size calculation in clinical research. Although this book is written from a viewpoint of clinical research and development, the principles and concepts presented in this book can also be applied to a non-clinical setting.

Table 2.2.1: Sample Sizes Based on One-Sided Test and Two-Sided Test At α-level of Significance

		One-sided Test		Two-sided Test	
α	δ^*	80%	90%	80%	90%
0.05	0.25σ	198	275	252	337
	0.50σ	50	69	63	85
	1.00σ	13	18	16	22
0.01	0.25σ	322	417	374	477
	0.50σ	81	105	94	120
	1.00σ	21	27	24	30

Some researchers from the academia and the pharmaceutical industry consider this false positive rate is acceptable and the evidence provided by the two clinical trials using the one-sided test procedure is rather substantial. Hence, the one-sided test procedure should be recommended. However, in practice a two-sided test may be preferred because placebo effect may be substantial in many drug products such as drug products regarding diseases in the central nervous system.

2.2.1 Remark

Dubey (1991) indicated that the FDA prefers a two-sided test over a one-sided test procedure in clinical research and development of drug products. In situations where (i) there is truly only concern with outcomes in one tail, and (ii) it is completely inconceivable that results could go in the opposite direction, one-sided test procedure may be appropriate (Dubey, 1991). Dubey (1991) provided situations where one-sided test procedure

Table 2.2.2: Comparison Between One-Sided Test and Two-Sided Test At α-level of Significance

Characteristic	One-sided Test	Two-sided Test
Hypotheses	Non-inferiority	Equality
	Superiority	Equivalence
One Trial	1/20	1/40
Two Trials	1/400	1/1600

Thus, O'Brien-Fleming boundary for the first interim analysis is equal to $(2.04)(\sqrt{5}) = 4.561$. The O'Brien-Fleming boundaries for the other 4 interim analyses can be similarly computed as 3.225, 2.633, 2.280, and 2.040, respectively. The O'Brien-Fleming boundaries are very conservative so that the early trial results must be extreme for any prudent and justified decision-making in recommendation of a possible early termination when very limited information is available. On the other hand, for the late phase of the trial when the accumulated information approaches the required maximum information, their boundaries also become quite close to the critical value when no interim analysis had been planned. As a result, the O'Brien-Fleming method does not require a significant increase in the sample size for what has already planned. Therefore, the O'Brien-Fleming group sequential boundaries have become one of the most popular procedures for the planned interim analyses of clinical trials.

2.4.2 Alpha Spending Function

The idea of the alpha spending function proposed by Lan and DeMets (1983) is to spend (i.e., distribute) the total probability of false positive risk as a continuous function of the information time. The implementation of the alpha spending function requires the selection and specification of the spending function in advance in the protocol. One cannot change and choose another spending function in the middle of trial. Geller (1994) suggested that the spending function should be convex and have the property that the same value of a test statistic is more compelling as the sample sizes increase. Since its flexibility and no requirement for total information and equal increment of information, there is a potential to abuse the alpha spending function by increasing the frequency of interim analyses as the results approach to the boundary. However, DeMets and Lan (1994) reported that alteration of the frequency of interim analyses has very little impact on the overall significance level if the O'Brien-Fleming-type or Pocock-type continuous spending function is used.

Pawitan and Hallstrom (1990) studied the alpha spending function with the use of the permutation test. The permutation test is conceptually simple and it provides an exact test for small sample sizes. In addition, it is valid for complicated stratified analysis in which the exact sampling distribution is in general unknown and large-sample approximation may not be adequate. Consider the one-sided alternative. For the kth interim analyses, under the assumption of no treatment effect, the null joint permutation distribution of test statistics $(Z_1, ..., Z_K)$ can be obtained by random permutation of treatment assignments on the actual data. Let $(Z^*_{1b}, ..., Z^*_{Kb}), b = 1, ..., B$, be the statistics computed from B treatment assignments and B be the total number of possible permutations. Given

$\alpha(s_1)$, $\alpha(s_2) - \alpha(s_1)$,..., $\alpha(s_k) - \alpha(s_{K-1})$, the probabilities of type I error allowed to spend at successive interim analyses, the one-sided boundaries $z_1, ...z_K$ can be determined by

$$\frac{\#\ of\ (Z_1^* > z_1)}{B} = \alpha(s_1),$$

and

$$\frac{\#\ of\ (Z_1^* > z_1 \text{ or } Z_2^* > z_2, ..., \text{ or} Z_k^* > z_k)}{B} = \alpha(s_k) - \alpha(s_{k-1}),$$

$k = 1, ..., K$. If B is very large, then the above method can be executed with a random sample with replacement of size B. The α spending function for an overall significance level of 2.5% for one-sided alternative is given by

$$\alpha(s) = \begin{cases} \frac{\alpha}{2}s, & \text{if } s < 1, \\ \alpha, & \text{if } s = 1. \end{cases}$$

In the interest of controlling the overall type I error rate at the α level of significance, sample size is necessarily adjusted according to the α spending function to account for the planned interim analyses. In some cases, sample size re-estimation without unblinding may be performed according to the procedure described in Section 1.3 of Chapter 1. More details can be found in Chapter 8.

2.5 Data Transformation

In clinical research, data transformation on clinical response of the primary study endpoint may be necessarily performed before statistical analysis for a more accurate and reliable assessment of the treatment effect. For example, for bioavailability and bioequivalence studies with healthy human subjects, the FDA requires that a log-transformation be performed before data analysis. Two drug products are claimed bioequivalent in terms of drug absorption if the 90% confidence interval of the ratio of means of the primary pharmacokinetic (PK) parameters, such as area under the blood or plasma concentration-time curve (AUC) and maximum concentration (Cmax), is entirely within the bioequivalence limits of (80%,125%). Let μ_T and μ_R be the population means of the test drug and the reference drug, respectively. Also, let X and Y be the PK responses for the test drug and the reference drug. After log-transformation, we assume that $\log X$ and $\log Y$ follow normal distributions with means μ_X^* and μ_Y^* and variance σ^2. Then

$$\mu_T = E(X) = e^{\mu_X^* + \frac{\sigma^2}{2}} \quad \text{and} \quad \mu_R = E(Y) = e^{\mu_Y^* + \frac{\sigma^2}{2}},$$

Table 2.5.1: Posterior Power Evaluation Under a Crossover Design

Data Type	Power
Raw Data	$1 - 2\Phi\left(\frac{0.2}{\mathrm{CV}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} - t_{\alpha,n_1+n_2-2}\right)$
Log-transformed Data	$1 - 2\Phi\left(\frac{0.223}{\sigma_e\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} - t_{\alpha,n_1+n_2-2}\right)$

BE limit for the log-transformed data is also symmetric about 0 after log transformation.

2.5.1 Remark

For the crossover design, since each subject serves as its own control, the inter-subject variation is removed from comparison. As a result, the formula for sample size calculation derived under a crossover design only involves the intra-subject variability. On the other hand, for the parallel design, formula for sample size calculation under a parallel design includes both the inter- and intra-subject variabilities. In practice, it is easy to get confused with the sample size calculation and/or evaluation of posterior power based on either raw data or log-transformed data under either a crossover design or a parallel design (Chow and Wang, 2001). As an example, posterior powers based on raw data and log-transformed data under a crossover design when the true mean difference is 0 are given in Table 2.5.1.

2.6 Practical Issues

2.6.1 Unequal Treatment Allocation

In a parallel design or a crossover design comparing two or more than two treatments, sample sizes in each treatment group (for parallel design) or in each sequence of treatments (for crossover design) may not be the same. For example, when conducting a placebo-control clinical trial with very ill patients or patients with severe or life-threatening diseases, it may not be ethical to put too many patients in the placebo arm. In this case, the investigator may prefer to put fewer patients in the placebo (if the

(Williams design), respectively. Section 3.6 discusses some practical issues regarding sample size calculation for comparing means in clinical research, including sample size reduction when switching from a two-sided test to a one-sided test or from a parallel design to a crossover design, sensitivity analysis with respect to change in variability, and a brief discussion regarding Bayesian approach.

3.1 One-Sample Design

Let x_i be the response from the ith sampled subject, $i = 1, ..., n$. In clinical research, x_i could be the difference between matched pairs such as the pre-treatment and post-treatment responses or changes from baseline to endpoint within a treatment group. It is assumed that x_i's are independent and identically distributed (i.i.d.) normal random variables with mean 0 and variance σ^2. Let

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.1.1}$$

be the sample mean and sample variance of x_i's, respectively. Also, let $\epsilon = \mu - \mu_0$ be the difference between the true mean response of a test drug (μ) and a reference value (μ_0). Without loss of generality, consider $\epsilon > 0$ ($\epsilon < 0$) an indication of *improvement* (*worsening*) of the test drug as compared to the reference value.

3.1.1 Test for Equality

To test whether there is a difference between the mean response of the test drug and the reference value, the following hypotheses are usually considered:

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

When σ^2 is known, we reject the null hypothesis at the α level of significance if

$$\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2},$$

where z_a is the upper ath quantile of the standard normal distribution. Under the alternative hypothesis that $\epsilon \neq 0$, the power of the above test is given by

$$\Phi\left(\frac{\sqrt{n}\epsilon}{\sigma} - z_{\alpha/2}\right) + \Phi\left(-\frac{\sqrt{n}\epsilon}{\sigma} - z_{\alpha/2}\right),$$

The power of this test is given by

$$1 - \mathcal{T}_{n-1}\left(t_{\alpha,n-1}\middle|\frac{\sqrt{n}(\epsilon-\delta)}{\sigma}\right).$$

The sample size needed to achieve power $1-\beta$ can be obtained by solving

$$\mathcal{T}_{n-1}\left(t_{\alpha,n-1}\middle|\frac{\sqrt{n}(\epsilon-\delta)}{\sigma}\right) = \beta.$$

By letting $\theta = (\epsilon-\delta)/\sigma$, Table 3.1.1 can be used to find the required sample size. From approximation (3.1.5), formula (3.1.6) can be used to calculate the required sample size when n is sufficiently large.

3.1.3 Test for Equivalence

The objective is to test the following hypotheses

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

The test drug is concluded to be equivalent to a gold standard on average if the null hypothesis is rejected at significance level α.

When σ^2 is known, the null hypothesis H_0 is rejected at significance level α if

$$\frac{\sqrt{n}(\bar{x}-\mu_0-\delta)}{\sigma} < -z_\alpha \quad \text{and} \quad \frac{\sqrt{n}(\bar{x}-\mu_0+\delta)}{\sigma} > z_\alpha.$$

The power of this test is given by

$$\Phi\left(\frac{\sqrt{n}(\delta-\epsilon)}{\sigma} - z_\alpha\right) + \Phi\left(\frac{\sqrt{n}(\delta+\epsilon)}{\sigma} - z_\alpha\right) - 1. \tag{3.1.7}$$

Although the sample size n can be obtained by setting the power in (3.1.7) to $1-\beta$, it is more convenient to use the following method. Note that the power is larger than

$$2\Phi\left(\frac{\sqrt{n}(\delta-|\epsilon|)}{\sigma} - z_\alpha\right) - 1. \tag{3.1.8}$$

Hence, the sample size needed to achieve power $1-\beta$ can be obtained by solving

$$\Phi\left(\frac{\sqrt{n}(\delta-|\epsilon|)}{\sigma} - z_\alpha\right) = 1 - \frac{\beta}{2}.$$

This leads to

$$n = \frac{(z_\alpha + z_{\beta/2})^2\sigma^2}{(\delta-|\epsilon|)^2}. \tag{3.1.9}$$

Test for Equivalence

Assume that one brand name drug for osteoporosis on the market has a responder rate of 60%. It is believed that a 20% difference in responder rate is of no clinical significance. Hence, the investigator wants to show the study drug is equivalent to the market drug in terms of responder rate. Assume that true responder rate for a study drug is 65%. By (4.1.4), at α=0.05, assuming that the true response rate is 60% (i.e., $p = 0.60$), the sample size required for achieving an 80% power is

$$n = \frac{(z_\alpha + z_{\beta/2})^2 p(1-p)}{(\delta - |p - p_0|)^2} = .$$

4.1.5 Remarks

For one-sample test for equality, there exist another approach, which is very similar to (4.1.1) but not exactly the same. This approach will reject the null hypothesis that $\epsilon = 0$ if

$$\left| \frac{\sqrt{n}\hat{\epsilon}}{\sqrt{p_0(1-p_0)}} \right| > z_{\alpha/2}. \tag{4.1.5}$$

Since (4.1.1) estimates the variance of $\sqrt{n}\hat{\epsilon}$ without any constraints, we refer to (4.1.1) as the unconditional method. One the other hand, since (4.1.5) estimates the variance of $\sqrt{n}\hat{\epsilon}$ conditional on the null hypothesis, we refer to (4.1.5) as the conditional method. Note that both (4.1.1) and (4.1.5) have asymptotic size α when n is sufficiently large. Then, which one should be used is always a dilemma because one is not necessarily more powerful than the other. For the purpose of completeness, the sample size calculation formula based on (4.1.5) is given below. The same idea can be applied to the testing problems of non-inferiority/superiority.

Under the alternative hypothesis ($\epsilon \neq 0$), the power of the test defined by (4.1.5) is approximately

$$\Phi\left(\frac{\sqrt{n}|\epsilon| - z_{\alpha/2}\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}} \right).$$

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

$$\frac{\sqrt{n}|\epsilon| - z_{\alpha/2}\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}} = z_\beta.$$

This leads to

$$n = \frac{[z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p(1-p)}]^2}{\epsilon^2}.$$

4.2 Two-Sample Parallel Design

Let x_{ij} be a binary response from the jth subject in the ith treatment group, $j = 1, ..., n_i$, $i = 1, 2$. For a fixed i, it is assumed that x_{ij}'s are i.i.d. with $P(x_{ij} = 1) = p_i$. In practice, p_i is usually estimated by the observed proportion in the ith treatment group:

$$\hat{p}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}.$$

Let $\epsilon = p_1 - p_2$ be the difference between the true mean response rates of a test drug (p_1) and a control (p_2). Without loss of generality, consider $\epsilon > 0$ ($\epsilon < 0$) an indication of *improvement* (*worsening*) of the test drug as compared to the control value. In what follows, formulas for sample size calculation to achieve a desired power under the alternative hypothesis are derived for testing equality, non-inferiority/superiority, and equivalence.

4.2.1 Test for Equality

To test whether there is a difference between the mean response rates of the test drug and the reference drug, the following hypotheses are usually considered:

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

We reject the null hypothesis at the α level of significance if

$$\left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} \right| > z_{\alpha/2}. \tag{4.2.1}$$

Under the alternative hypothesis that $\epsilon \neq 0$, the power of the above test is approximately

$$\Phi\left(\frac{|\epsilon|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_{\alpha/2} \right).$$

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by the following equation:

$$\frac{|\epsilon|}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} - z_{\alpha/2} = z_\beta.$$

This leads to

$$\begin{aligned} n_1 &= \kappa n_2 \\ n_2 &= \frac{(z_{\alpha/2}+z_\beta)^2}{\epsilon^2} \left[\frac{p_1(1-p_1)}{\kappa} + p_2(1-p_2) \right]. \end{aligned} \tag{4.2.2}$$

This leads to

$$
\begin{aligned}
n_1 &= \kappa n_2 \\
n_2 &= \tfrac{1}{\epsilon^2}[z_{\alpha/2}\sqrt{(1+1/\kappa)p(1-p)} + z_\beta\sqrt{p_1(1-p_1)/\kappa + p_2(1-p_2)}]^2.
\end{aligned}
$$

4.3 Two-Sample Crossover Design

In this section, we consider an $a \times 2m$ replicated crossover design comparing mean response rates of a test drug and a reference drug. Let x_{ijkl} be the lth replicate of a binary response ($l = 1, ..., m$) observed from the jth subject ($j = 1, ..., n$) in the ith sequence ($i = 1, ..., a$) under the kth treatment ($k = 1, 2$). Assume that $(x_{ij11}, ..., x_{ij1m}, ..., x_{ijk1}, ..., x_{ijkm}), i = 1, 2$, $j = 1, ..., n$ are i.i.d. random vectors with each component's marginal distribution specified by $P(x_{ijkl} = 1) = p_k$. Note that the observations from the same subject can be correlated with each other. By specifying that $P(x_{ijkl} = 1) = p_k$, it implies that there is no sequence, period, and crossover effects. The statistical model incorporates those effects are more complicated for binary data compared with continuous data. Its detailed discussion is beyond the scope of this book.

Let $\epsilon = p_2(\text{test}) - p_1(\text{reference})$,

$$\bar{x}_{ijk\cdot} = \frac{1}{m}(x_{ijk1} + \cdots + x_{ijkm}) \quad \text{and} \quad d_{ij} = \bar{x}_{ij1\cdot} - \bar{x}_{ij2\cdot}.$$

An unbiased estimator of ϵ is given by

$$\hat{\epsilon} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} d_{ij}.$$

According to the central limit theorem, $\hat{\epsilon}$ is asymptotically normally distributed as $N(0, \sigma_d^2)$, where $\sigma_d^2 = \text{var}(d_{ij})$ and can be estimated by

$$\hat{\sigma}_d^2 = \frac{1}{a(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{n}(d_{ij} - \bar{d}_{i\cdot})^2,$$

where

$$\bar{d}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} d_{ij}.$$

Without loss of of generality, consider $\epsilon > 0$ ($\epsilon < 0$) as an indication of *improvement* (*worsening*) of the test drug as compared to the reference drug.

Table 5.3.1: Sample Sizes and Critical Values for Two-Stage Designs
($p_1 - p_0 = 0.15$)

				Optimal Design		Minimax Design	
p_0	p_1	α	β	r_1/n_1	r/N	r_1/n_1	r/N
0.05	0.20	0.10	0.10	0/12	3/37	0/13	3/32
		0.05	0.20	0/10	3/29	0/13	3/27
		0.05	0.10	1/21	4/41	1/29	4/38
0.10	0.25	0.10	0.10	2/21	7/50	2/27	6/40
		0.05	0.20	2/18	7/43	2/22	7/40
		0.05	0.10	2/21	10/66	3/31	9/55
0.20	0.35	0.10	0.10	5/27	16/63	6/33	15/58
		0.05	0.20	5/22	19/72	6/31	15/53
		0.05	0.10	8/37	22/83	8/42	21/77
0.30	0.45	0.10	0.10	9/30	29/82	16/50	25/69
		0.05	0.20	9/27	30/81	16/46	25/65
		0.05	0.10	13/40	40/110	27/77	33/88
0.40	0.55	0.10	0.10	16/38	40/88	18/45	34/73
		0.05	0.20	11/26	40/84	28/59	34/70
		0.05	0.10	19/45	49/104	24/62	45/94
0.50	0.65	0.10	0.10	18/35	47/84	19/40	41/72
		0.05	0.20	15/28	48/83	39/66	40/68
		0.05	0.10	22/42	60/105	28/57	54/93
0.60	0.75	0.10	0.10	21/34	47/71	25/43	43/64
		0.05	0.20	17/27	46/67	18/30	43/62
		0.05	0.10	21/34	64/95	48/72	57/84
0.70	0.85	0.10	0.10	14/20	45/59	15/22	40/52
		0.05	0.20	14/19	46/59	16/23	39/49
		0.05	0.10	18/25	61/79	33/44	53/68
0.80	0.95	0.10	0.10	5/7	27/31	5/7	27/31
		0.05	0.20	7/9	26/29	7/9	26/29
		0.05	0.10	16/19	37/42	31/35	35/40

Table 5.3.3: Sample Sizes and Critical Values for Optimal Flexible Two-Stage Designs ($p_1 - p_0$=0.15)

p_0	p_1	α	β	r_i/n_i	R_j/N_j
0.05	0.20	0.10	0.10	0/15-16,1/17-22	2/30-31,3/32-37
		0.05	0.20	0/10-12,1/13-17	3/27-34
		0.05	0.10	1/17-24	4/41-46,5/47-48
0.10	0.25	0.10	0.10	2/19-25,3/26	6/44-45,7/46-51
		0.05	0.20	1/13-15,2/16-20	6/40,7/41-45,8/46-47
		0.05	0.10	2/21-24,3/25-28	9/57-61,10/62-64
0.20	0.35	0.10	0.10	6/28-31,7/32-35	15/62,16/63-65,
					17/66-68,18/69
		0.05	0.20	4/18-21,5/22-24,6/25	17/62-64,18/65-69,
		0.05	0.20	6/31,7/32-34,8/35-38	22/82-85,23/86-89
0.30	0.45	0.10	0.10	9/31,10/32-33	27/75-77,28/78-80
				11/34-37,12/38	29/81-82
		0.05	0.20	7/23,8/24-25,	27/73,28/74-76,
				9/26-29,10/30	29/77-78,30/79-80
		0.05	0.20	11/35-36,12/37-39,	36/98-99,37/100-102,
				13/40-42	38/103-104,39/105
0.40	0.55	0.10	0.10	12/30-31,13/32-33,	37/80-81,38/82-84,
				14/34-35,15/36-37	39/85-86,40/87
		0.05	0.20	11/25-26,12/27-29,	37/78,38/79-80,
				13/30-31,14/32	39/81-82,40/83-85
		0.05	0.10	16/38-39,17/40-41,	49/104-105,50/106-107,
				18/42-44,19/45	51/108-109,52/110-111
0.50	0.65	0.10	0.10	15/30,16/31-32,17/33-34,	44/78-79,45/80-81,
				18/35-36,19/37	46/82-83,47/84,48/85
		0.05	0.20	12/23,13/24-25,14/26-27,	45/77-78,46/79-80,
				15/28-29,16/30	47/81-82,48/83,49/84
		0.05	0.10	21/40,22/41-42,23/43-44,	59/103-104,60/105-106,
				24/45-46,25/47	61/107,62/108-109,63/110
0.60	0.75	0.10	0.10	16/27,17/28,18/29-30,	44/67,45/68,46/69-70,
				19/31-32,20/33,21/34	47/71,48/72,49/73-74
		0.05	0.20	14/22-23,15/24,16/25	46/68,47/69,48/70-71
		0.05	0.10	20/32-33,21/34,22/36-36,	61/90-91,62/92,63/93-94,
				23/37/24/38-39	64/95,65/96-97
0.70	0.85	0.10	0.10	13/19,14/20,15/21,	40/53,41/54,42/55,43/56,
				16/22-23,17/24,18/25-26	44/57-58,45/59,46/60
		0.05	0.20	9/13,10/14,11/15,12/16-17,	44/56-57,45/58,46/59,
				13/18,14/19,15/20	47/60,48/61-62,49/63
		0.05	0.10	17/24,18/26,19/26,	57/73-74,58/75,59/76-77,
				20/27-28,21/29,22/30,	60/78,61/79,62/80
				23/31	
0.80	0.95	0.10	0.10	8/10,9/11,10/12-13,	24/28,25/29,26/30,27/31,
				11/14,12/15,13/16,14/17	28/32,29/33,30/34-35
		0.05	0.20	7/9,8/10,9/11,10/12,	25/28,26/29,27/30,
				11/13,12/14,13/15,14/16	28/31-32,29/33,30/34,
					31/35
		0.05	0.10	10/12,11/13-14,12/15,	35/40,36/41,37/42,38/43,
				13/16,14/17,15/18,16/19	39/44,40/45-46,41/47

Table 5.3.5: Sample Sizes and Critical Values for Minimax Flexible Two-Stage Designs ($p_1 - p_0$=0.15)

p_0	p_1	α	β	r_i/n_i	R_j/N_j
0.05	0.20	0.10	0.10	0/16-22,1/23	2/26-28,3/29-33
		0.05	0.20	0/10-17	2/23,2/24-30
		0.05	0.10	0/22-27,1/28-29	3/33-34,4/35-40
0.10	0.25	0.10	0.10	1/25-27,2/28-32	5/37,6/38-42,7/43-44
		0.05	0.20	1/22-24,2/25-29	6/33-37,7/38-40
		0.05	0.10	2/25-29,3/30-32	8/49-52,9/53-56
0.20	0.35	0.10	0.10	6/37-39,7/40-42,8/43-440	14/54-55,15/56-59,16/60-61
		0.05	0.20	6/28,6/29-31,7/32-35	14/50-51,15/52-54,16/55-57
		0.05	0.10	8/41-45,9/46-48	19/71-72,20/73-74,21/75-78
0.30	0.45	0.10	0.10	11/43,12/44-46	23/64,24/65-67,
				13/47-48,14/49-50	25/68-69,26/70-71
		0.05	0.20	10/36,11/37,	23/60,24/61-63,
				12/38-39,13/40-43	25/64-65,26/66-67
		0.05	0.10	15/50-52,16/53-55,	32/85-86,33/87-89,
				17/56-57	34/90-91,35/92
0.40	0.55	0.10	0.10	16/43-44,17/45-46,	32/69-70,33/71,
				18/47,19/48-49,20/50	34/72-73,35/74-75,36/76
		0.05	0.20	13/34-35,14/36,	32/65-66,33/67-68,
				15/37-39,16/40-41	34/69-70,35/71,36/72
		0.05	0.10	23/60-61,24/62-63,	43/91,44/92-93,
				25/64-65,26/66-67	45/94,46/95-96,47/97-98
0.50	0.65	0.10	0.10	19/41,20/42-43,	38/67,39/68-69,40/70-71,
				21/44-45,22/57-57,23/48	41/72,42/73-74
		0.05	0.20	16/33,17/34-35,	38/64-65,39/66,
				18/36-37,19/38-39,20/40	40/67-68,41/69,42/70-71
		0.05	0.10	26/53,27/54-55,28/56,	52/89-90,53/91-92,
				29/57,30/58-59,31/60	54/93-94,55/95,56/96
0.60	0.75	0.10	0.10	22/38-39,23/40,24/41-42,	40/60,41/61,42/62-63,
				25/43,26/44-45	43/64,44/65-66,45/67
		0.05	0.20	18/31,19/32,20/33-34,	40/57-58,41/59,
				21/35,22/36-37,23/38	42/60-61,43/62,44/63-64
		0.05	0.10	23/39,24/40-41,25/42-43,	54/80,55/81,56/82,
				26/44,27/45,28/46	57/83-84,58/85,59/86-87
0.70	0.85	0.10	0.10	19/28-29,20/30,21/31,	36/46-47,37/48,38/49,
				22/32,23/33-34,24/35	39/50-51,40/52,41/53
		0.05	0.20	18/25,19/26,20/27-28,	36/45,37/46-47,38/48,
				21/29,22/30,23/31,24/32	39/49,40/50-51,41/52
		0.05	0.10	26/38,27/39,28/40,29/41,	48/62,49/63,50/64,51/65,
				30/42-43,31/44/32/45	52/66,53/67,54/68-69
0.80	0.95	0.10	0.10	9/12,10/13,11/14,12/15,	23/26,24/27-28,25/29,
				13/16,14/17,15/18,16/19	26/30,27/31,28/32,29/33
		0.05	0.20	6/8,7/9,8/10,9/11,	23/26,24/27,25/28,26/29,
				10/12-13,11/14,12/15	27/30,28/31,29/32,30/33
		0.05	0.10	22/26,23/27,24/28,25/29,	31/35,32/36,33/37,34/38,
				26/30,27/31,28/32,29/33	35/39-40,36/41,37/42

where

$$\beta = P(\text{reject } H_a \mid p_1),$$

computing the values of r_2, n_2, r_3, and n_3 that minimize the null expected sample size $EN(p_0)$ subject to the error constraints α and β, where

$$EN(p) = n_1 + n_2\{1 - \beta_1(p)\} + n_3\{1 - \beta_1(p) - \beta_2(p)\},$$

and β_i are the probability of making type II error evaluated at stage i. Ensign et al. (1994) use the value of

$$\beta - (1 - p_1)^{n_1}$$

as the type II error rate in the optimization along with type I error

$$\alpha = P(\text{reject } H_0 \mid p_0)$$

to obtain r_2, n_2, r_3, and n_3. Repeating this, n_i can then be chosen to minimize the overall $EN(p_0)$.

Tables 5.3.7 and 5.3.8 provide sample sizes for optimal three-stage designs based on the method proposed by Ensign et al. (1994) for a variety of design parameters. The tabulated results include the optimal size n_i and the critical value r_i of the ith stage. For example, the result in Table 5.3.7 corresponding to a design with $p_0 = 0.25$ and $p_1 = 0.40$ gives $0/6, 7/26, 24/75$ for achieving an 80% power at the 5% level of significance. In other words, at the first stage, six subjects are treated. If there is no response, then the trial is terminated. Otherwise, accrual continues to a total of 26 subjects at the second stage. If there are no more than 7 subjects respond, then stop the trial. Otherwise, proceed to the third stage by recruiting 49 additional subjects. We would conclude that the test drug is effective if there are more than 24 responses for the subjects in the three stages combined.

Note that the optimal three-stage designs proposed by Ensign et al. (1994) restrict the rejection region in the first stage to be zero response, and the sample size to at least 5. As an alternative, Chen (1997b) also extended Simon's two-stage to a three-stage design without these restrictions. As a result, sample sizes can be obtained by computing the values of r_1, n_1, r_2, n_2, r_3, and n_3 that minimize the expected sample size

$$\begin{aligned} EN &= n_1 + (1 - PET_1)n_2 + (1 - PET_{all})n_3 \\ PET_1 &= B(r_1; n_1, p) = \sum_{x \le r_1} b(x; n_1, p) \\ PET_{all} &= PET_1 + \sum_{x=r_1+1}^{\min(n_1, r_2)} b(d; n, p) B(r_2 - x; n_2, p). \end{aligned}$$

Table 5.3.7: Sample Sizes n_i and Critical Values r_i for Optimal Three-Stage Designs - Ensign et al. (1994) ($p_1 - p_0 = 0.15$)

p_0	p_1	α	β	Stage 1 r_1/n_1	Stage 2 $r_2/(n_1+n_2)$	Stage 3 $r_3/(n_1+n_2+n_3)$
0.05	0.20	0.10	0.10	0/12	1/25	3/38
		0.05	0.20	0/10	2/24	3/31
		0.05	0.10	0/14	2/29	4/43
0.10	0.25	0.10	0.10	0/11	3/29	7/50
		0.05	0.20	0/9	3/25	7/43
		0.05	0.10	0/13	3/27	10/66
0.15	0.30	0.10	0.10	0/12	4/28	11/55
		0.05	0.20	0/9	5/27	12/56
		0.05	0.10	0/13	6/35	16/77
0.20	0.35	0.10	0.10	0/11	7/34	16/63
		0.05	0.20	0/6	6/28	18/67
		0.05	0.10	0/9	10/44	23/88
0.25	0.40	0.10	0.10	0/8	8/32	23/76
		0.05	0.20	0/6	7/26	24/75
		0.05	0.10	0/9	11/41	30/95
0.30	0.45	0.10	0.10	0/7	13/41	28/79
		0.05	0.20	0/7	9/27	31/84
		0.05	0.10	0/9	14/43	38/104
0.35	0.50	0.10	0.10	0/9	12/34	33/81
		0.05	0.20	0/5	12/31	37/88
		0.05	0.10	0/8	17/45	45/108
0.40	0.55	0.10	0.10	0/11	16/38	40/88
		0.05	0.20	0/5	14/32	40/84
		0.05	0.10	0/10	19/45	49/104
0.45	0.60	0.10	0.10	0/6	15/34	40/78
		0.05	0.20	0/5	12/25	47/90
		0.05	0.10	0/6	20/42	59/114
0.50	0.65	0.10	0.10	0/5	16/32	46.84
		0.05	0.20	0/5	12/25	47/90
		0.05	0.10	0/6	20/42	59/114
0.55	0.70	0.10	0.10	0/7	19/34	46/75
		0.05	0.20	0/5	15/26	48/76
		0.05	0.10	0/5	23/40	64/96
0.60	0.75	0.10	0.10	0/5	21/34	47/71
		0.05	0.20	0/5	13/21	49/72
		0.05	0.10	0/5	14/23	90/98
0.65	0.80	0.10	0.10	0/5	17/26	47/66
		0.05	0.20	0/5	12/18	49/67
		0.05	0.20	0/5	8/13	74/78
0.70	0.85	0.10	0.10	0/5	14/20	45/59
		0.05	0.20	0/5	14/19	46/59
		0.05	0.10	0/5	12/17	68/72
0.75	0.90	0.10	0.10	0/5	16/21	36/44
		0.05	0.20	0/5	10/13	40/48
		0.05	0.10	0/5	8/11	55/57
0.80	0.95	0.10	0.10	0/5	5/7	27/31
		0.05	0.20	0/5	7/9	26/29
		0.05	0.10	0/5	8/10	44/45

Table 5.3.10: Sample Sizes n_i and Critical Values r_i for Optimal Three-Stage Designs - Chen (1997b) ($p_1 - p_0 = 0.20$)

p_0	p_1	α	β	Stage 1 r_1/n_1	Stage 2 $r_2/(n_1+n_2)$	Stage 3 $r_3/(n_1+n_2+n_3)$
0.05	0.25	0.10	0.10	0/9	1/18	2/26
		0.05	0.20	0/8	1/13	2/19
		0.05	0.10	0/10	1/17	3/30
0.10	0.30	0.10	0.10	0/10	2/19	4/26
		0.05	0.20	0/6	2/17	5/29
		0.05	0.10	1/13	3/23	7/45
0.15	0.35	0.10	0.10	1/12	3/21	7/33
		0.05	0.20	1/9	4/21	8/35
		0.05	0.10	2/15	5/27	11/51
0.20	0.40	0.10	0.10	1/10	6/26	11/43
		0.05	0.20	1/8	5/22	11/38
		0.05	0.10	3/17	7/30	14/50
0.25	0.45	0.10	0.10	3/16	7/25	13/41
		0.05	0.20	2/10	6/20	16/48
		0.05	0.10	4/18	10/33	19/58
0.30	0.50	0.10	0.10	3/13	9/28	17/46
		0.05	0.20	3/11	7/21	18/46
		0.05	0.10	4/16	11/32	23/60
0.35	0.55	0.10	0.10	6/18	13/33	20/48
		0.05	0.20	3/10	9/23	21/47
		0.05	0.10	6/18	15/38	27/62
0.40	0.60	0.10	0.10	7/18	9/26	22/46
		0.05	0.20	3/9	10/23	23/46
		0.05	0.10	6/16	17/38	32/66
0.45	0.65	0.10	0.10	5/13	13/27	26/50
		0.05	0.20	3/8	10/20	29/54
		0.05	0.10	6/15	17/34	34/63
0.50	0.70	0.10	0.10	4/10	12/24	26/45
		0.05	0.20	4/9	13/23	29/49
		0.05	0.10	7/15	19/34	38/65
0.55	0.75	0.10	0.10	5/11	12/21	27/43
		0.05	0.20	6/11	14/23	28/43
		0.05	0.10	7/14	16/27	36/56
0.60	0.80	0.10	0.10	6/11	14/22	29/43
		0.05	0.20	5/9	12/48	28/40
		0.05	0.10	6/11	19/29	38/55
0.65	0.85	0.10	0.10	5/9	13/19	25/34
		0.05	0.20	5/8	13/18	27/36
		0.05	0.20	6/10	16/23	35/47
0.70	0.90	0.10	0.10	5/8	11/15	22/28
		0.05	0.20	3/5	10/13	25/31
		0.05	0.10	6/9	16/21	31/39
0.75	0.95	0.10	0.10	3/5	6/8	16/19
		0.05	0.20	1/2	9/11	19/22
		0.05	0.10	6/8	13/16	24/28

tively, in Sections 6.2 and 6.3. Test statistics and the corresponding sample size calculation for categorical shift is discussed in Sections 6.4. Section 6.5 considers testing for carry-over effect in a 2×2 crossover design. Some practical issues are presented in Section 6.6.

6.1 Tests for Goodness-of-Fit

In practice, it is often of interest to study the distribution of the primary study endpoint under the study drug with some reference distribution, which may be obtained from historical (control) data or literature review. If the primary study endpoint is a categorical response that is non-binary, Pearson's chi-square test is usually applied.

6.1.1 Pearson's Test

For the ith subject, let X_i be the response taking values from $\{x_1, ..., x_r\}$, $i = 1, ..., n$. Assume that X_i's are i.i.d. Let

$$p_k = P(X_i = x_k),$$

where $k = 1, ..., r$. p_k can be estimated by $\hat{p}_k = n_k/n$, where n_k is the frequency count of the subjects with response value k. For testing goodness-of-fit, the following hypotheses are usually considered:

$$H_0 : p_k = p_{k,0} \text{ for all } k \quad \text{vs.} \quad p_k \neq p_{k,0} \text{ for some } k,$$

where $p_{k,0}$ is a reference value (e.g., historical control), $k = 1, ..., r$. Pearson's chi-square statistic for testing goodness-of-fit is given by

$$T_G = \sum_{k=1}^{r} \frac{n(\hat{p}_k - p_{k,0})^2}{p_{k,0}}.$$

Under the null hypothesis H_0, T_G is asymptotically distributed as a central chi-square random variable with $r - 1$ degrees of freedom. Hence, we reject the null hypothesis with approximate α level of significance if

$$T_G > \chi^2_{\alpha,r-1},$$

where $\chi^2_{\alpha,r-1}$ denotes the αth upper quantile of a central chi-square random variable with $r - 1$ degrees of freedom. The power of Pearson's chi-square test can be evaluated under some local alternatives. (A local alternative typically means that the difference between treatment effects in terms of

adverse event. It is expected that the sample size will be approximately evenly distributed across the four centers (i.e., $\pi_h = 25\%, h = 1, 2, 3, 4$). Suppose based on a pilot study, the values of $p_{h,ij}$'s within each country are estimated as follows:

		Binary Response		
center	Treatment	0	1	Total
1	Study Drug	0.35	0.15	0.50
	Placebo	0.25	0.25	0.50
	Total	0.60	0.40	1.00
2	Study Drug	0.30	0.20	0.50
	Placebo	0.20	0.30	0.50
	Total	0.50	0.50	1.00
3	Study Drug	0.40	0.10	0.50
	Placebo	0.20	0.30	0.50
	Total	0.60	0.40	1.00
4	Study Drug	0.35	0.15	0.50
	Placebo	0.15	0.35	0.50
	Total	0.50	0.50	1.00

By (6.3.1), δ is given by 0.3030. Hence the sample size needed in order to achieve an 80% ($\beta = 0.20$) power at the 5% ($\alpha = 0.05$) level of significance is given by

$$n = \frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2} = \frac{(1.96 + 0.84)^2}{0.3030^2} \approx 86.$$

6.4 Test for Categorical Shift

In clinical trials, it is often of interest to examine any change in laboratory values before and after the application of the treatment. When the response variable is categorical, this type of change is called a categorical shift. In this section, we consider testing for categorical shift.

6.4.1 McNemar's Test

For a given laboratory test, test results are usually summarized as either normal (i.e., the test result is within the normal range of the test) or abnormal (i.e., the test result is outside the normal range of the test). Let x_{ij} be the binary response ($x_{ij} = 0$: normal and $x_{ij} = 1$: abnormal) from the ith ($i = 1, ..., n$) subject in the jth ($j = 1$: pre-treatment and $j = 2$: post treatment) treatment. The test results can be summarized in the following

2×2 table:

	Post-Treatment		
Pre-Treatment	Normal	Abnormal	
Normal	n_{00}	n_{01}	$n_{0\cdot}$
Abnormal	n_{10}	n_{11}	$n_{1\cdot}$
	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot\cdot}$

where

$$\begin{aligned} n_{00} &= \sum_{i=1}^{n}(1-x_{i1})(1-x_{i2}) \\ n_{01} &= \sum_{i=1}^{n}(1-x_{i1})x_{i2} \\ n_{10} &= \sum_{i=1}^{n} x_{i1}(1-x_{i2}) \\ n_{11} &= \sum_{i=1}^{n} x_{i1}x_{i2}. \end{aligned}$$

Define

$$\begin{aligned} p_{00} &= P(x_{i1}=0, x_{i2}=0) \\ p_{01} &= P(x_{i1}=0, x_{i2}=1) \\ p_{10} &= P(x_{i1}=1, x_{i2}=0) \\ p_{11} &= P(x_{i1}=1, x_{i2}=1) \\ p_{1+} &= P(x_{i1}=1) = p_{10}+p_{11} \\ p_{+1} &= P(x_{i2}=1) = p_{01}+p_{11}. \end{aligned}$$

It is then of interest to test whether there is a categorical shift after treatment. A categorical shift is defined as either a shift from 0 (normal) in pre-treatment to 1 (abnormal) in post-treatment or a shift from 1 (abnormal) in pre-treatment to 0 (normal) in post-treatment. Consider

$$H_0 : p_{1+} = p_{+1} \quad \text{versus} \quad H_a : p_{1+} \neq p_{+1},$$

which is equivalent to

$$H_0 : p_{10} = p_{01} \quad \text{versus} \quad H_a : p_{10} \neq p_{01}.$$

The most commonly used test procedure to serve the purpose is McNemar's test, whose test statistic is given by

$$T_{MN} = \frac{n_{10}-n_{01}}{\sqrt{n_{10}+n_{01}}}.$$

abnormal (two categories), it is often classify them into three categories (i.e., below, within, and above the normal range). Let $x_{ij} \in \{1, ..., r\}$ be the categorical observation from the ith subject under the jth treatment ($j = 1$: pre-treatment and $j = 2$: post-treatment). In practice, the data are usually summarized by the following $r \times r$ contingency table:

	Post-Treatment				
Pre-Treatment	1	2	$\cdots$	r	
1	n_{11}	n_{12}	$\cdots$	n_{1r}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2r}	$n_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
r	n_{r1}	n_{r2}	$\cdots$	n_{rr}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot r}$	$n_{\cdot\cdot}$

Let

$$p_{ij} = P(x_{k1} = i, x_{k2} = j),$$

which is the probability that the subject will shift from i pre-treatment to j post-treatment. If there is no treatment effect, one may expect that $p_{ij} = p_{ji}$ for all $1 \leq i, j, \leq r$. Hence, it is of interest to test the hypotheses

$$H_0 : p_{ij} = p_{ij} \text{ for all } i \neq j \quad \text{versus} \quad H_a : p_{ij} \neq p_{ij} \text{ for some } i \neq j.$$

In order to test the above hypotheses, the test statistic proposed by Stuart and Maxwell is useful (see, e.g., Stuart, 1955). We refer to the test statistic as Stuart-Maxwell test, which is given by

$$T_{SM} = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}.$$

Under the null hypothesis H_0, T_{SM} follows a standard chi-square distribution with $r(r-1)/2$ degrees of freedom. Hence, for a given significance level of α, the null hypothesis should be rejected if $T_{SM} > \chi^2_{\alpha, r(r-1)/2}$.

Under the local alternative given by

$$\lim_{n\to\infty} n \sum_{i<j} \frac{(p_{ij} - p_{ji})^2}{p_{ij} + p_{ji}} = \delta,$$

T_{SM} is asymptotically distributed as a non-central chi-square random variable with $r(r-1)/2$ degrees of freedom and the non-centrality parameter δ. For a given degrees of freedom ($r(r-1)/2$) and a desired power $(1-\beta)$, δ can be obtained by solving

$$F_{r(r-1)/2}(\chi_{\alpha, r(r-1)/2} | \delta) = \beta, \tag{6.4.2}$$

Hence, the test statistic is given by

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

Under the null hypothesis H_0, T is asymptotically distributed as a standard normal random variable. We reject the null hypothesis at approximate α level of significance if

$$|T| > z_{\alpha/2}$$

.

Under the alternative hypothesis that $\gamma \neq 0$, the power of the this test procedure can be approximated by

$$\Phi\left(\gamma\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]^{-1/2} - z_{\alpha/2}\right).$$

For a given power of $1-\beta$ and assuming that $n = n_1 = n_2$, the sample size needed can be obtained by solving

$$\gamma\left[\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}\right]^{-1/2} - z_{\alpha/2} = z_\beta,$$

which leads to

$$n = \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\gamma^2}.$$

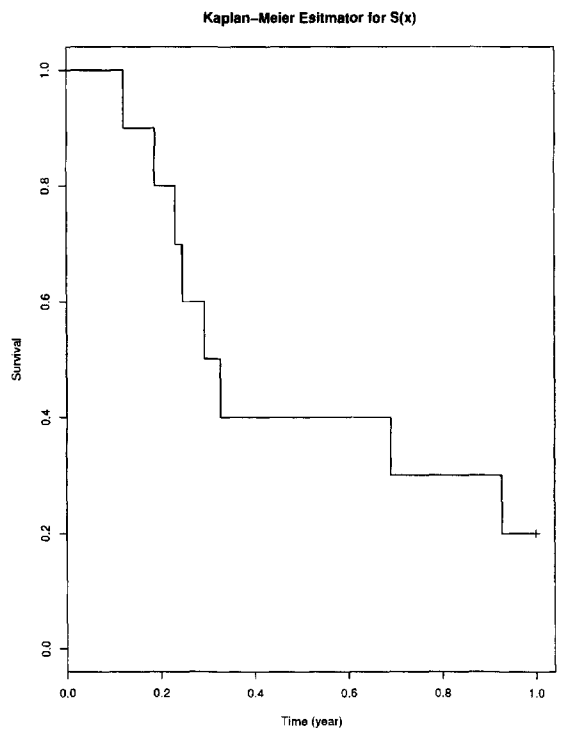
6.5.2 An Example

Consider a single-center, open, randomized, active-controlled, two-sequence, two-period crossover design with the primary efficacy endpoint of nocturnal hypoglycaemia. The objective of the study is to compare the study drug with a standard therapy on the marketplace in terms of the proportions of the patients who will experience nocturnal hypoglycaemia. As a result, the investigator is interested in conducting a larger trial to confirm whether such a effect truly exists. However, based on the results from a small-scale pilot study, no evidence of statistical significance in the possible carry-over effect was detected. According to the pilot study, the following parameters were estimated: $\gamma = 0.89$, $\sigma_1 = 2.3$, and $\sigma_2 = 2.4$. The sample size needed in order to achieve an 80% ($\beta = 0.2$) power at the 5% ($\alpha = 0.05$) level of significance is given by

$$n = \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\gamma^2} = \frac{(1.96 + 0.84)^2(2.3^2 + 2.4^2)}{0.89^2} \approx 110.$$

Hence, a total of 110 subjects are required in order to have the desired power for the study.

Figure 7.1.1: Kaplan-Meier Estimator for $S(x)$



distribution with the density function given by

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

When $\gamma > 0$, the entry distribution is convex, which implies fast patient entry at beginning. When $\gamma < 0$, the entry distribution is concave, which implies lagging patient entry. For convenience, we define $g(z) = 1/T_0$ when $\gamma = 0$, which implies uniform patient entry. Let t_{ij} be the time-to-event (i.e., the time from the patient entry to the time observing event) for the jth subject in the ith treatment group, $i = 1, 2$, $j = 1, ..., n_i$. It is assumed that t_{ij} follows an exponential distribution with hazard rate λ_i. The information observed from the sample is $(x_{ij}, \delta_{ij}) = (\min(t_{ij}, T - a_{ij}), I\{t_{ij} \le T - a_{ij}\})$.

Under the assumption that $n_1 = kn_2$, we have

$$n_2 = \frac{(z_\alpha + z_\beta)^2}{(\epsilon - \delta)^2}\left[\frac{\sigma^2(\lambda_1)}{k} + \sigma^2(\lambda_2)\right]. \tag{7.2.5}$$

7.2.3 Test for Equivalence

The objective is to test the following hypotheses:

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

The null hypothesis is rejected and the test drug is concluded to be equivalent to the control in average if

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

and

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

Under the alternative hypothesis ($|\epsilon| < \delta$), the power of the above testing procedure is approximately

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

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by solving the following equation

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

As a result, the sample size needed for achieving a power of $1 - \beta$ is given by $n_1 = kn_2$ and

$$n_2 = \frac{(z_\alpha + z_{\beta/2})^2}{(\delta - |\epsilon|)^2}\left(\frac{\sigma^2(\lambda_1)}{k} + \sigma^2(\lambda_2)\right). \tag{7.2.6}$$

7.2.4 An Example

Suppose that the sponsor is planning a trial among the patients with either Hodgkin's disease (HOD) or non-Hodgkin's lymphoma (NHL). The patients will be given either an allogeneric (allo) transplant from an HLA-matched

and

$$\pi(t) = \frac{p_2(1 - F_1(t))(1 - H_2(t))}{p_1(1 - F_0(t))(1 - H_1(t)) + p_2(1 - F_1(t))(1 - H_2(t))}.$$

Then L is asymptotically distributed as a normal random variable with variance 1 and mean given by

$$\frac{n^{1/2} \int_0^\infty \log(\lambda_2(t)/\lambda_1(t))\pi(t)(1 - \pi(t))V(t)dt}{\left[\int_0^\infty \pi(t)(1 - \pi(t))V(t)dt\right]^{1/2}}. \tag{7.3.5}$$

Under the assumption of proportional hazard, $\log(\lambda_2(t)/\lambda_1(t)) = b$ is a constant. Assume that $H_2(t) = H_1(t)$. Let $d = \int_0^\infty V(t)dt$, which is the probability of observing a event. In practice, mostly commonly $F_1(t) \approx F_0(t)$. In such a situation, it can be noted that $\pi(t) \approx p_2$, then the (7.3.5) becomes

$$\log(b)(np_1p_2d)^{1/2}.$$

Therefore, the two-sided sample size formula with significance level α and power $1 - \beta$ is given by

$$n = \frac{(z_{\alpha/2} + z_\beta)^2}{b^2 p_1 p_2 d}. \tag{7.3.6}$$

7.3.2 Test for Non-Inferiority/Superiority

We still assume that $z_i = 0$ for treatment 1 and $z_i = 1$ for treatment 2. The problem of testing non-inferiority and superiority can be unified by the following hypotheses:

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

where δ is the superiority or non-inferiority margin. When $\delta > 0$, the rejection of the null hypothesis indicates superiority over the reference value. When $\delta < 0$, the rejection of the null hypothesis implies non-inferiority against the reference value. When $b = \delta$, the test statistic

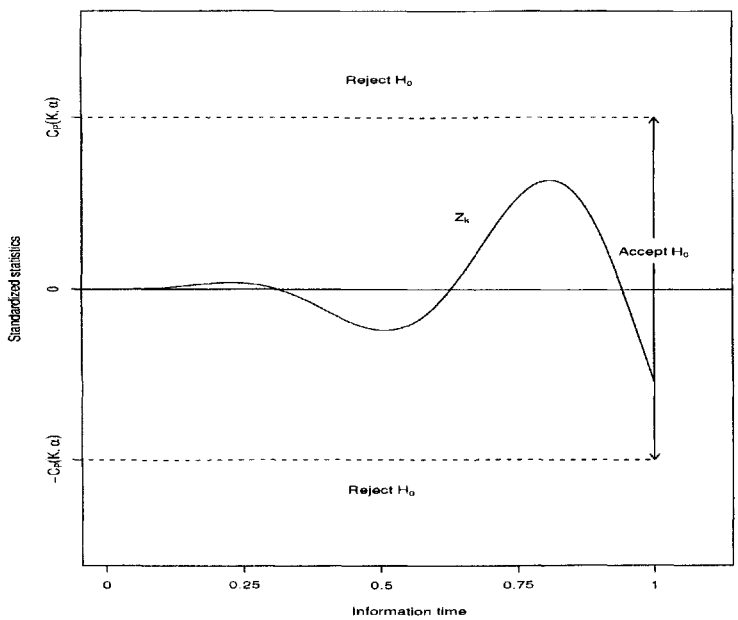
$$L = \frac{\sum_{k=1}^d \left(I_k - \frac{Y_{1i}e^\delta}{Y_{1i}e^\delta + Y_{2i}}\right)}{\left[\sum_{k=1}^d \left(\frac{Y_{1i}Y_{2i}e^\delta}{(Y_{1i}e^\delta + Y_{2i})^2}\right)\right]^{-1/2}}$$

follows a standard normal distribution when the sample size is sufficiently large. Hence, the null hypothesis should be rejected if $L > z_\alpha$. Under the alternative hypothesis, L is asymptotically distributed as a normal random variable with variance 1 and mean given by

$$\frac{n^{1/2} \int_0^\infty (\log(\lambda_2(t)/\lambda_1(t)) - \delta)\pi(t)(1 - \pi(t))V(t)dt}{\left[\int_0^\infty \pi(t)(1 - \pi(t))V(t)dt\right]^{1/2}}. \tag{7.3.7}$$

the treatment effects. Some discussion on sample size determination for this approach can be found in Emrich (1989) and Dixon and Simon (1988).

Figure 8.1.1: Pocock Type Stopping Rule



On the other hand, under the alternative hypothesis (i.e., $\theta \neq 0$), the power of the above test procedure can be determined by the total number of planned interim analysis (K), type I error rate (α), and type II error rate (β), and the proportion between σ^2 and δ^2 (i.e., σ^2/δ^2), where $\delta = |\mu_1 - \mu_2|$. As discussed in the previous chapters, if there are no interim analyses planned (i.e., $K = 1$), then the sample size is proportional to σ^2/δ^2. As a result, it is sufficient to specify the ratio $R_P(K, \alpha, \beta)$ of the maximum sample size of the group sequential test to the fixed sample size. The values of $R_P(K, \alpha, \beta)$ are given in Table 8.1.2. The maximum sample size needed for a group sequential trial with K interim analyses can be obtained by first calculating the fixed sample size without interim analyses, and then multiplying it by $R_P(K, \alpha, \beta)$.

8.1.2 An Example

Suppose that an investigator is interested in conducting a clinical trial with 5 interim analyses for comparing a test drug (T) with a placebo (P). Based on information obtained from a pilot study, data from the test drug and

If Pocock's test is used, then by Table 8.1.2, we have

$$n_{\text{max}} = n_{\text{fixed}} R_P(5, 0.05, 0.20) = 385 \times 1.229 \approx 474.$$

Hence, at each interim analysis, the sample size per treatment group is $474/5 = 94.8 \approx 95$.

On the other hand, if O'Brien and Fleming's method is employed, Table 8.2.2 gives

$$n_{\text{max}} = n_{\text{fixed}} R_B(5, 0.05, 0.20) = 385 \times 1.028 \approx 396.$$

Thus, at each interim analysis, the sample size per treatment group is $396/5 = 79.2 \approx 80$.

Alternatively, if Wang and Tsitis' test with $\Delta = 0.25$ is considered,Table 8.3.2 leads to

$$n_{\text{max}} = n_{\text{fixed}} R_{WT}(5, 0.05, 0.20, 0.25) = 385 \times 1.072 \approx 413.$$

As a result, at each interim analysis, the sample size per treatment group is given by $413/5 = 82.6 \approx 83$.

8.6 Time-to-Event Data

To apply the repeated significance test procedures to time-to-event data, for simplicity, we only consider Cox's proportional hazard model.

8.6.1 The Procedure

As indicated in Chapter 7, under the assumption of proportional hazards, the log-rank test is usually used to compare the two treatment groups. More specifically, let $h(t)$ be the hazard function of treatment group A and $e^{\theta}h(t)$ be the hazard function of treatment group B. Let d_k denote the total number of uncensored failures observed when the kth interim analysis is conducted, $k = 1, ..., K$. For illustration purposes and without loss of generality, we assume that there are no ties. Let $\tau_{i,k}$ be the survival times of these subjects, $i = 1, ..., d_k$. Let $r_{iA,k}$ and $r_{iB,k}$ be the numbers of subjects who are still at risk at the kth interim analysis at time $\tau_{i,k}$ for treatment A and B, respectively. The log-rank test statistic at the kth interim analysis is then given by

$$S_k = \sum_{i=1}^{d_k} \left(\delta_{iB,k} - \frac{r_{iB,k}}{r_{iA,k} + r_{iB,k}} \right),$$

μ_i and variance σ_{Bi}^2 and e_{ijk} is normally distributed as a normal random variable with mean 0 and variance σ_{Wi}^2. For a given subject with a fixed A_{ij}, x_{ijk} is normally distributed as a normal random variable with mean A_{ij} and variance $A_{ij}^2\sigma_{Wi}^2$. Hence, the CV for this subject is given by

$$CV_i = \frac{|A_{ij}\sigma_{Wi}|}{|A_{ij}|} = \sigma_{Wi}.$$

As it can be seen, the conditional random effects model assumes the CV is constant across subjects.

Define

$$\begin{aligned}
\bar{x}_{i\cdot\cdot} &= \frac{1}{nm}\sum_{j=1}^{n_i}\sum_{k=1}^{m} x_{ijk} \\
M_{i1} &= \frac{m}{n_i - 1}\sum_{j=1}^{n_i}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot})^2 \\
M_{i2} &= \frac{1}{n_i(m_i - 1)}\sum_{j=1}^{n_i}\sum_{k=1}^{m}(x_{ijk} - \bar{x}_{ij\cdot})^2.
\end{aligned}$$

It can be verified that

$$\begin{aligned}
E(\bar{x}_{i\cdot\cdot}) &= \mu_i \\
E(M_{i1}) &= (\mu_i^2 + \sigma_{BT}^2)\sigma_{Wi}^2 + m\sigma_{Bi}^2 = \tau_{i1}^2 \\
E(M_{i2}) &= (\mu_i^2 + \sigma_{BT}^2)\sigma_{Wi}^2 = \tau_{i2}^2.
\end{aligned}$$

It follows that

$$CV_i = \sigma_{Wi} = \sqrt{\frac{E(M_{i2})}{E^2(\bar{x}_{i\cdot\cdot}) + (M_{i1} - M_{i2})}}.$$

Hence, an estimator of CV_i is given by

$$\widehat{CV}_i = \sqrt{\frac{M_{i2}}{\bar{x}_{i\cdot\cdot}^2 + (M_{i1} - M_{i2})/m}}.$$

By Taylor's expansion,

$$\begin{aligned}
\widehat{CV}_i - CV_i &\approx \frac{1}{\tau_{i1}\tau_{i2}}(M_{i2} - \tau_{i2}^2) - \frac{\mu_i\tau_{i2}}{\tau_{i1}^3}(\bar{x}_{i\cdot\cdot} - \mu_i) \\
&\quad - \frac{\tau_{i2}}{2m\tau_{i1}^3}(M_{i1} - M_{i2} - (\tau_{i1}^2 - \tau_{i2}^2)) \\
&= k_0(\bar{x}_{i\cdot\cdot} - \mu_i) + k_1(M_{i1} - \tau_{i1}^2) + k_2(M_{i2} - \tau_{i2}^2),
\end{aligned}$$

Test for Similarity

For testing similarity, consider the following hypotheses:

$$H_0 : |CV_T - CV_R| \geq \delta \quad \text{versus} \quad H_a : |CV_T - CV_R| < \delta.$$

The two drug products are concluded to be similar to each other if the null hypothesis is rejected at a given significance level. The null hypothesis is rejected at α level of significance if

$$\frac{\widehat{CV}_T - \widehat{CV}_R + \delta}{\sqrt{\hat{\sigma}_T^{*2}/n_T + \hat{\sigma}_R^{*2}/n_R}} > z_\alpha \quad \text{versus} \quad \frac{\widehat{CV}_T - \widehat{CV}_R - \delta}{\sqrt{\hat{\sigma}_T^{*2}/n_T + \hat{\sigma}_R^{*2}/n_R}} < -z_\alpha.$$

Under the alternative hypothesis that $|CV_T - CV_R| < \delta$, the power of the above test procedure is approximately

$$2\Phi\left(\frac{\delta - |CV_T - CV_R|}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} - z_\alpha\right) - 1.$$

Hence, under the assumption that $n = n_1 = n_2$, the sample size needed in order to achieve $1 - \beta$ power at the α level of significance can be obtained by solving

$$\frac{\delta - |CV_T - CV_R|}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} - z_\alpha = z_{\beta/2}.$$

This gives

$$n = \frac{(z_\alpha + z_{\beta/2})^2(\sigma_T^{*2} + \sigma_R^{*2})}{(\delta - |CV_T - CV_R|)^2}.$$

An Example

Consider the same example as described in the previous subsection. Suppose it is found that the variability of the CV increases as the mean increases. In this case, the conditional random effects model is useful for comparing the two treatments. Again, we assume that CV of the test drug and the reference drug are given by 50% and 70%, respectively. Suppose it is also estimated from other studies that $\sigma_T^* = 0.30$ and $\sigma_R^* = 0.35$. Assume that 10% difference in CV is of no clinical importance. The sample size needed per treatment group in order to establish non-inferiority can be obtained as follows:

$$n = \frac{(1.64 + 0.84)^2(0.30^2 + 0.35^2)}{(0.10 + 0.70 - 0.50)^2} \approx 15.$$

As a result, 15 subjects per treatment group are needed in order to have an 80% power at the 5% level of significance.

9.3.1 Parallel Design with Replicates

Under model (9.1.1), define

$$s_{Bi}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot})^2, \tag{9.3.1}$$

where

$$\bar{x}_{i\cdot\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{x}_{ij\cdot}$$

and $\bar{x}_{ij\cdot}$ is given in (9.1.3). Note that $E(s_{Bi}^2) = \sigma_{Bi}^2 + \sigma_{Wi}^2/m$. Therefore,

$$\hat{\sigma}_{Bi}^2 = s_{Bi}^2 - \frac{1}{m}\hat{\sigma}_{Wi}^2$$

are unbiased estimators for the inter-subject variances, where $\hat{\sigma}_{Wi}^2$ is defined in (9.1.2).

Test for Equality

For testing equality in inter-subject variability, the following hypotheses are usually considered:

$$H_0 : \frac{\sigma_{BT}}{\sigma_{BR}} = 1 \quad \text{versus} \quad H_a : \frac{\sigma_{BT}}{\sigma_{BR}} \neq 1.$$

Testing the above hypotheses is equivalent to testing the following hypotheses:

$$H_0 : \sigma_{BT}^2 - \sigma_{BR}^2 = 0 \quad \text{versus} \quad H_a : \sigma_{BT}^2 - \sigma_{BR}^2 \neq 0.$$

Let $\eta = \sigma_{BT}^2 - \sigma_{BR}^2$. An intuitive estimator of η is given by

$$\begin{aligned} \hat{\eta} &= \hat{\sigma}_{BR}^2 - \hat{\sigma}_{BT}^2 \\ &= s_{BT}^2 - s_{BR}^2 - \hat{\sigma}_{WT}^2/m + \hat{\sigma}_{WR}^2/m. \end{aligned}$$

A $(1-\alpha)\times 100\%$ confidence interval for η is given by $(\hat{\eta}_L, \hat{\eta}_U)$, where

$$\hat{\eta}_L = \hat{\eta} - \sqrt{\Delta_L}, \quad \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

and

$$\begin{aligned} \Delta_L = {} & s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi^2_{\alpha/2, n_T - 1}}\right)^2 + s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi^2_{1-\alpha/2, n_R - 1}}\right)^2 \\ & + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m-1)}{\chi^2_{1-\alpha/2, n_T(m-1)}}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m-1)}{\chi^2_{\alpha/2, n_R(m-1)}}\right)^2 \end{aligned}$$

Hence, the sample size needed in order to achieve an 80% power ($1-\beta = 0.80$) in establishing non-inferiority with a non-inferiority margin 0.10 ($\delta = 1.10$) at the 5% level of significance ($\alpha = 0.05$) can be obtained as

$$\begin{aligned}\sigma^{*2} &= 2\left[\left(0.30^2 + \frac{0.20^2}{3}\right)^2 + 1.1^4 \times \left(0.40^2 + \frac{0.30^2}{3}\right)^2\right.\\ &\quad \left. + \frac{0.20^4}{3^2(3-1)} + \frac{1.1^4 0.30^4}{3^2(3-1)}\right]\\ &= 0.129.\end{aligned}$$

Hence,

$$n = \frac{0.129(1.64+0.84)^2}{(0.30^2 - 1.1^2 \times 0.40^2)^2} \approx 74.$$

9.3.2 Replicated Crossover Design

Under model (9.1.4), estimators of inter-subject variances can be defined by

$$s_{BT}^2 = \frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\bar{x}_{ijT.} - \bar{x}_{i.T.})^2,$$

$$s_{BR}^2 = \frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\bar{x}_{ijR.} - \bar{x}_{i.R.})^2,$$

where

$$\bar{x}_{i.k.} = \frac{1}{n_i}\sum_{j=1}^{n_i}\bar{x}_{ijk..}$$

Note that $E(s_{Bk}^2) = \sigma_{Bk}^2 + \sigma_{Wk}^2/m$ for $k = T, R$. Therefore, unbiased estimators for the inter-subject variance are given by

$$\hat{\sigma}_{BT}^2 = s_{BT}^2 - \frac{1}{m}\hat{\sigma}_{WT}^2 \tag{9.3.2}$$

$$\hat{\sigma}_{BR}^2 = s_{BR}^2 - \frac{1}{m}\hat{\sigma}_{WR}^2. \tag{9.3.3}$$

Test for Equality

For testing the equality in inter-subject variability, the following hypotheses are considered:

$$H_0 : \frac{\sigma_{BT}}{\sigma_{BR}} = 1 \quad \text{versus} \quad H_a : \frac{\sigma_{BT}}{\sigma_{BR}} \neq 1.$$

An Example

Consider the same example discussed in the previous subsection. Under the standard 2×2 crossover design, it is assumed that $\rho = 1$, $\sigma_{BT} = 0.30$, $\sigma_{BR} = 0.40$, $\sigma_{WT} = 0.20$, and $\sigma_{WR} = 0.30$. The objective is to establish non-inferiority with a margin $\delta = 1.1$. It follows that

$$\begin{aligned}\sigma^{*2} &= 2[(0.30^2 + 0.20^2)^2 + 1.1^4(0.4^2 + 0.3^2)^2 \\ &\quad -2 \times 1.1^2 \times 0.30^2 \times 0.4^2] = 0.147.\end{aligned}$$

As a result, the sample size needed in order to achieve an 80% ($1 - \beta = 0.80$) power at the 5% ($\alpha = 0.05$) level of significance is given by

$$n_s = \frac{(0.153)(1.64 + 0.84)^2}{(0.30^2 + 0.20^2 - 1.1^2 \times (0.4^2 + 0.3^2))^2} \approx 31.$$

Since $n_s = n_1 + n_2 - 2$, approximately 17 subjects should be assigned to each sequence to achieve an 80% ($1 - \beta = 0.80$) power.

9.4.4 Replicated $2 \times 2m$ Crossover Design

We can use similar argument for test of inter-subject variabilities under model 9.1.4 with the estimators

$$\hat{\sigma}^2_{Tk} = s^2_{Bk} + \frac{m-1}{m}\hat{\sigma}^2_{Wk}, \quad k = T, R.$$

Test for Equality

For testing the equality in total variability, consider the following hypotheses:

$$H_0 : \frac{\sigma^2_{TT}}{\sigma^2_{TR}} = 1 \quad \text{versus} \quad H_a : \frac{\sigma^2_{TT}}{\sigma^2_{TR}} \neq 1.$$

Testing the above hypotheses is equivalent to testing the following hypotheses:

$$H_0 : \sigma^2_{TT} - \sigma^2_{TR} = 0 \quad \text{versus} \quad H_a : \sigma^2_{TT} - \sigma^2_{TR} \neq 0.$$

Let $\hat{\eta} = \hat{\sigma}^2_{TT} - \hat{\sigma}^2_{TR}$. In Lee et al. (2002b), a $(1 - \alpha)\times 100\%$ confidence interval of η is given by $(\hat{\eta}_L, \hat{\eta}_U)$, where

$$\hat{\eta}_L = \hat{\eta} - \sqrt{\Delta_L}, \quad \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U},$$

design, subjects are randomly assigned to one of the two sequences of formulations. In the first sequence, n_1 subjects receive treatments in the order of TR (T = test formulation, R = reference formulation) at two different dosing periods, whereas in the second sequence, n_2 subjects receive treatments in the order of RT at two different dosing periods. A sufficient length of washout between dosing periods is usually applied to wear off the possible residual effect that may be carried over from one dosing period to the next dosing period. Let y_{ijk} be the original or the log-transformation of the PK response of interest from the ith subject in the kth sequence at the jth dosing period. The following statistical model is considered:

$$y_{ijk} = \mu + F_l + P_j + Q_k + S_{ikl} + e_{ijk}, \tag{10.2.1}$$

where μ is the overall mean; P_j is the fixed effect of the jth period ($j = 1, 2$, and $P_1 + P_2 = 0$); Q_k is the fixed effect of the kth sequence ($k = 1, 2$, and $Q_1 + Q_2 = 0$); F_l is the fixed effect of the lth formulation (when $j = k$, $l = T$; when $j \neq k$, $l = R$; $F_T + F_R = 0$); S_{ikl} is the random effect of the ith subject in the kth sequence under formulation l and (S_{ikT}, S_{ikR}), $i = 1, ..., n_k$, $k = 1, 2$, are independent and identically distributed bivariate normal random vectors with mean 0 and an unknown covariance matrix

$$\begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix};$$

e_{ijk}'s are independent random errors distributed as $N(0, \sigma_{Wl}^2)$; and S_{ikl}'s and e_{ijk}'s are mutually independent. Note that σ_{BT}^2 and σ_{BR}^2 are between-subject variances and σ_{WT}^2 and σ_{WR}^2 are within-subject variances, and that $\sigma_{TT}^2 = \sigma_{BT}^2 + \sigma_{WT}^2$ and $\sigma_{TR}^2 = \sigma_{BR}^2 + \sigma_{WR}^2$ are the total variances for the test and reference formulations, respectively. Under model (10.2.1), the ABE parameter δ defined in Section 10.1 is equal to $\delta = F_T - F_R$. According to the 2000 FDA guidance, ABE is claimed if the following null hypothesis H_0 is rejected at the 5% level of significance:

$$H_0 : \delta \leq \delta_L \quad \text{or} \quad \delta \geq \delta_U \quad \text{versus} \quad H_1 : \delta_L < \delta < \delta_U, \tag{10.2.2}$$

where δ_L and δ_U are given bioequivalence limits. Under model (10.2.1),

$$\hat{\delta} = \frac{\bar{y}_{11} - \bar{y}_{12} - \bar{y}_{21} + \bar{y}_{22}}{2} \sim N\left(\delta, \ \frac{\sigma_{1,1}^2}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right), \tag{10.2.3}$$

where $\bar{y}_{jk}$ is the sample mean of the observations in the kth sequence at the jth period and $\sigma_{1.1}^2$ is

$$\sigma_{a,b}^2 = \sigma_D^2 + a\sigma_{WT}^2 + b\sigma_{WR}^2 \tag{10.2.4}$$

Chow, Shao, and Wang (2002b) obtained the following approximate 95% upper confidence bound for λ. When $\sigma_{TR}^2 \geq \sigma_0^2$,

$$\hat{\lambda}_U = \hat{\delta}^2 + \hat{\sigma}_{TT}^2 - (1+\theta_{PBE})\hat{\sigma}_{TR}^2 + t_{0.05,n_1+n_2-2}\sqrt{V}, \tag{10.3.2}$$

where V is an estimated variance of $\hat{\delta}^2 + \hat{\sigma}_{TT}^2 - (1+\theta_{PBE})\hat{\sigma}_{TR}^2$ of the form

$$V = \left(2\hat{\delta}, 1, -(1+\theta_{PBE})\right) C \left(2\hat{\delta}, 1, -(1+\theta_{PBE})\right)',$$

and C is an estimated variance-covariance matrix of $(\hat{\delta}, \hat{\sigma}_{TT}^2, \hat{\sigma}_{TR}^2)$. Since $\hat{\delta}$ and $(\hat{\sigma}_{TT}^2, \hat{\sigma}_{TR}^2)$ are independent,

$$C = \begin{pmatrix} \frac{\hat{\sigma}_{1,1}^2}{4}\left(\frac{1}{n_1}+\frac{1}{n_2}\right) & (0,0) \\ (0,0)' & \frac{(n_1-1)C_1}{(n_1+n_2-2)^2} + \frac{(n_2-1)C_2}{(n_1+n_2-2)^2} \end{pmatrix},$$

where $\hat{\sigma}_{1,1}^2$ is defined by (10.2.5), C_1 is the sample covariance matrix of $((y_{i11}-\bar{y}_{11})^2, (y_{i21}-\bar{y}_{21})^2)$, $i = 1, ..., n_1$, and C_2 is the sample covariance matrix of $((y_{i22}-\bar{y}_{22})^2, (y_{i12}-\bar{y}_{12})^2)$, $i = 1, ..., n_2$.

When $\sigma_{TR}^2 < \sigma_0^2$, the upper confidence bound for λ should be modified to

$$\hat{\lambda}_U = \hat{\delta}^2 + \hat{\sigma}_{TT}^2 - \hat{\sigma}_{TR}^2 - \theta_{PBE}\sigma_0^2 + t_{0.05,n_1+n_2-2}\sqrt{V_0}, \tag{10.3.3}$$

where

$$V_0 = \left(2\hat{\delta}, 1, -1\right) C \left(2\hat{\delta}, 1, -1\right)'.$$

The confidence bound $\hat{\lambda}_U$ in (10.3.2) is referred to as the confidence bound under the reference-scaled criterion, whereas $\hat{\lambda}_U$ in (10.3.3) is referred to as the confidence bound under the constant-scaled criterion. In practice, whether $\sigma_{TR}^2 \geq \sigma_0^2$ is usually unknown. Hyslop, Hsuan, and Holder (2000) recommend using the reference-scaled criterion or the constant-scaled criterion according to $\hat{\sigma}_{TR}^2 \geq \sigma_0^2$ or $\hat{\sigma}_{TR}^2 < \sigma_0^2$, which is referred to as the estimation method. Alternatively, we may test the hypothesis of $\sigma_{TR}^2 \geq \sigma_0^2$ versus $\sigma_{TR}^2 < \sigma_0^2$ to decide which confidence bound should be used; i.e., if $\hat{\sigma}_{TR}^2(n_1+n_2-2) \geq \sigma_0^2\chi_{0.95,n_1+n_2-2}^2$, then $\hat{\lambda}_U$ in (10.3.2) should be used; otherwise $\hat{\lambda}_U$ in (10.3.3) should be used, where $\chi_{\alpha,r}^2$ denotes the αth upper quantile of the chi-square distribution with r degrees of freedom. This is referred to as the test method and is more conservative than the estimation method.

Based on an asymptotic analysis, Chow, Shao, and Wang (2002b) derived the following formula for sample size determination assuming that $n_1 = n_2 = n$:

$$n \geq \frac{\zeta(z_{0.05}+z_\beta)^2}{\lambda^2}, \tag{10.3.4}$$

in which n_1 subjects in sequence 1 receive treatments at three periods in the order of TRR, while n_2 subjects in sequence 2 receive treatments at three periods in the order of RTR. The statistical model under this design is still given by (10.4.1). An unbiased estimator of δ is

$$\hat{\delta} = \frac{\bar{x}_{T1} - \bar{x}_{R1} + \bar{x}_{T2} - \bar{x}_{R2}}{2} \sim N\left(\delta, \frac{\sigma^2_{1,0.5}}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

where $\sigma^2_{a,b}$ is given by (10.2.4); an unbiased estimator of $\sigma^2_{1,0.5}$ is

$$\hat{\sigma}^2_{1,0.5} = \frac{(n_1 - 1)s^2_{d1} + (n_2 - 1)s^2_{d2}}{n_1 + n_2 - 2} \sim \frac{\sigma^2_{1,0.5}\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 1};$$

an unbiased estimator of σ^2_{WR} is

$$\hat{\sigma}^2_{WR} = \frac{(n_1 - 1)s^2_{R1} + (n_2 - 1)s^2_{R2}}{2(n_1 + n_2 - 2)} \sim \frac{\sigma^2_{WR}\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 2};$$

and estimators $\hat{\delta}$, $\hat{\sigma}^2_{1,0.5}$, and $\hat{\sigma}^2_{WR}$ are independent, since $x_{iT1} - x_{iR1}, x_{iT2} - x_{iR2}, z_{iR1}$, and z_{iR2} are independent. Chow, Shao, and Wang (2002a) obtained the following approximate 95% upper confidence bound for γ. When $\sigma^2_{WR} \geq \sigma^2_0$,

$$\hat{\gamma}_U = \hat{\delta}^2 + \hat{\sigma}^2_{1,0.5} - (1.5 + \theta_{IBE})\hat{\sigma}^2_{WR} + \sqrt{U},$$

where U is the sum of the following three quantities:

$$\left[\left(|\hat{\delta}| + t_{0.05,n_1+n_2-2}\frac{\hat{\sigma}_{1,0.5}}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)^2 - \hat{\delta}^2\right]^2,$$

$$\hat{\sigma}^4_{1,0.5}\left(\frac{n_1 + n_2 - 2}{\chi^2_{0.95,n_1+n_2-2}} - 1\right)^2,$$

and

$$(1.5 + \theta_{IBE})^2\hat{\sigma}^4_{WR}\left(\frac{n_1 + n_2 - 2}{\chi^2_{0.05,n_1+n_2-2}} - 1\right)^2. \tag{10.4.6}$$

When $\sigma^2_{WR} < \sigma^2_0$,

$$\hat{\gamma}_U = \hat{\delta}^2 + \hat{\sigma}^2_{1,0.5} - 1.5\hat{\sigma}^2_{WR} - \theta_{IBE}\sigma^2_0 + \sqrt{U_0},$$

where U_0 is the same as U except that the quantity in (10.4.6) should be replaced by

$$1.5^2\hat{\sigma}^4_{WR}\left(\frac{n_1 + n_2 - 2}{\chi^2_{0.05,n_1+n_2-2}} - 1\right)^2.$$

Step 2′. Let $m = 30$ and find a smallest integer n_* such that (10.5.5) holds. The required sample sizes are $m = 30$ and $n = n_*$.

Table 10.5.1 contains selected m_* and n_* according to Steps 1-3 or Steps 1 and 2′ with $1 - \beta = 80\%$ and the simulated power p of the *in vitro* bioequivalence test using these sample sizes.

An Example

Suppose an investigator is interested in conducting a clinical trial with a parallel design with no replicates to establish *in vitro* bioequivalence between a generic drug product (test) and a brand name drug product (reference) in terms of *in vitro* bioavailability. Based on data obtained from pilot studies, the mean difference can be assumed to be 0% ($\delta = 0.00$). Also, it is assumed the intra-subject standard deviation of test and reference are given by 50% and 50%, respectively. It is further assumed that the inter-subject standard deviation of the test and the reference are given by 50% and 50%, respectively. According to Table 10.5.1, 47 subjects per treatment group are needed in order to yield an 80% power at the 5% level of significance.

modified.

An Example

To illustrate the use of sample size formula derived above, we consider the same example concerning a study of osteoporosis in post-menopausal women described in Chapter 3 for testing one-sample. Suppose a clinical trial is planned to investigate the effect of a test drug on the prevention of the progression to osteoporosis in women with osteopenia. Suppose that a pilot study with 5 subjects was conducted. According to the data from the pilot study, it was estimated that $p_2 = 0.30$, $p_3 = 0.40$, and $p_4 = 0.05$. Hence, the sample size needed in order to achieve an 80% power for detection of such a clinically meaningful improvement can be estimated by

$$\begin{aligned} n &= \frac{(z_{\alpha/2}/\sqrt{12} + z_\beta\sqrt{p_3 + 4p_4 - 4p_2^2})^2}{(1/4 - p_2)^2} \\ &= \frac{(1.96/\sqrt{12} + 0.84\sqrt{0.4 + 4 \times 0.05 - 4 \times 0.3^2})^2}{(0.25 - 0.3)^2} \\ &\approx 383. \end{aligned}$$

Thus, a total of 383 subjects are needed in order to have an 80% power to confirm the observed post-treatment improvement.

11.3 Two-Sample Location Problem

Let x_i, $i = 1, ..., n_1$, and y_j, $j = 1, ..., n_2$, be two independent random samples, which are respectively from a control population and a treatment population in a clinical trial. Suppose that the primary objective is to investigate whether there is a shift of location, which indicates the presence of the treatment effect. Similar to the one-sample location problem, the hypotheses of interest are given by

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

where θ represents the treatment effect. Consider the following model:

$$x_i = e_i, \quad i = 1, ..., n_1,$$

and

$$y_j = \theta + e_{n_1+j}, \quad j = 1, ..., n_2,$$

where the e_i's are random errors having mean 0. It is assumed that (i) each e_i comes from the same continuous population and (ii) the $n_1 + n_2$ e_i's are mutually independent. To test the above hypotheses, the Wilcoxon

12.1 Dose Response Studies

When studying the dose response relationship of an investigational drug, a randomized, parallel-group trial involving a number of dose levels of the investigational drug and a control is usually conducted. Ruberg (1995a, 1995b) indicated that some questions dictating design and analysis are necessarily addressed. These questions include (i) Is there any drug effect? (ii) What doses exhibit a response different from control? (iii) What is the nature of the dose response relationship? and (iv) What is the optimal dose? The first two questions are usually addressed by the analysis of variance, and the last two questions are related to the identification of the so-called minimum effective dose (MED). In this section, formulas for sample size calculation are derived based on the concept of dose response study using the analysis of variance and a commonly used test (Williams' test) for MED.

12.1.1 Dose Response Relationship

In a dose response study, suppose there is a control group and K dose groups. The null hypothesis of interest is then given by

$$H_0 : \mu_0 = \mu_1 = \cdots = \mu_K, \tag{12.1.1}$$

where μ_0 is mean response for the control group and μ_i is mean response for the ith dose group. The rejection of hypothesis (12.1.1) indicates that there is a treatment effect. The dose response relationship can then be examined under appropriate alternative hypotheses. Under a specific alternative hypothesis, the required sample size per dose group can then be obtained. Spriet and Dupin-Spriet (1992) identified the following eight alternative hypotheses (H_a) for dose responses:

(1) $\mu_0 < \mu_1 < \cdots < \mu_{K-1} < \mu_K$;

(2) $\mu_0 < \cdots < \mu_i = \cdots = \mu_j > \cdots > \mu_K$;

(3) $\mu_0 < \cdots < \mu_i = \cdots = \mu_K$;

(4) $\mu_0 = \cdots = \mu_i < \cdots < \mu_K$;

(5) $\mu_0 < \mu_1 < \cdots = \mu_i = \cdots = \mu_K$;

(6) $\mu_0 = \mu_1 \cdots = \mu_i < \cdots < \mu_{K-1} < \mu_K$;

(7) $\mu_0 = \mu_1 < \cdots < \mu_i = \cdots = \mu_K$;

(8) $\mu_0 = \cdots = \mu_i < \cdots < \mu_{K-1} = \mu_K$.

12.2 ANOVA with Repeated Measures

In clinical research, it is not uncommon to have multiple assessments in a parallel-group clinical trial. The purpose of such a design is to evaluate the performance of clinical efficacy and/or safety over time during the entire study course. Clinical data collected from such a design is usually analyzed by means of the so-called analysis of variance (ANOVA) with repeated measures. In this section, formulas for sample size calculation under the design with multiple assessments is derived using the method of analysis of variance with repeated measures.

12.2.1 Statistical Model

Let B_{ij} and H_{ijt} be the illness score at baseline and the illness score at time t for the jth subject in the ith treatment group, $t = 1, ..., m_{ij}$, $j = 1, ..., n_i$, and $i = 1, ..., k$. Then, $Y_{ijt} = H_{ijt} - B_{ij}$, the illness scores adjusted for baseline, can be described by the following statistical model (see, e.g., Longford, 1993; Chow and Liu, 2000):

$$Y_{ijt} = \mu + \alpha_i + S_{ij} + b_{ij}t + e_{ijt},$$

where α_i is the fixed effect for the ith treatment ($\sum_{i=1}^{k} \alpha_i = 0$), S_{ij} is the random effect due to the jth subject in the ith treatment group, b_{ij} is the coefficient of the jth subject in the ith treatment group and e_{ijt} is the random error in observing Y_{ijt}. In the above model, it is assumed that (i) S_{ij}'s are independently distributed as $N(0, \sigma_S^2)$, (ii) b_{ij}'s are independently distributed as $N(0, \sigma_\beta^2)$, and (iii) e_{ijt}'s are independently distributed as $N(0, \sigma^2)$. For any given subject j within treatment i, Y_{ijt} can be described by a regression line with slope b_{ij} conditioned on S_{ij}, i.e.,

$$Y_{ijt} = \mu_{ij} + b_{ij}t + e_{ijt}, \quad t = 1, ..., m_{ij},$$

where $\mu_{ij} = \mu + \alpha_i + S_{ij}$. When conditioned on S_{ij}, i.e., it is considered as fixed, unbiased estimators of the coefficient can be obtained by the method of ordinary least squares (OLS), which are given by

$$\hat{\mu}_{ij} = \frac{\sum_{i=1}^{k} Y_i \sum_{i=1}^{k} t_i^2 - \sum_{i=1}^{k} t_i \sum_{i=1}^{k} Y_i t_i}{m_{ij} \sum_{i=1}^{k} t_i^2 - (\sum_{i=1}^{k} t_i)^2}$$

$$\hat{b}_{ij} = \frac{m_{ij} \sum_{i=1}^{k} Y_i t_i - \sum_{i=1}^{k} Y_i \sum_{i=1}^{k} t_i}{m_{ij} \sum_{i=1}^{k} t_i^2 - (\sum_{i=1}^{k} t_i)^2}.$$

The sum of squared error for the jth subject in the ith treatment group, denoted by SSE_{ij}, is distributed as $\sigma^2 \chi^2_{m_{ij}-2}$. Hence, an unbiased estimator

of σ^2, the variation due to the repeated measurements, is given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} SSE_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(m_{ij}-2)}.$$

Since data from subjects are independent, $\sum_{i=1}^{k}\sum_{j=1}^{n_i} SSE_{ij}$ is distributed as $\sigma^2\chi^2_{n^*}$, where $n^* = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(m_{ij}-2)$. Conditioning on S_{ij} and b_{ij}, we have

$$\hat{\mu}_{ij} \sim N\left(\mu_{ij}, \frac{\sigma^2\sum_{i=1}^{k} t_i^2}{m_{ij}\sum_{i=1}^{k} t_i^2 - (\sum_{i=1}^{k} t_i)^2}\right),$$
$$\hat{b}_{ij} \sim N\left(b_{ij}, \frac{\sigma^2 m_{ij}}{m_{ij}\sum_{i=1}^{k} t_i^2 - (\sum_{i=1}^{k} t_i)^2}\right).$$

Thus, unconditionally,

$$\hat{\mu}_{ij} \sim N\left(\mu+\alpha_i, \sigma_S^2 + \sigma^2\frac{\sum_{i=1}^{k} t_i^2}{m_{ij}\sum_{i=1}^{k} t_i^2 - (\sum_{i=1}^{k} t_i)^2}\right),$$
$$\hat{b}_{ij} \sim N\left(\beta_{ij}, \sigma_\beta^2 + \sigma^2\frac{m_{ij}}{m_{ij}\sum_{i=1}^{k} t_i^2 - (\sum_{i=1}^{k} t_i)^2}\right).$$

When there is an equal number of repeated measurements for each subject, i.e., $m_{ij} = m$ for all i, j, the inter-subject variation, σ_S^2, can be estimated by

$$\hat{\sigma}_S^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mu_{ij}-\mu_{i.})^2}{\sum_{i=1}^{k}(n_i-1)} - \hat{\sigma}^2\frac{\sum_{i=1}^{k} t_i^2}{m\sum_{i=1}^{k} t_i^2 - (\sum_{i=1}^{k} t_i)^2},$$

where

$$\hat{\mu}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{\mu}_{ij}.$$

An estimator of the variation of b_{ij} can then be obtained as

$$\hat{\sigma}_\beta^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(b_{ij}-b_{i.})^2}{\sum_{i=1}^{k}(n_i-1)} - \hat{\sigma}^2\frac{m}{m\sum_{i=1}^{k} t_i^2 - (\sum_{i=1}^{k} t_i)^2},$$

where

$$\hat{b}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{b}_{ij}.$$

1. the primary study endpoint is the illness scores of mice;

2. the scientifically meaningful difference in illness score is considered to be 1.0 or 1.5 (note that changes of 1.0 and 1.5 over the time period between Day 19 and Day 34 are equivalent to 0.067 and 0.1 changes in slope of the illness score curve);

3. the standard deviation for each treatment group is assumed to be same (various standard deviations such as 0.75, 0.80, 0.85, 0.90, 0.95, 1.0, 1.05, 1.1, 1.15, 1.2, or 1.25 are considered);

4. the null hypothesis of interest is that there is no difference in the profile of illness scores, which is characterized between baseline (at dosing) and a specific time point after dosing, among treatment groups;

5. the probability of committing a type II error is 10% or 20%, i.e., the power is respectively 90%, or 80%;

6. the null hypothesis is tested at the 5% level of significance;

7. Bonferroni adjustment for α significant level for multiple comparisons were considered.

Under the one-way analysis of variance model with repeated measures, the formula for the sample size calculation given in the previous subsection can be used.

Table 12.2.1: Clinical Assessment of Induction of EAE in Mice

Stage 0:	Normal
Stage 0.5:	Partial limp tail
Stage 1:	Complete limp tail
Stage 2:	Impaired righting reflex
Stage 2.5:	Righting reflex is delayed (Not weak enough to be stage 3)
Stage 3:	Partial hind limb paralysis
Stage 3.5:	One leg is completely paralyzed, and one leg is partially paralyzed
Stage 4:	Complete hind limb paralysis
Stage 4.5:	Legs are completely paralyzed and moribund
Stage 5:	Death due to EAE

Table 12.4.1: Sensitivity Analysis of Reproducibility Probability $\hat{P}_\Delta$

	n							
ΔT	10	20	30	40	50	60	100	∞
1.96	0.407	0.458	0.473	0.480	0.484	0.487	0.492	0.500
2.02	0.429	0.481	0.496	0.504	0.508	0.511	0.516	0.524
2.08	0.448	0.503	0.519	0.527	0.531	0.534	0.540	0.548
2.14	0.469	0.526	0.542	0.550	0.555	0.557	0.563	0.571
2.20	0.490	0.549	0.565	0.573	0.578	0.581	0.586	0.594
2.26	0.511	0.571	0.588	0.596	0.601	0.604	0.609	0.618
2.32	0.532	0.593	0.610	0.618	0.623	0.626	0.632	0.640
2.38	0.552	0.615	0.632	0.640	0.645	0.648	0.654	0.662
2.44	0.573	0.636	0.654	0.662	0.667	0.670	0.676	0.684
2.50	0.593	0.657	0.675	0.683	0.688	0.691	0.697	0.705
2.56	0.613	0.678	0.695	0.704	0.708	0.711	0.717	0.725
2.62	0.632	0.698	0.715	0.724	0.728	0.731	0.737	0.745
2.68	0.652	0.717	0.735	0.743	0.747	0.750	0.756	0.764
2.74	0.671	0.736	0.753	0.761	0.766	0.769	0.774	0.782
2.80	0.690	0.754	0.771	0.779	0.783	0.786	0.792	0.799
2.86	0.708	0.772	0.788	0.796	0.800	0.803	0.808	0.815
2.92	0.725	0.789	0.805	0.812	0.816	0.819	0.824	0.830
2.98	0.742	0.805	0.820	0.827	0.831	0.834	0.839	0.845
3.04	0.759	0.820	0.835	0.842	0.846	0.848	0.853	0.860
3.10	0.775	0.834	0.849	0.856	0.859	0.862	0.866	0.872
3.16	0.790	0.848	0.862	0.868	0.872	0.874	0.879	0.884
3.22	0.805	0.861	0.874	0.881	0.884	0.886	0.890	0.895
3.28	0.819	0.873	0.886	0.892	0.895	0.897	0.901	0.906
3.34	0.832	0.884	0.897	0.902	0.905	0.907	0.911	0.916
3.40	0.844	0.895	0.907	0.912	0.915	0.917	0.920	0.925
3.46	0.856	0.905	0.916	0.921	0.924	0.925	0.929	0.932
3.52	0.868	0.914	0.925	0.929	0.932	0.933	0.936	0.940
3.58	0.879	0.923	0.933	0.937	0.939	0.941	0.943	0.947
3.64	0.889	0.931	0.940	0.944	0.946	0.947	0.950	0.953
3.70	0.898	0.938	0.946	0.950	0.952	0.953	0.956	0.959
3.76	0.907	0.944	0.952	0.956	0.958	0.959	0.961	0.965
3.82	0.915	0.950	0.958	0.961	0.963	0.964	0.966	0.969
3.88	0.923	0.956	0.963	0.966	0.967	0.968	0.970	0.973
3.94	0.930	0.961	0.967	0.970	0.971	0.972	0.974	0.977

Under the null hypothesis, Chang et al. (1994) showed that T is asymptotically distributed as a standard normal random variable. Hence, we would reject the null hypothesis if $|T| > z_{\alpha/2}$. Assume that $n = n_T = n_C$ and under the alternative hypothesis, it can be shown that

$$\bar{x} \to_{a.s.} (\mu_T + \mu_C)/2 = \mu_*$$
$$\hat{p} \to_{a.s.} (p_T + p_C)/2 = p_*.$$

Without loss of generality, we assume $p_T\mu_T > p_C\mu_C$ under the alternative hypothesis. Thus, the power of the above test can be approximated by

$$1 - \Phi\left(\frac{z_{\alpha/2}\sqrt{2\mu_*^2(p_*(1-p_*)) + 2p_*(\sigma_T^2 + \sigma_C^2)} - (\mu_T p_T - \mu_R p_R)}{\sqrt{p_T(\sigma_T^2 + \mu_T^2(1-p_T)) + p_R(\sigma_R^2 + \mu_R^2(1-p_R))}}\right).$$

Hence, the sample size needed in order to achieve a desired power of $1-\beta$ can be obtained by solving

$$\frac{z_{\alpha/2}\sqrt{2\mu_*^2 p_*(1-p_*) + 2p_*(\sigma_T^2 + \sigma_C^2)} - \sqrt{n}(\mu_T p_T - \mu_R p_R)}{\sqrt{p_T(\sigma_T^2 + \mu_T^2(1-p_T)) + p_R(\sigma_R^2 + \mu_R^2(1-p_R))}} = -z_\beta.$$

This leads to

$$n = \frac{1}{(\mu_T p_T - \mu_R p_R)^2}\left[z_{\alpha/2}\sqrt{2\mu_*^2 p_*(1-p_*) + 2p_*(\sigma_T^2 + \sigma_C^2)} + z_\beta\sqrt{p_T(\sigma_T^2 + \mu_T^2(1-p_T)) + p_R(\sigma_R^2 + \mu_R^2(1-p_R))}\right]^2.$$

It should be noted that the above formula is slightly different from the one derived by Chang et al. (1994), which is incorrect.

An Example

Consider a clinical trial with $\mu_T = 0.20$, $\mu_C = 0.30$, $p_T = 0.10$, $p_R = 0.20$ and $\sigma_T^2 = \sigma_C^2 = 0.15$. The sample size needed in order to have an 80% ($\beta = 0.20$) power for detecting such a difference at the 5% ($\alpha = 0.05$) level of significance is given by

$$\begin{aligned} n &= \frac{1}{(0.2\times0.1 - 0.3\times0.2)^2}\left[1.96\sqrt{2\times0.25^2\times0.15\times0.85 + 2\times0.15\times0.3} \right. \\ &\quad \left. + 0.84\sqrt{0.1(0.15^2 + 0.2\times0.9) + 0.2(0.15^2 + 0.3\times0.8)}\right]^2 \\ &\approx 527. \end{aligned}$$

As a result, 527 subjects per treatment group are required for achieving an 80% power for detecting such a difference in the burden-of-illness score.

Upper Quantiles of the χ^2 Distribution

	α				
df	0.100	0.050	0.025	0.010	0.005
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2103	10.5966
3	6.2514	7.8147	9.3484	11.3449	12.8382
4	7.7794	9.4877	11.1433	13.2767	14.8603
5	9.2364	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5345	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5894
10	15.9872	18.3070	20.4832	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7568
12	18.5493	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3620	24.7356	27.6882	29.8195
14	21.0641	23.6848	26.1189	29.1412	31.3193
15	22.3071	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1565
19	27.2036	30.1435	32.8523	36.1909	38.5823
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6706	35.4789	38.9322	41.4011
22	30.8133	33.9244	36.7807	40.2894	42.7957
23	32.0069	35.1725	38.0756	41.6384	44.1813
24	33.1962	36.4150	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9279
26	35.5632	38.8851	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1945	46.9629	49.6449
28	37.9159	41.3371	44.4608	48.2782	50.9934
29	39.0875	42.5570	45.7223	49.5879	52.3356
30	40.2560	43.7730	46.9792	50.8922	53.6720

Upper Quantiles of the F Distribution ($\alpha = 0.050$)

	df1							
df2	9	10	11	12	16	20	25	30
2	19.3848	19.3959	19.4050	19.4125	19.4333	19.4458	19.4558	19.4624
3	8.8123	8.7855	8.7633	8.7446	8.6923	8.6602	8.6341	8.6166
4	5.9988	5.9644	5.9358	5.9117	5.8441	5.8025	5.7687	5.7459
5	4.7725	4.7351	4.7040	4.6777	4.6038	4.5581	4.5209	4.4957
6	4.0990	4.0600	4.0274	3.9999	3.9223	3.8742	3.8348	3.8082
7	3.6767	3.6365	3.6030	3.5747	3.4944	3.4445	3.4036	3.3758
8	3.3881	3.3472	3.3130	3.2839	3.2016	3.1503	3.1081	3.0794
9	3.1789	3.1373	3.1025	3.0729	2.9890	2.9365	2.8932	2.8637
10	3.0204	2.9782	2.9430	2.9130	2.8276	2.7740	2.7298	2.6996
11	2.8962	2.8536	2.8179	2.7876	2.7009	2.6464	2.6014	2.5705
12	2.7964	2.7534	2.7173	2.6866	2.5989	2.5436	2.4977	2.4663
13	2.7144	2.6710	2.6347	2.6037	2.5149	2.4589	2.4123	2.3803
14	2.6458	2.6022	2.5655	2.5342	2.4446	2.3879	2.3407	2.3082
15	2.5876	2.5437	2.5068	2.4753	2.3849	2.3275	2.2797	2.2468
16	2.5377	2.4935	2.4564	2.4247	2.3335	2.2756	2.2272	2.1938
17	2.4943	2.4499	2.4126	2.3807	2.2888	2.2304	2.1815	2.1477
18	2.4563	2.4117	2.3742	2.3421	2.2496	2.1906	2.1413	2.1071
19	2.4227	2.3779	2.3402	2.3080	2.2149	2.1555	2.1057	2.0712
20	2.3928	2.3479	2.3100	2.2776	2.1840	2.1242	2.0739	2.0391
21	2.3660	2.3210	2.2829	2.2504	2.1563	2.0960	2.0454	2.0102
22	2.3419	2.2967	2.2585	2.2258	2.1313	2.0707	2.0196	1.9842
23	2.3201	2.2747	2.2364	2.2036	2.1086	2.0476	1.9963	1.9605
24	2.3002	2.2547	2.2163	2.1834	2.0880	2.0267	1.9750	1.9390
25	2.2821	2.2365	2.1979	2.1649	2.0691	2.0075	1.9554	1.9192
26	2.2655	2.2197	2.1811	2.1479	2.0518	1.9898	1.9375	1.9010
27	2.2501	2.2043	2.1655	2.1323	2.0358	1.9736	1.9210	1.8842
28	2.2360	2.1900	2.1512	2.1179	2.0210	1.9586	1.9057	1.8687
29	2.2229	2.1768	2.1379	2.1045	2.0073	1.9446	1.8915	1.8543
30	2.2107	2.1646	2.1256	2.0921	1.9946	1.9317	1.8782	1.8409

Fleming, T.R. (1990). Evaluation of active control trials in AIDS. *Journal of Acquired Immune Deficiency Syndromes*, 3 (suppl. 2), 582-587.

Fleming, T.R. and Harrington, D.R. (1991). *Counting Process and Survival Analysis.* John Wiley and Sons, New York, NY.

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

Frick, M.H., Elo, O., Haapa, K., Heinonen, O.P., Heinsalmi, P., Helo, P. Huttunen, J.K., Kaitaniemi, P., Koskinen, P., Manninen, V., Maenpaa, H., Malkonen, M., Manttari, M., Norola, S., Pasternack, A., Pikkarainen, J., Romo, M., Sjoblom, T., and Nikkila, E.A. (1987). Helsikini heart study: primary-prevention trial with gemfibrozil in middle-aged men with dyslipidemia. *New England Journal of Medicine*, 317, 1237-1245.

Friedman, L.M., Furberg, C.D., and DeMets, D.L. (1981). *Fundamentals of Clinical Trials.* John Wiley and Sons, New York, NY.

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

Gail, M.H. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subjects. *Biometrics*, 71, 431-444.

Geller, N.L. (1994). Discussion of interim analysis: the alpha spending function approach. *Statistics in Medicine*, 13, 1353-1356.

George, S.L. and Desu, M.M. (1973). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases*, 27, 15-24.

Gilbert, G.S. (1992). *Drug Safety Assessment in Clinical Trials.* Marcel Dekker, Inc., New York, NY.

Glantz, S.A. (1987). *Primer of Biostatistics*, 2nd Edition, McGraw-Hill, Inc., New York, NY.

Gould, A.L. (1995). Planning and revision the sample size for a trial. *Statistics in Medicine*, 14, 1039-1051.

Graybill, F. and Wang, C.M. (1980). Confidence intervals on nonnegative linear combinations of variances. *Journal of American Statistical Association*, 75, 869-873.

Green, S. J. and Dahlberg, S. (1992). Planned versus attained designs in phase II clinical trials. *Statistics in Medicine*, 11, 853-862.

GUSTO (1993). An investigational randomized trial comparing four thrombolytic strategies for acute myocardial infarction. The GUSTO Investigators. *The New England Journal of Medicine*, 329, 673-682.

Haaland, P.D. (1989). *Experiemntal Design in Biotechnology.* Marcel Dekker, Inc., New York, NY.

Haybittle, J.L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology*, 44, 793-797.

Harris, E.K. and Albert, A. (1990). *Survivorship Analysis for Clinical Studies.* Marcel Dekker, Inc., New York, NY.

Heart Special Project Committee (1988). Organization, review and administration of cooperative studies (Greenberg report): A report from the Heart Special Project Committee to the National Advisory Council. *Controlled Clinical Trials*, 9, 137-148.

Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparison Procedure.* John Wiley and Sons, New York, NY.

Hollander, M. and Wolfe, D.A. (1973). *Nonparametric Statistical Methods.* John Wiley and Sons, New York, NY.

Howe, W.G. (1974). Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variables. *Journal of American Statistical Association*, 69, 789-794.

Hughes, M.D. (1993). Stopping guidelines for clinical trials with multiple treatments. *Statistics in Medicine*, 12, 901-913.

Hughes, M.D., and Pocock, S.J. (1988). Stopping rules and estimation problems in clinical trias. *Statistics in Medicine*, 7, 1231-1242.

Huque, M.F. and Dubey, S. (1990). Design and analysis for therapeutic equivalence clinical trials with binary clinical endpoints. Proceedings of Biopharmaceutical Section of the American Statistical Association, Alexandria, VA, 91-97.

Hyslop, T., Hsuan, F. and Holder, D.J. (2000). A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine*, 19, 2885-2897.

ICH (1998a). Ethnic Factors in the Acceptability of Foreign Clinical Data. Tripartite International Conference on Harmonization Guidline, E5.

Thall, P.F., Simon, R.M., and Estey, E.H. (1996). New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *Journal of Clinical Oncology*, 14, 296-303.

Ting, N., Burdick, R.K., Graybill, F.A., Jeyaratnam, S., Lu, T.C. (1990). *Journal of Statistical Computation and Simulation*, 35, 135-143.

Tsai, A.A., Rosner, G.L., and Mehta, C.R. (1984). Exact confidence interval following a group sequential test. *Biometrics*, 40, 797-803.

Tsiatis, A.A, Rosner, G.L., and Mehta, C.R. (1984). Exact confidence interval for lowing a group sequential test. *Biometrics*, 40, 797-803.

Tu, D. (1997). Two one-sided tests procedures in establishing therapeutic equivalence with binary clinical endpoints: fixed sample performances and sample size determination. *Journal of Statistical Computation and Simulation*, 59, 271-290.

Tygstrup, N., Lachin, J.M., and Juhl, E. (1982). *The Randomized Clinical Trials and Therapeutic Decisions.* Marcel Dekker, Inc., New York, NY.

Wang, H. and Chow, S.C. (2002). A practical approach for parallel trials without equal variance assumption. *Statistics in Medicine*, 21, No 20, To appear.

Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43, 193-200.

Wiens, B.L., Heyse, J.F., and Matthews (1996). Similarity of three treatments, with application to vaccine development. Proceedings of the Biopharmaceutical Section of the American Statistical Association, 203-206.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80-83.

Wilcoxon, F. and Bradley, R.A. (1964). A note on the paper "Two sequential two-sample grouped rank tests with application to screening experiments." *Biometrics*, 20, 892-895.

Williams, D.A. (1971). A test for differences between treatment means when several doses are compared with a zero dose control. *Biometrics*, 27, 103-118.

Williams, D.A. (1972). The comparison of several dose levels with a zero dose control. *Biometrics*, 28, 519-531.

Index

Adequate and well-controlled clinical study, 2
Adjustment for covariates, 39
Adjustment for dropout, 39
Alpha spending function, 35, 202
ANOVA
    Mean, 68
    Proportion, 97
    Repeated measures, 302
Area under the curve (AUC), 255
Asymptotic tests, 109
Average bioequivalence, 256, 257

Binomial, 183
Bioavailability, 255
Bioequivalence, 255
Bioequivalence criteria, 256
    ABE, 256
    IBE, 256
    PBE, 256
Bioequivalence test, 255
Bootstrap, 290

Carryover effect, 156
Categorical shift, 151
    McNemar's test, 151
    Stuart-Maxwell test, 153
Center imbalance, 43
Clinically meaningful difference, 12
Cochran-Mantel-Heanszel test, 149
Comparing inter-subject variabilities, 231
Comparing intra-subject CV, 222
Comparing intra-subject variabilities, 214
Comparing means, 47
    ANOVA, 68
    One-Sample, 48
    Two-Sample Parallel, 55
    Williams Design, 72
Comparing total variabilities, 239
Comparing variabilities, 213
Conditional method, 86, 171, 205
Conditional power, 207
    Mean, 207
    Proportion, 208
Conditional random effects model, 226
Confounding, 26
Contingency table, 143
    Carryover effect test, 156
    Categorical shift, 151
    Goodness-of-fit, 144
    Independence, 146
Cox's proportional hazards model, 172
Crossover design, 30, 31
CV, 222
    Conditional random effect, 226
    Simple random, 223

Data transformation, 36
Dose Response, 294
Dropout, 172

Early stopping, 45
Equality
    CV
        Conditional random effect, 228

Simple random, 224
Inter-subject
Replicated crossover, 235
Replicated paralellel, 232
Intra-subject
Replicated crossover, 219
Replicated parallel, 215
Mean
One-Sample, 48
Two-Sample Crossover, 64
Two-Sample Parallel, 56
Williams Design, 73
Proportion
One-sample, 82
Two-sample crossover, 94
Two-sample parallel, 87
Williams design, 100
Relative risk
Crossover, 107
Parallel, 104
Time-to-event
Cox's model, 173
Exponential model, 167
Log-rank test, 177
Total
Non-replicated crossover, 246
Non-replicated parallel, 240
Replicated crossover, 249
Replicated parallel, 243
Equivalence
Mean
One-Sample, 52
Two-Sample Crossover, 66
Two-Sample Parallel, 59
Williams Design, 74
Proportion
One-sample, 84
Two-sample crossover, 95
Two-sample parallel, 88
Williams design, 101
Relative risk
Crossover, 109
Parallel, 105
Time-to-event
Cox's model, 176
Exponential model, 169
Exact p-value, 116
Exact test, 109, 115
Binomial test, 115
Fisher's exact test, 119
Flexible two-stage, 126
Optimal three-stage, 127
Optimal two-stage, 122
Exponential model, 164

Fixed alternative, 159, 183
Contingency table, 159
Time-to-event, 183
Flexible designs for multiple-arm trials, 139
Food and Drug Administration, 2
Food and Drug and Cosmetics Act, 4
Fundamental Bioequivalence Assumptio, 255

Goodness-of-fit, 143, 144
Group sequential, 185
Conditional power, 207
Mean, 185
Proportion, 199
Sample size re-estimation, 204
Time-to-event, 200
Group sequential boundary, 33
Inner wedge, 195
O'Brien-Fleming, 190
Mean, 190
Proportion, 200
Pocock, 186
Mean, 186
Proportion, 200
Wang-Tsiatis, 191
Mean, 191
Proportion, 200

Hazard function, 163
Historical control, 183