GROUP SEQUENTIAL PROCEDURES FOR MULTIPLE COMPARISON OF RATES OF CHANGE

THÈSE N° 1823 (1998)

PRÉSENTÉE AU DÉPARTEMENT DE MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Bernard CERUTTI

D.E.A. de statistique et modèles aléatoires en économie et finance, Universités Panthéon-Sorbonne et Denis Diderot, France
de nationalité française

acceptée sur proposition du jury:

Prof. P. Nuesch, directeur de thèse
Prof. A.C. Davison, rapporteur
Dr J. Eisele, rapporteur
Prof. D. Reboussin, rapporteur
Prof. T. Zoubeidi, rapporteur

Lausanne, EPFL
1998

Le Professeur Toufik Zoubeidi de l'Université des Emirats Arabes Unis à Al Ain, définit, encadra et me permit de mener à bien ce travail. Qu'il soit assuré de l'expression de ma plus profonde gratitude. Ma reconnaissance s'adresse aussi aux Professeurs Anthony Davison de l'Ecole Polytechnique Fédérale de Lausanne, et David Reboussin de l'Ecole de Médecine Bowman Gray à Winston-Salem, pour l'attention extrêmement sérieuse qu'ils prêtèrent à ce document, leurs critiques pertinentes et enrichissantes.

Merci au Professeur Peter Nuesch pour sa relecture décisive et attentionnée, ainsi que pour la confiance qu'il m'accorda au cours de ces dernières années; au Docteur Jeffrey Eisele de Novartis à Bâle, qui enrichit grandement la qualité et la précision du premier chapitre; au Docteur Jean-Marie Helbling pour son accueil, ses conseils, et sa confiance dont je bénéficiai bien avant de quitter la perle des lagunes; au Fonds National Suisse de la Recherche Scientifique pour son soutien financier.

Je suis fortement redevable envers tous les membres des unités de statistique. Mes pensées vont aussi à ceux qui m'accueillirent lorsque je vins à Lausanne.

I miei ringraziamenti vanno pure ad Alessandra Brazzale che rilesse la versione completa del manoscritto, sacrificando parte del suo tempo libero durante una fine settimana.

My gratitude to Judith Munroe for her excellent teaching of the English language.

And to the matchless scientific, computer, and personal advisors: Alex Amiguet, Eva Cantoni, Enrico and Valérie Chavez, Jane Doe, Anne-Catherine Favre, Olivier Renaud, Armin Röhrl, and Jacques Zuber.

My especial thanks to Jeff Henson, Sarah Minogue, and her father — enrolled under pressure I presume — for their useful translations, corrections, and suggestions.

Demain, comme aujourd'hui, je leur serai reconnaissant.

B.C.

Le 18 juin 1998

Abstract

Group sequential methods find a particular field of application in clinical trials because patient recruitment is by nature sequential. It is important to minimize the number of patients exposed to an inferior treatment. Moreover, the experimenter is required to monitor the data on a few occasions during the trial to check for toxicity or other unsuspected harmful side-effects. These intermediate analyses may reveal a significant superiority of one treatment, resulting in an early termination of the trial, but only a sequential design allows for such early termination.

The availability of fast and efficient algorithms for estimating parameters of multivariate distributions (e.g. Expectation Maximization or Newton-Raphson algorithms) and for computing multivariate normal probabilities opens new frontiers by allowing complex but very flexible experimental designs for the comparison of several treatments. Using these powerful tools, together with the concept of spending function, flexible procedures for multiple comparison of treatments have been designed.

In 1991, Lee and DeMets (*Journal of the American Statistical Association* **86**, 757–62) proposed a group sequential procedure for comparing the rates of change of two treatments. Our work generalizes their model, a linear mixed effects model with repeated measurements, to several treatments. We derive group sequential procedures for

- comparing the rates of change of several treatments to a control while controlling the overall significance level, or more generally, comparing general contrasts of changes of several treatments while controlling the overall significance level, and
- studying and comparing data-dependent allocation rules and strategies for assigning patients to treatments.

These procedures and algorithms are applied to real and simulated data.

The importance of this work lies in the fact that quite often in clinical trials, we compare the effects of more than one treatment or the effects of different doses of the same treatment. Here we provide flexible procedures for multiple comparisons to the medical experimenter. With these procedures, the number and times of interim analyses need not be specified in advance, missing observations can be handled, and the experimenter

can freely choose which treatment effects to test at any time he wishes. The overall significance level of the tests is controlled. Moreover algorithms for the execution of these procedures are provided.

This study tries to open avenues of further theoretical and numerical studies in multiple comparisons, a field which has a long history in theoretical and applied statistics and which recently has seen vigorous new developments to which the above-described research contributes.

Résumé

Le champ d'application des méthodes séquentielles sur données groupées dans le domaine des études cliniques est très vaste, car le recrutement des patients s'effectue en général de manière séquentielle. Il est alors important de pouvoir minimiser le nombre de patients qui se verront administrer un traitement d'efficacité moindre.

Au cours d'une étude clinique, une variable réponse principale est mesurée lors de chaque visite de contrôle prévue par le protocole d'expérience, et visée par l'attaché de recherche clinique. Le praticien doit être en mesure de contrôler régulièrement l'absence de toxicité ou de tout autre effet secondaire dangereux. Ces analyses intermédiaires peuvent aussi apporter l'évidence de la supériorité d'un traitement, et ainsi mettre un terme à l'étude en cours. Seul un plan d'expérience de nature séquentielle autorise un tel arrêt anticipé, qui soit rigoureusement justifiable d'un point de vue statistique.

Aujourd'hui, des algorithmes rapides et efficaces, par exemple de type Expectation Maximization ou Newton–Raphson, permettent d'estimer les paramètres de distributions multivariées. De même, il est aisé d'évaluer l'intégrale de la densité de variables aléatoires normales sur un compact étoilé de $\mathbb{R}^p$. En utilisant de pair le concept de fonction de dépense de l'erreur, on peut alors développer des procédures flexibles pour comparer plusieurs traitements entre eux de manière séquentielle.

En 1991, Lee and DeMets (*Journal of the American Statistical Association* **86**, 757–62) proposèrent une procédure séquentielle sur données groupées pour comparer les courbes de croissance respectives de deux traitements. Le présent document généralise leur procédure, pour comparer plusieurs traitements, dans le cadre des modèles linéaires à effets mixtes. Des procédures séquentielles sur données groupées sont proposées pour

- Comparer les taux de croissance de plusieurs traitements face à un contrôle, ou de manière générale, de plusieurs traitements entre eux, tout en garantissant un seuil global de signification désiré.
- Etudier et comparer des règles d'allocations pour optimiser l'affectation de patients nouvellement incorporés dans l'expérience.

Le comportement de ces procédures est étudié par simulation, ou par application sur des jeux de données existants.

En pratique, il arrive fréquemment que le clinicien compare plusieurs traitements, ou plusieurs doses du même traitement. Pour ce faire, des procédures très souples et les algorithmes correspondants lui sont ici fournis. Il n'est nullement impératif de spécifier à l'avance ni le nombre, ni les dates des analyses intermédiaires. Les données manquantes, en tout cas de manière aléatoire, sont traitées sans difficulté. Enfin, le praticien peut librement choisir lors de chaque étape intermédiaire quels contrastes doivent être testés. Bien entendu, le seuil global de signification souhaité reste sous contrôle.

Cette étude essaye d'offrir de nouvelles perspectives au domaine des comparaisons multiples, qui a déjà un long passé statistique, aussi bien théorique que pratique. Il connaît aujourd'hui une forte croissance, à laquelle le présent document voudrait pouvoir contribuer.

Contents

Chapter 1

Introduction

This introductory chapter begins with a presentation of some of the elements making up a clinical experiment. Since our work finds a particular field of applications in clinical experiments, a brief discussion of them is found in Section 1.1.

A study performed on patients may last for years, at great human and financial cost. How can we try to stop it at the right time? For example, the experiment may be stopped either because the superiority of a new therapy becomes evident, or because there is no longer any hope of a positive conclusion. Stopping at the right time is the fundamental objective of sequential methods which are introduced in Section 1.2.

Our work is concerned with the situation where we measure the evolution of one or several variables in certain individuals over a period of time. Section 1.3 is dedicated to modelling this type of study.

The manner of combining this type of model with sequential methods is briefly presented in Section 1.4. Finally, some relevant references are given in Section 1.5, and a chapter outline is found in the last section.

The reader who is familiar with the fields mentioned above may easily switch to the next chapter. He or she might need only the model presented in Section 1.3.2.

1.1 Clinical experiments

Experimental studies are not new, although they used to be carried out more or less voluntarily. The terrifying example below (Contente Domingues and Guerreiro, 1988), reported from a voyage from Brazil to the Portuguese Indies, is an example of observational study.

> Appelée scorbut par les Hollandais, & par les Portugais le mal de gencives. Nos Français l'appellent le mal de terre, & on ne sait pourquoi, car elle prend

à la mer, & se guérit en terre. C'est une maladie fort commune le long du voyage, & est contagieuse, même à l'approcher, & sentir l'haleine d'un autre. Elle vient ordinairement à cause des grandes longueurs du voyage, & longue demeure sur mer sans prendre terre, & aussi faute de se laver, nettoyer & changer de linge & d'habits, avec l'air marin, l'eau de mer, la corruption des eaux douces, & des vivres, & se laver en eau de mer, sans après se laver d'eau douce, puis le froid, & dormir la nuit au serein, tout cela cause ce mal.

Ceux qui en sont surpris en deviennent enflés comme hydropiques, & l'enflure est dure comme du bois, principalement aux cuisses & jambes, les joues & la gorge, & tout cela est couvert de sang meurtri de couleur livide & plombée, comme de tumeurs & contusions qui rendent les muscles & les nerfs raides & perclus. Outre ce, les gencives sont ulcérées & noires, la chair toute enlevée, & les dents disloquées, & branlantes, comme si elles ne tenaient qu'à bien peu de chose, & même la plus grande partie en tombe. Avec cela une haleine si puante & infecte qu'on ne peut s'en approcher; car on sent cela d'un bout du navire à l'autre. On ne perd pas l'appétit, mais l'incommodité des dents est telle, qu'on ne saurait manger, sinon choses liquides dont alors il se trouve peu ès navire, & cependant on devient si avide, qu'il semble qu'on aurait pas assez de tous les vivres du monde pour s'assouvir.

En somme, que l'incommodité en est bien plus grande que la douleur, que l'on sent seulement en bouche et gencives. de sorte que bien souvent on meurt en parlant buvant & mangeant, sans avoir eu connaissance de sa mort. Outre cela, cette maladie rend si opiniâtre & bizarre, que tout déplaît. Il y en a qui en meurent en peu de jours, d'autres durent plus longtemps sans mourir. Ils ont la couleur blême & jaunâtre: & quand ce mal veut prendre, les cuisses & les jambes sont couvertes de petites pustules & taches comme morsures de puces, qui est le sang meurtri qui sort par les pores du cuir: & les gencives commencent à s'altérer, & devenir chancreuses. Ils sont sujets aussi à syncopes, évanouissements & défaillements de nerfs.

Comme nous étions en l'île de S. Laurent, il en mourut trois ou quatre des nôtres, de cette maladie, & comme on leur ouvrit la tête, on leur trouva tout le cerveau noir, gâté & putréfié. les poumons deviennent secs, & retirés comme du parchemin approché du feu. le foie & la rate grossissent démesurément, & sont noirs & couverts d'apostumes pleines de matière la plus puante au monde. Lors que l'on a cette maladie une plaie ne se guérit & dessèche jamais, aussi devient comme une gangrène & putréfiée. Quand on est sur mer, & que cette maladie prend, on a beau user de remèdes, car tout y est inutile, & n'y en a point d'autre que de prendre terre quelque part si on peut, afin d'avoir des rafraîchissements d'eaux douces & fraîches, & de fruits, sans quoi, l'on ne peur jamais guérir, quoi qu'on y fasse. C'est une

chose terrible de voir les gros morceaux de chair pourrie qu'il faut couper des gencives.

Actually, some authors attribute the first recorded controlled trial to Lind (1753). He investigated in 1747 the effect of oranges and limes on scurvy, on HMS Salisbury.

The term clinical experiment means any medical experiment in which human subjects are involved. Very strong ethical constraints apply when experimentation concerns people. For instance, it would be desirable for the experiment to be such that no new patient is submitted to a treatment that has already been recognised as inferior.

1.1.1 Description of a study

Generally, we distinguish two types of studies:

- **Observational studies**: in such studies, the environment determines who is exposed to the factor, who is the object of the study, and who is not. Though such studies may suggest links, they may not necessarily imply a causal relation. Observational studies may be divided into three categories:
  - **cross-sectional**: the different measures taken on patients are made at the same instant. This kind of study does not allow the establishment of a causal relation: for example, let us suppose that we measure each patient's blood pressure, and at the same instant, the presence or absence of heart disease. We can certainly demonstrate a link, but in no way can we determine if hyper-tension is a consequence of the heart disease, or if it is the cause.
  - **case-control**: here, we consider a sample having a given characteristic (for example a group of patients suffering from lung cancer), and we measure the presence or absence of certain characteristics (is the patient a smoker?). The term "retrospective study" is well suited, to the extent that the data are collected after the factors of interest have manifested themselves. Collecting information afterwards gives nevertheless an additional source of imprecision.
  - **prospective**: in this kind of trial, we begin the study by taking a sample having a given characteristic (the patient is a smoker), and another without this characteristic (the patient does not smoke). The trial is achieved by measuring the frequency of appearance of the primary factor (lung cancer). Prospective studies are sometimes called "cohort studies".

  If the prospective study is the most convincing one for highlighting a causal relation, it is nevertheless often difficult to set up, costly, and sometimes impossible to undertake: it may take decades before the disease is diagnosed, or the disease may be so rare that a huge sample is required.

- **Experimental studies**: here, the investigator determines who is exposed and who is not. Such studies allow direct inference about causal relations. In the medical field, an experimental study is called a clinical trial (Friedman, Furberg, and DeMets, 1985). The researcher usually splits the patients randomly into several groups. To the extent possible, he or she tries to maintain all the relevant factors constant or controlled.

  Generally, a clinical study compares the effects of medicines with that of a placebo, or a new therapy with a standard therapy. It may however compare a medical treatment to surgery, or in other fields, different learning methods.

1.1.2 Therapeutic trials in the harmonisation era

Clinical trials pose many ethical dilemmas. For example, one should ensure that patients do not receive a treatment that seems to be less effective. But one must also maintain a certain scientific objectivity: a study which rules out a treatment when there is doubt about its effectiveness would result in a decision without any scientific value. However, there may still be scientific value if there was information on side-effects.

Unfortunately, the issues at stake are often contradictory:

- ethics and statistical validity;
- premature stopping and loss of credibility within medical circles; and
- individual and collective interests.

Because of these conflicts, the need has progressively been felt for international common standards, able to deal with the establishment of protocols, ethical rules, and the operating mode of steering committees.

The principle of the International Conferences of Harmonization was born in 1989, arising from the findings of a marked disparity of legal requirements among countries. The International Conferences of Harmonization have a triple goal:

- to create a dialogue between the regulators — the Food and Drug Administration in the USA, the Committee for Proprietary Medicinal Products in Europe, and the Ministry of Health and Welfare in Japan — and the pharmaceutical industry;
- to save resources; and
- to give a set of common, practical recommendations for a more uniform interpretation and application of technical guidelines.

The progress which has been accomplished is already significant and for instance, has led to the adoption of directives E3 (Structure and Content of Clinical Study Reports), E6 (Good Clinical Practice), and E9 (Statistical Considerations in the Design of Clinical Trials). Nonetheless, disparities remain: 80 statisticians work at the FDA, none at the beginning of 1997 at the French Ministry of Health!

After these introductory paragraphs, we consider some statistical aspects of the problem.

1.1.3 The different phases of clinical development

Clinical development is usually divided into four phases. During Phase I, the new medicine is administered to healthy volunteers, in order to check for unforeseen serious side effects, which would immediately end the trial. Different doses are tested, for studying the tolerance level of the new product. In the case where the treatment may be toxic, as in oncology, or in the treatment of cardiovascular diseases, the healthy volunteers are replaced by patients who usually do not hold much hope for remission.

During Phase II, the medicine is distributed to persons suffering from the target sickness, so as to give an idea of its effectiveness. At the end of the first two phases, the investigators are in a position to propose a complete rule for prescribing. This rule includes the mode of administration, the interval between doses, dosage, the associated treatments, and the policy to follow in the case of non-effectiveness.

During Phase III, the new medicine is compared with alternatives (placebo, previous treatment). This phase is crucial for obtaining authorisation for sale of the drug. The article by Thall, Simon, and Ellenberg (1988) is a good reference in phase II/III studies.

Phase IV is designed to ensure continuous monitoring, often after commercialisation, and provides additional security that may be used for marketing purposes. Many pharmaceutical companies call Phase III Phase IV, from the time that a production licence is granted.

1.1.4 Control of error rates

The rules established by the official monitoring bodies require a level of significance low enough that new products, often more expensive than the old ones, are not placed on the market unless they are demonstrably better. However, the relative importance of the first and second type of errors depends on the trial: a Type II error is important in oncology — screening trials a larger Type I error rate may be acceptable in order to increase power — while a Type I error is more relevant for a costly flu medicine.

In certain practical studies, the most important aspect can be the sign of the parameters being studied. One can then speak of directional decisions and introduce the notion of

a Type III error (see Hochberg and Tamhane 1987, p. 39). This error is the probability of any misclassification of signs of non-zero parameter values.

1.2 Scope of sequential methods

Sequential analysis concerns the studies in which the number of observations needs not be determined in advance. It is characterised by three elements:

- a stopping rule,
- an allocation rule, and
- a decision rule.

The purpose of sequential analysis is the simultaneous optimisation of the sample size necessary and the quality of the decision rule. The use of standard theory, i.e. work on a fixed-size sample, used to be protected for a long time by an array of passionate defenders. In addition, when a problem is tractable by both fixed-sample and sequential procedures, the former are conceptually and mathematically easier to carry out. This has long contributed to the under-use of sequential methods.

We recall some situations and justifications for sequential analysis:

- intrinsically sequential analyses such as monitoring processes or the so-called "secretary problem";
- situations where only sequential analysis can provide a solution;
- economical use of human, animal, and material resources;
- taking ethical factors into account; and
- reinforcement of a fixed sample size procedure.

Sequential methods have a role to play in each phase of clinical development:

- in the course of Phase I, the first patients receive a small fraction of the dose determined through animal testing. Then increasing doses are administered. The approach is entirely sequential: when should it end? If it should continue, with how many additional patients? However, formal sequential design for Phase I is rare in practice: it is in fact difficult to formulate the precise objective of such a study, and what reduces statistical power is the very small number of persons involved. However, there are now formal Bayesian sequential designs for Phase I (see O'Quigley, Pepe, and Fisher, 1990).

- As far as Phase II is concerned, we can compare a treatment with a standard. We can also compare several new treatments, in order to single out the most promising for Phase III.

- The greatest contribution of statistics in clinical trials comes in Phase III: large scale comparison of treatments. Unfortunately this assertion should be tempered, because Phase III trials are sometimes in practice not well thought out.

  Large samples and the assurance of high power are essential in order to convince health authorities and the medical establishment of the efficacy of the new therapy. Generally, a new treatment is compared with a placebo or with other treatments. The null hypothesis is the lack of difference between their effects. Measurements are taken over time, and the objective of the sequential approach is to stop the trial as soon as a difference between the effects is ascertained. The investigator must balance a premature stopping, which lacks credibility, and a too-conservative method, which would result in the continuation of the trial, even though it would be preferable to stop it.

  We mention here the particular case of the bioequivalence trial, in which a new better treatment (which has fewer side effects, is more easily administered, or costs less) is compared with previous treatments. Its purpose is to show that there is no difference between the effects of the two treatments. In such a situation, we seek a premature stopping in favour of the hypothesis H_0, of equal effects of all treatments. Power rather than the level of significance is the thing of central interest. We refer to Jennison and Turnbull (1993), who propose a double triangle type sequential test for the equivalence between two treatments, or Betensky (1997), who evaluates, at each step, the probability of rejecting H_0 at the anticipated end of the experiment, given the current data.

- Statistical models are seldom used in the Phase IV. We however refer as an example to the "yellow card" that is used in the UK. British general practitioners report to the Department of Public Health all problems encountered following the prescription of a given medicine, after agreement of volunteer patients.

1.3 Longitudinal data, linear mixed effects model

1.3.1 Longitudinal data

In univariate statistics, we usually take a single measure on each individual. In a multivariate model, we have a vector of variables for each individual. On the other hand, when we speak of longitudinal data, we possess a vector of measures for each individual, but these now represent the same physical quantity measured at a sequence

of observation times. There may be also several vectors of measures that concern several variables.

Longitudinal data combine elements of multivariate statistics and time series. However, they differ from standard multivariate models in the sense that the time aspect leads to stronger dependence between the measurements than with more conventional multidimensional data. They also differ from conventional time series models because they consist of several short time series, one for each individual.

One of the most common examples of longitudinal data is growth curves. The example introduced below is used in the next chapters.

Example 1.3.1 *These data are taken from Crowder and Hand (1990, pp. 74-78). The purpose is to study the effect of a protein diet on the growth of chickens, for a period of three weeks. There are four groups of chickens of sizes 16, 10, 10, and 9, corresponding to an inspection, and three groups with dietary patterns with protein replacement rates of 10%, 20%, and 40% (see Figure 1.1). The animals which had to be withdrawn from the experiment were not included in the data used, because the objective is inference on healthy individuals. Twelve observations are available for each individual, with no missing values.*

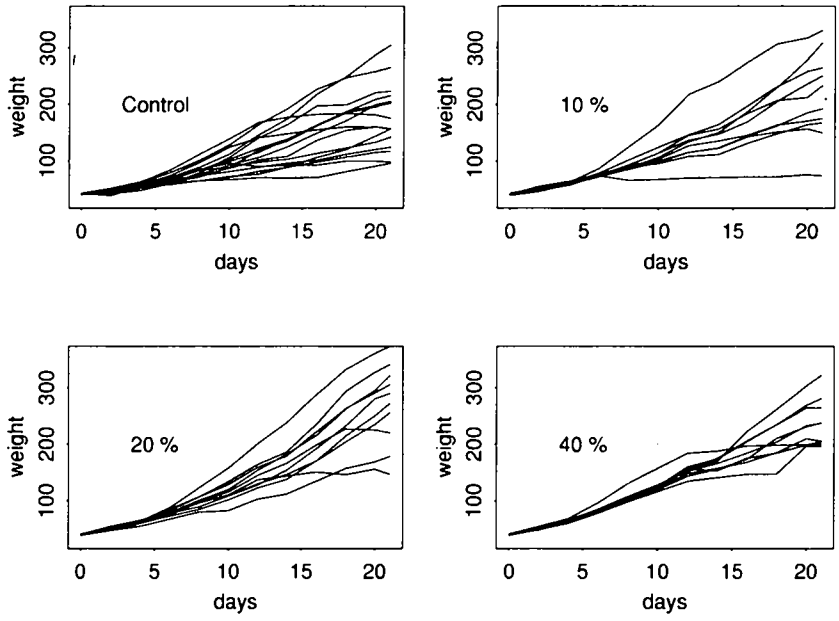

Figure 1.1: *Growth of chickens with different protein diets. The data are taken from Crowder and Hand (1990).*

1.3.2 The linear mixed effects model

When we study rates of change, we have to focus on individual histories, looked on as sample paths or realizations of stochastic process. A clinical trial is often composed of measures, over time, on patients who enter the trial in a staggered manner. In addition, measurements are often not taken on all of the patients on the same date, or may be missing. The many reasons include sickness, vacations, and relocation.

It is therefore opportune to allow the experimental design to be different for each patient. Let Y_i be a $n_i \times 1$ vector containing the measurements for subject i; then a straightforward model is

$$Y_i = X_i \beta + e_i, \tag{1.3.1}$$

where X_i is a known $n_i \times p$ matrix, β is a $p \times 1$ vector of unknown regression parameters, and the e_i are independently distributed as $\mathcal{N}_{n_i}(0, \Sigma_i)$. We give below a few possible structures for Σ_i, listed by Jennrich and Schluchter (1986). The number of parameters in Σ_i is $n(\theta)$:

- $\Sigma_i = \sigma^2 I$: the observations are independent and identically distributed. This is the standard regression model, for which $n(\theta) = 1$;
- $\Sigma_i = \sigma_b^2 E + \sigma^2 I$, where the matrix E is made of 1s; this structure is a mixed-model ANOVA (Analysis of Variance) structure, sometimes called a compound symmetry model, and $n(\theta) = 2$;
- $\Sigma_i = Z_i D Z_i' + \sigma^2 I$: here, Z is a $n_i \times q$ known matrix, and D is a $q \times q$ unknown matrix. This model is called a mixed effects model, and $n(\theta) = 1 + q(1+q)/2$;
- $\Sigma_{i[j,k]} = \sigma_l^2$, where $l = |j - k| + 1$; this structure is called a banded, or general-autoregressive structure, and $n(\theta) = n_i$.
- $\Sigma_{i[j,k]} = \sigma^2 \rho^{|j-k|}$: this corresponds to the standard autoregressive structure of order one. There are two unknown parameters, σ^2 and ρ.
- Finally, if Σ_i is unstructured, we have $n(\theta) = n_i(n_i + 1)/2$.

The parameter β in (1.3.1) stands for a mean treatment effect. But clinical or pre-clinical trials usually involve human beings or animals; for each individual, there is likely to be a consistent departure from the mean treatment effect. This lasting characteristic of the individual is ascribed to many unknown or uncontrolled factors, which are naturally modelled by a normal random variable, usually called random effect. The mixed effect model — the third case in the list above — includes both fixed and random effects. Since it is often used, mentioned in literature, and also very flexible, we have worked in the framework of the linear mixed effects model, which is discussed with further details in the following.

Statistical model

Let h be the treatment number ($h = 1, \ldots, T$). Then, for the ith patient assigned to the hth treatment, we have

$$Y_i^h = X_i^h \beta^h + Z_i^h b_i^h + e_i^h, \tag{1.3.2}$$

where Y_i^h is a $n_i^h \times 1$ vector of observations taken on subject i, assigned to treatment h. The matrices X_i^h and Z_i^h are $n_i^h \times p$ and $n_i^h \times q$ matrices of known covariates. The $p \times 1$ vector β^h is the fixed effect. Vectors b_i^h and e_i^h are independent errors distributed as $\mathcal{N}_q(0, \sigma^2 D)$, and as $\mathcal{N}_{n_i^h}(0, \sigma^2 I)$.

We define

$$V_i^h = \text{var}(Y_i^h) = \sigma^2 (I + Z_i^h D Z_i^{h\prime}), \tag{1.3.3}$$

$$\text{and } W_i^h = (V_i^h)^{-1}. \tag{1.3.4}$$

We generally take $p = 2$, corresponding to a linear model containing a slope and an intercept. In the absence of additional information, we take $q = 2$ and $Z_i^h = X_i^h$.

Let us suppose that we are at the instant of performing the kth interim test ($k = 1, \ldots, K$). Then

$$Y_i^h(k) = X_i^h(k) \beta^h + Z_i^h(k) b_i^h + e_i^h(k), \tag{1.3.5}$$

$$X_i^h(k) = \begin{pmatrix} 1 & X_{i,1}^h \\ 1 & X_{i,2}^h \\ \vdots & \vdots \\ 1 & X_{i,n_i^h(k)}^h \end{pmatrix}, \beta^h = \begin{pmatrix} \beta_0^h \\ \beta_1^h \end{pmatrix} \text{ and } b_i^h = \begin{pmatrix} b_{0,i}^h \\ b_{1,i}^h \end{pmatrix},$$

where $n_i^h(k)$, $h = 1, \ldots, T$, $i = 1, \ldots, m^h(k)$, is the number of measures taken on patient i, assigned to treatment h, up to the kth analysis, and $m^h(k)$ is the number of patients assigned to treatment h, up to the kth analysis.

The linear mixed effects model (1.3.5) is convenient, because it clearly distinguishes

- the treatment effect β^h,
- the reaction of each patient b_i^h, and
- the variation within the measures taken on a single patient e_i^h.

In addition, there are no constraints on the number of subjects assigned to each treatment, or on times of measurements.

Follow-up remarks

One of the assumptions of (1.3.2) is the independence of the effects b_i^h, violation of which may lead to inconsistent estimators of the fixed effects.

Using random effects models is far from innocuous. Their unjustified use leads to poor quality estimators and over-large confidence intervals. On the other hand, not taking the random effects into account when they are present leads to underestimation of the heterogeneity of the data. Specification tests have been developed to check the need for random effects (Hausman, 1978).

This remark about the need for random effects is important in sequential analysis: use of stopping rules can bias estimators of fixed effects, and lead to an artificial heterogeneity which wrongly suggests using a random effects model (Hughes, Freedman, and Pocock, 1992).

1.4 Group sequential procedures

1.4.1 Reluctance to use fully sequential procedures

In many clinical trials, patients arrive sequentially. A fully sequential plan is an experimental design where an interim test is performed after every measurement.

Siegmund (1977, 1985) showed that sequential procedures, such as the Sequential Probability Ratio Test (SPRT), used with adequate allocation rules, were generally worthy of consideration. In spite of his work, the implementation of such procedures remains sporadic.

Practitioners are generally reticent about sequential procedures, and prefer to use a single final test. O'Brien and Fleming (1979) give some of their practical objections:

- a feeling of uncertainty and scepticism with regard to the nearly optimal procedures;
- the number of sequential methods already proposed is a deterrent;
- sequential methods bring about complex designs; and
- unforeseen reasons may lead to serious modifications of the design; but changes in the connected sequential plan are then intricate.

Pocock (1993) mentions other drawbacks:

- lack of credibility due to the small scale of the experiment;

- lack of realism about the difference of effects between treatments;
- larger confidence intervals;
- bias due to stopping rules;
- partial appraisal of costs and benefits, due to the procedure's rapidity;
- the communication of the intermediate results may be a source of pressures and "recommendations" for the investigators; and
- the higher risk of misleading conclusion.

As regards the error rate, Armitage, McPherson, and Rowe (1969) addressed the problem of the drastic increase of Type I error when a test of level α is repeated over time. Consider for instance normal random values $\mathcal{N}(\mu = 0, \sigma^2 = 1)$. If we tested fifty times the null hypothesis $H_0 : \mu = 0$, at level $\alpha = 0.01$, the overall Type I error would be 0.53! Nevertheless, Armitage (1975) stressed that the repeated-test approach should be followed up.

1.4.2 A suitable compromise

We are able to argue against the objections mentioned in the previous section. In reality, every investigator monitors "first hand" the evolution of the data as soon as new observations are collected. Is not this method sequential?

Grouping observations is a very good alternative to a fully sequential method. The investigator periodically examines his or her data. Thus, he or she is able to stop the trial in the case of strong evidence against the null hypothesis. In addition, a clinical trial often takes place in several centres. In such a situation, we do not hold out much hope that every new measure will be transmitted to the central monitoring centre without delay, and difficulty in scheduling meetings can also cause delays; but it is easier to wait until a given number of new measurements is transmitted, and perform an interim test.

Data are also collected in almost the same manner as for a non-sequential experiment. Finally, the gain achieved by a periodic inspection of the data is obtained virtually without loss of power (O'Brien and Fleming, 1979), or increase in Type I error.

The advantages of group sequential methods are well illustrated by Pocock (1982). He gives an example of a group sequential procedure that allows a 40% reduction (under the alternative hypothesis) of the Average Sample Number (ASN) compared to a non-sequential procedure, and an increase of only 19% under the null hypothesis.

A fully sequential design also leads to a reduction of 40% of the ASN under the alternative hypothesis, but to an increase of more than 35% under the null hypothesis.

Experience has shown (Pocock, 1982) that the number of interim tests is often rather low (less than or equal to five), unless there is strong a priori evidence against the null hypothesis.

1.4.3 Some group sequential procedures

We present here well-known group sequential procedures. They deal with two-sided tests: even when we compare the effect of a treatment to a placebo, we must be able to avoid unforeseen harmful effects.

We consider normal observations with known variance. Since one will work with grouped data, this does not cause any problems even for moderate sample sizes. The general framework is defined by the assumptions below.

Assumptions 1.4.1 *We consider two treatments, A and B. Then,*

- *at each step, two groups of n_A and n_B patients are assigned respectively to treatments A and B;*
- *observations are normal $\mathcal{N}(\mu, \sigma^2)$, with $\mu = \mu_A$, or $\mu = \mu_B$, and we set $\delta = \mu_A - \mu_B$;*
- *the maximum number of interim tests is N;*
- *the variance σ^2 is known; and*
- *we test $H_0 : \delta = 0$ vs $H_0 : \delta \neq 0$.*

Pocock's procedure

Under Assumptions 1.4.1, Pocock (1977) proposes group sequential tests, when $n = n_A = n_B$. After the ith measurement, we consider the statistic

$$\bar{d}_i = \sum_{j=1}^{i} \frac{\bar{x}_{A,j} - \bar{x}_{B,j}}{i} \sim \mathcal{N}\left(\mu_A - \mu_B, \frac{2\sigma^2}{i\,n}\right), \tag{1.4.1}$$

where $\bar{x}_{A,j}$ is the average of the measures collected from treatment A between the $(j-1)$th interim test and the jth test. We reject H_0 if

$$\mathrm{pr}_i = 2\left[1 - \Phi\left(\sqrt{i\,n}\,\frac{\bar{d}_i}{\sqrt{2\sigma^2}}\right)\right] < \alpha', \tag{1.4.2}$$

where Φ is the cumulative distribution function of a standardised normal random variable. The level α', which depends on N, is chosen to guarantee a fixed significance level α. We write

$$\alpha' = \mathrm{pr}_P(2, N, \alpha), \tag{1.4.3}$$

because the interim significance level α', depends on N, on α, and on the number of present treatments, that is, two. If we call $\mathrm{b}_P(2, N, \alpha)$ the corresponding boundary, we reject the null hypothesis if

$$Z_i^2 = \frac{in}{2\sigma^2}\bar{d}_i^{\,2} \geq \mathrm{b}_P^2(2, N, \alpha). \tag{1.4.4}$$

For $\alpha = 0.05$ or $\alpha = 0.01$, Pocock (1977) provides a table indicating the value to choose for α'. The power, $1 - \beta$, is expressed as a function of the ratio

$$\frac{\sqrt{n}\delta}{\sqrt{2}\sigma}. \tag{1.4.5}$$

With fixed N, α in $\{0.01, 0.05\}$, and $1 - \beta$ in $\{0.5, 0.75, 0.9, 0.95, 0.99\}$, tables give the value for (1.4.5), in other words n.

Even a restricted value of N allows a substantial reduction of the ASN under the alternative hypothesis. This reduction is very close to, and sometimes better than the reduction obtained by a fully sequential procedure.

If σ^2 is unknown, Φ is replaced in (1.4.2) by the distribution function of a Student's distribution with $2(in - 1)$ degrees of freedom. Pocock (1982) observes that the desired level α is still controlled, and that the loss of power is slight.

By normal approximation, we are able to use the tables of Pocock (1982) for exponential data, binary data, and for the application of the Wilcoxon test.

If we try to optimise the intermediate significant levels, when power and the Type I error rate are fixed, it turns out that these levels are almost constant. Thus, unless power is inferior to 90% or the sample size proves inadequate, results show that a constant significant level α' given by (1.4.3) is highly workable.

O'Brien and Fleming's procedure

O'Brien and Fleming's procedure (1979) is a little more supple — there are no more restrictions $n_A = n_B$ — and aims to improve the power. An interim test is performed after n_A new observations for treatment A, and n_B for treatment B. We consider the statistic

$$Z_i = \frac{1}{\sqrt{i}} \sum_{j=1}^{i} \frac{(\bar{x}_{A,j} - \bar{x}_{B,j})}{\sqrt{\mathrm{var}(\bar{x}_{A,j} - \bar{x}_{B,j})}} = \sum_{j=1}^{i} \frac{z_j}{\sqrt{i}}.$$

Under H_0, Z_i and z_j are standard normal. O'Brien and Fleming (1979) stop the experiment in favour of the alternative at the first test i, $1 \leq i \leq N$, such that

$$\frac{i}{N} Z_i^2 \geq \mathrm{b}_{OBF}^2(2, N, \alpha), \tag{1.4.6}$$

where $\mathrm{b}_{OBF}(2, N, \alpha)$ is obtained by simulation. For $N \leq 5$, $\mathrm{b}_{OBF}^2(2, N, \alpha)$ may be replaced by the quantile $1 - \alpha$, of a χ_1^2.

In addition, we are allowed, at step i, to take a proportion k_i of n_A and n_B, in which case the stopping rule (1.4.6) is replaced by

$$\frac{\sum_{j=1}^{i} k_i}{\sum_{j=1}^{N} k_i} Z_i^2 \geq \mathrm{b}_{OBF}^2(2, N, \alpha).$$

If sample sizes are large, the procedure of O'Brien and Fleming is also usable with continuous observations, or binary observations.

Whereas the stopping boundaries of Pocock (1977) are constant, those of O'Brien and Fleming (1979) become progressively smaller, with the aim of increasing power. Moreover, this may answer more easily two questions that must be solved before any decision for early termination can or should be reached (DeMets, 1984): Could the current trends likely be reversed if the trial continued to the end? What would the impact of early termination have on acceptability of the results to medical and scientific colleagues?

Lan and DeMets' procedure

The method of Lan and DeMets (1983) offers great flexibility. In fact, N, the number of intermediate steps, does not have to be specified in advance. Thus, we are able to deal with many practical situations:

- depending on the trial, we are able to change the frequency of the tests; and
- the recruitment of patients may be slower than expected, so the experiment runs longer than the estimated duration, and therefore, N increases.

The context of use is very general, that is,

- two treatments A and B; and
- the boundaries are usually computed by normal approximations.

We assume without loss of generality that the experiment is performed on a time interval of duration one. If a test is performed at time t, $0 < t \leq 1$, we compute the interim significant level by using an Error Spending Function (ESF) $\alpha^*(\cdot)$. If the random variable τ represents the moment when we stop the experiment, we have at least asymptotically

$$\begin{aligned} \text{pr}\left(|B(t_1)| > b_1\right) &= \text{pr}\left(\tau \in [0, t_1]\right) = \alpha^*(t_1), \\ \text{pr}\left(|B(t_j)| < b_j,\ j = 1, \ldots, i-1, \text{ and } |B(t_i)| > b_i\right) &= \text{pr}\left(\tau \in (t_{i-1}, t_i]\right) \\ &= \alpha^*(t_i) - \alpha^*(t_{i-1}), \end{aligned}$$

where $B(\cdot)$ is a standard Brownian motion and b_i a boundary. In its most general form, Lan and DeMets' method does not rely on normal approximations, because we are able to replace $B(\cdot)$ by a continuous stochastic process; but the boundaries b_i may be more difficult to compute.

To avoid bias due to the experimenter, the choice of the ESF $\alpha^*(\cdot)$ must be made before the trial begins. How is this choice made? Some examples are given below, where α is the Type I error rate. The first one is

$$\alpha^*_{OBF}(t) = 2 - 2\Phi\left(\frac{q_{z_{\alpha/2}}}{\sqrt{t}}\right). \tag{1.4.7}$$

This ESF leads to intermediate significance levels which are fairly close to those of the method of O'Brien and Fleming (1979), for uniform analysis patterns, when the increase of information is proportional to time. A procedure that uses this ESF is very conservative at the beginning of the experiment. It is particularly suitable for studies on long term effects. Another ESF is

$$\alpha^*_P(t) = \alpha\ \log(1 + (e-1)t). \tag{1.4.8}$$

This function is similar to the procedure of Pocock (1977), when the number of observations between the tests remains constant. Roughly speaking, $\alpha^*_P(\cdot)$ makes for frequent early stopping, but also loss of power. Finally, Lan and DeMets (1983) propose

$$\alpha^*_{LDM}(t) = \alpha t^s, \quad \text{with } s > 0. \tag{1.4.9}$$

For example, if $s = 1$, the procedure is more conservative than (1.4.8), but less conservative than (1.4.7).

Parameter t should represent the information accumulated throughout the experiment rather than calendar time, represented by t_{cal}. These two notions are addressed by Lan and DeMets (1989). Let us consider an experiment which will last $T_{cal} = 5$ years, and involves a thousand patients. At the end of the second year, measures have been taken on five hundred patients. We have $t_{cal} = 0.4$, whereas $t = 0.5$.

Generally, there exists a monotone transformation g, such that $t = g(t_{cal})$, which allows us to go from t_{cal} to t. If g is known,

$$\alpha^*(t) = \alpha^* \left[g(t_{cal}) \right].$$

If the entry of patients into the experiment is proportional to time, then we have simply $g(t_{cal}) = t_{cal}/T_{cal}$. If the maximum duration is fixed, we may replace the unknown function g by an estimator $\hat{g}$, which is often $\hat{t} = \hat{g}(t_{cal}) = t_{cal}/T_{cal}$. Finally, if the maximum number of patients is fixed, then t is simply the ratio of the number of patients involved in the experiment at time t_{cal} to the maximum number of patients expected.

When t is estimated, and we are at step k, we consider

$$Z(t) = \frac{B(t)}{\sqrt{t}},$$

and we solve

$$\text{pr}\left(|Z(t_j)| \geq c(\hat{t}_j) \text{ for at least one } j \in \{1, \ldots, k\} \right) = \alpha^*(\hat{t}_j).$$

We simply note that

$$\begin{aligned} &\text{pr}\left(|Z(t_j)| \geq c(\hat{t}_j) \text{ for at least one } j \in \{1, \ldots, k\} \right) \\ = \ &\text{pr}\left(|Z(t_j/t_k)| \geq c(\hat{t}_j) \text{ for at least one } j \in \{1, \ldots, k\} \right); \end{aligned}$$

thus the ratio t_j/t_k is known even if t_j and t_k are unknown, because it is enough to establish the ratio of the number of patients involved in the experiment for steps j and k. Hence we are able to evaluate $\alpha^*(\hat{t}_j)$, and therefore $c(\hat{t}_j)$, $j = 1, \ldots, k$. The simulations made by Lan and DeMets (1989) show that underestimating t leads to a slight increase of power, but does not prolong the experiment.

Falissard and Lellouch's procedure

There is sometimes a contradiction as regards the final decision between a sequential procedure and a "classical" procedure. The objective of Falissard and Lellouch (1992) is to remedy this contradiction.

Their idea relies on the rejection of the null hypothesis if, and only if, r consecutive tests have rejected H_0.

Let $S_i = \sum_{j=1}^{i} \bar{d}_i$, where $\bar{d}_i$ is defined by (1.4.1). The null hypothesis H_0 is rejected if, and only if :

$$\exists i \in \{1, \ldots, N - r + 1\} \ : \ |S_{i+l}| > q_{z_{\alpha'/2}} \sqrt{i + l}, \quad l = 0, \ldots, r - 1. \tag{1.4.10}$$

The probability $\text{pr}_{N,r,\alpha}$ that we obtain r consecutive rejections under H_0, must be as close as possible to the target level α. The value α' used in (1.4.10) is computed using an iterative method of linear interpolation.

Whitehead's restricted procedure

Whitehead (1997) uses the statistics Z_i^* and V_i^*, namely the efficient score for δ (under Assumptions 1.4.1 with $n_A = n_B$), and the Fisher information about δ contained in Z_i^*. The procedure stops whether lower or upper lines are reached, that is if

- $Z_i^* = a + dV_i^*, \quad V_i^* \leq L$, or
- $Z_i^* = -a - dV_i^*, \quad V_i^* \leq L,$

where L is chosen so that $L \geq V_{fix}^*$, i.e. the information required for a non-sequential test. If $d = 0$, the procedure is O'Brien and Fleming's procedure. However, overshoots due to discrete monitoring should be considered. The famous Christmas tree adjustment (Whitehead, 1997) leads to the following boundaries:

- $Z_i^* = a + dV_i^* - 0.583\sqrt{V_i^* - V_{i-1}^*}, \quad V_i^* \leq L$, and
- $Z_i^* = -a - dV_i^* + 0.583\sqrt{V_i^* - V_{i-1}^*}, \quad V_i^* \leq L.$

Conclusion

Of course, none of the procedures described above is the best, if we consider simultaneously power, ASN, duration, control of the error rate α, or strength in the face of deviations. Indeed, none of them is able to eliminate every contradiction: the same data may lead to different decisions. Even the procedure of Falissard and Lellouch (1992) could lead to different results, if we use different values of N and r.

The methods we have described are well known. There exist many others; for example, Peto, Pike, Armitage, Breslow, Cox, Howard, Mantel, McPherson, Peto, and Smith (1976) propose the use of very low significance levels for every intermediate step, i.e. to concentrate the Type I error rate on the final test.

Table 1.1 illustrates several methods that have been described, with uniform analysis pattern. In the framework of Assumptions 1.4.1, we consider a procedure made of five equally-spaced inspection times, $t = 0.2, 0.4, 0.6, 0.8$, and 1, that is, $N = 5$. The boundaries are displayed for standardised statistics. There is a decreasing level of difficulty for early trial termination when we go from O'Brien and Fleming to Pocock's procedure. We can also check that α_{OBF}^*, and α_P^*, lead to good approximation of O'Brien and Fleming, and Pocock's procedure. Optimum boundaries (Geller and Pocock, 1987) are computed when the power γ equals 0.5, 0.75, or 0.8; these boundaries are optimum in the sense that they minimize the ASN for a given value of γ. Another interesting issue that can be observed in Table 1.1 is how α_{LDM}^* with $s = 1.5$ leads to almost optimal boundaries for large value of γ.

t	0.2	0.4	0.6	0.8	1.0
α_P^*	2.44	2.43	2.41	2.40	2.39
Pocock	2.41	2.41	2.41	2.41	2.41
α_{LDM}^* with $s = 1$	2.58	2.49	2.41	2.34	2.28
α_{LDM}^* with $s = 1.5$	2.85	2.59	2.43	2.29	2.18
α_{LDM}^* with $s = 2$	3.09	2.72	2.47	2.28	2.11
Peto	3.29	3.29	3.29	3.29	1.97
α_{OBF}^*	4.38	3.36	2.68	2.29	2.03
O'Brien and Fleming	4.56	3.23	2.63	2.28	2.04
Restricted procedure	4.43	3.13	2.55	2.21	1.98
Optimal for $\gamma = 0.5$	3.66	2.88	2.57	2.37	2.04
Optimal for $\gamma = 0.75$	2.99	2.54	2.41	2.35	2.16
Optimal for $\gamma = 0.8$	2.86	2.48	2.39	2.36	2.20

Table 1.1: *Boundary values for $Z(t)$, that is the standardized statistic for the difference of means (see Assumptions 1.4.1). The global error rate is $\alpha = 0.05$, and there are five equally-spaced inspection times.*

We often use in our work the ESFs α_{OBF}^*, α_P^*, and α_{LDM}^* defined by (1.4.7), (1.4.8), and (1.4.9), because of their flexibility. In addition, they cover most practical applications of the error spending function method to group sequential designs. The restricted procedure has not been taken into account, since Stallard and Facey (1996) advise against it, in a situation we have often dealt with; that is, monitoring with O'Brien and Fleming boundaries and a bounded number of intermediate analyses.

1.5 General bibliography

The body of literature relevant to the medical field is extremely vast, with about two million new articles published every year. Among the works designed for practitioners and oriented towards statistics, the book by Wassertheil-Smoller (1995) is an excellent introduction to biostatistics and to epidemiology. Senn wrote two general books respectively about cross-over trials (1993), and statistical issues in drug development (1997). Friedman et al. (1985) and Meinert and Tonascia (1986) wrote some textbooks specifically on design of large clinical trials. Pocock (1993) describes the statistical and ethical aspect in the monitoring of clinical studies. He is also the author of a reflection on the life of a statistician in the medical academic community (1995) which is very instructive.

The main theoretical developments in sequential analysis may be found in the books by Wald (1947), Ghosh (1970), Siegmund (1985), and Wetherill and Glazebrook (1986).

Finally, the volume edited by Ghosh and Sen (1991), to which numerous authors contributed, is a good reference, with a complete bibliography.

The application of sequential methods to clinical trials is presented by DeMets and Lan (1984), and by Whitehead (1997), who includes important mathematical developments. The procedures presented are directly usable, thanks to the PEST3 program (Brunier and Whitehead, 1994). The textbook edited by Peace (1992) gives many examples of applications of sequential statistics on real data. Instructions and guidelines to be adopted for monitoring intermediate analyses are given by Geller and Pocock (1987).

Beyond all doubt, one of the best references about longitudinal data is the book by Diggle, Liang, and Zeger (1994). The program OSWALD (Smith, Robertson, and Diggle, 1996) makes for easy analysis of longitudinal data.

Evaluating sample sizes is not the main purpose of our work. A good review of the problem in the field of clinical trial is found in Lachin (1981). We shall also mention Kim and DeMets (1992) as a good reference for the influence of a plan for sequential analysis on sample size.

Finally, Bayesian methods for clinical trials are presented by Berry (1985, 1987), and practical aspects are discussed by Hughes (1993a).

1.6 Chapter outline

In 1991, Ghosh and Basu (Ghosh and Sen, 1991) pointed out that no group sequential works had been done hitherto in comparing more than two treatments. Since then, theory in this field has improved, and many new procedures have been developed. In Chapter 2, we propose new procedures for multiple comparisons, in the fields of the linear mixed effects models. Their properties are discussed, and so are the parameter estimates, and the connection with existing methods.

In Chapter 3, we present the main results of various simulation studies. The choice of ESF and the small-sample case are dealt with. We also briefly discuss the confidence interval approach.

An introductory section about sequential allocation rules is found at the beginning of Chapter 4. Then, we characterize an asymptotically optimal allocation rule for the two-treatment case. The main features of this procedure are illustrated by simulation studies. Finally a generalization is proposed as a conjecture.

Chapter 2

Sequential Multiple Comparison

Whereas Chapter 1 discusses sequential methods for comparing two treatments, this chapter presents sequential procedures for comparison of multiple treatments. In Section 2.1, we see that the step from a two-treatment comparison towards three or more treatments is actually more difficult to make than it may seem at first sight.

In Section 2.2, we make a brief presentation of some existing group sequential procedures used for multiple comparison.

New general sequential procedures for comparing several treatments to a control are presented in Section 2.3. Their main advantages are strong control of Type I error, flexibility and wide range of use. Their properties are examined in Section 2.4, where we propose a method of computing relevant p-values.

Parameter estimation for model (1.3.2) is finally discussed in Section 2.5. We devote particular attention to the situation in which scale parameters are unknown.

2.1 Multiple hypothesis testing

Studying the results of a clinical experiment involving three or more treatments is not as straightforward as might be thought. There is a large number of possible outcomes, regarding subsets of hypotheses to be simultaneously tested.

Several questions must be raised in order to lay down a methodology that answers the experimenter's objectives. The planning of a clinical trial is a complex issue, which can be successfully concluded only if one has a very precise objective in mind (Senn, 1997). This remark could be considered obvious, but is not so easy to implement in practice. Trying to answer several questions by performing a single experiment may bring about a design which answers none effectively.

Let us consider an experiment with a placebo and three treatments, whose effects are μ_1, μ_2, μ_3 and μ_4. Here are some examples of null hypotheses, and the exact questions they are expected to answer:

- suppose we want to compare increasing doses of a treatment with a placebo. Three hypotheses may be of primary concern: the difference between placebo and treatment groups, i.e.

$$H_{01} : \frac{\mu_2 + \mu_3 + \mu_4}{3} = \mu_1,$$

and the dose effects,

$$H_{02} : \mu_2 \geq \mu_3,$$
$$H_{03} : \mu_3 \geq \mu_4;$$

- in a dose finding setup, we may ask whether a dose increase goes along with an increase in efficacy. In that case the null hypotheses would be

$$\begin{aligned} H_{01} &: \mu_1 = \mu_2, \\ H_{02} &: \mu_2 = \mu_3, \\ H_{03} &: \mu_3 = \mu_4; \end{aligned} \tag{2.1.1}$$

- no treatment differs from the control, i.e.

$$\begin{aligned} H_{01} &: \mu_1 = \mu_2, \\ H_{02} &: \mu_1 = \mu_3, \\ H_{03} &: \mu_1 = \mu_4; \text{ and} \end{aligned} \tag{2.1.2}$$

- equivalence of all treatments:

$$\begin{aligned} H_{01} &: \mu_1 = \frac{\mu_2 + \mu_3 + \mu_4}{3}, \\ H_{02} &: \mu_2 = \frac{\mu_1 + \mu_3 + \mu_4}{3}, \\ H_{03} &: \mu_3 = \frac{\mu_1 + \mu_2 + \mu_4}{3}, \\ H_{04} &: \mu_4 = \frac{\mu_1 + \mu_2 + \mu_3}{3}. \end{aligned} \tag{2.1.3}$$

In the last situation, the fourth hypothesis, H_{04}, should be omitted, since it is simply a linear combination of H_{01}, H_{02} and H_{03}. The set (2.1.3) is equivalent to (2.1.2). However, it may be more informative if one wants to focus on every simple hypothesis H_{0i}, because each H_{0i} deals with the entire set of treatments. Moreover, the corresponding statistic has a lower variance than one of a pairwise comparison, at least in the standard situation of independent observations, with equal variance in each group.

More details about multiple testing are given by Bauer (1991), who makes a rather complete overview in sequential and non-sequential situations. He also proposes (Bauer and Köhne, 1994) a methodology which allows the experimenter to make appropriate adjustments, or to change the study protocol.

2.1.1 Hypothesis statement

Let T be the number of treatments. Parameters of interest are called μ_i, $i \in \{1, \ldots, T\}$. We always consider location parameters.

Definition 2.1.1 *If pairwise comparisons are made, we use the notation:*

- H_{0ij} *is the hypothesis of equivalence of treatments* i *and* j, $i \in \{1, \ldots, T\}$, $i \neq j$, *that is*

$$H_{0ij} : \mu_i = \mu_j;$$

- $\mathcal{J}$ *is the subset of all hypotheses* H_{0ij} *tested during the trial; for example,*
  - *comparison of* $T - 1$ *treatments to a control:* $\mathcal{J}_1 = \{H_{012}, H_{013}, \ldots, H_{01T}\}$,
  - *all pairwise comparisons:* $\mathcal{J}_2 = \{H_{0ij},\ i, j \in \{1, \ldots, T\}, i \neq j\}$.

If all treatments are compared to all others, we may use the hypotheses given by the following definition.

Definition 2.1.2 *If hypothesis* H_{0i}, $i \in \{1, \ldots, T\}$ *corresponds to*

$$\mu_i = \frac{\mu_1 + \cdots + \mu_{i-1} + \mu_{i+1} + \cdots + \mu_T}{T - 1},$$

then equivalence of all treatments is tested with

$$\bigcap_{i=1}^{T} H_{0i}.$$

2.1.2 Control of error rates

When we test several hypotheses simultaneously, the event of making a Type I error is more complex than in a simple case. In fact, it depends on whether the whole set, or each single hypothesis, is considered first. Let $\mathcal{H}$, be a set of hypotheses to be tested, and $\mathcal{H}^* \subseteq \mathcal{H}$, the subset of the true null hypotheses in $\mathcal{H}$.

Definition 2.1.3 *A procedure controls the overall level of significance α, i.e. the Family-wise Error Rate (FWE), in a weak sense, if the probability of rejection of at least one hypothesis in $\mathcal{H}^*$ is lower than α, when all the hypotheses in $\mathcal{H}$ are true; that is*

$$\mathrm{pr}_{\mathcal{H}=\mathcal{H}^*}(\exists H \in \mathcal{H}^* : H \text{ is rejected }) \leq \alpha.$$

Definition 2.1.4 *A procedure controls strongly the overall level of significance α if the probability of rejection of at least one hypothesis in $\mathcal{H}^*$ is lower than α, under all circumstances; that is*

$$\sup_{\mathcal{H}^* \subseteq \mathcal{H}} \mathrm{pr}\,(\exists H \in \mathcal{H}^* : H \text{ is rejected }) \leq \alpha.$$

Consideration given to Type I error should not overshadow power, which may often be the main concern of the experimenters. Many large trials have surprisingly little concern for power because they are in fact often overpowered.

Generally speaking, multiple comparisons lead to high-order cumbersome integrations, especially when these comparisons are repeated over time. A simpler approach consists in dividing the overall level of significance by the number of hypotheses to be tested. Then, every single hypothesis can be tested separately. However, this method, usually called Bonferroni's method, is rather conservative. Actually, it could be used if correlations between statistics do not exceed 0.5 (Pocock, Geller, and Tsiatis, 1987), and the power of such a procedure is satisfactory only if one simple hypothesis is false. Unfortunately, this is rarely the case in practice.

Bauer (1991) introduces a more subtle Bonferroni-type approximation, which consists in ignoring the first steps of the sequential procedure. There are other interesting refinements of the Bonferroni method (Benjamini and Hochberg, 1995).

Another alternative consists of using a univariate test that concentrates all the statistics connected with the simple hypotheses in one statistic. This method may be more powerful, but is not always appropriate. It is sometimes a bone of contention between experimenters, since it could bring together very disparate elements.

To conclude, there are many approaches regarding multiple testing. Most of the methods presented in this chapter deal with the comparison of several treatments to a single control, that is, the hypotheses defined in (2.1.2). However, they could be easily used for other sets of hypotheses like (2.1.1) or (2.1.3).

2.2 Group sequential testing in clinical trial

Three group sequential testing procedures for comparing several treatments are presented in this section, viz. comparison with a control, equivalence of T treatments, and all pairwise comparisons. The results presented in this section can be essentially ascribed to Follmann, Proschan, and Geller (1994). We shall also mention Hughes (1993b) who deals with more practical details. But his objectives are slightly different since the controlled error is different too. This section concludes with several remarks about the features of such procedures, and about the use of Bonferroni's method.

2.2.1 Comparison to a control

Let m be the number of patients in each group at time t, and M, the maximum value of m. This section deals with the test of $H_0 : H_{012} \cap \cdots \cap H_{01T}$. The hypotheses $H_{012}, \ldots, H_{01T}$, are sequentially tested, and every hypothesis $H_{01j} : \mu_1 = \mu_j$, $j = 2, \ldots, T$, has a corresponding statistic Z_{1j}, where

$$Z_{1j} = \frac{\hat{\mu}_j - \hat{\mu}_1}{\sqrt{\frac{\sigma_j^2+\sigma_1^2}{m}}}, \quad j = 2, \ldots, T,$$

and σ_i^2 is the variance of any observation taken on treatment i, $i = 1, \ldots, T$.

Experiment process

For $t \leq t'$, covariances are given by

$$\operatorname{cov}(Z_{1j}\{t\}, Z_{1k}\{t'\}) = \begin{cases} \frac{\sigma_1^2}{\sqrt{\sigma_1^2+\sigma_j^2}\sqrt{\sigma_1^2+\sigma_k^2}}\sqrt{\frac{t}{t'}}, & j \neq k, \\ \sqrt{t/t'}, & j = k. \end{cases}$$

Scale parameters, when unknown, are replaced by their usual estimators.

Procedure 2.2.1 *At time t, which corresponds to step number k, (usually $t = k/K = m/M$), the hypothesis H_{01j}, $j = 2, \ldots, T$ is rejected if*

$$|Z_{1j}(t)| > \mathrm{b}_P(T, K, \alpha) \text{ for Pocock's procedure,}$$
$$|Z_{1j}(t)| > t^{-1/2}\mathrm{b}_{OBF}(T, K, \alpha) \text{ for O'Brien and Fleming's procedure,}$$

where $\mathrm{b}_P(T, K, \alpha)$ and $\mathrm{b}_{OBF}(T, N, \alpha)$ are tabulated (Follmann et al., 1994) if $\sigma_1^2 = \cdots = \sigma_T^2$. Otherwise, simulation is suggested. Let $c = \mathrm{b}_P(T, K, \alpha)$, or $c = \mathrm{b}_{OBF}(T, N, \alpha)$ depending on the chosen method,

- *if* $\exists\, j : Z_{1j}(t) > c$, *the experiment is stopped;*
- *if* $\exists\, j : Z_{1j}(t) < -c$, *the treatment* j *is rejected and the trial is carried on, without treatment* j. *In this situation, the value of* T *is changed in* $\mathrm{b}_P(\cdot, K, \alpha)$, *or* $\mathrm{b}_{OBF}(\cdot, K, \alpha)$, *and is replaced by the number of the remaining treatments.*

To sum up, the trial is stopped, if the final interim analysis is reached, if at least one treatment has been shown to be superior to the control, or if all treatments have been rejected as inferior to the control.

Follmann et al. (1994) prove that Procedure 2.2.1 strongly controls the overall level of significance α.

2.2.2 Pairwise comparisons

All the hypotheses H_{0ij} are tested, where $i = 1, \ldots, T$, $j = 1, \ldots, T$ and $i \neq j$. For any hypothesis H_{0ij}, we use the statistic

$$Z_{ij} = \frac{\hat{\mu}_j - \hat{\mu}_i}{\sqrt{\frac{\sigma_j^2 + \sigma_i^2}{m}}}. \tag{2.2.1}$$

Comments made in the previous section about scale parameters also hold here. The values $\mathrm{b}_P(T, K, \alpha)$ and $\mathrm{b}_{OB}(T, K, \alpha)$ are tabulated for $\sigma_1^2 = \cdots = \sigma_T^2$ (Follmann et al., 1994). Rejection of H_{0ij} usually brings about withdrawal of the identified inferior treatment. Then any other simple hypothesis connected with the rejected treatment is cancelled. Finally, the experiment is carried on; we just change T in $\mathrm{pr}_P(\cdot, K, \alpha)$, or $\mathrm{pr}_{OBF}(\cdot, K, \alpha)$, and replace it by the number of the remaining treatments. The overall level of significance α, is always controlled in a weak sense, and strongly controlled when $T = 3$.

2.2.3 Equivalence of T treatments

Let us consider the following situation: the overall null hypothesis $H_{02} \cap H_{03} \cap \cdots \cap H_{0T}$ (Definition 2.1.2) is sequentially tested, until rejection, without exceeding step K, where K is the maximal number of interim analyses. We use the statistics $Z_{\xi_2}(t), \ldots, Z_{\xi_T}(t)$, where

$$Z_{\xi_i}(t) = \frac{1}{\sqrt{V_i}} \left(\hat{\mu}_i - \frac{\hat{\mu}_1 + \cdots + \hat{\mu}_{i-1} + \hat{\mu}_{i+1} + \cdots + \hat{\mu_T}}{T-1} \right),$$

and

$$V_i = \frac{1}{m} \left[\sigma_i^2 + \frac{\sigma_1^2 + \cdots + \sigma_{i-1}^2 + \sigma_{i+1}^2 + \cdots + \sigma_T^2}{T-1} \right].$$

We define

$$\hat{\Sigma}_T(t),$$

which is the estimate of the variance matrix of $[Z_{\xi_2}(t), \ldots, Z_{\xi_T}(t)]$. The equivalence of all treatments is tested using the statistic

$$\chi^2(t) = [Z_{\xi_2}(t), \ldots, Z_{\xi_T}(t)]\, \hat{\Sigma}_T(t)^{-1}\, [Z_{\xi_2}(t), \ldots, Z_{\xi_T}(t)]'. \tag{2.2.2}$$

The distribution of $\chi^2(t)$ is easily simulated, since $\chi^2(t)$ is asymptotically distributed as χ^2_{T-1}. This result is obviously exact in the normal case. Some tables supplied by Jennison and Turnbull (1991) give values for $b_{OBF}(T, K, \alpha)$ and $b_P(T, K, \alpha)$. In the situation where $\sigma_1^2 = \cdots = \sigma_T^2 = \sigma^2$, (2.2.2) is written

$$\begin{aligned} F(t) &= \frac{m}{\hat{\sigma}^2} \sum_{i=1}^{T} (\hat{\mu}_i - \hat{\mu})^2, \\ \text{where } \hat{\mu} &= \frac{\hat{\mu_1} + \cdots + \hat{\mu_T}}{T}, \end{aligned}$$

and $F(t)/(T-1)$ is distributed like a $F_{T-1,T(m-1)}$ random variable. The withdrawal of treatment during the trial is not allowed. Whereas non-rejection of the null hypothesis indicates the equivalence of all treatments , it is legitimate to ask which of the treatments are different, in case of rejection.

Senn (1997) argues that it is analogous to an opinion-poll interviewer stopping an individual in the street, giving a list of several government policies and asking him or her, "do you disagree with any of these?" The answer "yes" would not be very informative. So, we may stress that both of the procedures previously introduced (comparison with a control and pairwise comparison) lead to more precise inference than when we use (2.2.2).

2.2.4 Further comments

The disadvantages of the procedures that we have presented may be summarized as follows:

- first, if variances are not identical for each treatment, the existing tables are no longer appropriate, unless one is willing to proceed as if they were. Follmann et al. (1994) suggest simulations. This might be considered non trivial, given the sequential nature of the context. We should point out that this situation is likely to be faced in the case of longitudinal data; and

- secondly, if for any unexpected reason, the subset of hypotheses to be tested, or more generally the contrasts' matrix should be modified, pre-established tables cannot be used.

Bonferroni's method is a simpler, but less powerful alternative to the procedures presented in Sections 2.2.1 and 2.2.2. Let card($\mathcal{H}$) be the number of hypotheses H_{0ij} to be tested. The overall significance level α is simply divided by card($\mathcal{H}$), and critical values are such that

$$\text{pr}\left(\mathcal{R}_{ij}(t_1) \cup \mathcal{R}_{ij}(t_2) \cup \cdots \cup \mathcal{R}_{ij}(t_r)\right) = \frac{\alpha^*(t_r)}{\text{card}(\mathcal{H})},$$

where $t_1, \ldots, t_r$ are the times of interim analyses, and $\mathcal{R}_{ij}(t)$ stands for the rejection of H_{0ij} at time t. The function $\alpha^*(\cdot)$ is an ESF (Error Spending Function).

The greater the number of treatments, the more conservative is Bonferroni's method when compared to tabulated values from Follmann et al. (1994). However, we should add that different variances between treatments can be dealt with more easily when Bonferroni critical values are used.

Nevertheless, this rather simple and conservative approach should not be used rashly, especially under circumstances when scale parameters are not equal among treatments. Let us consider a single step trial, in which several treatments are compared to a control at level 5% (Follmann et al., 1994). If the control variance was high, the correlations between all statistics used would be close to one; so the critical value for any comparison would be close to 1.96, whereas Bonferroni method would lead to far higher values, e.g. 2.50 in the four-treatment case!

2.3 A sequential procedure for multiple comparison with a control

Taking into account the remarks of Section 2.2.4, we now propose some general group sequential procedures for comparing several treatments with a control. This work is a generalization of Lee and DeMets' method (1991) for comparing two treatments. Apart from its flexibility, it allows the experimenter to withdraw any apparently different treatment during the trial, while continuing to monitor the remaining treatments.

We have chosen the case of multiple comparison with a control, but the algorithm that we have written is also able to deal with any type of contrast matrix, provided the associated variance matrix is positive definite.

The framework used in the previous sections remains valid, but slight adaptations are necessary, due to the longitudinal nature of data. We consider the linear mixed effects

model (1.3.5), and suppose that the main objective is to infer about fixed effects β_1^h, $h = 1, \ldots, T$, that is, trends in treatment effects. The covariance structure between the standardized statistics (2.2.1), which is demonstrated by Follmann et al. (1994),

$$\mathrm{cov}(Z_{ij}\{t\}, Z_{kl}\{t'\}) = \begin{cases} 0, & \{i,j\} \cap \{k,l\}, \\ \frac{\sigma_i^2}{\sqrt{\sigma_i^2+\sigma_j^2}\sqrt{\sigma_i^2+\sigma_l^2}}\sqrt{\frac{t}{t'}}, & i = k,\ j \neq l, \\ \sqrt{t/t'}, & i = k,\ j = l, \end{cases}$$

for $t \leq t'$, and $i, j, k, l \in \{1, \ldots, T\}$, still holds when we deal with longitudinal data (Reboussin, 1995).

2.3.1 Setup

The control is labelled 1. At every interim analysis k, we compute $\hat{\beta}_1^h(k)$, $h = 1, \ldots, T$. We test the overall null hypothesis

$$H_0 : \{\beta_1^1 = \beta_1^2\} \cap \cdots \cap \{\beta_1^1 = \beta_1^T\}, \tag{2.3.1}$$

versus

$$H_1 : \exists\, h \in \{2, \ldots, T\} : \beta_1^h \neq \beta_1^1. \tag{2.3.2}$$

The statistics used are

$$S^{(h)}(k) = \hat{\beta}_1^1(k) - \hat{\beta}_1^h(k), \quad h = 2, \ldots, T, \tag{2.3.3}$$

where $\hat{\beta}_1^h(k)$ is the ML (Maximum Likelihood) estimator of trend for treatment h, at the kth interim analysis. Let $\mathcal{H} = \{H_{01h} : \beta_1^1 = \beta_1^h, h = 2, \ldots, T\}$, the set of hypotheses to be tested. During the trial, it may occur that results from other experiments make one of the treatments seem less attractive, due to an apparent inferiority, high toxicity, other undesirable effects, or unexpected financial cost, etc. Under such conditions, the experimenter may want to test this treatment more often, in order to drop it from the trial earlier. Also the poor number of new measures on one of the treatments may lead the researcher to ignore temporarily one of the component hypotheses of $\mathcal{H}$. To handle such situations, our procedure allows the experimenter to choose the subset of hypotheses $\mathcal{H}(k) \subseteq \mathcal{H}$ to be tested at any kth interim analysis.

Since we are not willing to fix the number of interim analyses, we use an ESF, which indicates the proportion of the experiment-wise error α to be spent at any step k. This proportion is usually called exit probability $\pi(k)$. For example, approximations to the method of Pocock (1.4.8), or O'Brien and Fleming (1.4.7) are used.

Because sequential multiple comparison procedures could have many different issues, we introduce specific notation, in order to make the procedures description easier.

Definition 2.3.1 *We define the following subsets and value:*

1. $\delta_0(k) : \{h : H_{01h} : \beta_1^1 = \beta_1^h$ *is not rejected at step* l, $l = 1, \ldots, k-1\}$, *which is the subset of hypotheses* H_{01h} *not rejected by the* k*th step.*
2. $\mathcal{H}(k) : \{h \in \delta_0(k) : H_{01h} : \beta_1^1 = \beta_1^h$ *is tested at step* $k\}$, *is the set of hypotheses* H_{01h} *tested at step* k.
3. $\delta_1(k) : \{h \in \{2, \ldots, T\} : H_{01h} : \beta_1^1 = \beta_1^h$ *is rejected at step* $k\}$, *is the set of hypotheses rejected at the* k*th step.*
4. $\eta = \sum_k \mathbb{I}_{\{\text{card}(\delta_1(k))\neq 0\}}$ *is the total number of significant interim analyses.*

At step k, we test differences between trends by using

$$S(k) = \left[S^{(i_1)}(k), S^{(i_2)}(k), \ldots, S^{(i_{\text{card}(\mathcal{H}\{k\})})}(k)\right]', \tag{2.3.4}$$

where $\{i_1, \ldots, i_{\text{card}(\mathcal{H}\{k\})}\} \subseteq \{2, \ldots, T\}$ is the subset of treatment numbers which are tested at step k, apart from the control.

2.3.2 An alternative to Fisher's Least Significant Difference

When the objective is solely to test H_0, we stop and reject H_0 at the first significant interim analysis. In the same time, we want to be able to identify the treatments h significantly different from the control, such that H_{01h} is rejected. In the standard situation of independent normal observations, we fulfil this purpose by using a procedure referred to in the literature as the Least Significant Difference test. This test consists of performing multiple t tests each at level α only if the preliminary F test is significant at level α. In our framework, the preliminary F test is replaced by a group sequential F test.

Fisher's Least Significant Difference does not control the FWE strongly. We want to propose an alternative which does. We use the statistics $S^{*(h)}(k)$, the standardized values of $S^{(h)}(k)$, $k = 1, \ldots, K$, $h = 2, \ldots, T$, that is,

$$S^{*(h)}(k) = \frac{\hat{\beta}_1^1(k) - \hat{\beta}_1^h(k)}{\sqrt{\text{var}(\hat{\beta}_1^1\{k\}) + \text{var}(\hat{\beta}_1^h\{k\})}}. \tag{2.3.5}$$

The null hypothesis is rejected at step k, if h exists such that $|S^{*(h)}(k)| > c(k)$, where $c(k)$ is an appropriate critical value. The event $A(k)$ stands for non-rejection of H_0 for the $k-1$ first interim tests, and rejection of H_0 at step k, that is

$$A(k) = \bigcap_{l=1}^{k-1} \bigcap_{h\in\mathcal{H}(l)} \{|S^{*(h)}(l)| \leq c(l)\} \bigcap \{\exists\, h \in \mathcal{H}(k) : |S^{*(h)}(k)| > c(k)\}. \tag{2.3.6}$$

Let $\pi(k)$ be the intermediate significance level at step k, i.e.

$$\pi(k) = \mathrm{pr}_{H_0}[A(k)]; \tag{2.3.7}$$

since the sets $A(k)$ are disjoint, we have at time t_k, corresponding to step k,

$$\alpha(t_k) = \mathrm{pr}_{H_0}\left[\bigcup_{k=l}^{k} A(l)\right] = \sum_{l=1}^{k} \mathrm{pr}_{H_0}[A(l)] = \sum_{l=1}^{k} \pi(l).$$

Of course, $\pi(l)$ depends on the chosen ESF. Then, we find $c(k)$ by solving

$$\pi(k) = \int_{\bar{R}(1)} \cdots \int_{\bar{R}(k-1)} \int_{R(k)} f_{r(k)}\left[(x_1, \ldots, x_{r(k)}), 0, \Sigma_S\right] dx_1 \ldots dx_{r(k)},$$

where Σ_S is the variance matrix of all the statistics $S^{*(h)}(l)$, $l = 1, \ldots, k$, $h \in \mathcal{H}(l)$, used for testing,

$$\begin{aligned} \bar{R}(l) &= [-c(l), c(l)]^{\mathrm{card}(\mathcal{H}\{l\})}, \quad l = 1, \ldots, k-1, \\ R(k) &= \{(y_1, \ldots, y_{\mathrm{card}(\mathcal{H}\{k\})}) \in \mathbb{R}^{\mathrm{card}(\mathcal{H}\{k\})} : \exists\, y_i \notin [-c(k), c(k)], \\ &\quad i \in [1, \ldots, \mathrm{card}(\mathcal{H}\{k\})]\}, \end{aligned}$$

$f_{r(k)}(\cdot, 0, \Sigma_S)$ is the density of the multivariate normal distribution $\mathcal{N}_{r(k)}(0, \Sigma_S)$, and $r(k) = \sum_{l=1}^{k} \mathrm{card}(\mathcal{H}\{l\})$.

Thus, we have to compute the integral of a multivariate normal density over a hyper-rectangle of $\mathbb{R}^p$, where p is equal to the total number of simple comparisons made during the first k steps. The value of p may be rather high, but this is no deterrent. This assertion could be considered a rather unrealistic argument pro domo sua, if efficient algorithms were not available. We use the procedure MULNOR (Schervish, 1984), and the algorithm of Lohr (1992), which is usually far faster than MULNOR if p is high, but slower if p is low.

The procedure proposed below follows general principles of group sequential testing, and allows a full control of the experiment-wise error α. If k corresponds to $t = 1$, the equality $\sum_{l=1}^{k} \pi(l) = \alpha$ holds. If the experiment is not stopped at step $l-1$ on behalf of H_1, new measurements are taken until time t_l. Then statistics $S(l)$ — see (2.3.4) — and the critical value $c(l)$ are computed. If $|S^{*(h)}(l)| > c(l)$, $h \in \mathcal{H}(l)$, H_0, defined by (2.3.1), is rejected. Finally, H_0 is not rejected if and only if we obtain $|S^{*(h)}(k)| \leq c(k)$, $\forall h \in \mathcal{H}(k)$.

We should point out that as soon as $\mathcal{H}(k) \neq \mathcal{H}(k')$, $k \neq k'$, the choice of t for the ESF may be not straightforward. We suggest the following processes, which would nevertheless need further investigation. At step k, we can use

- the amount of information which corresponds to the remaining treatments, that is, $t = \sum_{h \in \delta_0(k)} \left[\text{var} \left(\hat{\beta}_1^h \{k\} \right) \right]^{-1} /$(targeted value of information);
- we can also take the same parameter t as above, but replace $c(k)$ by $c^h(k)$; for example, we choose a lower boundary for a statistic whose variance has greatly decreased since step $k-1$, and a higher boundary in an opposed situation; and
- finally Bonferroni's method might be used, with separate ESFs. This allows different parameters t, one for every comparison.

Procedure 2.3.1 *Alternative to Fisher's Least Significant Difference.*

0 Set
  1 the experiment-wise error rate α,
  2 a time limit,
  3 the error spending function $\alpha(\cdot)$.
1 $k \leftarrow 1$.
2 1 Collect measures without overshooting the time limit, and
  2 choose the subset $\mathcal{H}(k)$.
3 Compute
  1 $\hat{\sigma}^2(k)$ *and* $\hat{D}(k)$, *estimates of* σ^2 *and* D,
  2 statistics $S^{(h)}(k)$ *where* $h \in \mathcal{H}(k)$, *and*
  3 $\pi(k) = \alpha(t_k) - \alpha(t_{k-1})$.
4 Then, compute $c(k)$ *by solving* $\text{pr}_{H_0}(A\{k\}) = \pi(k)$.
5 1 For any $h \in \mathcal{H}(k)$ *reject* H_{01h} *if* $|S^{*(h)}(k)| > c(k)$;
  2 if time limit is reached or $\delta_0(k) = \emptyset$ *or* $\delta_1(k) \neq \emptyset$, *go to 6, else set* $k \leftarrow k+1$ *and go to 2.*
6 1 If $\exists h$: H_{01h} *is rejected then decide* H_1, *else decide* H_0;
  2 stop.

Basically, the procedure stops as soon as the overall null hypothesis H_0 is rejected.

2.3.3 Multiple comparison using conditional distributions

When the overall null hypothesis H_0 is rejected, we can also continue the experiment and compare each treatment to the control. Then, the trial is stopped until either every H_{01h} is rejected, or the planned end of the experiment has been reached. For example, when $H_0 : \{\beta_1^1 = \beta_1^2\} \cap \cdots \cap \{\beta_1^1 = \beta_1^T\}$, has been rejected, information about treatments which have not been declared different from the control may be well appreciated. For instance, a cheaper treatment or a treatment with fewer side effects could be at stake.

Definition 2.3.2 *Let*

1. τ_l *be the instant (information) corresponding to the* l*th significant test, if it exists, otherwise* $\tau_l = +\infty$*; and*

2. $\mathcal{F}(t)$ *the sigma-algebra generated by the set of statistics* $S^{(h)}(\cdot)$*, used until time* t*.*

We see that τ_l is a stopping time for the increasing family of sigma-algebras $\mathcal{F}(t)$, $t \in \{t_1, t_2, \dots\}$, where t_k is the time of step k.

To simplify the procedure when a significant step τ_l is reached, we propose using conditional distributions of statistics $S^{(h)}$, $h \in \delta_0(\tau)$, on $\mathcal{F}(\tau_l)$. We can distinguish two advantages of conditioning:

- As soon as a simple null hypothesis is rejected, we can reduce the order of the integral used to obtain critical values. Nevertheless, this argument is less relevant if the independent increment simplification can be used (Reboussin, 1995); and

- Proschan, Follmann, and Geller (1994) point out that unfortunately, the effect of dropping treatments is that recruitment to the remaining treatments per unit time will increase because roughly the same number of patients will be allocated to fewer treatments. They add that it is not clear what the effect of this assumption violation is on the joint distribution of the test statistics over study time. By conditioning, we start a new experiment, so to speak. Thus, the question above becomes irrelevant.

Let $Y(k)$ be the set of observations available until step k. We proceed as if scale parameters were known. It seems easier to compute the distribution of the statistics, conditionally on the observations taken until the last stopping time. However, Proposition 2.3.1 shows that, except under special circumstances — for example identical designs and equal numbers of patients assigned to each treatment — conditional expectation of $\hat{\beta}_1^l - \hat{\beta}_1^h$ depends on β^l and β^h. Therefore, we choose to use the distributions of $S^{(h)}(k)$ conditional on $\mathcal{F}(\tau_l)$, rather than on $Y(\tau_l)$, $k = l+1, \dots, K$.

Proposition 2.3.1 *Let* $k' > k$ *some positive integers. The expectancy of* $\hat{\beta}^h(k')$*, estimation of* β^h *at step* k'*, conditionally on* $Y(k)$*, equals*

$$\beta^h + \left[\sum_{i=1}^{m^h(k')} X_i^h(k')' W_i^h(k') X_i^h(k')\right]^{-1} \sum_{i=1}^{m^h(k)} X_i^h(k)' W_i^h(k) \left[Y_i^h(k) - X_i^h(k)\beta^h\right].$$

□ Proof:
Application of Theorem 2.1 of Seber (1984, pp. 18–19), which deals with conditional distributions of multi-normal random variables.
■

Because we condition after every significant step, we need to introduce new error rates.

Definition 2.3.3 *Let* $\alpha, \alpha_2, \ldots, \alpha_{T-1}$ *denote the following probabilities, which are constant by design:*

$$\begin{cases} \mathrm{pr}_{H0}(\tau_1 \leq 1) = \alpha, \\ \\ \mathrm{pr}_{H0}(\tau_2 \leq 1 \mid \mathcal{F}\{\tau_1\}) = \alpha_2, \\ \ldots \\ \mathrm{pr}_{H0}(\tau_{T-1} \leq 1 \mid \mathcal{F}\{\tau_{T-2}\}) = \alpha_{T-1}. \end{cases}$$

Remembering that η is the total number of significant interim analyses, we can deduce the following properties.

Proposition 2.3.2 *Using the notation of Definitions 2.3.1 and 2.3.3, we have*

$$\begin{cases} \mathrm{pr}_{H0}(\eta \geq 1) = \alpha, \\ \\ \mathrm{pr}_{H0}(\eta = l) = \alpha \left(\prod_{i=2}^{l} \alpha_i\right)(1 - \alpha_{l+1}), \quad \textit{for } l \leq T-2, \\ \\ \mathrm{pr}_{H0}(\eta = T-1) = \alpha \prod_{i=2}^{T-1} \alpha_i. \end{cases}$$

□ Proof: First, the FWE is given by $\mathrm{pr}_{H0}(\eta \geq 1) = 1 - \mathrm{pr}_{H0}(\eta = 0) = 1 - \mathrm{pr}_{H0}(\tau_1 > 1) = 1 - (1 - \alpha) = \alpha$. Moreover, we have $\mathrm{pr}_{H0}(\eta = 1) = \mathrm{pr}_{H0}(\tau_1 \leq 1, \tau_2 > 1) = \mathrm{pr}_{H0}(\tau_1 \leq 1) - \mathrm{pr}_{H0}(\tau_1 \leq 1, \tau_2 \leq 1) = \alpha - \mathrm{E}_{H_0}\left(\mathrm{E}\left[\mathbf{1}_{\{\tau_2 \leq 1\}} \mid \mathcal{F}(\tau_1)\right] \mathbf{1}_{\{\tau_1 \leq 1\}}\right) = \alpha - \alpha_2 \,\mathrm{pr}_{H0}[\tau_1 \leq 1] = \alpha(1 - \alpha_2)$. By induction, for $l = 1, \ldots, T-2$, we have

$$\begin{aligned} \mathrm{pr}_{H0}(\eta = l) &= \mathrm{pr}_{H0}(\tau_1 \leq \cdots \leq \tau_l < 1, \tau_{l+1} > 1) \\ &= \mathrm{pr}_{H0}(\tau_1 \leq \cdots \leq \tau_l < 1) - \mathrm{E}_{H_0}\left(\mathbf{1}_{\{\tau_{l+1} \leq 1 \mid \mathcal{F}(\tau_l)\}} \mathbf{1}_{\{\tau_1 \leq \cdots \leq \tau_l < 1\}}\right) \\ &= \mathrm{pr}_{H0}(\tau_1 < \cdots < \tau_l < 1)\,(1 - \alpha_{l+1}) \\ &= \alpha\,\alpha_2 \cdots \alpha_l\,(1 - \alpha_{l+1}). \end{aligned}$$

Finally, for $T-1$, $\mathrm{pr}_{H0}(\eta = T-1) = \mathrm{pr}_{H0}[\tau_1 < \tau_2 < \cdots < \tau_{T-1} < 1] = \alpha\,\alpha_2 \ldots \alpha_{T-1}$.
■

Let $\tilde{A}(k)$ be the event

$$\tilde{A}(k) = \bigcap_{l=k'}^{k-1} \bigcap_{h \in \mathcal{H}(l)} \left[\{|S^{*(h)}(l)| \leq c(l)\} \bigcap \{\exists\, h' \in \mathcal{H}(k) : |S^{*(h')}(k)| > c(k)\} \mid \mathcal{F}(\kappa\{k\})\right],$$

that is, a simple hypothesis is rejected at step k, given all the statistics $S^{*(\cdot)}(\cdot)$ until the latest significant step, and $k' = \kappa(k) + 1$, where $\kappa(k) = 0 \vee [\max(l = 1, \ldots, k : \delta_1(l) \neq \emptyset)]$. We are now able to propose the following procedure:

Procedure 2.3.2 *Multiple comparison using conditional distributions*

0 Set
  1 the experiment-wise error rate α,
  2 α_i *i=2,...,T-1, and*
  3 the error spending function $\alpha(\cdot)$.
1 $k \leftarrow 1$ *and* $i \leftarrow 1$.
2 1 Collect measures without exceeding the time limit;
  2 choose the time for testing, and the subset $\mathcal{H}(k)$.
3 Compute:
  1 $\hat{\sigma}^2(k)$ *and* $\hat{D}(k)$, *estimates of* σ^2 *and* D,
  2 the statistics $S^{(h)}(k)$ *for* $h \in \mathcal{H}(k)$,
  3 $\pi(k) = \alpha(t_k) - \alpha(t_{k-1})$.
4 Compute $c(k)$ *by solving* $\mathrm{pr}_{H_0}(A\{k\}) = \pi(k)$,
  or $\mathrm{pr}_{H_0}(\tilde{A}\{k\}) = \pi(k)$ *if* $i > 1$.
5 1 For any $h \in \mathcal{H}(k)$ *reject* H_{01h} *if* $|\tilde{S}^{*(h)}(k)| > c(k)$;
  2 if time limit is reached or $\delta_0(k) = \emptyset$ *go to 6,*
    else if $\delta_1(k) \neq \emptyset$ *then*
      stop measurement on treatment h, for $h \in \delta_1(k)$,
      set $\alpha \leftarrow \alpha_{i+1}$, *and then,* $\alpha(t_k) \leftarrow 0$;
      set $i \leftarrow i + 1$;
  3 set $k \leftarrow k + 1$ *and go to 2.*
6 1 If $\exists h$: H_{01h} *is rejected then decide* H_1 *else* H_0;
  2 stop.

To sum up, we drop any rejected treatment at a given step τ. Then, we continue the experiment with the remaining treatments, using conditional distribution on $\mathcal{F}(\tau)$ and a new FWE that is conditional on $\mathcal{F}(\tau)$. We may add that this conditional approach can be also applied to non-longitudinal data.

2.3.4 Generalization of Follmann, Proschan, and Geller's procedure

If one is not willing to use the conditional distributions on previous observed statistics, we propose a simple alternative to Procedure 2.3.2. This alternative could nevertheless be slower from a computational point of view. As soon as any treatment is rejected

at step k, the value $\pi(1) + \cdots + \pi(k)$ is updated, by removing from (2.3.6) every hypothesis connected to every treatment that has just been rejected. This is in fact a straightforward method derived from Follmann et al. (1994).

Properties of Procedures 2.3.1, 2.3.2, and 2.3.3 are discussed in the following section.

Procedure 2.3.3 *Generalization of Follmann, Proschan, and Geller's procedure*

0 Set
  1 the experiment-wise error rate α,
  2 the error spending function $\alpha(\cdot)$, *and*
  3 $k \leftarrow 1$.
1 1 Collect measures without overshooting the time limit;
  2 choose the time for testing, and the subset $\mathcal{H}(k)$.
2 Compute:
  1 $\hat{\sigma}^2(k)$ *and* $\hat{D}(k)$, *estimates of* σ^2 *and* D,
  2 the statistics $S^{(h)}(k)$ *with* $h \in \mathcal{H}(k)$, *and*
  3 $\pi(k) = \alpha(t_k) - \alpha(t_{k-1})$.
3 Compute $c(k)$ *by solving* $\mathrm{pr}_{H_0}(A\{k\}) = \pi(k)$.
4 1 For any $h \in \mathcal{H}(k)$ *reject* H_{01h} *if* $|\tilde{S}^{*(h)}(k)| > c(k)$;
  2 if time limit is reached or $\delta_0(k) = \emptyset$ *go to 6;*
    else if $\delta_1(k) \neq \emptyset$ *then*
      stop measurements on treatment h *for* $h \in \delta_1(k)$;
      update $A(l)$ *and* $\pi(l)$, *for* $l = 1, \ldots, k$;
      set $\alpha(k+1) = \alpha - \pi(1) - \cdots - \pi(k)$.
5 Set $k \leftarrow k+1$ *and go to 1.*
6 1 If $\exists h$: H_{01h} *is rejected then decide* H_1 *else decide* H_0;
  2 Stop.

Basically, we drop the rejected treatments at step k; then we update $\sum_{l=1}^{k} \pi(l)$ by removing every hypothesis connected to the dropped treatments, and we continue the experiment using the remaining FWE, that is the updated value of $\alpha - \sum_{l=1}^{k} \pi(l)$.

2.4 Properties

2.4.1 Control of error rates

Many trials must be able to demonstrate strict control of Type I error rate in order to convince regulatory agencies. It seems advisable that any error about a simple

hypotheses H_{01h} should be controlled under all configurations, that is, the FWE should be strongly controlled. In this section, we show that Procedures 2.3.1, 2.3.3, and under particular circumstances Procedure 2.3.2, control strongly the FWE.

Proposition 2.4.1 *The FWE of Procedures 2.3.1 and 2.3.3 is α, and it is strongly controlled.*

□ Proof:
Procedure 2.3.1 can be considered as a particular case of Procedure 2.3.3 when we stop the experiment at the first significant step. In addition, its proof is similar to that of Proposition 2.4.2 and as a consequence, omitted.
■

Proposition 2.4.2 *The FWE of Procedure 2.3.2 is α. If every step $\alpha \leftarrow \alpha_{i+1}$ is replaced by $\alpha \leftarrow \alpha - \alpha(\tau_i)$ in Procedure 2.3.2, then the FWE is strongly controlled.*

□ Proof: we shall first show that the procedure is well defined, that is, the relation $\pi(k+1) \leq \alpha(t_{k+1})$ must hold for any step k, where $\pi(k+1)$ is the interim significance level at step $k+1$, and $\alpha(\cdot)$ an error spending function. If $\delta_1(l) = \emptyset \ \forall l = 1, \ldots, k$, i.e. there has been no rejection up to step k,

$$\begin{aligned}
\alpha(t_{k+1}) &= \mathrm{pr}_{H_0}\left(\tau_1 \leq 1 \bigcap \tau_1 \leq t_k\right) \\
&= \mathrm{pr}_{H_0}\left(\tau_1 \leq 1 \bigcap \tau_1 = t_{k+1}\right) + \mathrm{pr}_{H_0}\left(\tau_1 \leq 1 \bigcap \tau_1 < t_{k+1}\right) \\
&= \mathrm{pr}_{H_0}\left(\tau_1 = t_{k+1}\right) + \mathrm{pr}_{H_0}\left(\tau_1 \leq 1 \bigcap \tau_1 < t_{k+1}\right) \\
&= \pi(k+1) + \mathrm{pr}_{H_0}\left(\tau_1 \leq 1 \bigcap \tau_1 < t_{k+1}\right) \\
&\geq \pi(k+1).
\end{aligned}$$

Now we suppose that there was a rejection at time τ_{l-1}, with $\tau_{l-1} \leq t_k$,

$$\begin{aligned}
\alpha(t_{k+1}) &= \mathrm{pr}_{H_0}\left(\tau_l \leq 1 \bigcap \tau_l \leq t_{k-1} \mid \mathcal{F}(\tau_l)\right) \\
&= \pi(k+1) + \mathrm{pr}_{H_0}\left(\tau_l \leq 1 \bigcap \tau_l \leq t_{k+1} \mid \mathcal{F}(\tau_l)\right) \\
&\geq \pi(k+1).
\end{aligned}$$

Finally we have

$$\mathrm{pr}_{H_0}\left[\bigcup_k A(k)\right] = \sum_k \mathrm{pr}_{H_0}\left[A(k)\right] = \sum_k \pi(k) \leq \alpha,$$

where $A(k)$ is defined by (2.3.6).

To show that the FWE is strongly controlled, we first suppose that the trial is stopped as soon as a simple hypothesis like H_{01h} is rejected. Let $\mathcal{J}$ be the set of all simple hypotheses which are tested during the experiment. We note that $\mathcal{J} = \mathcal{J}^c \cup \mathcal{J}^*$, where $\mathcal{J}^*$ is the subset of true hypotheses. In the same way, $\mathcal{H}(k) = \mathcal{H}^c(k) \cup \mathcal{H}^*(k)$ is the set of corresponding indices. For instance, if $\mathcal{J} = \{H_{012}\}$, then $\mathcal{H} = \{2\}$. By definition, strong control of FWE is

$$\text{pr}\left(\cup_k \cup_{h\in\mathcal{H}^*(k)} |S^{*(h)}(k)| > c^*\{k\}\right) = \alpha,$$

where $c^*(k)$ is a relevant boundary. Let $\tilde{k}_1$ be the first step when a hypothesis that belongs to $\mathcal{J}$ is rejected. For $k \leq \tilde{k}_1$, $|S^{*(h)}|$ is compared to $c'(k) \geq c^*(k)$, since $\mathcal{H}^*(k) \subseteq \mathcal{H}(k)$. Hence,

$$\begin{aligned}\text{pr}\left(\cup_{k\leq\tilde{k}_1} \cup_{h\in\mathcal{H}^*(k)} |S^{*(h)}(k)| > c'\{k\}\right) &\leq \text{pr}\left(\cup_{k\leq\tilde{k}_1} \cup_{h\in\mathcal{H}^*(k)} |S^{*(h)}(k)| > c^*\{k\}\right)\\ &\leq \text{pr}\left[\cup_k \cup_{h\in\mathcal{H}(k)} |S^{*(h)}| > c^*(k)\right] = \alpha.\end{aligned}$$

Then, we should consider two situations:

- there is $h \in \mathcal{H}^*(\tilde{k}_1)$ such that $h \in \delta_1(\tilde{k}_1)$; in this case, the context is identical to the particular case we dealt with above, so

$$\text{pr}\left(\cup_{k\leq\tilde{k}_1} \cup_{h\in\mathcal{H}^*(k)} |S^{*(h)}(k)| > c\{k\}\right) \leq \alpha;$$

- there were i rejections, at steps $\tilde{k}_1, \ldots, \tilde{k}_i$; in addition, $\mathcal{H}^*(k) \subseteq \delta_0(k)$, $k = 1, \ldots, \tilde{k}_{i-1}$ and $\mathcal{H}^*(\tilde{k}_i) \cap \delta_1(\tilde{k}_i) \neq \emptyset$. For $k = \tilde{k}_{i-1} + 1, \ldots, \tilde{k}_i$, $R(k)$ stands for all the hypotheses tested, and not rejected, since step $\tilde{k}_{i-1} + 1$, up to step $k - 1$, with at least one hypothesis rejected at the kth step. Let $R^*(k) \subseteq R(k)$, the subset which is connected with the hypotheses in $\mathcal{J}^*$; then

$$\begin{aligned}\text{pr}\left(R^*\{k\}\right) &= \text{E}\left(\mathbf{1}_{\{R^*(k)\}}\right) = \text{E}\left[\text{E}\left(\mathbf{1}_{\{R^*(k)\}} \mid \mathcal{F}\{\tau_{i-1}\}\right)\right]\\ &= \text{E}\left[\text{E}_{H_i,\, H_i\in\mathcal{J}^*}\left(\mathbf{1}_{\{R^*(k)\}} \mid \mathcal{F}\{\tau_{i-1}\}\right)\right];\end{aligned}$$

using the same arguments about critical values as those previously mentioned, we eventually prove that

$$\text{E}_{H_j,\, H_j\in\mathcal{J}^*}\left[\mathbf{1}_{\{R^*(k)\}} | \mathcal{F}(\tau_{i-1})\right] \leq \pi(k), \quad k = \tilde{k}_{i-1} + 1, \ldots, \tilde{k}_i.$$

Then, it follows that

$$\text{pr}\left(\cup_k \cup_{h\in\mathcal{H}^*(k)} |S^{*(h)}(k)| > c\{k\}\right) \leq \alpha.$$

■

2.4.2 Marginal Type I error rate

Suppose that Procedure 2.3.1 or 2.3.3 is used, and that $A^{(h)}(k)$ stands for rejection of H_{01h}, when the kth intermediate test is performed. We have

$$A^{(h)}(k) = \bigcap_{l=1}^{k-1} \bigcap_{h \in \mathcal{H}(l)} \left[|S^{*(h)}(l)| \leq c(l)\right] \bigcap \left[|S^{*(h)}(k)| > c(k)\right], \tag{2.4.1}$$

where $S^{*(h)}(l)$ is defined by (2.3.5). The marginal Type I error connected with H_{01h}, is given by

$$\alpha^{(h)} = P_{H_{01h}}\left[\bigcup_{k=1, k:h \in \mathcal{H}(k)}^{K} A^{(h)}(k)\right] = \sum_{k=1, k:h \in \mathcal{H}(k)}^{K} P_{H_{01h}}\left[A^{(h)}(k)\right],$$

where $A^{(h)}(k)$ is defined by (2.4.1).

If Procedure 2.3.2 is used, and if there are i steps for which at least one simple hypothesis is rejected, corresponding to the stopping times $\tau_1, \ldots, \tau_i$, we have

$$\begin{aligned}
\alpha^{(h)} \leq & \sum_{k=1}^{\tau_1} \mathrm{pr}_{H_{01h}}\left[A^{(h)}(k)\right] \\
& + \sum_{j=1}^{i-1} \sum_{k=\tau_j+1}^{\tau_{j+1}} \mathrm{pr}_{H_{01h}}\left[\left(\bigcap_{l=\tau_j+1}^{k-1} |S^{*(h)}(l)| \leq c\{l\}\right) \bigcap \left(|S^{*(h)}(k)| > c\{k\}\right) \mid \mathcal{F}(\tau_j)\right] \\
& + \sum_{k \geq \tau_i+1} \mathrm{pr}_{H_{01h}}\left[\left(\bigcap_{l=\tau_i+1}^{k-1} |S^{*(h)}(l)| \leq c\{l\}\right) \bigcap \left(|S^{*(h)}(k)| > c\{k\}\right) \mid \mathcal{F}(\tau_i)\right],
\end{aligned}$$

where k and l are such that $h \in \mathcal{H}(k)$, and $h \in \mathcal{H}(l)$.

2.4.3 Coherence and consonance

Two notions are inherent in the field of multiple comparisons:

- coherence, that is, if H_A implies H_B, then non-rejection of H_A must imply non-rejection of H_B; and
- consonance, i.e. if a multiple hypothesis H_A is rejected, then at least one of the simple hypotheses in H_A must be rejected.

We refer to Hochberg and Tamhane (1987) for any further detail.

However, coherence and consonance can hardly be discussed when we deal with Procedures 2.3.1–2.3.3, since they offer a number of different endpoints. Most of the difficulty may be caused by the experimenter's freedom to choose the subset of hypotheses to be tested at each step, and by the sequential nature of these procedures.

As an example, we consider Procedure 2.3.1, and we suppose that the whole set of simple null hypotheses is tested at every step. In that case, we are dealing with a union-intersection procedure. The hypothesis H_0 is rejected if and only if at least one hypothesis H_{01h} is rejected. Such a procedure is coherent and consonant: if H_0 is not rejected, then every H_{01h}, $h = 2, \ldots, T$, is not rejected; moreover, H_0 is rejected only if at least one hypothesis H_{01h}, $h = 2, \ldots, T$, is rejected.

2.4.4 p-values

In some cases a simple accept-reject dichotomy may not seem sufficiently informative. Siegmund (1985) considers attained significance levels as means of making more informative inferential statements. He gives the following definition of p-value in sequential procedures.

Definition 2.4.1 *The p-value is the weakest attained significance level for which the null hypothesis H_0 is rejected at a certain time t.*

Suppose that the overall null hypothesis H_0 is rejected at time $\tau = t$. Then, the attained significance level is $\mathrm{pr}_{H_0}(\tau \leq t)$, which is to be compared to α. At each step k, we are able to compute the significance level which is compared to $\pi(k)$, defined by (2.3.7),

$$\mathrm{pv}(k) = \mathrm{pr}\left(\bigcap_{l=1}^{k-1}\left[\bigcap_{h\in\mathcal{H}(l)} |S^{*(h)}(l)| \leq c(l)\right] \bigcup_{j\in\mathcal{H}(k)} \left(|S^{*(j)}(k)| > S_{obs}\{k\}\right)\right),$$

where is $S_{obs}(k)$ is the maximum amongst the observed values $|S^{*(j)}(k)|$, $j \in \mathcal{H}(k)$. Then, the p-value at kth step would be

$$p\text{-value}(k) = \sum_{l=1}^{k-1} \pi(l) + \mathrm{pv}(k). \tag{2.4.2}$$

We can also estimate the level of significance for any single hypothesis H_{01h} by computing

$$\mathrm{pv}_h(k) = \sum_l \mathrm{pr}\left(\left[\bigcap_{l'\in\{l^*\leq l-1: h\in\mathcal{H}(l^*)\}} R(l')\right] \bigcap \left(|S^{*(h)}(l)| > d^{(h)}\{l\}\right)\right), \tag{2.4.3}$$

where $l \in \{l^* \leq k : h \in \mathcal{H}(l^*)\}$, $R(l') = \{|S^{*(h)}(l')| \leq c(l')\}$, $d^{(h)}(r) = |S^{(h)}_{obs}(r)|$ for $r = \max\{l : h \in \mathcal{H}(l)\}$, and $d^{(h)}(l) = c(l)$ otherwise.

2.5 Parameter estimation

In this section, we discuss how the parameters of (1.3.5) are estimated when we use Procedures 2.3.1–2.3.3.

2.5.1 Fixed and random effect estimation

Fixed effect estimation

We consider the linear mixed effects model (1.3.5), where

$$\Theta = (D, \sigma^2), \tag{2.5.1}$$

is supposed to be known, and k is the interim analysis' number. We rewrite formula (1.3.3), which leads to

$$\text{var}\left[Y_i^h(k)\right] = \sigma^2(I + Z_i^h(k)DZ_i^h(k)') = V_i^h(k) = \left[W_i^h(k)\right]^{-1}.$$

The ML and least squares (LS) estimator of β^h (Lee and DeMets, 1991) are given by

$$\hat{\beta}^h(k) = \left[\sum_{i=1}^{m^h(k)} X_i^h(k)'W_i^h(k)X_i^h(k)\right]^{-1} \sum_{i=1}^{m^h(k)} X_i^h(k)'W_i^h(k)Y_i^h(k),$$

where $m^h(k)$ is the number of patients who have been assigned to treatment h until step k. If we assume that

$$\left(\sum_{i=1}^{m^h(k)} X_i^h(k)'W_i^h(k)X_i^h(k)\right)^{-1}$$

exists, then $\hat{\beta}^h$ is the best unbiased estimator of β^h.

Random effect estimation

The estimator of random effect (Laird and Ware, 1982) is

$$\hat{b}_i^h(k) = DZ_i^h(k)'W_i^h(k)\left[Y_i^h(k) - X_i^h(k)\hat{\beta}^h\right]\sigma^2.$$

In fact, the random variable $\hat{b}_i$ is not a ML estimator. However, $\hat{b}_i$ is optimal in the sense of having the smallest mean squared error, within the class of all the linear unbiased estimators of b_i. This result is ascribed to Harville (1976), who generalized the Gauss-Markov Theorem to include linear mixed effects models.

2.5.2 Scale parameter estimation

In a simple regression model with independent normal errors, $\hat{\Theta}_{ML}$ is biased downwards. However, one may avoid this drawback by using a Restricted Maximum Likelihood (REML) estimator. The REML estimator maximizes the likelihood of Θ — in our situation, scale parameter Θ is (2.5.1) — on the space spanned by $u'Y$, with $\mathrm{E}(u'Y) = 0$, and $I - X(X'X)^{-1}X' = uu'$, where X is the design matrix. In other words, we maximize the likelihood of $u'Y$. This concept was introduced by Patterson and Thompson (1971), and by Harville (1974) for linear mixed effects models.

A Bayesian justification for REML estimators relies upon the fact that the estimation of Θ does not require estimating the prior distribution of Θ, given β. Hence there is no inaccuracy induced by estimation at this step. Moreover, a bad estimate of the prior distribution of β would have no consequences for $\hat{\Theta}$.

In fact, the distinction between ML and REML estimators is important only when we use a model which tends to saturation, i.e. if the number of parameters to be estimated is close to the number of observations. As Diggle et al. (1994) point out, theoretical results are harder to find, because the two methods are asymptotically equivalent when either or both the number of observations and the number of individuals tend to infinity, for a fixed number of parameters.

If the number of parameters is high, the REML estimator is clearly preferred. In Procedures 2.3.1–2.3.3, we use REML estimators, because the number of patients assigned to a treatment in a trial may be rather modest.

We use an EM (Expectation Maximization) algorithm to estimate Θ. If we roughly sum up the situation, it seems that the EM algorithm is slower than the Newton-Raphson procedure, but convergence is more secure (Lindstrom and Bates, 1988).

Proposition 2.5.1 *We define* $N = \sum_{h=1}^{T}\sum_{i=1}^{m^h} n_i^h$, *and* $M = \sum_{h=1}^{T} m^h$. *For the model (1.3.2), REML estimators of* σ^2 *and* D *are obtained using the following iterative process:*

1. $D_{\omega+1} = D_\omega + \frac{1}{M}\sum_{h=1}^{T}\sum_{i=1}^{m^h} A_i^h$, *and*

2. $(\sigma^2)_{\omega+1} = (\sigma^2)_\omega + \frac{1}{N}\sum_{h=1}^{T}\sum_{i=1}^{m^h} B_i^h$, *with*

- $A_i^h = \sigma_\omega^{-2}\,\hat{b}_i^h\,\hat{b}_i^{h\prime} - D_\omega Z_i^{h\prime} P_i^h Z_i^h D_\omega\,\sigma_\omega^2$,
- $B_i^h = [\hat{r}_i^h - Z_i^h\,\hat{b}_i^h]'[\hat{r}_i^h - Z_i^h\,\hat{b}_i^h] - \sigma_\omega^4\,\mathrm{tr}(P_i^h)$,
- $P_i^h = (\hat{V}_i^h)^{-1} - (\hat{V}_i^h)^{-1}X_i^h(X^{h\prime}\hat{V}^h X^h)^{-1}X_i^{h\prime}(\hat{V}_i^h)^{-1}$,
- $\hat{r}_i^h = Y_i^h - X_i^h\hat{\beta}^h$,

where ω is the iteration number.

The choice of D_0 and σ_0^2 is not of primary importance. For example, we can set $D_0 = I$, and choose the ordinary least squares estimate for σ_0^2.

□ Proof:
For discussion of the EM algorithm, one should refer to Tanner (1996). The proof of the above result is given by Lindstrom and Bates (1988). If one wants ML estimators instead of REML estimators, $P_i^h = (\hat{V}_i^h)^{-1} - (\hat{V}_i^h)^{-1}X_i^h(X^{h\prime}\hat{V}^h X^h)^{-1}X_i^{h\prime}(\hat{V}_i^h)^{-1}$ is to be replaced by $P_i^h = (\hat{V}_i^h)^{-1}$.
■

We shall add that $\hat{\beta}^h(\hat{\Theta}_{REML})$ is a consistent estimator of β^h. Kackar and Harville (1984) proved that under mild conditions, $\hat{\beta}^h(\hat{\Theta}_{REML})$ is unbiased. Moreover, $\hat{\Theta}_{REML}$ is consistent (and so is $\hat{\Theta}_{ML}$), and $\hat{\beta}^h$ is a continuous function of $\hat{\Theta}_{REML}$. Thus, $\hat{\beta}^h(\hat{\Theta}_{REML}) \xrightarrow{p} \hat{\beta}^h$ by Slutsky–Fréchet Theorem. Since $\hat{\beta}^h = \hat{\beta}^h_{ML}$ is consistent, we eventually have $\hat{\beta}^h(\hat{\Theta}_{REML}) \xrightarrow{p} \beta^h$. Nevertheless, this remark does not constitute a proof of convergence, because the convergence of the EM algorithm to $\hat{\Theta}_{REML}$ is not certain.

2.5.3 Statistics for trend differences

In this section we discuss the estimation of the trend differences in (2.3.3), which are needed for testing the hypotheses (2.3.1).

Proposition 2.5.2 *Under the linear mixed effects model (1.3.5), when* Θ *is known, the trend estimators* $\{\hat{\beta}_1^h(1), \ldots, \hat{\beta}_1^h(K),\ h = 1, \ldots, T\}$, *are jointly distributed as*

$$\mathcal{N}_{TK}\left(\begin{pmatrix} M^1X^1\beta_1^1 \\ M^2X^2\beta_1^2 \\ \vdots \\ M^TX^T\beta_1^T \end{pmatrix}, \Sigma_\beta\right),$$

where

$$\Sigma_\beta = \begin{pmatrix} M^1V^1M^{1\prime} & 0 & \cdots & \cdots \\ 0 & M^2V^2M^{2\prime} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & M^TV^TM^{T\prime} \end{pmatrix} = \begin{pmatrix} \Sigma_{\beta^1} & 0 & \cdots & \cdots \\ 0 & \Sigma_{\beta^2} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \Sigma_{\beta^T} \end{pmatrix},$$

and $M^h = \begin{pmatrix} M^h(1) \\ \vdots \\ M^h(K) \end{pmatrix}$,

$$M^h(k) = (0\,1)\left[X^h(k)'W^h(k)X^h(k)\right]^{-1}X^h(k)'W^h(k)J^h(k),$$

$$J^h(k) = \left[\text{diag}\left(Q_1^h(k), Q_2^h(k), \ldots, Q_{m^h(k)}^h(k)\right) : O[\Omega(k)]\right],\ O[\Omega(k)]\ \textit{is the null}$$

matrix with dimension $\Omega(k)$ *equal to* $\left(\sum_{i=1}^{m^h(k)} n_i^h\{k\}, \sum_{i=m^h(k)+1}^{m^h(K)} n_i^h\{k\}\right)$,

and $Q_i^h(k) = \left(I\left[n_i^h(k), n_i^h(k)\right] : O[n_i^h(k), n_i^h(K) - n_i^h(k)]\right)$.

□ Proof:
Appendix B of Lee and DeMets (1991). We just have to take $h = 1, \ldots, T$, instead of $h = 1, 2$.
■

Situations when Θ is unknown lead to similar results asymptotically using the same arguments as in the previous section. The estimation of Σ_β in Proposition 2.5.2 seems to be rather intricate. But fortunately, this is a misleading impression. The proposition below shows that, given $\hat{\Theta}$, derivation of $\hat{\Sigma}_\beta$ is simple.

Proposition 2.5.3 *Let $\Sigma_{\beta^h\,[i,j]}(k)$, denote the (i,j)th element of the matrix Σ_{β^h} at step $k \leq K$. If we suppose that Θ is known,*

$$\Sigma_{\beta^h\,[K,K]}(K) = \Sigma_{\beta^h\,[k,K]}(K) = \left(X^h(K)W^h(K)X^h(K)'\right)^{-1}_{[2,2]},$$
$$\forall k = 1,\ldots,K, \qquad \forall h = 1,\ldots,T.$$

□ Proof:
The results are obtained by straightforward developments of $\left(M^hV^hM^{h'}\right)_{[K,K]}$ and $\left(M^hV^hM^{h'}\right)_{[k,K]}$, which are defined in Proposition 2.5.2.
■

By using ESFs, we have the opportunity of keeping the value of K undefined, since the overall significance level is not affected by K. Hence, Propositions 2.5.2 and 2.5.3 can be used at any step and for any K,

$$\hat{\Sigma}_{\beta^h}(K) = \begin{pmatrix} \hat{\Sigma}_{\beta^h}(K-1) & \hat{\Sigma}_{\beta^h\,[K,K]}(K)\,\vec{1}_{K-1} \\ \hat{\Sigma}_{\beta^h\,[K,K]}(K)\,\vec{1}'_{K-1} & \hat{\Sigma}_{\beta^h\,[K,K]}(K) \end{pmatrix},$$

where $\hat{\Sigma}_{\beta^h}(K-1)$ is the variance estimator of $\left(\hat{\beta}^h_1\{1\},\ldots,\hat{\beta}^h_1\{K-1\}\right)'$.

Proposition 2.5.4 *Under H_0, the statistics $S(k)$, defined by (2.3.4), satisfy*

$$(S(1)',\ldots,S(K)')' \sim \mathcal{N}_{r(K)}\left(0,\, P\{K\}C\{K\}\Sigma_\beta\{K\}C\{K\}'P\{K\}'\right),$$

$$\textit{where}\quad r(K) = \sum_{k=1}^{K} \mathrm{card}(\mathcal{H}\{k\}),$$

P and C are contrast matrices of dimensions $_{[r(K),r(K)]}$ and $_{[r(K),TK]}$ respectively.

□ Proof:
Straightforward computation. The result will be illustrated by the example below.
■

Example 2.5.1 *Let us consider a particular case, e.g. when $T = 3$ and $K = 2$. Suppose that at any step, treatments 1 and 2, as well as 1 and 3, are to be compared. Hence $C_{(4,6)}$ and $P_{(4,4)}$ equal*

$$C = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix} \quad \textit{and} \quad P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Practical difficulties may arise while $\hat{\Sigma}^h(\kappa)$ is being computed. We actually estimate this matrix by using successive values of $\hat{\Theta}$. It could happen that the matrix is no longer positive definite, which is a major drawback. This difficulty may be solved, either by the use of $\hat{\Theta}$ estimated at step 1, until the end of trial, or by the estimation of $\hat{\Sigma}^h(\kappa-1)$ a posteriori, given $\hat{\Theta}$ computed at step κ.

2.6 Discussion

In this chapter, we proposed three flexible procedures for comparing several treatments to a control. Procedures 2.3.1 and 2.3.3, and Procedure 2.3.2 with further condition about the error rates of Definition 2.3.3, control strongly the FWE. In fact, the corresponding programs are drawn up such that any contrast can be tested at any interim analysis, provided that the variance matrix of the corresponding statistics is still positive definite. Procedure 2.3.1 controls the FWE strongly in every situation. For Procedures 2.3.2 and 2.3.3, strong control depends on the chosen contrasts.

Allowing the experimenter to choose the subset of hypotheses to be tested may be considered as an open door for ambiguity. It all comes down whether every choice is firmly argued. In large Phase III trials, this decision would have to be made by committee, and the more options there are, the harder and more ambiguous the outcome. There may be good reason not to test a particular arm at a particular analysis. But if that results is rejecting an arm which otherwise would have been kept, or keeping one that would have been rejected, the validity of the trial's outcome will rest on those reasons. If that leads to disagreement in the target audience, then the price of this freedom is very high.

Of course, one can always contemplate using these procedures like a useful online and unofficial data control. Further properties of Procedures 2.3.1-2.3.3 were studied by simulation. Results are found in the next chapter.

Chapter 3

Application and Simulation Studies

In this chapter, we study the performance of Procedures 2.3.1 to 2.3.3. In Section 3.1, we first deal with the computation of intermediate significance levels (2.3.7). The use of error spending functions has been mentioned several times, and we discuss here the choice of suitable functions and time indicators.

Results of simulation studies are presented in Section 3.2. A few suggestions for controlling the overall significance level in the small sample case are made in Section 3.3.

To conclude, an alternative approach to Procedures 2.3.1 – 2.3.3 is presented in Section 3.4. This approach based on simultaneous confidence intervals use the results of Liu (1995, 1996), who works on independent normal populations with a common known variance.

3.1 Error spending function and time

3.1.1 Choice of the error spending function

O'Brien and Fleming's procedure (1979) is conservative (see Section 1.4.3) with decreasing boundaries. On the contrary, Pocock's procedure (1977) is less conservative, and has constant boundaries. Their approximate ESFs, α^*_{OBF} (1.4.7) and α^*_P (1.4.8), have of course the same features. Figure 3.1 displays the variation of some ESFs through time; one may also refer to Table 1.1.

Because there is no need to define the number of intermediate analyses in advance, it is simpler to use some ESFs like α^*_P, α^*_{OBF}, or α^*_{LDM} (1.4.9), than to use O'Brien and Fleming or Pocock's procedures. There are also some situations in which an ESF is the only possibility to be considered: Geller and Pocock (1987) mentioned a practical case, for which the experimental design was drawn up after the first intermediary test had been performed.

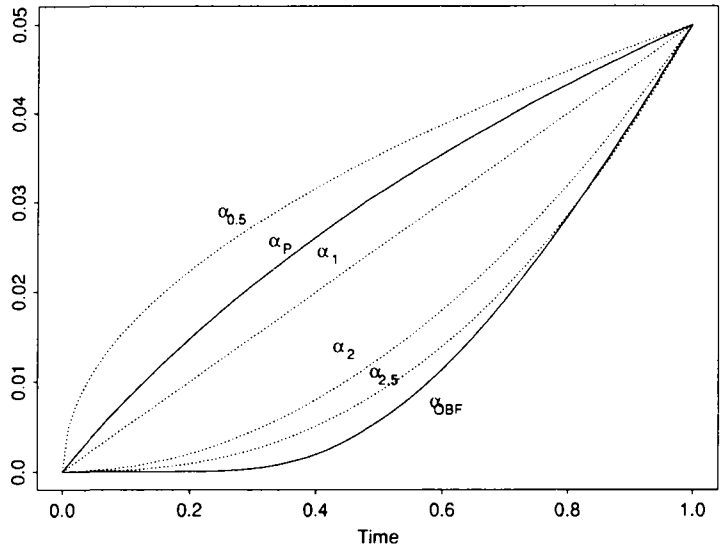

Figure 3.1: *Error spending functions* α_P^*, α_{OBF}^*, *and* α_{LDM}^* *with* $s = 0.5, 1, 2,$ *and* 2.5.

We introduce another ESF that we write α_C^*, defined by

$$\alpha_C^*(t_k) = \alpha - \sum_{l=0}^{k-1} \pi(l), \tag{3.1.1}$$

with

$$\pi(k) = \alpha_C^*(t_k) \left[\frac{t_k - t_{k-1}}{t_K - t_{k-1}} \right]^s, \quad 0 \leq t_k \leq t_K, \text{ and } \pi(0) = 0,$$

where t_k is the time corresponding to the kth test, and α is the overall significant level. The function α_C^* is very adaptable, because it allows the user to change t_K, that is to revalue the duration of the experiment. The functions α_C^* and $\alpha_{LDM}^* = \alpha t^s$ are very similar, but α_C^* is more sensitive to the choice of s. Unfortunately, the shape of α_C^* depends on the timing of the analyses, which may not be advisable. However, taking for example $s = 2$, we point out an interesting feature: when the number of interim analyses increases, α_C^* becomes very conservative. This leads to a method similar to Peto et al. (1976), whereas α_{LDM}^* is far less conservative.

The ESFs α_C^* and α_{LDM}^* are convex for $s > 1$, as advised by Kim and DeMets (1987). Indeed, it is highly recommended to be rather conservative when a new experiment is started. There are two main justifications of this:

- first, errors of measures' transcription often occur at the beginning of a new trial (Geller and Pocock, 1987); such errors can be detected only if the trial has not been already stopped! In addition,

- it should be borne in mind that the early stopping of a trial may make it lose some scientific credibility. Here again, we come up against the underlying antagonism between individual and collective ethics (Pocock, 1993). And finally,

- early benefit may be outweighed by later harm.

We shall add that the value $s = 2$ has often turned out to be rather satisfactory when we have used α_C^* or α_{LDM}^*.

3.1.2 Time indicator

Every ESF is a function of a parameter, usually written t (information), the meaning of which has been briefly discussed in Section 1.4.3. However, there are other indicators that can be contemplated; the simplest is of course the calendar time.

Lan, Reboussin, and DeMets (1994) use Fisher information corresponding to the estimators under interest, which turns out to be an element of outstanding importance in the sequential analysis of trials. However, the overall information at the planned end of the experiment is unlikely to be known in advance. Thus, it should be estimated: since it is directly related to sample size — and so to the cost of the study — the available estimates are sometimes very good. Kim and DeMets (1992) discuss the relation of sample size and information.

The algorithm LANDEM (Reboussin, DeMets, Kim, and Lan, 1992) allows the experimenter to start the trial by considering calendar time, and then to go on with a parameter t proportional to the increase in information. This increase is easy to compute, because it only relies on the covariance between successive estimators. However, it could happen that the Family-wise Error Rate (FWE) α is used completely before the planned end of the experiment. Of course, LANDEM also allows simple use of calendar time, which avoids difficulty with estimated information.

Finally, we can consider as a time indicator either the number of patients or the number of observations. Both of them are more suitable than calendar time, but the time based on the number of observations is very flexible, even if it is not a panacea. The experimenter has only to set an upper limit for the number of measurements, which represents the end of the experiment. Whereas Lan and DeMets (1989) note that it is hardly thinkable to end an experiment with a number of patients fixed in advance — consider for instance a rare pathology — it is much easier to reach exactly a number of measures fixed in advance.

Example 3.1.1 *We consider the linear mixed effects model (1.3.5), and choose the same design for every patient, that is,*

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 9 & 12 & 15 & 18 \end{pmatrix}', \text{ with } \sigma^2 = 1 \text{ and } D = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix}.$$

Figure 3.2 shows different time indicator functions, namely, calendar, observations, and Fisher information. A great difference is observed among the behaviours of those functions.

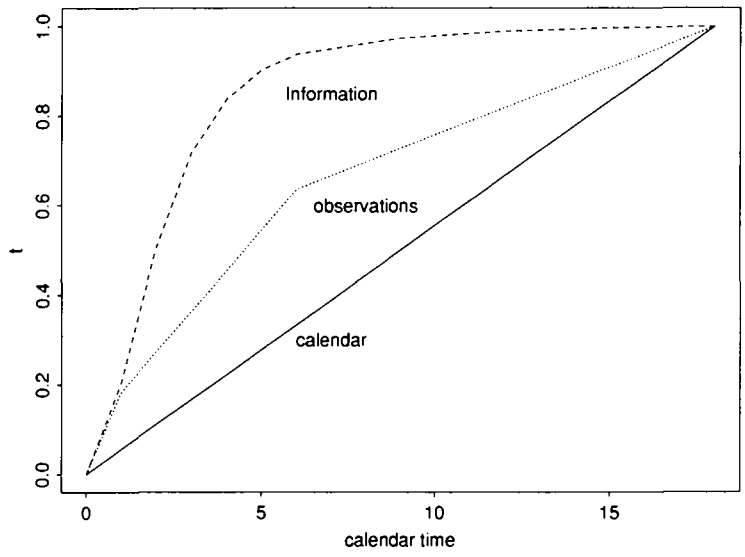

Figure 3.2: *Example 3.1.1; comparison of several time indicators.*

Seeing Figure 3.2, we may ask whether taking an ESF based on observations is really too conservative. Nothing is certain. It is indicated in Section 3.1.1, that one must act prudently when a new trial is started.

In fact, any interim level of significance $\pi(k)$ is a combination of both a time indicator and an ESF. Thus there are plenty of possibilities. Suppose for example that we choose Fisher information with a given ESF. We can also choose the number of observations as time indicator, take a less conservative ESF, and eventually have almost the same $\pi(k)$.

Example 3.1.2 *(Example 1.3.1 continued) We choose the model (1.3.5), and we suppose that the experimenter wants to test sequentially the equality of growth trend between the control and the 10% and 20% group. Procedure 2.3.1 is used with intermediate tests at the* $10th$, $16th$ *and* $21st$ *days. We use the ESFs* α_P^*, α_{OBF}^*, α_{LDM}^*, *and* α_C^* *defined by (1.4.8), (1.4.7), (1.4.9) and (3.1.1). Three time indicators are considered, viz. calendar, observations, and information. Table 3.1 gives stopping dates (rejection of the overall null hypothesis of equality of growth trends), and the intermediate levels of significance* π_j, $j = 1, 2, 3$.

Every procedure used leads to the rejection of the null hypothesis, more or less belatedly. Such a result is satisfactory, but cannot be guaranteed in general. Table 3.1 shows that calendar time and observations lead to more conservative procedures. We can also point out the strong disparity among the values of $\pi(1)$*: between 0.26% and 4.91%.*

Time	Calendar				Observations				Fisher Information	
	day	$\pi(j)$ ×100			day	$\pi(j)$ ×100			day	$\pi(j)$ ×100
α_P^*	10	2.99			10	3.10			10	4.91
α_{OBF}^*	16	0.45	2.02		10	0.56	0.18		10	4.68
α_{LDM}^*, $s = 1$	10	2.38			10	2.50			10	4.86
α_{LDM}^*, $s = 2$	16	1.13	1.77		16	1.25	1.56		10	4.72
α_{LDM}^*, $s = 4$	16	0.26	1.43		21	0.31	1.27	3.42	10	4.46
α_C, $s = 1$	10	2.38			10	2.50			10	4.86
α_C, $s = 2$	16	1.13	1.15		16	1.25	0.94		10	4.72
α_C, $s = 3$	21	0.54	0.72	3.74	21	0.62	0.55	3.83	10	4.59

Table 3.1: *Example 3.1.2; stopping dates — rejection of the overall null hypothesis of equality — and intermediate levels of significance* $\pi(j)$ *until the experiment is stopped.*

We often use a time indicator based on observations for the simulation studies, because it has turned out to be very easy and flexible to handle, and does not require estimation of Fisher information when scale parameters are unknown.

Anyway, it seems unthinkable to lay down general rules here, since they will be inextricably linked to the particular trial. However, when Fisher information is available, or can be estimated accurately, it should be used as time indicator, whatever the chosen ESF is.

3.2 Simulation studies

3.2.1 Comparison with non-sequential methods

It is useful to compare the performance of Procedures 2.3.1, 2.3.2, and 2.3.3 with some non-sequential methods that include a unique test at the end of the experiment. A multiple hypothesis test with normal statistics can be used. This is nothing but Procedure 2.3.1 when a unique final test is performed.

Otherwise, we can use a χ^2 test (Crowder and Hand, 1990, pp. 73–74), which combines all the simple hypotheses in a unique one. This test is based on the result below.

Proposition 3.2.1 *Let C be a contrast matrix (as in Example 2.5.1). The statistic*

$$\left(C\hat{\beta}_1 - C\beta_1\right)' \left(C\,\Sigma_\beta\,C'\right)^{-1} \left(C\hat{\beta}_1 - C\beta_1\right),$$

has a non-central χ^2 distribution, with degrees of freedom equal to $\mathrm{rank}(C)$*, and with a non-centrality parameter*

$$\delta = (C\beta_1)' \left(C\Sigma_\beta C'\right)^{-1} (C\beta_1), \text{ where } \beta_1' = (\beta_1^1, \ldots, \beta_1^T)'.$$

□ Proof:
This proposition follows from Muirhead's Theorem 1.4.5 (1982, p. 31) which deals with quadratic forms obtained from a multi-normal random variable. We have only to see that $\left(C\Sigma_\beta C'\right)^{-1}$ is symmetric and non-singular (in our situation, C is a full rank matrix), and that $\left(C\Sigma_\beta C'\right)^{-1}\left(C\Sigma_\beta C'\right) = I$ is of course idempotent.
■

In practice, Σ_β is be replaced by $\hat{\Sigma}_\beta$ and we use asymptotic results, that is, there is a convergence in distribution.

Suppose we want to test simultaneously several hypotheses $H_{01}, \ldots, H_{0l}$, and that each of them has a corresponding normal statistic Z_i, $i = 1, \ldots, l$. We call multi-normal test a test whose zone of non-rejection is

$$\{z \in \mathbb{R}^l : |z_i| \leq c_i, \forall i\}. \tag{3.2.1}$$

Zones of rejection for the multi-normal test, and for the χ^2 test are identical in the two-treatment case, but different in other situations (respectively a hyper-rectangle and a hyper-ellipsoid). Power performances are also very different, as we see in the following illustration:

Example 3.2.1 *We consider a three-treatment situation,* $h = 1, 2, 3$, *with 50 patients per group. The design for every patient is*

$$X_i^h = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 20 \end{pmatrix}, \quad h = 1, 2, 3, \quad i = 1, \ldots, 50, \quad \text{with } D = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}, \quad \text{and } \sigma^2 = 1.$$

We test $H_0 : \{\beta_1^1 = \beta_1^2\} \cap \{\beta_1^1 = \beta_1^3\}$, *versus* $H_1 : \exists h \in \{2, 3\} : \beta_1^h \neq \beta_1^1$, *at level 5%, when* D *and* σ^2 *are known. Thus this test is exact. Figure 3.3 shows the power difference between the multi-normal test and the* χ^2 *test. We single out two areas. When one treatment is superior to the control but the other is inferior, the* χ^2*-test is clearly more powerful. On the other hand, if both treatments are superior, or inferior to the control, the multi-normal test is slightly more powerful (see Figure 3.3).*

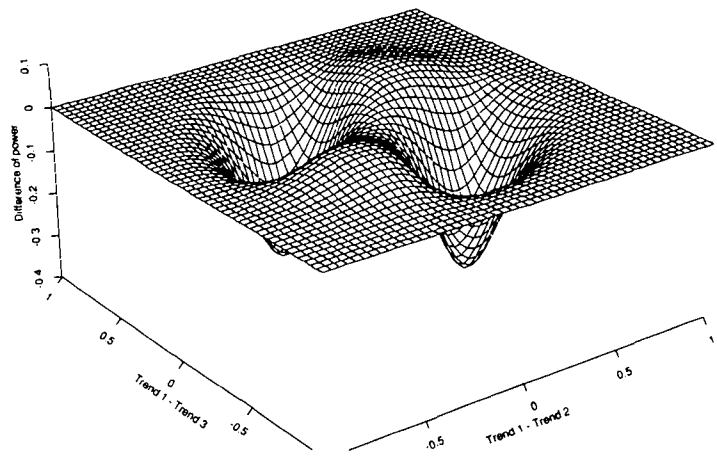

Figure 3.3: *Example 3.2.1; theoretical difference of power* ($\alpha = 5\%$) *between multi-normal test and the* χ^2*-test.*

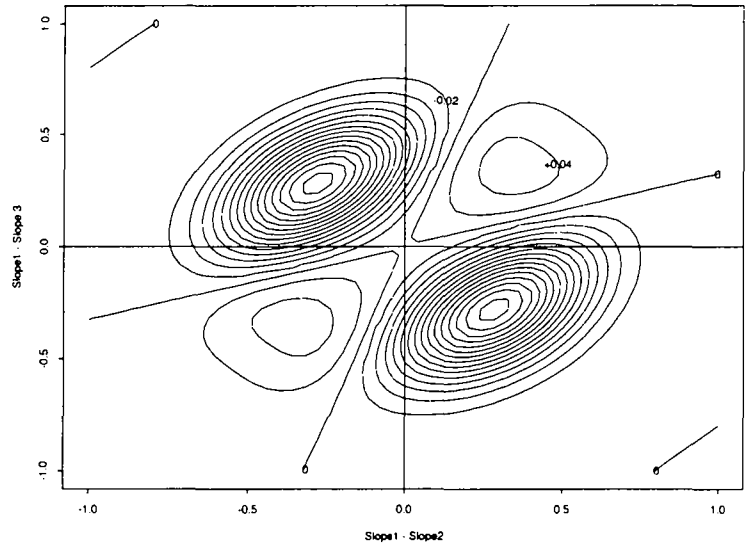

Figure 3.4: *Example 3.2.1; theoretical difference of power ($\alpha = 5\%$) between multi-normal and the χ^2-test.*

It is misleading to conclude that the χ^2-test is advantageous. Consider the prevalent situation in which new products are in advanced stages of development, and all expected to be, for instance, superior to the control. The first and third quadrants of Figure 3.4 are consequently of primary interest; and the area where the multi-normal test is superior covers more than 70% of the surface of these two quadrants.

3.2.2 The results of Lee and DeMets

In the linear mixed effects model, the useful results of Lee and DeMets (1995) for the two-treatment case should be borne in mind.

They first study the difference between their procedure and another one, that they call ad-hoc method. Instead of (1.3.5), it uses the model $Y_i^h(k) = X_i^h(k)\beta^h + e_i^h(k)$, with $e_i \sim \mathcal{N}(0, \sigma^2 I)$. Actually, it appears that their results are similar, except for staggered entries: the simplest method is then less powerful. Moreover, it is more sensitive to the choice of the ESF, especially if we have staggered entries. This sensitivity even makes for under or over-estimation of the FWE, but it is reduced when conservative ESFs are used.

Lee and DeMets (1995) also investigate the effects brought about by different situations: small numbers of patients, small numbers of observations, irregularity in the design, and

errors with random variance. Apart from the situation of small numbers of patients — which we discuss in Section 3.3 — they obtain satisfactory results, which let them infer a certain "robustness" of their procedure to violations of the typical assumptions, as small samples or non-constant σ^2 for the error.

Lastly, they discuss the use of information as a time indicator. It is very interesting to note that as soon as the second step is reached, critical values obtained using observations as the time indicator are close to the values computed using information. This fact is undoubtedly another point in favour of using observations as time indicator, in addition to its flexibility. Even though observations could be an imperfect estimator of Fisher information, results are eventually not significantly affected. A potential slight loss of power is compensated by a small saving in observations, and vice versa.

3.2.3 Simulations with non staggered entries

We carry out simulation studies, mostly with three treatments. By taking different numbers of patients per group, and observations per patient, we are able to cover a large spectrum of power values. We also check that the experiment-wise error is controlled.

Example 3.2.2 *We set for the model (1.3.5)*

$$X_i^h = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \end{pmatrix}, \forall i\,, \forall h \quad \textit{with } D = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}, \quad \textit{and } \sigma^2 = 1.$$

We take groups of sizes 10, 20, 30, or 50, and make one, two, three, or four interim tests of $H_0 : H_{012} \cap H_{013} = \{\beta_1^1 = \beta_1^2\} \cap \{\beta_1^1 = \beta_1^3\}$. *We use the ESF* α_c^* *(3.1.1), with* $s = 2$. *The above information is summarized in Table 3.2.*

Measurements per patient	Number of interim steps K			
n	1	2	3	4
5	4	2, 4	2, 3, 4	1, 2, 3, 4
10	9	4, 9	3, 6, 9	2, 4, 7, 9
20	19	9, 19	6, 12, 19	4, 9, 14, 19

Table 3.2: *Example 3.2.2; times (calendar) of interim tests, depending on the chosen numbers of steps and measurements on every patient.*

The values displayed in Tables 3.4 – 3.7 were obtained by simulation (2,500 trials), using Procedure 2.3.1. We mention γ, the theoretical power of the corresponding non sequential multi-normal test. The value between brackets corresponds to the χ^2-test mentioned in Proposition 3.2.1. The empirical power is denoted $\hat{\gamma}$. The values $\hat{\gamma}_2$ and $\hat{\gamma}_3$ stand for the proportions of H_{01j}, $j = 2, 3$, rejected. The symbol * displayed next to an empirical power value indicates that it falls outside of an interval centred on the corresponding theoretical value γ, more or less a two-standard deviation band based on a Bernoulli approximation.

From Table 3.4, it can be verified that the FWE is controlled, except for the ten-patient per group case, where it is systematically overshot. It may also be checked that in Table 3.5, the proportion H_{013} rejected is always below 5%. Finally, it turns out that the proportion of times the hypotheses H_{01j}, $j = 2, 3$, are rejected is far more sensitive to the number of interim analyses than the power is.

Example 3.2.3 *We consider Example 3.2.2 again, but with*

$$D = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \ \beta_1^1 = 1, \ \beta_1^2 = 1.4 \text{ and } \beta_1^3 = 1.3;$$

there are 50 patients per group, and ten observations per patient; we take several ESFs and time indicators. Two schemes were chosen for the interim analyses. The first scheme is almost regular in terms of observations (calendar: 2,4,6, and 9), and the other one in term of Fisher information (calendar: 1,2,4, and 9). Values of the different time indicators are found in Table 3.3.

Time indicator	Step for testing				
	1	2	4	6	9
calendar	0.11	0.22	0.44	0.67	1
observations	0.20	0.30	0.50	0.70	1
information	0.30	0.58	0.83	0.93	1

Table 3.3: *Example 3.2.3; value of the time indicator if we use calendar, the number of observations, or Fisher information.*

The conservatism of any procedure using the ESF α^*_{OBF} is clearly exemplified in Table 3.8. We are also able to note that α^*_C is more sensitive to s than α^*_{LDM}.

As we expect, the use of α^*_{OBF} leads to a rather good power performance. Moreover, α^*_{OBF} has the best ratio *power/observations* used, if we use Fisher information. In Table 3.8, we see that any ESF based on the number of observations is more conservative.

patients m	obs. n	steps K	% obs. used	γ ×100	$\hat{\gamma}$ ×100	$\hat{\gamma}_2$ ×100	$\hat{\gamma}_3$ ×100
50	20	1	100.0	5.0	4.8	2.6	2.6
		2	99.4	(5.0)	4.8	2.4	2.6
		3	99.2		4.8	2.4	2.6
	10	1	100.0		5.0	2.6	2.9
		2	99.2		5.0	2.5	2.8
		3	99.0		5.1	2.5	2.8
	5	1	100.0		5.4	3.1	2.7
		2	99.3		5.2	2.7	2.6
		3	99.1		5.0	2.6	2.4
30	20	1	100.0		5.6	3.4	2.8
		2	99.2		5.6	3.3	2.7
		3	99.1		5.6	3.3	2.7
	10	1	100.0		5.2	3.0	2.6
		2	99.4		5.2	2.9	2.5
		3	99.2		5.3	3.0	2.5
	5	1	100.0		5.4	3.2	2.5
		2	99.1		5.6	3.2	2.8
		3	98.8		5.6	3.1	2.7
20	20	1	100.0		5.2	2.8	3.0
		2	99.4		5.2	2.7	3.0
		3	98.7		5.2	2.6	2.9
	5	1	100.0		5.4	3.1	2.7
		2	99.3		5.2	2.7	2.6
		3	99.1		5.0	2.6	2.4
10	20	1	100.0		6.1*	3.4	3.0
		2	99.0		6.1*	3.3	3.0
		3	98.8		6.0*	3.4	2.9
	5	1	100.0		6.0*	3.3	3.4
		2	98.9		6.7*	3.4	3.8
		3	98.7		6.6*	3.4	3.7

Table 3.4: *Example 3.2.2 with* $\beta_1^1 = \beta_1^2 = \beta_1^3 = 1$; *percentage of the maximum number of observations used, theoretical FWE,* γ, *empirical FWE,* $\hat{\gamma}$, *and proportions of* H_{012} *and* H_{013} *rejected,* $\hat{\gamma}_2$ *and* $\hat{\gamma}_3$. *We used Procedure 2.3.1.*

patients m	obs. n	steps K	% obs. used	γ ×100	$\hat{\gamma}$ ×100	$\hat{\gamma}_2$ ×100	$\hat{\gamma}_3$ ×100
50	20	1	100.0	43.0	42.9	41.4	2.6
		2	69.4	(53.1)	42.9	41.3	2.0
		3	62.0		42.9	41.3	1.1
		4	59.4		42.9	41.3	1.7
	10	1	100.0	42.6	42.6	41.2	2.6
		2	89.8	(52.6)	42.4	41.0	2.0
		3	87.2		42.5	41.2	1.7
		4	87.8		42.5	41.1	1.7
30	20	1	100.0	27.0	27.2	25.6	2.7
		2	93.7	(34.0)	27.2	25.5	2.2
		3	92.3		27.3	25.5	2.1
		4	92.4		27.2	25.4	2.0
	10	1	100.0	26.8	27.0	25.5	2.7
		2	94.0	(33.7)	27.0	25.5	2.2
		3	92.2		27.0	25.5	1.8
		4	92.6		27.0	25.5	1.4
	5	1	100.0	24.8	25.0	23.3	2.7
		2	95.6	(31.3)	25.0	23.1	2.4
		3	94.6		25.0	23.1	2.4
		4	95.1		24.5	22.6	2.3
20	20	1	100.0	19.0	19.0	16.8	3.0
		2	96.2	(23.8)	19.0	16.8	2.7
		3	95.1		19.0	16.8	2.6
		4	95.2		19.0	16.8	2.4
	5	1	100.0	17.6	18.5	16.6	3.0
		2	97.3	(22.0)	17.9	15.8	2.7
		3	96.5		17.7	15.6	2.6
		4	96.7		17.9	15.6	2.8
10	20	1	100.0	11.5	12.4	9.9	3.7
		2	97.7	(13.8)	12.4	9.8	3.4
		3	97.1		12.4	9.8	3.1
		4	97.0		12.4	9.7	3.0
	5	1	100.0	10.86	11.9	9.1	3.4
		2	98.4	(13.0)	11.7	8.8	3.4
		3	97.7		11.4	8.8	3.4
		4	97.5		11.4	9.2	3.6

Table 3.5: *Example 3.2.2 with* $\beta_1^1 = \beta_1^3 = 1$ *and* $\beta_1^2 = 1.4$*; percentage of the maximum number of observations used, theoretical power,* γ*, empirical power,* $\hat{\gamma}$*, and proportions* H_{012} *and* H_{013} *rejected,* $\hat{\gamma}_2$ *and* $\hat{\gamma}_3$*. We used Procedure 2.3.1.*

patients m	obs. n	steps K	% obs. used	γ $\times 100$	$\hat{\gamma}$ $\times 100$	$\hat{\gamma}_2$ $\times 100$	$\hat{\gamma}_3$ $\times 100$
50	20	1	100.0	49.0	49.1	41.0	23.8
		2	85.9	(44.5)	49.1	38.9	19.5
		3	83.0		49.2	38.2	17.9
	10	1	100.0	48.6	48.1	40.8	24.2
		2	87.0	(44.1)	48.1	38.2	19.8
		3	84.8		48.1	37.2	18.2
	5	1	100.0	45.4	43.7	36.5	20.8
		2	91.9	(41.0)	43.3*	34.3	16.9
		3	90.0		43.2*	33.2	15.6
30	20	1	100.0	32.0	32.2	25.6	14.8
		2	92.4	(28.3)	32.2	24.0	12.4
		3	90.8		32.2	23.8	11.7
	10	1	100.0	31.7	30.3	23.7	14.6
		2	93.4	(28.0)	30.3	22.5	13.0
		3	92.0		30.3	21.6	12.1
	5	1	100.0	29.5	30.0	23.3	14.7
		2	94.6	(26.0)	30.3	22.4	13.1
		3	93.5		30.3	22.0	12.3
20	20	1	100.0	22.8	22.2	16.7	10.4
		2	94.7	(20.0)	22.2	15.9	9.1
		3	93.4		22.2	15.7	8.6
	5	1	100.0	21.1	21.4	15.4	9.8
		2	96.9	(18.5)	21.0	14.9	8.8
		3	96.0		21.2	14.6	8.3
10	20	1	100.0	13.6	14.7	9.7	7.3
		2	96.9	(12.1)	14.7	9.5	6.5
		3	96.1		14.7	9.3	6.3
	5	1	100.0	12.8	14.8*	10.4	7.2
		2	97.7	(11.4)	14.8*	9.8	6.9
		3	97.0		14.7*	9.3	6.7

Table 3.6: *Example 3.2.2 with* $\beta_1^1 = 1$, $\beta_1^2 = 1.4$, *and* $\beta_1^3 = 1.3$; *percentage of the maximum number of observations used, theoretical power,* γ, *empirical power,* $\hat{\gamma}$, *and proportions* H_{012} *and* H_{013} *rejected,* $\hat{\gamma}_2$ *and* $\hat{\gamma}_3$. *We used Procedure 2.3.1.*

patients m	obs. n	steps K	% obs. used	γ $\times 100$	$\hat{\gamma}$ $\times 100$	$\hat{\gamma}_2$ $\times 100$	$\hat{\gamma}_3$ $\times 100$
50	20	1	100.0	85.6	85.7	60.8	79.7
		2	66.0	(82.8)	85.7	48.6	74.8
		3	57.6		85.7	43.3	71.7
		4	54.4		85.7	38.7	69.2
	10	1	100.0	85.2	86.0	61.7	79.5
		2	67.2	(82.4)	85.9	47.3	74.2
		3	62.0		86.0	43.0	71.4
		4	61.4		85.8	37.8	68.8
	5	1	100.0	82.1	82.8	56.8	75.0
		2	78.8	(78.9)	82.2	44.3	68.4
		3	76.0		82.2	41.7	67.1
		4	75.3		82.0	36.3	64.1
30	20	1	100.0	64.4	65.0	38.5	55.0
		2	78.7	(59.8)	65.1	31.6	51.6
		3	73.9		65.1	27.5	50.1
		4	72.7		65.0	26.2	48.8
	10	1	100.0	63.9	63.4	37.7	53.8
		2	80.6	(59.3)	63.4	30.2	49.5
		3	77.4		63.4	27.3	48.4
		4	77.7		63.6	26.8	46.4
	5	1	100.0	60.2	59.9	35.5	50.4
		2	87.1	(55.6)	59.8	29.5	46.2
		3	84.5		59.8	28.0	44.6
		4	84.4		59.7	26.2	43.3
20	20	1	100.0	47.4	47.4	25.5	38.2
		2	86.3	(42.7)	47.4	21.3	34.9
		3	83.2		47.4	19.2	34.6
		4	83.0		47.4	18.3	33.8
	5	1	100.0	43.9	44.8	24.2	35.9
		2	91.3	(39.3)	44.2	19.8	33.6
		3	89.7		44.2	18.7	32.8
		4	90.0		44.3	18.2	32.5
10	20	1	100.0	26.5	30.4*	17.1	21.8
		2	92.6	(23.2)	30.4*	15.0	20.4
		3	90.8		30.0*	14.1	19.8
		4	90.6		30.4*	13.6	19.5
	5	1	100.0	24.6	25.9*	14.1	18.2
		2	96.0	(21.4)	26.1*	13.24	17.2
		3	94.7		26.2*	12.80	16.5
		4	94.8		26.1*	12.60	16.1

Table 3.7: *Example 3.2.2 with* $\beta_1^1 = 1$, $\beta_1^2 = 1.5$, *and* $\beta_1^3 = 1.6$; *percentage of the maximum number of observations used, theoretical power,* γ, *empirical power,* $\hat{\gamma}$, *and proportions* H_{012} *and* H_{013} *rejected,* $\hat{\gamma}_2$ *and* $\hat{\gamma}_3$. *We use Procedure 2.3.1.*

ESF	% obs. used	$\hat{\gamma}$ ×100	$\hat{\gamma}_2$ ×100	$\hat{\gamma}_3$ ×100	% error	% rejection H_0 at step 1	2	3	4	R ×100
time indicator : observation; tests at time 1, 2, 4 and 9 (calendar)										
α_P^*	79.1	45.6 *	34.1	17.2	0.1	18.8	19.5	34.4	27.3	57.7
α_{OBF}^*	90.7	48.1	38.6	19.7	0.1	0.6	3.6	32.7	63.1	53.1
α_C^* 0.5	76.7	42.3 *	31.6	16.0	0.2	26.2	25.8	32.0	16.0	55.1
α_{LDM}^* 0.5	79.1	43.7 *	32.6	16.9	0.2	25.3	17.7	30.4	26.5	55.3
α_{LDM}^* 1	80.8	47.0	35.2	18.2	0.1	13.4	18.4	34.6	33.6	58.1
α_{LDM}^* 2	86.0	47.8	37.7	18.9	0.1	4.0	12.4	34.9	48.7	55.5
α_C^* 2	90.1	47.8	37.9	19.4	0.1	4.0	6.7	25.5	63.7	53.0
time indicator : information; tests at time 1, 2, 4 and 9 (calendar)										
α_P^*	75.3	42.0 *	31.4	15.8	0.2	25.7	32.6	30.9	10.9	55.8
α_{OBF}^*	78.3	47.8	36.5	18.5	0.1	1.2	29.4	47.4	21.9	61.1
α_C^* 0.5	75.3	38.8 *	28.6	14.8	0.2	32.8	35.6	25.3	6.3	51.6
α_{LDM}^* 0.5	76.1	40.0 *	29.6	15.2	0.2	31.8	28.5	28.6	11.2	52.6
α_{LDM}^* 1	75.4	44.0 *	32.8	16.8	0.1	19.6	32.7	34.7	13.1	58.3
α_{LDM}^* 2	77.2	46.6 *	35.2	18.0	0.1	8.1	31.2	41.0	19.7	60.3
α_C^* 2	80.0	47.6	36.1	17.8	0.1	8.0	22.7	39.5	29.8	59.4
time indicator : observation; tests at time 2, 4, 6 and 9 (calendar)										
α_P^*	75.9	46.6 *	34.8	17.8	0.1	47.9	28.6	12.7	10.7	61.3
α_{OBF}^*	86.0	48.1	37.1	17.7	0.1	4.3	32.8	32.2	30.7	55.9
α_C^* 0.5	73.9	45.4 *	34.6	18.4	0.1	57.1	29.4	9.4	4.0	61.4
α_{LDM}^* 0.5	75.0	45.9 *	34.6	18.4	0.1	56.5	23.4	10.5	9.5	61.1
α_{LDM}^* 1	77.7	47.6	35.6	18.1	0.1	38.3	30.6	15.7	15.4	61.3
α_{LDM}^* 2	82.2	48.0	36.5	17.4	0.1	18.8	33.9	23.2	24.1	58.4
α_C^* 2	85.6	48.0	36.7	17.8	0.1	18.8	21.7	20.0	39.5	56.1
time indicator : information; tests at time 2, 4, 6 and 9 (calendar)										
α_P^*	73.3	43.9 *	33.7	18.6	0.1	65.8	25.9	5.6	2.6	59.9
α_{OBF}^*	76.4	47.7	36.2	18.4	0.1	30.8	47.7	13.8	7.7	62.4
α_C^* 0.5	72.7	42.8 *	56.9	19.2	0.1	71.6	24.3	3.5	0.6	58.8
α_{LDM}^* 0.5	73.4	43.1 *	32.8	18.5	0.1	71.1	21.0	5.2	2.7	58.7
α_{LDM}^* 1	73.7	45.1 *	34.4	18.4	0.1	59.0	28.8	8.2	4.0	61.2
α_{LDM}^* 2	75.4	47.0	35.2	18.2	0.1	41.0	40.0	12.1	7.0	62.4
α_C^* 2	76.9	47.4	35.4	17.7	0.1	40.7	33.7	11.3	14.3	61.6

Table 3.8: *Example 3.2.3; percentage of the maximum number of observations used, empirical power, proportion of H_{012} and H_{013} rejected, percentage of wrong decision about trends' order when H_0 is rejected. Distribution of the stopping times (H_0 rejected); R is the ratio empirical power / percentage of observation used. The trends are $\beta_1^1 = 1$, $\beta_1^2 = 1.4$ and $\beta_1^3 = 1.3$. The symbol * next to an empirical power value indicates that it falls outside an interval centred on the corresponding theoretical value γ, more or less a two-standard deviation band based on a Bernoulli approximation. We used Procedure 2.3.1.*

This naturally leads to an increase in the number of observations used. However, we note that the gain is hardly significant in terms of power, especially for the ESFs that are conservative, like α^*_{OBF}, or α^*_{LDM} with $s > 1$.

When possible, we also compare Procedures 2.3.1, 2.3.2, and 2.3.3, with the method of Proschan et al. (1994) described in Procedure 2.2.1.

Method	$\hat{\gamma}$ ×100	$\hat{\gamma}_2$ ×100	$\hat{\gamma}_3$ ×100	$\hat{\gamma}^*_2$ ×100	$\hat{\gamma}^*_3$ ×100	% obs. used
50, 50 and 50 patients						
1 step	5.0	2.6	2.9			100.0
Proschan $_{OBF}$	5.8	3.4	3.4	3.0	3.2	99.9
Proschan $_{OBF}$ Bonf.	5.4	3.3	3.2	2.8	2.9	99.0
Procedure 2.3.2 α^*_{OBF}	5.2	2.4	3.1	2.4	3.0	99.0
50, 10 and 30 patients						
1 step	5.7	3.2	2.8			100.0
Proschan $_{OBF}$	6.7*	3.8	3.3	3.7	3.1	92.2
Proschan $_{OBF}$ Bonf.	6.2*	3.7	3.2	3.5	2.8	99.3
Procedure 2.3.2 α^*_{OBF}	5.8	3.0	3.2	3.0	3.0	99.2

Table 3.9: *Example 3.2.4 with* $\beta_1^1 = \beta_1^2 = \beta_1^3 = 1$; *empirical power,* $\hat{\gamma}$, *proportion of* H_{012} *and* H_{013} *rejected,* $\hat{\gamma}_j$, *proportion of* H_{012} *and* H_{013} *rejected,* $\hat{\gamma}^*_j$, *if the trial was stopped after the first rejection (Procedure 2.3.1), and proportion of the maximum number of observations used.*

Example 3.2.4 *Tables 3.9 and 3.10 are obtained by simulations, in the same framework as Example 3.2.3, with four interim analyses at dates 1, 2, 4 and 9. We use information time. In Table 3.9, we have* $\beta_1^1 = \beta_1^2 = \beta_1^3 = 1$, *whereas in Table 3.10* $\beta_1^1 = 1$, $\beta_1^2 = 1.4$, *and* $\beta_1^3 = 1.3$. *Two cases were considered: 50 patients per group in the first one, and 50 patients for the first group, ten for the second and 30 for the third group in the second case.*

In the second situation, one can establish that the hypotheses of equality of variance of the statistics $S^{(h)}(k)$, *defined by (2.3.3), at a certain step* k *are clearly not respected. Nevertheless, we still employ the Proschan et al. (1994) tables of critical values. Four Procedures are used, namely a multi-normal test — see (3.2.1) — with a single step at the end of trial, the procedure of Proschan et al. (1994) with O'Brien and Fleming's method, using either exact boundaries, or Bonferroni values, and Procedure 2.3.2 with* $\alpha = \alpha_2 = 0.05$.

Finally, we consider a four-treatment example, with $\beta_1^1 = 1$, $\beta_1^2 = 1.7$, $\beta_1^3 = 1$ *and* $\beta_1^4 = 1.4$, *in the same framework as Example 3.2.4. We perform three intermediate tests, using information time. Results are displayed in Table 3.11.*

Method	$\hat{\gamma}$ ×100	$\hat{\gamma}_2$ ×100	$\hat{\gamma}_3$ ×100	$\hat{\gamma}_2^*$ ×100	$\hat{\gamma}_3^*$ ×100	% error	% obs. used
50, 50 and 50 patients							
1 step	48.1	24.2	40.8			0.1	100.0
Proschan $_{OBF}$	49.5	26.8	43.0	19.5	37.6	0.1	87.7
Proschan $_{OBF}$ Bonf.	48.5	26.2	42.2	19.0	37.1	0.1	87.5
Procedure 2.3.2 α^*_{OBF}	46.9	20.6	36.0	18.6	36.0	0.2	88.5
50, 10 and 30 patients							
1 step	28.4	13.9	18.7			0.1	100.0
Proschan $_{OBF}$	29.8	16.1	19.7	13.7	17.8	0.2	95.4
Proschan $_{OBF}$ Bonf.	28.5	15.4	19.0	13.1	17.1	0.1	95.6
Procedure 2.3.2 α^*_{OBF}	26.8	12.2	16.8	12.2	16.3	0.1	96.0

Table 3.10: *Example 3.2.4 with* $\beta_1^1 = 1$, $\beta_1^2 = 1.3$ *and* $\beta_1^3 = 1.4$; *empirical power,* $\hat{\gamma}$, *proportion of* H_{012} *and* H_{013} *rejected,* $\hat{\gamma}_j$, *proportion of* H_{012} *and* H_{013} *rejected,* $\hat{\gamma}_j^*$, *if the trial was stopped after the first rejection (Procedure 2.3.1), percentage of wrong decision about trends' order when* H_0 *is rejected, and proportion of the maximum number of observations used.*

Method	$\hat{\gamma}$ ×100	$\hat{\gamma}_2$ ×100	$\hat{\gamma}_3$ ×100	$\hat{\gamma}_4$ ×100	$\hat{\gamma}_2^*$ ×100	$\hat{\gamma}_3^*$ ×100	$\hat{\gamma}_4^*$ ×100	% obs. used
1 step	47.8	43.4	1.7	15.8				100.0
Proschan $_{OBF}$	49.7	45.3	3.0	18.1	43.6	1.7	11.3	92.4
Procedure 2.3.3 α^*_{OBF}	46.5	42.5	2.5	17.0	40.8	1.4	10.1	92.3
Procedure 2.3.2 α^*_{OBF}	46.5	40.8	1.4	10.9	40.8	1.4	10.1	96.0
Procedure 2.3.3 α^*_{LDM}	43.5	39.2	2.9	15.5	37.7	2.0	9.9	91.9
Procedure 2.3.3 α^*_P	41.5	37.5	3.0	14.8	36.0	2.2	9.6	91.9
Procedure 2.3.2 α^*_P	41.5	36.0	2.2	10.6	36.0	2.2	9.6	92.5

Table 3.11: *Example 3.2.4 with* $\beta_1^1 = 1$, $\beta_1^2 = 1.7$, $\beta_1^3 = 1$, *and* $\beta_1^4 = 1.4$; *empirical power,* $\hat{\gamma}$, *proportion of* H_{012}, H_{013}, *and* H_{014} *rejected,* $\hat{\gamma}_j$, *proportion of* H_{012}, H_{013}, *and* H_{014} *rejected,* $\hat{\gamma}_j^*$, *if the trial was stopped after the first rejection (Procedure 2.3.1), and proportion of the maximum number of observations used.*

It could be argued that the overshooting the FWE in Table 3.9 is due to the fact that there are only ten individuals in the second group. We perform the same simulations, using the true scale parameters; these simulation actually show that all FWEs are controlled, except for Follmann et al.'s (1994) method, where the FWE is still outside

a two-standard deviation band from the targeted value. Finally, we can check that the experiment-wise error rate in Table 3.11 is strongly controlled (see $\hat{\gamma}_3$), and that Procedure 2.3.2 is more conservative than 2.3.3, especially if we consider treatment 4, which is closer to the control than treatment 2 is.

3.2.4 Staggered entries

Actually, non-staggered entry is a quite unusual situation in large clinical trials, and the utility of sequential testing after subject recruitment has ended is greatly reduced.

Example 3.2.5 *We carry out simulations (2,500 trials) with 50 patients per group, using both observations and information as time indicator, and several ESFs. The context is similar to Example 3.2.4, but we perform three interim analyses at dates 2, 4 and 13. We use information time or the time based on observations. In Table 3.12, we have* $\beta_1^1 = \beta_1^2 = \beta_1^3 = 1$, *whereas in Table 3.13, we have* $\beta_1^1 = 1$, $\beta_1^2 = 1.4$, *and* $\beta_1^3 = 1.3$. *When the experiment starts, 16 individuals (per group) enter the trial at calendar date 0 (ten measures from 0 to 9), 17 at date 2 (ten measures from 2 to 11), and finally 17 at date 4 (10 measures from 4 to 13).*

ESF	$\hat{\gamma}$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	% rejection step			number	obs.
	×100	×100	×100	1	2	3	patients	used
Time indicator: observations								
α_P^*	5.3	2.8	2.8	23	20	57	148.2	98.1
α_{OBF}^*	5.0	2.6	2.9	0	0	100	150.0	100.0
α_{LDM}^* $s = 0.5$	5.5	3.0	2.9	39	27	34	147.0	97.0
α_{LDM}^* $s = 1$	4.9	2.6	2.6	15	13	72	148.9	98.8
α_{LDM}^* $s = 2$	5.0	2.4	2.9	1	8	91	149.8	99.7
Time indicator: information								
α_P^*	5.4	3.0	2.8	41	27	32	147.0	96.9
α_{OBF}^*	4.9	2.5	2.8	0	12	88	149.7	99.6
α_{LDM}^* $s = 0.5$	4.7	2.7	2.3	58	25	17	145.7	95.8
α_{LDM}^* $s = 1$	5.4	2.7	3.0	33	27	40	147.5	97.3
α_{LDM}^* $s = 2$	5.1	2.7	2.8	12	10	78	149.1	99.1

Table 3.12: *Example 3.2.4 with* $\beta_1^1 = \beta_1^2 = \beta_1^3 = 1$; *empirical FWE,* $\hat{\gamma}$, *proportion,* $\hat{\gamma}_j$, *of* H_{012} *and* H_{013} *rejected, stopping time's distribution, number of patients involved, proportion of the maximum number of observations used. We used Procedure 2.3.1.*

ESF	$\hat{\gamma}$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	errors	% rejection step			number	obs.	ζ
	×100	×100	×100	×100	1	2	3	patients	used	
Time indicator: observations										
α_P^*	46.0	19.2	36.0	0.1	9	26	65	139.9	87.7	32.9
α_{OBF}^*	48.2	24.0	40.8	0.1	1	0	99	149.5	99.5	32.2
α_{LDM}^* $s = 0.5$	42.6	16.9	32.8	0.2	15	32	53	136.7	84.3	31.1
α_{LDM}^* $s = 1$	46.8	20.5	37.4	0.1	6	20	74	142.5	90.7	32.8
α_{LDM}^* $s = 2$	48.0	22.5	40.1	0.1	1	8	91	147.5	96.7	32.6
Time indicator: information										
α_P^*	41.8	16.4	32.1	0.2	16	34	50	136.1	83.7	30.7
α_{OBF}^*	47.7	21.7	38.8	0.1	1	19	80	145.0	93.0	32.9
α_{LDM}^* $s = 0.5$	37.0	14.0	27.8	0.3	22	38	40	134.7	84.9	27.5
α_{LDM}^* $s = 1$	43.9	17.8	34.2	0.2	12	32	56	137.5	84.9	31.9
α_{LDM}^* $s = 2$	47.3	20.9	38.2	0.1	4	18	78	143.7	92.0	32.9

Table 3.13: *Example 3.2.4 with* $\beta_1^1 = 1$, $\beta_1^2 = 1.3$ *and* $\beta_1^3 = 1.4$; *Empirical power,* $\hat{\gamma}$, *proportion* $\hat{\gamma}_j$ *of* H_{012} *and* H_{013} *rejected, proportion of wrong decision about trends' order when* H_0 *is rejected, stopping times' distribution, number of patients involved, proportion of the maximum number of observations used, and ratio* ζ = (*power* × 100/*number of patients involved*) × 100.

3.2.5 Discussion

In the staggered-entry situation, which is prevalent in practice, we see in Table 3.12 that the FWE is well-controlled. Moreover, Table 3.13 allows us to say that staggered entries have a slight effect on power, that is, there is a small increase in power because the procedures tend to be more conservative, whatever the time indicator or ESF used.

We can conclude from Tables 3.5 – 3.7, and 3.8 that conservative ESFs as α_{OBF}^*, or α_{LDM}^* and α_C^* with $s \geq 2$, do not reduce power in a significant way when we make a moderate number of interim tests. According to Tables 3.9 – 3.11, it seems that the boundaries of Follmann et al. (1994) are not as resistant as our procedures for controlling the FWE when scale parameters are unknown. In this particular example, the only possible explanation should result in the way of computing the boundaries (simulation or numerical integration), since there is no theoretical difference between the methods. The conservative behaviour of Procedure 2.3.2 may need further investigation too.

The only relevant drawback is the difficulty in controlling the FWE when few patients are recruited in each group. This problem is dealt with in the next section.

3.3 Small samples

Small sample sizes have great importance when ethical and cost considerations must be taken into account. As we note in Table 3.4, and Lee and DeMets (1995) point out for the two-treatment case, the FWE is not well controlled when the total number of patients involved in the experiment is small.

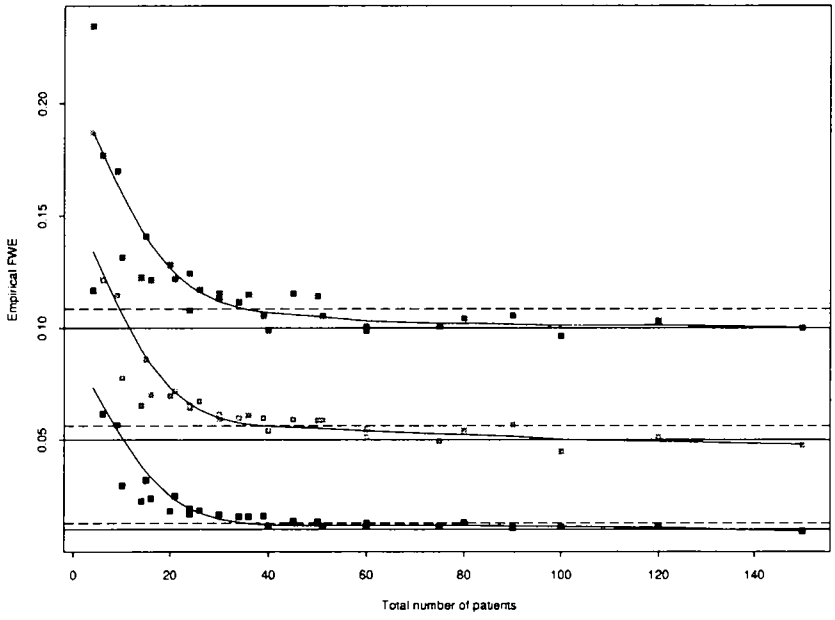

Figure 3.5: *Empirical FWEs computed by simulation from trials with theoretical overall significance levels of 1%, 5% and 10%. Lines were obtained using the smoothing spline technique. Dotted lines show the upper 95% confidence bounds around the theoretical FWEs.*

Because scale parameters are in fact estimated, the statistics $S^h(k)$ at (2.3.3) are normally distributed, but only asymptotically. In group sequential procedures, this additional problem has been addressed by Jennison and Turnbull (1995). Establishing theoretical results seems difficult to contemplate in a situation where variance matrices are functions of parameters computed by the EM (Expectation Maximization) algorithm. However, we are still able to make the analogy with the simple case of independent normal random variables, which would lead us to conclude that $S^h(k)$ is no longer normally distributed, but distributed like a Student random variable whose degrees of freedom are linked to the number of individuals in the experiment.

If we take the same designs and scale parameters as in Example 3.2.1, Figure 3.5 shows FWE values obtained by simulations from the two-treatment or the three-treatment case. Theoretical FWEs were supposed to be 1%, 5% or 10%. Whatever the theoretical FWE, a threshold between 30 and 35 individuals seems to be necessary in order to state that the FWE is actually controlled.

Such simulations may be used as a correction table, once the scale parameters have been estimated. For instance, if we consider an experiment with three groups of five individuals, and if we want to make a 5% level test, we should set the FWE to 1% when the program is used.

3.3.1 Adjustment of variance

From a theoretical point of view, the estimation of $\text{var}(\hat{\beta}^h)$ has several drawbacks. In spite of the use of REML (Restricted Maximum Likelihood) estimators for D and σ^2 in (1.3.5), two points have still to be dealt with:

- the effects of the variability of $\hat{D}$ and $\hat{\sigma}^2$ on $\text{var}\left[\hat{\beta}^h(\hat{D},\hat{\sigma^2})\right]$ are not allowed for; and
- it cannot be asserted that $\text{var}\left[\hat{\beta}^h(\hat{D},\hat{\sigma^2})\right]$ is unbiased. This could lead to exaggerated to too conservative zones of rejection.

Let us set $\hat{\Psi}^h = \text{var}(\hat{\beta}^h)$. Kenward and Roger (1997) suggest that $\hat{\Psi}^h$ is replaced by $\hat{\Psi}^{h*} = \hat{\Psi}^h + 2\hat{\Lambda}^h$, where $\hat{\Lambda}^h$ is a correction matrix obtained by Taylor series expansion around the scale parameters. Note that when the design is balanced, $\hat{\Lambda}^h = 0$. If C is a contrast matrix with l rows, it can be proved under some regularity conditions that the random variable

$$\frac{m}{m+l-1}\frac{1}{l}\left(\hat{\beta}^h - \beta^h\right)' C' \left(C\hat{\Phi}^{h*}C'\right) C \left(\hat{\beta}^h - \beta^h\right) \tag{3.3.1}$$

is exactly $F_{l,m}$ distributed, where m is an estimated number of degrees of freedom. We carried out simulations in the situation of a non-sequential test, with two treatments and very unbalanced designs. We studied how to choose the number of individuals in each group, so as to control the FWE (supposed to be 5%). It turned out that the use of $\hat{\Psi}^{h*}$ instead of $\hat{\Psi}^h$ allowed a reduction from 18 to 15 individuals. So, the scope of this method seems indeed rather reduced in our situation. In addition, highly unbalanced design is not common in practice.

3.3.2 Use of Student statistics

Considering ordinary instead of generalized least squares, is a simple way to lower the influence of the scale parameter on the estimation of $\hat{\beta}^h$. But such methods always lead to greater variances and less powerful tests (Giesbrecht and Burns, 1985). On the contrary, approximate t-tests turn out to be a useful tool for testing linear contrasts among fixed effects. In our situation, we want to make sequential multiple comparisons; but the problem is more intricate, because there are potentially many correlated contrasts at a given step k and also over time.

Let us consider the simulations of Section 3.3.1. We are able to keep the empirical FWE in a two-standard deviation band around 5%, just by roughly using boundaries for $S^{(2)}(1)$ based on a t distribution, with m degrees — the same as in (3.3.1) — of freedom.

That is the reason why we suggest, as an empirical rule, that boundaries should be computed using a multivariate t distribution — see for example Sutradhar (1986) for a complete characterization — in lieu of multi-normal distributions. This may be done for Procedures 2.3.1 and 2.3.3, when small samples are studied.

Example 3.3.1 *We consider the model (1.3.2), with two groups of six individuals,* $\sigma^2 = 1$, $D = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$, *and the design* $\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 9 \end{pmatrix}'$. *We use Procedure 2.3.1, and compute the boundaries using a multivariate t distribution. Table 3.14 shows the empirical powers obtained, and the proportion of observations used. We use different values of* β_1^h, *and different ESFs.*

In Table 3.14 we see that the use of a multivariate Student distribution makes for slightly over-conservative procedures; it actually turns out that tail thickness of the statistics distribution is over-estimated. Nevertheless, it is better to guarantee that the FWE is controlled by the targeted value, than to overshoot it, sometimes considerably. In addition, loss of power seems to be rather small, and we shall add that use of Student boundaries lead to more robust procedures, even if the word robust may be not intrinsically correct.

If robustness were investigated, a better process would be to consider multivariate t models for both error and random effects (Little, 1988). Even if the t distribution is not, of course, a panacea for all robustness problems, Lange, Little, and Taylor (1989) note that this approach combines conceptual simplicity — since it is based on a parametric model — with generality, because it can be applied in a wide range of settings.

case	ESF	$\hat{\gamma}$ ×100	% obs. used	R_γ ×100	R_{obs} ×100
$\beta_1^1 = \beta_1^2 = \beta_1^3 = 1$	α_P^*	2.9	98.9	52	101
	α_{OBF}^*	3.1	99.7	56	101
	α_{LDM}^* $s = 0.5$	3.0	98.7	54	100
	α_{LDM}^* $s = 1$	2.8	99.2	51	101
	α_{LDM}^* $s = 2$	3.1	99.6	56	101
$\beta_1^1 = 1, \beta_1^2 = \beta_1^3 = 1.1$	α_P^*	12.8	95.0	72	103
	α_{OBF}^*	13.4	98.4	72	102
	α_{LDM}^* $s = 0.5$	12.2	94.7	70	103
	α_{LDM}^* $s = 1$	12.8	95.7	70	103
	α_{LDM}^* $s = 2$	13.1	97.5	70	103
$\beta_1^1 = 1, \beta_1^2 = \beta_1^3 = 1.4$	α_P^*	41.4	88.9	80	119
	α_{OBF}^*	42.9	92.1	81	107
	α_{LDM}^* $s = 0.5$	40.0	80.8	78	109
	α_{LDM}^* $s = 1$	42.2	84.8	80	111
	α_{LDM}^* $s = 2$	42.6	88.9	81	109
$\beta_1^1 = 1, \beta_1^2 = \beta_1^3 = 1.6$	α_P^*	61.7	71.1	85	112
	α_{OBF}^*	64.0	87.8	87	115
	α_{LDM}^* $s = 0.5$	61.4	69.5	85	111
	α_{LDM}^* $s = 1$	62.7	73.8	86	113
	α_{LDM}^* $s = 2$	63.7	80.6	87	114
$\beta_1^1 = 1, \beta_1^2 = \beta_1^3 = 2$	α_P^*	91.8	49.6	96	109
	α_{OBF}^*	92.6	67.3	97	123
	α_{LDM}^* $s = 0.5$	92.6	52.0	97	116
	α_{LDM}^* $s = 1$	92.1	51.2	97	110
	α_{LDM}^* $s = 2$	92.6	57.7	97	116

Table 3.14: *Example 3.3.1; empirical power and proportion of observations used. The value R_γ is the ratio between empirical power and empirical power if scale parameters are known; R_{obs} is the ratio between observations used and observations used if scale parameters are known.*

To conclude, we consider the two-treatment case. In a simpler situation than that of model (1.3.5), which corresponds to D known, Jennison and Turnbull (1997a) recently proved that the sequence

$$(\theta\{k\}, \varsigma\{k\}) = \left(\hat{\beta}_1^2(k) - \hat{\beta}_1^1(k), \sigma^2 \sum_{h=1}^{2} \sum_{i=1}^{m^h(k)} U_i^h(k)' W_i^h(k) U_i^h(k) \right),$$

is Markov, and obtained conditional distributions of $(\theta\{k\}, \varsigma\{k\})$ given $(\theta\{k-1\}), \varsigma\{k-1\})$. In the above formula, W_i^h is defined by (1.3.4), and $U_i^h(k) = Y_i^h(k) - X_i^h(k)\hat{\beta}^h(k)$.

When the scale parameter σ^2 is unknown, their work settles the basis for group sequential t-tests (besides χ^2 and F-tests). Thus, such a development will allow for more accurate models for group sequential tests on small samples.

3.4 Procedure based on confidence intervals

3.4.1 Advantages of the confidence interval approach

Whitehead (1997) distinguishes different ways of using Repeated Confidence Intervals (RCIs). They could be used as a sequential design, that is, the experiment is stopped as soon as a Confidence Interval (CI) does not contain zero. In that case, interim coverage percentages must be planned in advance. Otherwise the procedure is much more flexible, but may be considered too subjective. On the other hand, we can compute RCIs without a stopping rule, like an indicator. It is then a control "on-line", which may be used at any time when data are being collected.

Jennison and Turnbull (1989) give a complete introduction to the RCI approach, with its application to different data. In the case of multiple comparison, the RCI approach is well illustrated by Liu, for the all-pairwise comparison case (1995), or when several treatments are compared to a control (1996). The context is however more restrictive than in Procedures 2.3.1 - 2.3.3. In fact, variances of the statistics used have to be equal at each interim step.

We now consider model (1.3.5). At a given interim step k, the program we have developed allows the user to compute simultaneous CIs for differences $\beta_1^l - \beta_1^h$, at level $1 - \pi(k)$, where $\pi(k)$ is given by (2.3.7). It is also possible to have them displayed, without considering the stopping rules of Procedures 2.3.1 to 2.3.3.

3.4.2 Illustration

Taking Example 1.3.1 on growth of chickens, we test the hypothesis $H_0 : \{\beta_1^1 = \beta_1^2\} \cap \{\beta_1^1 = \beta_1^3\}$ vs $H_1 : \{\beta_1^1 \neq \beta_1^2\} \cup \{\beta_1^1 \neq \beta_1^3\}$. We use the ESF α_C^* with $s = 2$, and we use observations as the time indicator. Three interim tests are planned at dates 10, 16, and 21 (days). According to Table 3.1, the experiment is stopped at $k = 2$ (10th day), and H_{03} is rejected.

The values of statistics $S^h(k)$, $k = 1, 2, 3$, $h = 1, 2$, are given in Table 3.15, as well as the corresponding CIs, and interim significance levels $\pi(k)$. We also give the marginal significance levels and p-values, computed with (2.4.3) and (2.4.2). The REML estimators of scale parameters are

$$\hat{\sigma}^2 = 78.60, \text{ and } \hat{D} = \begin{pmatrix} 0.71 & -0.26 \\ -0.26 & 0.10 \end{pmatrix}.$$

day	$S^{(2)}(k)$	CI	$\text{pv}_2(k)$	$S^{(3)}(k)$	CI	$\text{pv}_3(k)$	$\pi(k)$	p-value(k)
10	−1.30	[−3.55 ; +0.95]	0.116	−2.15	[−4.41 ; +0.10]	0.009	0.0125	0.0179
16	−1.24	[−4.27 ; +1.32]	0.290	−3.10	[−6.13 ; −0.07]	0.010	0.0094	0.0070
21	−1.66	[−4.64 ; +1.32]		−4.47	[−7.45 ; −1.50]		0.0281	

Table 3.15: *Example 1.3.1 of growth of chickens; statistics of difference of trends, RCIs and marginal significance levels* $\text{pv}_h(k)$, *and* p*-value.*

We shall add that this example is used for a purely illustrative purpose. The fourth treatment was not tested, because its average trend is far larger than the others. In fact, the mixed effects linear models (1.3.5) may be inadequate, since a convex movement is observed when residuals are plotted, and the corresponding variogram shows correlations of order 2 and 4. Crowder and Hand (1990) suggest another model, without mixed effects, but with a three-dimensional fixed effect (intercept, time, and time squared). It eventually does not turn out to be very satisfying either. Keeping model (1.3.5), and taking square roots of observations seems to give better results. The final conclusion is the same, viz. rejection of $H_0 : H_{02} \cap H_{03}$ at $t = 16$, but the inopportune phenomena described above are almost completely removed.

3.5 Conclusion

Generally speaking, Procedures 2.3.1 – 2.3.3 give satisfactory results and performance. Because the correlations over time amongst statistics are accurately allowed for in every situation, we obtained stable results, whatever the chosen ESF or time indicator. We wrote the corresponding program to deal with as many situations as possible, while allowing the user maximum flexibility regarding interim dates, hypotheses to test, time indicators, and choice of ESF.

A summary file is made for every analysis, from which an S routine immediately displays some features, including

- a plot of measurements by treatment,
- a plot of all measurements with mean by time, and the possibility to single out a given treatment;
- residual analysis: plot, boxplot, histogram, boxplot over time, and variogram; and
- a mixed effects boxplot.

This chapter cannot be concluded without mentioning that the theory of longitudinal data analysis was expanded to generalized linear models by Liang and Zeger (1986). For an example with mixed effects, see Zeger and Karim (1991).

Important simplifications can be made in the field of group sequential testing where statistics with independent increments can be used. Such a structure was obtained in the two-treatment case by Reboussin, Lan, and DeMets (1992), and then more generally by Reboussin (1995). In the field of generalized linear models, this is done by Gange and DeMets (1996), who suggest a Wald statistic, and Lee, Kim, and Tsiatis (1996), who use a score test or a Wald test. Both of these works result in the use of sequential χ^2-tests, and show that the ESF approach can be easily considered. An overview of the subject which covers general parametric regression and censored survival data can be found in Jennison and Turnbull (1997b).

To sum up, all the above methods use a statistic U with a structure like

$$\mathrm{cov}(U\{k\}, U\{k'\}) = \mathrm{var}(U\{k'\}) \text{ for } k' \leq k,$$

where k and k' are the numbers of the intermediate analyses. For the corresponding standardized statistics $U^*(k)$, that lead to

$$\mathrm{cov}(U^*\{k\}, U^*\{k'\}) = \sqrt{\mathrm{var}(U\{k\})/\mathrm{var}(U\{k'\})} \text{ for } k' \leq k.$$

In our framework we have

$$\mathrm{cov}(S\{k\}, S\{k'\}) = \mathrm{var}(S\{k\}) \text{ for } k' \leq k,$$

where $S(k)$ is defined by (2.3.4), if the same contrasts are tested at every step. Of course, we are always able to bring together the hypotheses so as to use a sequential χ^2-test; but in such a case, we fall back on a two step procedure, like Fisher Least Significant Difference; but we can hardly control the FWE strongly.

That is the reason why we have privileged the multiple-testing approach: the computation of boundaries is no longer a deterrent, due to a possible simplification, and today's computational availabilities and algorithms. It is actually surprising that relying on the algorithm of Schervish (1984) is still considered an obstacle, while far faster procedures exist.

Chapter 4

Allocation Rule

The ability to set adequate and ethical allocation rules is a great asset of sequential methods. The crucial importance of allocation rules is well illustrated by the lively discussions (Ware, 1989) about the ECMO (treatment of persistent pulmonary hypertension of newborns) experiment. In this chapter, we aim to introduce an allocation rule adapted to linear mixed effects models. Some standard sequential allocation rules are presented in Section 4.1, which also includes a group sequential procedure.

In a standard situation, a single measurement is made on every patient. But when we study longitudinal data, new patients join the experiment while some measures are still being taken on other patients who have already joined the trial. Section 4.2 introduces further hypotheses about the linear mixed effects model (1.3.2). These flexible hypotheses simplify the drawing up of the new allocation rule, which is introduced in Section 4.3.

We present results of simulation studies in Section 4.4. Finally, we propose some conjectures for generalization in Section 4.5.

4.1 Sequential allocation rules

We usually use an allocation rule when we have to assign a newcomer in the trial. But there are often two simultaneous targets:

- to single out the best treatment as soon as possible, and
- to reduce the ITN, i.e. the number of patients assigned to the inferior treatment.

A simple rule was proposed by Anscombe (1963):

Rule 4.1.1 *Assume that the experiment deals with N patients and two treatments.*

1. *As a first step, n patients are assigned to each treatment and we identify the best treatment, or so-called best treatment;*
2. *then, $N - 2n$ patients are assigned to the treatment which is looked on as best.*

4.1.1 Robbins and Siegmund's procedure

Robbins and Siegmund (1974) propose a fully sequential rule, which relies upon the SPRT (Sequential Probability Ratio Test).

Assumptions 4.1.1 *We assume the following context:*

- *there are two treatments;*
- *measures are normally distributed $\mathcal{N}(\mu_1, 1)$, and $\mathcal{N}(\mu_2, 1)$;*
- *a random variable X is observed for a patient assigned to treatment 1, respectively Y for treatment 2;*
- *we define $\delta = \mu_2 - \mu_1$, and $\theta = (\mu_2 + \mu_1)/2$; and*
- *we want to test $H_0 : \delta = -\delta^*$ vs $H_1 : \delta = \delta^*$.*

When we have observed $x_1, \ldots, x_m$, and $y_1, \ldots, y_n$, the procedure stops for (M, N), the first pair (m, n) for which $L_{m,n} \notin (B, A)$, with $0 < B < 1 < A$, and

$$L_{m,n} = e^{2\delta^* \frac{mn}{m+n}(\bar{y}_n - \bar{x}_m)}, \tag{4.1.1}$$

where $\bar{x}_m = \frac{1}{m}\sum_{i=1}^{m} x_i$, and $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$. For $\delta \neq 0$ we have

$$\mathrm{pr}_\delta(\text{ make a wrong decision }) \leq \left(1 + e^{-\frac{\delta \log B}{\delta^*}}\right)^{-1}. \tag{4.1.2}$$

The inequality in (4.1.2) may be replaced by an equality, if we are willing to neglect the boundaries' overshoot beyond A or B. Thus, the probability of making a wrong decision does not depend either on the allocation rule, or on θ.

In addition, if $b = -\log B/(2\delta^*)$,

$$\mathrm{E}_\delta\left(\frac{MN}{M+N}\right) \simeq \frac{b}{\delta}\left(\frac{e^{2b\delta} - 1}{e^{2b\delta} + 1}\right), \quad \text{if } \delta \neq 0,$$

$$\text{and} \quad \mathrm{E}_0\left(\frac{MN}{M+N}\right) \simeq b^2, \quad \text{if } \delta = 0.$$

Furthermore, we have

$$\mathrm{E}_\delta\,(M+N) \geq 4\mathrm{E}_\delta\left(\frac{MN}{M+N}\right),$$

in other words, the average number of patients is minimized if we use pairwise allocation. Moreover, under no circumstances can the ITN be reduced by more than fifty percent, in comparison with pairwise allocation.

Robbins and Siegmund propose a rather intuitive allocation rule, because it does not minimize any cost function.

Rule 4.1.2 *from Robbins and Siegmund:*

1. *choose* $c \geq b$;
2. *having observed* $x_1, \ldots, x_m$, *and* $y_1, \ldots, y_n$, *assign a new patient to treatment 1 (otherwise 2) if*

$$\frac{1}{c}\frac{mn}{m+n}(\bar{y}_n - \bar{x}_m) < \frac{n-m}{m+n}.$$

Some cost functions are introduced later by Siegmund (1985):

- a cost $g(\delta)$ for an allocation to treatment 1, and
- a cost $h(\delta)$ for an allocation to treatment 2, where

$$g(\delta) = h(-\delta) = \begin{cases} 1 + \frac{d}{\sqrt{2}}\delta & \text{if } \delta \geq 0, \\ 1 & \text{if } \delta < 0. \end{cases} \tag{4.1.3}$$

The expected cost is then

$$g(\delta)\mathrm{E}_\delta[M] + h(\delta)\mathrm{E}_\delta[N]. \tag{4.1.4}$$

It may seem logical to add a cost due to a wrong decision. However, it is shown from (4.1.2) that power and Type I error are essentially invariant to the choice of the allocation rule. Therefore, the minimization of cost reduces to minimization of (4.1.4).

Rule 4.1.3 *of Siegmund (1985):*

1. *choose g and h, that is, d in (4.1.3),*
2. *having observed $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, assign a new patient to treatment 1 (otherwise to treatment 2), if*

$$\frac{m}{n} < \sqrt{\frac{h(\bar{y}_n - \bar{x}_m)}{g(\bar{y}_n - \bar{x}_m)}}.$$

For large samples, we have

$$\mathrm{E}_\delta(M) \simeq (1 + \sqrt{h/g})\mathrm{E}_\delta\left(\frac{MN}{M+N}\right), \quad \text{and} \quad \mathrm{E}_\delta(N) \simeq (1 + \sqrt{g/h})\mathrm{E}_\delta\left(\frac{MN}{M+N}\right).$$

Thus, the expected cost (4.1.4) reaches its lower bound, which equals

$$\left(\sqrt{h} + \sqrt{g}\right)^2 \mathrm{E}_\delta\left(\frac{MN}{M+N}\right).$$

Using a few simulations, Siegmund (1985) showed the merits of this simple rule, but he went on to point out that results are very sensitive to the choice of d in (4.1.3).

4.1.2 Procedure of Louis

Louis (1975) introduces also cost functions $g(\delta)$ and $h(\delta)$, as Siegmund's procedure does. He defines them by

$$g(\delta) = h(-\delta) = \begin{cases} \gamma \text{ if } \delta \geq 0, \\ 1 \text{ if } \delta < 0, \end{cases}$$

where $\gamma > 1$. The procedure stops for (M, N), the first pair (m, n) such that $L_{m,n} \notin \left(\frac{1}{A}, A\right)$, with $1 < A$ and $L_{m,n}$ defined by (4.1.1). Hence, the expected cost is

$$\mathrm{E}_\delta(N + M) + (\gamma - 1)\mathrm{E}_\delta\left(N\,\mathbf{I}_{\{\delta<0\}} + M\,\mathbf{I}_{\{\delta>0\}}\right). \tag{4.1.5}$$

Rule 4.1.4 *Method of Louis. Every new individual is assigned to*

$$\begin{cases} \text{treatment 1 if } \frac{m}{m+n} < q_\gamma(L_{m,n}), \\ \text{treatment 1 with probability } q_\gamma(L_{m,n}) \text{ if } \frac{m}{m+n} = q_\gamma(L_{m,n}) \text{ (2 otherwise), and} \\ \text{treatment 2 if } \frac{m}{m+n} > q_\gamma(L_{m,n}), \end{cases}$$

where $L_{m,n}$ *is given by (4.1.1),* $\gamma > 1$, *and* $q_\gamma(L_{m,n}) = \frac{\sqrt{\gamma L_{m,n}+1}}{\sqrt{\gamma L_{m,n}+1}+\sqrt{L_{m,n}+\gamma}}$.

Rule 4.1.4 is asymptotically optimal, for continuous time, and for the criterion (4.1.5), that is, results obtained are better than those derived from Rule 4.1.2. This behaviour is of course predictable, since the rule introduced by Louis is specially designed to minimize the expected cost defined above.

4.1.3 Further comments

Chèvre (1992) compares Rules 4.1.2, 4.1.4, and another rule, proposed by Zoubeidi (1989), which uses a Bayesian approach. If we keep the same Type I error rate, we are not surprised to note that none of these allocation rules is uniformly best.

However, it is more surprising to find that the rule proposed by Robbins and Siegmund (1974), which is in fact the most intuitive, often gives the best results: if Type I and Type II errors are kept constant, Rule 4.1.2 often minimizes the ASN (Average Sample Number) and the ITN. Indeed, unlike the rule proposed by Zoubeidi (1989) or Rule 4.1.4, Rule 4.1.2 does not satisfy any precise optimality criterion.

This eventually allows one to conclude that the chosen cost function is not of primary importance when trying to set up an allocation rule.

4.1.4 Grouped data

We consider the situation of Assumptions 1.4.1, with $\sigma^2 = 1$, $\delta = \mu_1 - \mu_2$, and the standardized statistic $(in/2)^{-1/2}\bar{d}_i$, where $\bar{d}_i$ is defined by (1.4.1). If we use an allocation rule, this statistic becomes

$$\tilde{d}_i = \sqrt{\frac{2np_i(1-p_i)}{i}} \sum_{j=1}^{i} (\bar{x}_j - \bar{y}_j),$$

$$\text{hence } \tilde{d}_i \sim \mathcal{N}(\delta\sqrt{\frac{in}{2}}, 1),$$

where p_i, $0 \leq p_i \leq 1$, is a random variable which belongs to the sigma-algebra generated by the whole set of observations available until step $i-1$. This value stands for the proportion of patients who joined the trial at step i, and were assigned to the first treatment.

Zoubeidi (1996) shows that under $H_0 : \delta = 0$, vectors $\left[(n/2)^{-1/2}\bar{d}_1, \ldots, (in/2)^{-1/2}\bar{d}_i\right]'$ and $\left(\tilde{d}_1, \ldots, \tilde{d}_i\right)'$ are identically and normally distributed. Thus, the procedures of

Pocock (1977), O'Brien and Fleming (1979), or Falissard and Lellouch (1992) may be used.

Without loss of generality, we assume that a "good" treatment corresponds to a high value of the location parameter, and vice versa for a "bad" treatment. We define a non-increasing function c, for which

$$c(\delta) \equiv \bar{c}, \quad \forall \delta : |\delta| < \epsilon.$$

Definition 4.1.1 *When we use an allocation rule $\mathcal{A}$, the random variable $\tau(\mathcal{A})$ represents the step number at which the experiment is stopped.*

The value of ϵ stands for the width of the indifference zone, and c is the cost of one measurement of treatment 1. Then, the expected sampling cost is

$$R(\mathcal{A}) = \mathrm{E}\left[2n \sum_{i=1}^{\tau(\mathcal{A})} c(\delta)p_i(\mathcal{A}) + c(-\delta)(1 - p_i(\mathcal{A})) + g(i)c(-|\delta|)\right], \tag{4.1.6}$$

where $g(x) \geq x - 1$, and $p_i(\mathcal{A})$ is the percentage of patients assigned to the first treatment. Such a cost is drawn up to favour the early stopping of the procedure, since the function $g(\cdot)$ penalizes a continuation of the experiment beyond step i. The cost brought about by every allocation which stops at step i is in fact always larger than any other allocation which would stop at step $i-1$. In addition, the term $c(\delta)p_i(\mathcal{A})+c(-\delta)[1-p_i(\mathcal{A})]$ forces the allocation rule to assign fewer patients to the inferior treatment. We say that an allocation rule $\mathcal{A}^*$ is asymptotically efficient if we have,

$$\lim_{n\to\infty} \frac{R(\mathcal{A}^*)}{\min_{\mathcal{A}} R(\mathcal{A})} = 1.$$

Rule 4.1.5 *When $n \to \infty$, an asymptotically efficient procedure (Zoubeidi, 1996) for the expected cost (4.1.6) is given by the following steps:*

1. *choose c, ϵ and the Type I error α;*
2. *stop the experiment, and decide H_1 at step i, if $|\tilde{d}_i| > \mathrm{b}(i, \alpha)$, where $\mathrm{b}(i, \alpha)$ is a suitable boundary (e.g. Pocock, or O'Brien and Fleming's method);*
3. *otherwise, assign $2np_i$ new patients to treatment 1, and $2n(1 - p_i)$ to treatment 2, with*

$$p_j = \begin{cases} \frac{1}{2} - \frac{1}{2}\sqrt{1 - \min(1, 4a)} & \text{if } \hat{\delta}_{i-1} \leq 0, \\ \frac{1}{2} + \frac{1}{2}\sqrt{1 - \min(1, 4a)} & \text{otherwise .} \end{cases}$$

The values $\hat{\delta}_{i-1}$ and a are defined by

$$\hat{\delta}_{i-1} = \frac{1}{i-1}\sum_{j=1}^{i-1}[p_j\bar{x}_j - (1-p_j)\bar{y}_j], \text{ and}$$

$$a = \frac{1}{2n}\left[\frac{\sqrt{i}\mathrm{b}(i,\alpha) - |\sum_{j=1}^{i-1}\sqrt{2np_j(1-p_j)}(\bar{x}_j - \bar{y}_j)|}{\hat{\delta}_{i-1}}\right]^2.$$

Simulations show that Rule 4.1.5 attains its objectives (especially reduction of the ITN). Of course, the price to pay, with respect to a pairwise allocation, is a slight increase in the total number of patients. If several error spending functions are used, we may consider the criterion RD, where

$$RD = \frac{\text{difference of numbers of patients between the two groups}}{\text{total number of patients}}. \qquad (4.1.7)$$

It turns out that the procedure of Pocock (1977) gives better results than the procedure of O'Brien and Fleming (1979), and that the total number of patients is smaller. However, power is also slightly reduced, by 1.5% for $\delta = 0.5$ up to 2.5% for $\delta = 0.1$.

4.1.5 Conclusion

A fairly complete list of references about sequential allocation rules is found in the chapter written by A. Basu, A. Bose, and J.K. Ghosh, in the handbook by Ghosh and Sen (1991).

Sequential allocation rules may have several objectives :

- to reduce experimenter bias (see for example Wei, 1978),
- to obtain confidence intervals of fixed width (Eisele, 1994), or
- to aim at a good randomization between different strata (Zoubeidi, 1994).

In this chapter, we will concentrate on the ability of singling out the best treatment as soon as possible, while reducing the number of patients assigned to the inferior treatment.

4.2 Longitudinal data

The allocation that we will present comes within the framework defined here. In (1.3.2), we assume that

$$X_i^h = Z_i^h, \ h = 1, \ldots, T, \text{ and } i = 1, \ldots, m^h.$$

For simplifying notation, we write

$$\begin{aligned} X^h(k) &= (X_1^h(k)', \ldots, X_{m^h(k)}^h(k)')', \\ Y^h(k) &= (Y_1^h(k)', \ldots, Y_{m^h(k)}^h(k)')', \\ V^h(k) &= \text{diag}(V_1^h(k), \ldots, V_{m^h(k)}^h(k)), \text{ and} \\ V^h(k) &= W^h(k)^{-1}. \end{aligned}$$

We suppose, without loss of generality, that measurements are taken at times $t = 0, 1, 2 \ldots$, and that every measure is taken with probability $\lambda > 0$. The event "the measure is taken" is assumed to be completely independent from any other variable connected with the experiment. Thus, the matrix $X_i^h(k)$ consists of the non-null rows of

$$T_i^h(k) = \text{diag}\left(\eta_{i,0}^h, \ldots, \eta_{i,n_{\max}(i,h,k)}^h\right) \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & n_{\max}(i,h,k) \end{pmatrix}, \tag{4.2.1}$$

where the independent random variables $\eta_{i,1}^h, \ldots, \eta_{i,n_{\max}(i,h,k)}^h$ are Bernoulli $\mathcal{B}(\lambda)$, and $n_{\max}(i,h,k)+1$ is the maximum number of observations that could be taken on patient i assigned to treatment h up to step k. The right hand side matrix of (4.2.1) equals the design matrix when $\lambda = 1$. It is also assumed that every patient for whom a single measurement is available is withdrawn from the analysis.

To sum up, two constraints have been added to (1.3.2). First, the matrix that connects observations and the random effect of every patient equals the design matrix. In practice, Lindstrom and Bates (1988) note that the matrix Z_i^h is often a subset of the columns of X_i^h. This allows us to conclude that this restriction $Z_i^h = X_i^h$ is not too constraining. Secondly, the structure of X_i^h is also a little more restrictive. We should also bear in mind that in practice, λ may be affected by the treatment group, that is $\lambda = \lambda_h$.

Since we deal with missing at random variables, and observed at random variables, Theorem 6.1 of Rubin (1976) allows us to ignore the process which has caused the absence of a certain number of measures. In such a case, the distribution of any statistic

based on the observed measurements has indeed the same distribution as its conditional distribution, given the $\eta_{i,j}^h$.

The above assumptions lead to the following propositions.

Proposition 4.2.1 *For all $h = 1, \ldots, T$, and for all $k = 1, \ldots, K$, we have*

$$\left[X^h(k)'W^h(k)X^h(k)\right]^{-1} = \sigma^2[(X^h\{k\}'X^h\{k\})^{-1} + D],$$

where the scale parameters σ^2 and D are defined in Section 1.3.2.

□ Proof: We omit the index k. Since $X^h = Z^h$, we have

$$\begin{aligned}\left(\sigma^2 X^{h'}W^hX^h\right)^{-1} &= \left(I - D\left[D + DX^{h'}X^hD\right]^{-1}DX^{h'}X^h\right)^{-1}(X^{h'}X^h)^{-1} \\ &= (X^{h'}X^h)^{-1} - D.\end{aligned}$$

■

Proposition 4.2.2 *For all $h = 1, \ldots, T$, $k = 1, \ldots, K$, $i = 1, \ldots, m^h(k)$, and for all $\lambda > 0$, if $n_{\max}(i,h,k) \to \infty$, then*

$$X_i^h(k)'W_i^h(k)X_i^h(k) \xrightarrow{\mathcal{L}^p \text{ and a.s.}} \sigma^{-2}D^{-1}.$$

□ Proof of Proposition 4.2.2: indices i, h and k are omitted.

$$T_i^h(k)'T_i^h(k) = T'T = \begin{pmatrix} \sum_{j=0}^{n_{\max}} \eta_j & \sum_{j=0}^{n_{\max}} j\eta_j \\ \sum_{j=0}^{n_{\max}} j\eta_j & \sum_{j=0}^{n_{\max}} j^2\eta_j \end{pmatrix}.$$

Then,

$$\| T'T \|_1 = \sum_{j=0}^{n_{\max}} (j + j^2)\eta_j \geq \sum_{j=0}^{n_{\max}} \eta_j,$$

where $\| A_{(n,n)} \|_1 = \max\{\sum_{i=1}^n |A_{[i,j]}| \,;\, 1 \leq j \leq n\}$. By the strong law of large numbers, we have

$$\begin{aligned}\lim_{n_{\max}\to\infty} \frac{\| X'X \|_1}{n_{\max}} &\geq \lambda \text{ a.s. and in } \mathcal{L}^p, \text{ hence} \\ \lim_{n_{\max}\to\infty} \| (X'X)^{-1} \|_1 &= 0 \text{ a.s. and in } \mathcal{L}^p.\end{aligned}$$

It follows that

$$\begin{aligned}\sigma^2 X'WX &= X'X[I - D(I + X'XD)^{-1}X'X] \\ &= D^{-1} - D^{-1}(X'X)^{-1}D^{-1} + o\left(D^{-1}(X'X)^{-1}D^{-1}\right),\end{aligned}$$

where $o(\cdot)$ is such that when a given matrix A_n equals $o(B_n)$,
then $\lim_{n\to\infty} \| A_n \|_1 / \| B_n \|_1 = 0$ a.s. and in $\mathcal{L}^p$. So we have $\sigma^2 X'WX \overset{\mathcal{L}^p \text{ and a.s.}}{\longrightarrow} D^{-1}$.
■

How fast can information about a single individual be gained? Let us take

$$D = d\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{4.2.2}$$

where d represent the importance of individual effects if they are compared with the standard errors. We give here the percentage of information about an individual i, when it is compared with its limit, which is given by Proposition 4.2.2. This percentage was computed with $\rho = 0.1, 0.2, 0.5, 0.8$, and 0.9, $d = 0.2, 0.5, 1, 2, 5, 10$, and 50 and a design matrix with $2, 3, 4, 5, 10, 20, 30, 40, 50$, and 100 rows. An extract is displayed in Table 4.1.

$n_{\max}$	d						
	0.2	0.5	1	2	5	10	50
2	27.5	39.3	50.0	62.2	77.9	86.9	96.9
3	42.2	53.7	63.8	74.3	86.2	92.2	98.3
4	52.2	62.7	71.9	80.9	90.2	94.6	98.8
5	58.9	68.7	77.0	84.9	92.6	96.0	99.1
10	73.8	81.9	88.0	92.8	96.7	98.3	99.6
20	82.8	89.6	93.8	96.5	98.5	99.2	99.8
30	86.7	92.6	95.8	97.7	99.0	99.5	99.9
40	89.0	94.2	96.8	98.3	99.3	99.6	99.9
50	90.6	95.3	97.4	98.6	99.4	99.7	99.9
100	94.5	97.5	98.7	99.3	99.7	99.9	100.0

Table 4.1: *Percentage of information (when $\rho = 0.5$) supplied by an individual, when compared with its limit given by Proposition 4.2.2.*

Unlike for ρ, large values for d and $n_{\max}$ increase the speed of convergence. In a standard case, for which $\rho = 0.5$ and $d = 1$, we note in Table 4.1 that eleven observations are enough to supply 88% of information about an individual. Thus, we will consider that between two intermediate steps, almost complete information is available about

the individuals previously involved in the experiment. However, this implies that measurements are frequent compared to interim analyses, especially when the experiment begins. One must be aware that there are situations where this is not true.

The effect of λ on the results mentioned above is also studied. Taking the same values for ρ, d, and $n_{\max}$ as we did above, we made 10,000 simulations with $\lambda = 0.3, 0.5, 0.8$, and 0.9. An excerpt is displayed in Table 4.2, when $\lambda = \rho = 0.5$. Even if measures are taken with probability 0.5, only 20% of information is lost in the worst situation. This allows wide use of Proposition 4.2.2, whenever there are missing values in the data.

d	0.2	%	0.5	%	1	%	2	%	5	%	10	%
$n_{\max}$												
2	0.22	82	0.32	82	0.41	82	0.51	82	0.66	85	0.77	89
3	0.34	80	0.44	81	0.52	81	0.61	82	0.74	85	0.82	89
5	0.50	84	0.58	84	0.65	84	0.73	85	0.82	89	0.88	92
10	0.68	91	0.74	90	0.80	91	0.86	92	0.92	95	0.95	99
20	0.77	94	0.84	93	0.89	95	0.93	96	0.97	98	0.98	99
30	0.81	94	0.88	95	0.92	96	0.95	98	0.98	99	0.99	99
50	0.86	95	0.92	96	0.95	98	0.97	99	0.99	99	0.99	100

Table 4.2: *Ratio of information available to the asymptotic value ($n_{\max} \to \infty$) when $\lambda = 0.5$ and $\rho = 0.5$; columns of percentages display information available when information available is compared to the case $\lambda = 1$.*

4.3 Allocation rule for two treatments

4.3.1 Scope

In this section, we present an asymptotically efficient allocation rule for the two-treatment case, in linear mixed effects models. This allocation rule has the following features:

- it essentially depends on the parameters of interest, that is, the difference of trends, rather than the trends themselves;
- the Type I error is controlled; and
- performance should be optimal or essentially optimal for a given cost function.

In a longitudinal data analysis, new information collected at step k results from two sources:

- patients who joined the trial at step $k-1$ (like the standard case), and
- patients who had already joined the trial before step $k-1$.

Using Proposition 4.2.2, we can nevertheless argue that once a given number of measures is taken on a patient, additional measurements become almost non-informative. If the design is regular enough, i.e. λ is close to one, we will therefore assume that complete information is available for every patient, whenever he or she joined the trial.

This assumption is well founded, especially if there is a small number of interim analyses. It just emphasizes that the important quantity to deal with in longitudinal data analysis is the number of patients, rather than the number of observations.

In the two-treatment case, hypotheses (2.3.1) and (2.3.2) reduce to

$$H_0 : \beta_1^2 = \beta_1^1 \text{ vs } H_1 : \beta_1^2 \neq \beta_1^1.$$

4.3.2 Statistics used

In Chapters 2 and 3, we use the statistic

$$S(k) = \hat{\beta}_1^1(k) - \hat{\beta}_1^2(k). \tag{4.3.1}$$

However, when an allocation rule is used, the number of patients in both groups, $m^h(k)$, $h = 1, 2$, is a random variable. Hereafter, we shall write

$$\tilde{S}(k) = \hat{\beta}_1^1(k) - \hat{\beta}_1^2(k),$$

so as to make a clear distinction. When the trial begins, $m^1(0)$ are assigned to treatment 1, and $m^2(0)$ to treatment 2. Values of $m^1(0)$ and $m^2(0)$ are fixed, or are possibly random variables, but independent from the observations. We usually take $m^1(0) = m^2(0)$. When the first step is reached,

$$m_+^h(1) \text{ patients join the trial, and are assigned to treatment } h,$$

so that $m^h(1) = m^h(0) + m_+^h(1)$, $h = 1, 2$. We also write

$$\begin{aligned} \text{var}\,(S\{k\}) &= (0\ 1)\left[\left(X^1\{k\}'W^1\{k\}X^1\{k\}\right)^{-1} + \left(X^2\{k\}'W^2\{k\}X^2\{k\}\right)^{-1}\right](0\ 1)', \\ &= L^1(k) + L^2(k), \\ &= G(k). \end{aligned}$$

Definition 4.3.1 *We denote by $\mathcal{F}_k$ the sigma-algebra generated by*

$$\{\tilde{S}(l) : l = 1, \ldots, k\}.$$

Lemma 4.3.1 *Distribution of $\tilde{S}$.*

- *The random variable $\tilde{S}(k)$, given $\mathcal{F}_{k-1}$, is normally distributed.*
- *Under H_0: $\beta_1^1 = \beta_1^2$, we have*

$$\mathrm{E}\left(\tilde{S}(k) \mid \mathcal{F}_{k-1}\right) = \frac{G(k)}{G(k-1)}\tilde{S}(k-1), \text{ and}$$

$$\mathrm{var}\left(\tilde{S}(k) \mid \mathcal{F}_{k-1}\right) = \left[1 - \frac{G(k)}{G(k-1)}\right] G(k).$$

□ Proof:
We use Theorem 2.1 of Seber (1984, pp. 18–19), which deals with the conditional distribution of multi-normal random variables.
■

Proposition 4.3.1 *We define*

$$S^*(1) = \left[\mathrm{var}(\tilde{S}\{1\})\right]^{-\frac{1}{2}} \left[\tilde{S}(1) - \mathrm{E}[\tilde{S}(1)]\right] \text{ and },$$

$$S^*(k) = \left[\mathrm{var}(\tilde{S}(k)|\mathcal{F}_{k-1})\right]^{-\frac{1}{2}} \left[\tilde{S}(k) - \mathrm{E}\left(\tilde{S}(k)|\mathcal{F}_{k-1}\right)\right].$$

Then, under H_0: $\beta_1^1 = \beta_1^2$, $(S^*\{1\}, \ldots, S^*\{k\})'$ is normally distributed, $\mathcal{N}_k(0, I)$.

□ Proof:
Since $m^h(0)$, $h = 1, 2$, are fixed, $S^*(1)$ is a zero-mean standardized normal random variable. Then, assume that the result is right for $k-1$, and let us take $(t_1, \ldots, t_k) \in \mathbb{R}^k$. The characteristic function of $(S^*(1), \ldots, S^*(k))$ is

$$\begin{aligned}
\mathrm{E}\left[e^{i\sum_{l=1}^{k} t_l S^*(l)}\right] &= \mathrm{E}\left[e^{i\sum_{l=1}^{k-1} t_l S^*(l)} \mathrm{E}\left(e^{it_k S^*(k)}|\mathcal{F}_{k-1}\right)\right] \\
&= \mathrm{E}\left[e^{i\sum_{l=1}^{k-1} t_l S^*(l)}\right] e^{-\frac{1}{2}t_k^2} \quad \text{by Lemma 4.3.1 ,} \\
&= e^{-\frac{1}{2}\sum_{l=1}^{k} t_l^2} \quad \text{by induction.}
\end{aligned}$$

■

Since the distribution of $S^*(1), \ldots, S^*(k)$ is independent of the allocation rule, we are able to use Pocock or O'Brien and Fleming's method, or any spending rate function. Indeed, the corresponding critical values will not be affected by the introduction of the allocation rule.

4.3.3 Sampling costs

A good allocation rule should essentially depend on the difference of trends, rather than the trends themselves.

Hence, we define

$$\delta = \beta_1^1 - \beta_1^2,$$

and assume that the larger the trend β_1^h, the better the treatment. Then, it seems natural to define the following sampling cost:

Definition 4.3.2 *We use a cost function $d(\cdot)$. This function is continuous, non-decreasing, and such that $d(\delta) \equiv \bar{d}$, for all δ : $|\delta| < \epsilon$. Thus, assigning a new patient to treatment 2 has a cost $d(\delta)$.*

Let τ denote the last interim analysis. For any allocation rule $\mathcal{A}$, we define the expected sampling costs

$$R(\mathcal{A}) = \mathrm{E}\left[\sum_{k=1}^{\tau(\mathcal{A})} r(k-1, \mathcal{A})\right], \tag{4.3.2}$$

where $r(k, \mathcal{A}) = d(\delta)[m^2(k) - m^2(k-1)] + d(-\delta)[m^1(k) - m^1(k-1)]$, and $r(0) = d(\delta)m^2(0) + d(-\delta)m^1(0)$.

4.3.4 The allocation problem

Having defined the expected sampling cost, we want to derive an asymptotically efficient allocation which belongs to the general class defined below.

Definition 4.3.3 *The class $\mathcal{C}_a$ is the class of allocation rules satisfying the following properties:*

- *at any step k, the proportion of patients assigned to every treatment is at least ζ, a fixed value $0 < \zeta \leq \frac{1}{2}$, set before the trial begins;*
- *the maximum number of interim analyses is K.*

Assumptions 4.3.1 *We assume that the trial is such that*

- $m_0 = m^1(0) = m^2(0)$ *patients are initially assigned to each treatment;*

- *at step k, $m_+^1(k) + m_+^2(k) = 2m$ new patients join the trial, of whom $m_+^1(k) = 2mp(k)$ are assigned to treatment 1, $m_+^2(k) = 2m[1 - p(k)]$ are assigned to treatment 2; and*
- *$n_i^h(k)$, the number of measures on patient i, who was assigned to treatment h, is written*

$$n_i^h(k) = g_i^h(k)n,$$

where the number of measurements n, is the maximum number of measures that we are able to take on a patient, and $0 \leq g_i^h(1) \leq \cdots \leq g_i^h(K) \leq 1$.

The first question that should be answered is when the trial should stop if there is a significant difference between the two treatments. Lemma 4.3.2 shows that asymptotically, the stopping time $\tau(\mathcal{A})$ should be smaller than or equal two, regardless of the timing of analysis.

Lemma 4.3.2 *For all $\mathcal{A} \in \mathcal{C}_a$, for all δ such that $|\delta| > \epsilon$, we have*

$$\lim_{m\to\infty, n\to\infty} \mathbf{1}_{\{\tau(\mathcal{A})\leq 2\}} = 1 \text{ in probability.}$$

□ Proof: to simplify notation, we write

$$\begin{aligned} v &= \sigma^{-2}\left(D_{[2,2]}\right)^{-1}, \\ p(2) &= p, \text{ and} \end{aligned}$$

we suppose that $n_{\max}(i, h, 1)$ is large. At step 2, we use the statistic

$$S^*(2) = \frac{1}{\sqrt{\frac{1}{G(2)} + \frac{1}{G(1)}}} \left(\frac{\tilde{S}(2)}{G(2)} - \frac{S(1)}{G(1)}\right).$$

By Propositions 4.2.2 and 4.2.1, we know that $S^*(2)$ converges almost surely to

$$\left[C - \frac{vm_0}{2}\right]^{-1/2} \left[C\tilde{S}^2(2) - \frac{vm_0}{2}S^2(1)\right], \tag{4.3.3}$$

$$\text{where } C = \frac{v(m_0 + 2mp)(m_0 + 2m(1-p))}{2m_0 + 2m}.$$

The expansion of $S^*(2)^2$ in m leads to

$$\begin{aligned} S^*(2)^2 &= 2vmp(1-p)\tilde{S}(2)^2 + v\left(-2m_0p(1-p) + \frac{3}{2}m_0\right)[\tilde{S}(2)]^2 \\ &\quad -vm_0\tilde{S}(2)S(1) + O\left(1/m\right), \\ \frac{S^*(2)^2}{m} &= 2vp(1-p)[\tilde{S}(2)]^2 + O(1/m) \text{ almost surely.} \end{aligned} \tag{4.3.4}$$

Because v is in fact unknown, it is replaced by a consistent estimator. In addition, $\tilde{S}(2)$ converges to δ in probability, because $\tilde{S}(2)$ is a consistent estimator of δ (Lee and DeMets, 1991). Since every continuous function g satisfies $g(X_n, Y_n) \xrightarrow{p} g(X, Y)$, if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ (Chow and Teicher, 1988, p. 73), we have

$$\lim_{m\to\infty, n\to\infty} \frac{S^*(2)^2}{m} \geq 2v\zeta(1-\zeta)\delta^2 > 0 \text{ in probability.}$$

Statistic $|S^*(2)|$ is compared with Q_2, where Q_2 is the $1 - \pi_2/2$ quantile of a random variable $\mathcal{N}(0,1)$, and π_2 is the interim significance level at step 2. Moreover, Q_2 is independent of ζ and δ. So, we have

$$\forall \pi_2 \, \exists m' : \forall m > m' \; S^*(2)^2 > Q_2^2 \text{ in probability .}$$

■

When there is a significant difference between the two treatments, experiments should stop at the same step, asymptotically. This is the purpose of the following lemma:

Lemma 4.3.3 *For all $\mathcal{A} \in \mathcal{C}_a$, for all $\delta : |\delta| > \epsilon$,*

$$\lim_{m\to\infty, n\to\infty} \mathbf{1}_{\{\tau(\mathcal{A})=\tau(\mathcal{A}_p)\}} = 1 \textit{ in probability,}$$

where $\mathcal{A}_p$ is the pairwise allocation rule.

□ Proof: by Lemma 4.3.2, for all $\mathcal{A} \in \mathcal{C}_a$, for all δ such that $|\delta| > \epsilon$, $\lim_{m\to\infty, n\to\infty} \mathbf{1}_{\{\tau(\mathcal{A})\leq 2\}} = 1$ in probability; hence,

$$\lim_{m\to\infty, n\to\infty} \mathbf{1}_{\{\tau(\mathcal{A})\leq 2\}} = \lim_{m\to\infty, n\to\infty} \mathbf{1}_{\{\tau(\mathcal{A}_p)\leq 2\}}.$$

Since the equality $\mathbf{1}_{\{\tau(\mathcal{A}_p)=1\}} = \mathbf{1}_{\{\tau(\mathcal{A})=1\}}$ holds — all the allocation rules in $\mathcal{A}$ are pairwise when the experiment starts — we deduce that the limits exist, and

$$\lim_{m\to\infty, n\to\infty} \mathbf{1}_{\{\tau(\mathcal{A})=i\}} = \lim_{m\to\infty, n\to\infty} \mathbf{1}_{\{\tau(\mathcal{A}_p)=i\}}, \; i = 1, 2, \text{ thus}$$

$$\lim_{m\to\infty, n\to\infty} \mathbf{1}_{\{\tau(\mathcal{A})=\tau(\mathcal{A}_p)\}} = 1 \text{ in probability.}$$

■

On the contrary, when there is no significant difference between treatments no allocation rule in $\mathcal{A}$ should stop before the planned end of the experiment. As the previous one, this property is asymptotic.

Lemma 4.3.4 *For ϵ small enough, for all $\mathcal{A} \in \mathcal{C}_a$, and for all $\delta : |\delta| < \epsilon$, we have for $k = 2, \ldots, K-1$,*

- $\lim_{m\to\infty, n\to\infty} \mathbb{1}_{\{\tau(\mathcal{A})=k\}} = \lim_{m\to\infty, n\to\infty} \mathbb{1}_{\{\tau(\mathcal{A}_p)=k\}} = 0$ *in probability, and*
- $\lim_{m\to\infty, n\to\infty} \mathbb{1}_{\{\tau(\mathcal{A})\geq K\}} = \lim_{m\to\infty, n\to\infty} \mathbb{1}_{\{\tau(\mathcal{A}_p)\geq K\}}$ *in probability.*

□ Proof: we choose $\epsilon < \sqrt{\frac{K}{mv(2K-1)}}\, q_{z_{1-\alpha/2}}$, where $q_{z_{1-\alpha/2}}$ is the $1-\alpha/2$ quantile of a random variable $\mathcal{N}(0,1)$. Under H_0, by Proposition 4.3.1, the statistics $S^*(k)$, $k = 1, \ldots, K$, are independent and normally distributed, with mean zero and variance one; thus, the boundaries Q_k for $|S^*(k)|$ are such that $q_{z_{1-\frac{\alpha}{2}}} \leq Q_k$, $k = 1, \ldots, K$. The value of Q_k depends on the chosen error spending function. Since we take an allocation rule in $\mathcal{C}_a$, we have obviously

$$\mathbb{1}_{\{\tau(\mathcal{A})=1\}} = \mathbb{1}_{\{\tau(\mathcal{A}_p)=1\}}.$$

At step 2, we use (4.3.3), and we obtain

$$\begin{aligned}
\lim_{m\to\infty, n\to\infty} S^*(2)^2 &= \lim_{m\to\infty, n\to\infty} (2vmp\{1-p\})\, \tilde{S}(2)^2 \text{ in probability,} \\
&\leq \lim_{m\to\infty, n\to\infty} \left(\frac{1}{2}vm\right) \epsilon^2 \text{ in probability,} \\
&\leq q^2_{z_{1-\frac{\alpha}{2}}} \text{ in probability.}
\end{aligned}$$

At step $k = 3, \ldots K$, if n and m tend to infinity, $\tilde{S}^2(k+1)$ and $\tilde{S}^2(k)$ converge in probability to δ; hence,

$$\lim_{m\to\infty, n\to\infty} S^*(k+1)^2 = \lim_{m\to\infty, n\to\infty} \left(\frac{1}{G(k+1)} - \frac{1}{G(k)}\right) \tilde{S}^2(k)^2 = \lim_{m\to\infty, n\to\infty}$$
$$v\left[\frac{(m_0+2mkp_k)(m_0+2mkq_k)}{2m_0+2mk} - \frac{(m_0+2m(k-1)p_{k-1})(m_0+2m(k-1)q_{k-1})}{2m_0+2m(k-1)}\right]\delta^2$$

in probability, where $p_k = 1-q_k$, is such that $m_0 + 2mkp_k = m^1(k+1)$, and $p_{k-1} = 1-q_{k-1}$, is such that $m_0 + 2m(k-1)p_{k-1} = m^1(k)$. If we rewrite p_k and q_k as $p_k = [(k-1)p_{k-1} + p^*]/k$, and $q_k = [(k-1)q_{k-1} + q^*]/k$, we have in probability

$$\begin{aligned}
\lim_{m\to\infty, n\to\infty} S^*(k+1)^2 &= \frac{2vm}{k}\left[-(k-1)p_{k-1}q_{k-1} + (k-1)(p^*q_{k-1} + p_{k-1}q^*) + p^*q^*\right]\delta^2, \\
&\leq vm\left[\frac{2K-1}{K}\right]\delta^2 \leq q^2_{z_{1-\frac{\alpha}{2}}} \leq Q^2_{k+1}.
\end{aligned}$$

We conclude that

$$\lim_{m\to\infty,n\to\infty} \mathbf{I}_{\{\tau(\mathcal{A})=k\}} = \lim_{m\to\infty,n\to\infty} \mathbf{I}_{\{\tau(\mathcal{A}_p)=k\}} = 0, \quad \forall k = 2,\dots,K-1,$$

and $\mathbf{I}_{\{\tau(\mathcal{A})=1\}} = \mathbf{I}_{\{\tau(\mathcal{A}_p)=1\}}$; finally

$$\lim_{m\to\infty,n\to\infty} \mathbf{I}_{\{\tau(\mathcal{A})\geq K\}} = \lim_{m\to\infty,n\to\infty} \mathbf{I}_{\{\tau(\mathcal{A}_p)\geq K\}}.$$

■

Finally, if the treatments are almost equivalent, that is $\delta \leq \epsilon$, the longer the experiment, the costlier it is.

Lemma 4.3.5 *Suppose that δ is such that $|\delta| \leq \epsilon$. Then, for all $\mathcal{A}_1 \in \mathcal{C}_a$, for all $\mathcal{A}_2 \in \mathcal{C}_a$, if $\tau(\mathcal{A}_1) \geq \tau(\mathcal{A}_2)$ stochastically, then $R(\mathcal{A}_1) \geq R(\mathcal{A}_2)$.*

□ Proof: since $|\delta| \leq \epsilon$, the cost $r(k,\mathcal{A}) = r(k)$ does not depend on the chosen allocation rule. Hence, we have

$$R(\mathcal{A}_1) - R(\mathcal{A}_2) = \mathrm{E}\left[\sum_{k=\tau(\mathcal{A}_2)+1}^{\tau(\mathcal{A}_1)} r(k-1)\right] \geq 0.$$

■

To sum up, if $|\delta| \leq \epsilon$, we should stop the procedure (see Lemma 4.3.5) as soon as possible, in order to minimize the expected cost. On the other hand, if $|\delta| > \epsilon$ and m is large, the procedure should stop at step 2 (see Lemma 4.3.2).

Thus, we want to establish an allocation rule that minimizes the expected sampling cost asymptotically, at each step, as if it were the last step before the end of the experiment. Suppose we are at step k, and that $|\delta| > \epsilon$. We have to find the optimal value $p(k) = p$, such that

$$\begin{aligned} p(1-p) &\leq 1/4, \\ |S^*(k-1)| &\leq Q_{k-1}, \\ |S^*(k)| &> Q_k. \end{aligned} \tag{4.3.5}$$

Hence,

$$|S^*(k)| - |S^*(k-1)| > Q_k - Q_{k-1},$$
$$|S^*(k)| > |S^*(k-1)| + Q_k - Q_{k-1},$$

$$\text{that is, } |S^*(k)| \geq a, \tag{4.3.6}$$

with

$$a = \max\left(0\,,\ |S^*(k-1)| + Q_k - Q_{k-1}\right), \tag{4.3.7}$$

and

$$S^*(k) \;=\; \left[\frac{1}{G(k)} - \frac{1}{G(k-1)}\right]^{-1/2}\left[\frac{\tilde{S}^2(k)}{G(k)} - \frac{\tilde{S}^2(k-1)}{G(k-1)}\right].$$

We test sequentially $H_0 : \beta_1^1 = \beta_1^2$ vs $H_0 : \beta_1^1 \neq \beta_1^2$. At step 1, the issue is the same, no matters which allocation rule is chosen in $\mathcal{C}_a$. Then, situations are different.

Stop at step 2: we have to find the best p, that is the proportion of patients assigned to treatment 1 just after the first interim analysis, such that the experiment stops at step 2. By expression (4.3.4), we know that $\lim\limits_{m\to\infty,n\to\infty} S^*(2)^2$ equals in probability

$$\lim_{m\to\infty,n\to\infty} 2vmp(1-p)\tilde{S}(2)^2 + v\left(-2m_0p(1-p) + \frac{3}{2}m_0\right)\tilde{S}(2)^2 - vm_0\tilde{S}(2)S(1).$$

Formula (4.3.6) leads to

$$2vmp(1-p)\tilde{S}(2)^2 + v\left(-2m_0p(1-p) + \frac{3}{2}m_0\right)\tilde{S}(2)^2 - vm_0\tilde{S}(2)S(1) \geq a^2,$$

that is,

$$p(1-p) \geq \frac{a^2 + vm_0\tilde{S}(2)S(1) - \frac{3}{2}m_0v\tilde{S}(2)^2}{2v\,(m-m_0)\,\tilde{S}(2)^2} = A. \tag{4.3.8}$$

Since $p(1-p) \leq \frac{1}{4}$, we finally have

$$p = \frac{1 \pm \sqrt{1 - \min(4A, 1)}}{2}.$$

In (4.3.8), we can distinguish two tendencies: the larger m, the more unbalanced is the allocation rule. On the contrary, the smaller is $\tilde{S}(2)^2$, the more balanced is the allocation rule.

Solution if $\hat{\delta} < 0$	$p = \frac{1-\sqrt{1-\min(4A,1)}}{2}$
Solution if $\hat{\delta} \geq 0$	$p = \frac{1+\sqrt{1-\min(4A,1)}}{2}$

Table 4.3: *Asymptotic optimal value for p — proportion of new patients assigned to treatment 1 at step 1 — such that the experiment stops at step 2. Value of A is given by (4.3.8), and $\hat{\delta} = S(1)$.*

We note from (4.3.8) that $\tilde{S}(2)$ is in fact unknown at step 1. It should be replaced by $S(1)$. However, we are able to choose $m_0 \to \infty$, such that $\lim_{m\to\infty, m_0\to\infty} m_0/m = 0$. Then, $S(1)$ converges to $\tilde{S}(2)$, and the developments we have made are still valid.

Stop at step k>2: we first remember that

$$\lim_{m\to\infty, n\to\infty} \tilde{S}^2(k) = \lim_{m\to\infty, n\to\infty} \tilde{S}^2(k-1) \text{ in probability,}$$

hence,

$$\lim_{m\to\infty, n\to\infty} S^*(k) = \lim_{m\to\infty, n\to\infty} \left(\frac{1}{G(k)} - \frac{1}{G(k-1)}\right)^{1/2} \tilde{S}^2(k-1).$$

Let $m(k)$ be the total number of patients who have joined the trial before step k. If $\tilde{S}^2(k-1) \neq 0$, (4.3.6) reduces to

$$\frac{vm^1(k)m^2(k)}{m(k)} - \frac{vm^1(k-1)m^2(k-1)}{m(k-1)} \geq \frac{a^2}{\tilde{S}^2(k-1)^2},$$

$$m^1(k)m^2(k) \geq \frac{m(k)}{v}\left[\frac{vm^1(k-1)m^2(k-1)}{m(k-1)} + \frac{a^2}{\tilde{S}^2(k-1)^2}\right],$$

which will be written

$$p(1-p) + bp + e > f, \tag{4.3.9}$$

where

$$b = \frac{m^2(k-1) - m^1(k-1)}{2m}, \tag{4.3.10}$$

$$e = \frac{m^1(k-1)m^2(k-1) + 2mm^1(k-1)}{4m^2}, \tag{4.3.11}$$

$$f = \frac{m(k)}{4m^2v}\left[\frac{vm^1(k-1)m^2(k-1)}{m(k-1)} + \frac{a^2}{[\tilde{S}^2(k-1)]^2}\right]. \tag{4.3.12}$$

We note that $e \geq 0$ and $f \geq 0$, whereas $b \in \mathbb{R}$. By combining (4.3.9) and the constraint $p(1-p) \leq \frac{1}{4}$, we have

$$\begin{aligned} p(1-p) + bp + e &\geq f \vee h_{min}(b), \text{ and} \\ p(1-p) + bp + e &\leq h_{max}(b), \end{aligned}$$

where

$$h_{max}(b) = \begin{cases} e & \text{if } b \in (-\infty, -1], \\ e + \left(\frac{1+b}{2}\right)^2 & \text{if } b \in (-1, +1], \\ e + b & \text{if } b \in (+1, +\infty), \end{cases}$$

and

$$h_{min}(b) = \begin{cases} e + b & \text{if } b \in (-\infty, 0], \\ e & \text{if } b \in (0, +\infty). \end{cases}$$

Thus, the constraints are summarized by

$$p(1-p) + bp + e \geq \min\{h_{\max}(b), \max[f, h_{\min}(b)]\} = f^*.$$

If $\min\{h_{\max}(b), \max[f, h_{\min}(b)]\}$ does not equal f, the unique solution is $p = 0$, $p = 1$, or $p = (b+1)/2$ whenever $-1 \leq b \leq 1$. In the other situations, we obtain $p = p_1$ and $p = p_2$ such that

$$p_1 = \frac{b + 1 - \sqrt{(b+1)^2 + 4(e-f)}}{2}, \text{ and}$$

$$p_2 = \frac{b + 1 + \sqrt{(b+1)^2 + 4(e-f)}}{2}.$$

Proposition 4.3.2 *Asymptotic solution to the problem of minimizing the expected cost. At step k, when $n \to \infty$ and $m \to \infty$, the optimal value of p, such that $2mp$ patients are assigned to treatment 1, is given by*

- *Table 4.3 if $k = 1$, and*

- *Table 4.4 otherwise.*

The only value unknown at step $k-1$ is Q_k, but the trial should stop if $|\delta| > \epsilon$. Hence, we replace in practice Q_k by $\tilde{Q}_k$, which is the boundary computed when we choose an intermediate significance level of $\alpha - \alpha(k-1)$. If $|\delta| \leq \epsilon$, all allocation rules (Lemma 4.3.4) are equivalent. Thus, the rule defined above is also used.

The choice of class $\mathcal{C}_a$ (see Definition 4.3.3) requires that $\zeta \leq p \leq 1-\zeta$. Therefore, we shall replace p, given by Proposition 4.3.2, by

$$\zeta \vee [p \wedge (1-\zeta)] .$$

Then, Proposition 4.3.2 leads to the allocation rule proposed below. Its asymptotic efficiency is proved at the end of the section.

Rule 4.3.1 *An asymptotically efficient procedure for the expected sampling cost (4.3.2), when $n \to \infty$ and $m \to \infty$, is defined by the following steps:*

1. *choose ϵ, the ESF (Error Spending Function) $\alpha(\cdot)$, the Type I error α, m_0, and m;*

2. *stop the trial and decide*

   - *H_1 at step k, if $|\tilde{S}(k)^*| > Q_k$, or*
   - *H_0 if $\alpha(k) = \alpha$ and $|\tilde{S}(k)^*| \leq Q_k$;*

3. *otherwise, assign $2mp(k+1)$ new patients to treatment 1, and $2m(1-p(k+1))$ to treatment 2, where $p(k+1)$ is given by Proposition 4.3.2.*

It may be not easy to give a straightforward illustration of Table 4.4. However, we point out that if b and $\tilde{S}$ are close to zero, the allocation is almost balanced, which is the least we wish! The fundamental property of Rule 4.3.1 is its asymptotic efficiency, which is stated in Proposition 4.3.3. The end of this section concentrates on its proof.

Proposition 4.3.3 *Let $\mathcal{A}_e \in \mathcal{C}_a$ be the procedure defined by Rule 4.3.1. Then, for all $\delta \in \mathbb{R}$,*

$$\frac{R(\mathcal{A}_e)}{\min_{\mathcal{A}\in\mathcal{C}_a} R(\mathcal{A})} \text{ converges to 1, as } m \to \infty, \text{ and } n \to \infty.$$

$b \geq 1$	$e \geq f$	$f \geq e + b$	$e + b > f > e$
Solution if $\hat{\delta} > 0$	$p = 1$	$p = 1$	$p = 1$
Solution if $\hat{\delta} < 0$	$p = 0$	$p = 1$	$p = [b + 1 - \sqrt{(b+1)^2 + 4(e-f)}]/2$
$1 > b \geq 0$	$e \geq f$	$f \geq e + \left(\frac{b+1}{2}\right)^2$	$e + \left(\frac{b+1}{2}\right)^2 > f > e$
Solution if $\hat{\delta} > 0$	$p = 1$	$p = \frac{b+1}{2}$	$p = [b + 1 + \sqrt{(b+1)^2 + 4(e-f)}]/2$ if $f > e + b$ $p = 1$ if $e + b \geq f$
Solution if $\hat{\delta} < 0$	$p = 0$	$p = \frac{b+1}{2}$	$p = [b + 1 - \sqrt{(b+1)^2 + 4(e-f)}]/2$
$0 > b \geq -1$	$e + b \geq f$	$f \geq e + \left(\frac{b+1}{2}\right)^2$	$e + \left(\frac{b+1}{2}\right)^2 > f > e + b$
Solution if $\hat{\delta} > 0$	$p = 1$	$p = \frac{b+1}{2}$	$p = [b + 1 + \sqrt{(b+1)^2 + 4(e-f)}]/2$
Solution if $\hat{\delta} < 0$	$p = 0$	$p = \frac{b+1}{2}$	$p = [b + 1 - \sqrt{(b+1)^2 + 4(e-f)}]/2$ if $f > e$ $p = 0$ if $f \leq e$
$-1 > b$	$e + b \geq f$	$f \geq e$	$e > f > e + b$
Solution if $\hat{\delta} > 0$	$p = 1$	$p = 0$	$p = [b + 1 + \sqrt{(b+1)^2 + 4(e-f)}]/2$
Solution if $\hat{\delta} < 0$	$p = 0$	$p = 0$	$p = 0$

Table 4.4: *Asymptotic optimal value for p — proportion of new patients assigned to treatment 1 at step $k > 1$ — such that the experiment stops at step $k + 1$. Values for b, e, and f are given by (4.3.10), (4.3.11), and (4.3.12).*

In order to demonstrate Proposition 4.3.3, we first need the following lemma.

Lemma 4.3.6 *For all $\delta \in \mathbb{R}$, and any procedure $\mathcal{A} \in \mathcal{C}_a$,*

$$\begin{aligned}\lim_{m\to\infty,n\to\infty} \frac{R(\mathcal{A})}{m} \geq\ & \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} \sum_{i=1}^{\tau(\mathcal{A}_p)} l_{\min}(i)\mathbf{1}_{\{\tau(\mathcal{A}_p)\leq 2\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right] \\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} \sum_{i=1}^{K} l(i,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_p)\geq K\}}\mathbf{1}_{\{|\delta|\leq\epsilon\}}\right] \\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_p)=1\}}\mathbf{1}_{\{|\delta|\leq\epsilon\}}\right],\end{aligned}$$

where $\mathcal{A}_p \in \mathcal{C}_a$, is the pairwise allocation, $l(i,\mathcal{A}) = \frac{r(i-1,\mathcal{A})}{m}$, and $l_{\min}(i) = \min_{\mathcal{A}\in\mathcal{C}_a} l(i,\mathcal{A}) = 2(1-\zeta)d(-|\delta|) + 2\zeta d(|\delta|)$.

□ Proof:

$$\begin{aligned}\frac{R(\mathcal{A})}{m} &= \mathrm{E}\left[\sum_{i=1}^{\tau(\mathcal{A})} l(i,\mathcal{A})\mathbf{1}_{\{|\delta|>\epsilon\}}\right] + \mathrm{E}\left[\sum_{i=1}^{\tau(\mathcal{A})} l(i,\mathcal{A})\mathbf{1}_{\{|\delta|\leq\epsilon\}}\right] \\ &= \mathrm{E}\left[\sum_{i=1}^{K} l(i,\mathcal{A})\mathbf{1}_{\{|\delta|>\epsilon\}}\mathbf{1}_{\{\tau(\mathcal{A})\geq i\}}\right] + \mathrm{E}\left[\sum_{i=1}^{K} l(i,\mathcal{A})\mathbf{1}_{\{|\delta|\leq\epsilon\}}\mathbf{1}_{\{\tau(\mathcal{A})\geq i\}}\right].\end{aligned}$$

$$\begin{aligned}\frac{R(\mathcal{A})}{m} \geq\ & \mathrm{E}\left[\sum_{i=1}^{K} l_{\min}(i)\mathbf{1}_{\{|\delta|>\epsilon\}}\mathbf{1}_{\{\tau(\mathcal{A}_p)\geq i\}}\mathbf{1}_{\{\tau(\mathcal{A}_p)\leq 2\}}\mathbf{1}_{\{\tau(\mathcal{A})=\tau(\mathcal{A}_p)\}}\right] && (4.3.13)\\ &+ \mathrm{E}\left[\sum_{i=1}^{K} l(i,\mathcal{A})\mathbf{1}_{\{|\delta|>\epsilon\}}\mathbf{1}_{\{\tau(\mathcal{A})\geq i\}}\mathbf{1}_{\{\tau(\mathcal{A})\leq 2\}}\mathbf{1}_{\{\tau(\mathcal{A})\neq\tau(\mathcal{A}_p)\}}\right] && (4.3.14)\\ &+ \mathrm{E}\left[\sum_{i=1}^{K} l(i,\mathcal{A})\mathbf{1}_{\{|\delta|>\epsilon\}}\mathbf{1}_{\{\tau(\mathcal{A})\geq i\}}\mathbf{1}_{\{\tau(\mathcal{A})>2\}}\right] && (4.3.15)\\ &+ \mathrm{E}\left[l(1,\mathcal{A}_p)\mathbf{1}_{\{|\delta|\leq\epsilon\}}\mathbf{1}_{\{\tau(\mathcal{A}_p)=1\}}\right] \\ &+ \mathrm{E}\left[\sum_{i=2}^{K-1}\sum_{j=1}^{i} l(j,\mathcal{A}_p)\mathbf{1}_{\{|\delta|\leq\epsilon\}}\mathbf{1}_{\{\tau(\mathcal{A})=i\}}\right] && (4.3.16)\\ &+ \mathrm{E}\left[\sum_{i=1}^{K} l(i,\mathcal{A}_p)\mathbf{1}_{\{|\delta|\leq\epsilon\}}\mathbf{1}_{\{\tau(\mathcal{A})\geq K\}}\right]. && (4.3.17)\end{aligned}$$

Indicator functions are measurable and positive, and so are the functions $l(\cdot,\cdot)$, which are continuous in δ and p, when i is fixed, hence measurable and positive. Moreover, we have

$$\begin{aligned} r(k,\mathcal{A}) &= d(\delta)[m^2(k)-m^2(k-1)]+d(-\delta)[m^1(k)-m^1(k-1)], \\ &= 2md(\delta)p(k,\mathcal{A})+2md(-\delta)\left[1-p(k,\mathcal{A})\right], \end{aligned}$$

thus,

$$\begin{aligned} l(i,\mathcal{A}) &= 2d(\delta)p(i,\mathcal{A})+2d(-\delta)\left[1-p(i,\mathcal{A})\right] \quad i\leq K, \\ \sum_{i=1}^{K} l(i,\mathcal{A}) &\leq 2Kd(|\delta|)\in \mathcal{L}^1. \end{aligned}$$

If we apply the Lebesgue dominate convergence theorem for random variables that converge in probability, when $n\to\infty$ and $m\to\infty$, we obtain

- (4.3.13) $\longrightarrow \mathrm{E}\left[\lim_{m\to\infty,n\to\infty}\sum_{i=1}^{K} l_{\min}(i)\mathbb{1}_{\{|\delta|>\epsilon\}}\mathbb{1}_{\{\tau(\mathcal{A}_p)\leq 2\}}\mathbb{1}_{\{\tau(\mathcal{A}_p)\geq i\}}\right]$
  by Lemma 4.3.3,
- (4.3.14) $\longrightarrow 0$ by Lemma 4.3.3,
- (4.3.15) $\longrightarrow 0$ by Lemma 4.3.2,
- (4.3.16) $\longrightarrow 0$ by Lemma 4.3.4,
- (4.3.17) $\longrightarrow \mathrm{E}\left[\lim_{m\to\infty,n\to\infty}\sum_{i=1}^{K} l(i,\mathcal{A}_p)\mathbb{1}_{\{|\delta|\leq\epsilon\}}\mathbb{1}_{\{\tau(\mathcal{A}_p)\geq K\}}\right]$ by Lemma 4.3.4.

It follows that

$$\begin{aligned} \lim_{m\to\infty,n\to\infty}\frac{R(\mathcal{A})}{m} \geq\ & \mathrm{E}\left[\lim_{m\to\infty,n\to\infty}\sum_{i=1}^{\tau(\mathcal{A}_p)} l_{\min}(i)\mathbb{1}_{\{|\delta|>\epsilon\}}\mathbb{1}_{\{\tau(\mathcal{A}_p)\leq 2\}}\right] \\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty}\sum_{i=1}^{K} l(i,\mathcal{A}_p)\mathbb{1}_{\{|\delta|\leq\epsilon\}}\mathbb{1}_{\{\tau(\mathcal{A}_p)\geq K\}}\right] \\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbb{1}_{\{|\delta|\leq\epsilon\}}\mathbb{1}_{\{\tau(\mathcal{A}_p)=1\}}\right]. \end{aligned}$$

■

□ Proof of Proposition 4.3.3:

$$\begin{aligned}
\frac{R(\mathcal{A}_e)}{m} &= \mathrm{E}\left[\sum_{i=1}^{\tau(\mathcal{A}_e)} l(i,\mathcal{A}_e)\right] \\
&= \mathrm{E}\left[l(1,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_p)\le 2\}}\mathbf{1}_{\{\tau(\mathcal{A}_p)=\tau(\mathcal{A}_e)\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right] && (4.3.18)\\
&+ \mathrm{E}\left[l(1,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_e)\le 2\}}\mathbf{1}_{\{\tau(\mathcal{A}_p)\ne\tau(\mathcal{A}_e)\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right] && (4.3.19)\\
&+ \mathrm{E}\left[l(1,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_e)\le 2\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right] && (4.3.20)\\
&+ \mathrm{E}\left[l(2,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_e)=2\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right] && (4.3.21)\\
&+ \mathrm{E}\left[l(2,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_e)=2\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right] && (4.3.22)\\
&+ \mathrm{E}\left[\sum_{i=3}^{K}\sum_{j=1}^{i} l(j,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_e)=i\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right] && (4.3.23)\\
&+ \mathrm{E}\left[\sum_{i=3}^{K-1}\sum_{j=1}^{i} l(j,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_e)=i\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right] && (4.3.24)\\
&+ \mathrm{E}\left[\sum_{j=1}^{K} l(j,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_e)\ge K\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right], && (4.3.25)
\end{aligned}$$

then, (4.3.22) is equal to

$$\begin{aligned}
\mathrm{E}\left[l(2,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_e)=2\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right] &= \mathrm{E}\left[l(2,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_e)=2\}}\mathbf{1}_{\{\tau(\mathcal{A}_e)\ne\tau(\mathcal{A}_p)\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right] \\
&+ \mathrm{E}\left[l(2,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_e)=2\}}\mathbf{1}_{\{\tau(\mathcal{A}_e)=\tau(\mathcal{A}_p)\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right].
\end{aligned}$$

Let $\mathcal{C}_a^* \subset \mathcal{C}_a$ be the class of allocations that use the same fixed p, $p \in [\zeta, 1-\zeta]$, at every step. Then, for all $\mathcal{A}^* \in \mathcal{C}_a^*$, the allocation rule $\mathcal{A}_e$ is such that

$$\begin{aligned}
\limsup_{m\to\infty}\left(\lim_{n\to\infty} l(2,\mathcal{A}_e)\mathbf{1}_{\{\tau(\mathcal{A}_p)=2\}}\right) &\le l(2,\mathcal{A}^*)\limsup_{m\to\infty}\left(\lim_{n\to\infty}\mathbf{1}_{\{\tau(\mathcal{A}_p)=2\}}\right) \\
&\le \min_{\mathcal{A}^*\in\mathcal{C}_a^*} l(2,\mathcal{A}^*)\limsup_{m\to\infty}\left(\lim_{n\to\infty}\mathbf{1}_{\{\tau(\mathcal{A}_p)=2\}}\right) \\
&\le l_{\min}(2)\limsup_{m\to\infty}\left(\lim_{n\to\infty}\mathbf{1}_{\{\tau(\mathcal{A}_p)=2\}}\right). \qquad (4.3.26)
\end{aligned}$$

By the same arguments as those of the proof of Lemma 4.3.6, it turns out that

- (4.3.18) $\longrightarrow \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{I}_{\{\tau(\mathcal{A}_p)=1\}}\mathbf{I}_{\{|\delta|\le\epsilon\}}\right]$ by Lemma 4.3.4,
- (4.3.19) $\longrightarrow 0$ by Lemma 4.3.4,
- (4.3.20) $\longrightarrow \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{I}_{\{|\delta|>\epsilon\}}\right]$ by Lemma 4.3.2,
- (4.3.21) $\longrightarrow 0$ by Lemma 4.3.4,
- (4.3.23) $\longrightarrow 0$ by Lemma 4.3.2,
- (4.3.24) $\longrightarrow 0$ by Lemma 4.3.4,
- (4.3.25) $\longrightarrow \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} \sum_{j=1}^{K} l(j,\mathcal{A}_p)\mathbf{I}_{\{\tau(\mathcal{A}_p)\ge K\}}\mathbf{I}_{\{|\delta|\le\epsilon\}}\right]$ by Lemma 4.3.4.

Thus, we have

$$\begin{aligned}
\limsup_{m\to\infty}\left(\lim_{n\to\infty}\frac{R(\mathcal{A}_e)}{m}\right) &= \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{I}_{\{\tau(\mathcal{A}_p)=1\}}\mathbf{I}_{\{|\delta|\le\epsilon\}}\right] \\
&+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{I}_{\{|\delta|>\epsilon\}}\right] \\
&+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} \sum_{j=1}^{K} l(j,\mathcal{A}_p)\mathbf{I}_{\{\tau(\mathcal{A}_p)\ge K\}}\mathbf{I}_{\{|\delta|\le\epsilon\}}\right] \\
&+ \limsup_{m\to\infty}\left(\lim_{n\to\infty}\mathrm{E}\left[l(2,\mathcal{A}_e)\mathbf{I}_{\{\tau(\mathcal{A}_p)=2\}}\mathbf{I}_{\{\tau(\mathcal{A}_e)=\tau(\mathcal{A}_p)\}}\mathbf{I}_{\{|\delta|>\epsilon\}}\right]\right) \\
&+ \limsup_{m\to\infty}\left(\lim_{n\to\infty}\mathrm{E}\left[l(2,\mathcal{A}_e)\mathbf{I}_{\{\tau(\mathcal{A}_e)=2\}}\mathbf{I}_{\{\tau(\mathcal{A}_e)\ne\tau(\mathcal{A}_p)\}}\mathbf{I}_{\{|\delta|>\epsilon\}}\right]\right).
\end{aligned}$$

By Lebesgue's dominated convergence theorem and Lemma 4.3.3, the last term is zero. Then, by Fatou's Lemma, we have

$$\begin{aligned}
\limsup_{m\to\infty}\left(\lim_{n\to\infty}\frac{R(\mathcal{A}_e)}{m}\right) &\le \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{I}_{\{\tau(\mathcal{A}_p)=1\}}\mathbf{I}_{\{|\delta|\le\epsilon\}}\right] \\
&+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{I}_{\{|\delta|>\epsilon\}}\right] \\
&+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} \textstyle\sum_{j=1}^{K} l(j,\mathcal{A}_p)\mathbf{I}_{\{\tau(\mathcal{A}_p)\ge K\}}\mathbf{I}_{\{|\delta|\le\epsilon\}}\right] \\
&+ \mathrm{E}\left[\limsup_{m\to\infty}\left(\lim_{n\to\infty} l\{2,\mathcal{A}_e\}\right)\limsup_{m\to\infty}\left(\lim_{n\to\infty}\mathbf{I}_{\{\tau(\mathcal{A}_p)=2\}}\right)\mathbf{I}_{\{|\delta|>\epsilon\}}\right].
\end{aligned}$$

Since $\limsup_{m\to\infty}\left(\lim_{n\to\infty}\mathbf{1}_{\{\tau(\mathcal{A}_p)=2\}}\right) = \lim_{m\to\infty,n\to\infty}\mathbf{1}_{\{\tau(\mathcal{A}_p)=2\}}$, we use (4.3.26) and Lemma 4.3.3; then,

$$\begin{aligned}\limsup_{m\to\infty}\left(\lim_{n\to\infty}\frac{R(\mathcal{A}_e)}{m}\right) &\le \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_p)=1\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right]\\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{1}_{\{|\delta|>\epsilon\}}\right]\\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty}\sum_{j=1}^{K} l(j,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_p)\ge K\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right]\\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l_{\min}(2)\mathbf{1}_{\{\tau(\mathcal{A}_p)=2\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right].\end{aligned}$$

Finally,

$$\begin{aligned}\limsup_{m\to\infty}\left(\lim_{n\to\infty}\frac{R(\mathcal{A}_e)}{m}\right) &\le \mathrm{E}\left[\lim_{m\to\infty,n\to\infty} l(1,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_p)=1\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right]\\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty}\sum_{j=1}^{K} l(j,\mathcal{A}_p)\mathbf{1}_{\{\tau(\mathcal{A}_p)\ge K\}}\mathbf{1}_{\{|\delta|\le\epsilon\}}\right]\\ &+ \mathrm{E}\left[\lim_{m\to\infty,n\to\infty}\sum_{i=1}^{\tau(\mathcal{A}_p)} l_{\min}(i)\mathbf{1}_{\{\tau(\mathcal{A}_p)\le 2\}}\mathbf{1}_{\{|\delta|>\epsilon\}}\right],\end{aligned}$$

$$\limsup_{m\to\infty}\left(\lim_{n\to\infty}\frac{R(\mathcal{A}_e)}{m}\right) \le \lim_{m\to\infty,n\to\infty}\frac{R(\mathcal{A})}{m}\quad \forall\mathcal{A}\in\mathcal{C}_a \text{ (by Lemma 4.3.6).}$$

■

4.4 Simulation studies

We use simulation studies to compare the asymptotically efficient Rule 4.3.1 with other allocation rules, especially the pairwise allocation. We investigate control of Type I error, power, average sample size, and the number of patients assigned to the inferior treatment.

Example 4.4.1 *We consider a trial involving two groups of patients. In (1.3.2), we choose* $\sigma^2 = 1$, $\beta_0^1 = \beta_0^2 = \beta_1^1 = 1$,

$$D = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}, \text{ and a common design } X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 19 \end{pmatrix}'.$$

When the trial begins, 20 patients are assigned to each treatment. Then, we use an allocation rule, and we assign 100 new patients at calendar times $t_{c1} = 2$, *and 100 at time* $t_{c2} = 5$. *Results obtained by simulations (2,500 trials) are displayed in Tables 4.5 and 4.6. Tables 4.7 and 4.8 are similar, but we take* $t_{c1} = 6$, *and* $t_{c2} = 12$. *We consider three situations:*

- *no staggered entries (labelled ns), that is, 120 patients are assigned to each treatment at the beginning;*
- *pairwise allocation (labelled* $\mathcal{A}_p$*); and*
- *allocation defined by Rule 4.3.1 (labelled* $\mathcal{A}_e$*), with* $\zeta = 0.1$.

Five different ESFs are used, namely, α_P^* *(1.4.8),* α_{OBF}^* *(1.4.7), and* α_C^* *(3.1.1), with* $s = 0.5, 1,$ *and* 2. *We set the Type I error at 0.05, and consider a time indicator based either on observation or information.*

Whereas the ASN and the ITN are very similar if we use information time or observation time, the loss of power could be much larger with observations than with information. We observe that Rule 4.3.1 gives a large reduction in allocation to the inferior treatment, with a slight increase in the ASN. In general, this reduction is attained at the expense of a loss of power.

As we note in Chapter 3, when we use Procedure 2.3.2, conditioning leads to conservative procedures: the ASN is often large. Thus, use of α_P^* seems advisable, even if the criterion *RD* defined by (4.1.7) is slightly better when we take α_{OBF}^*. We add that Rule 4.3.1 is stable, whatever the ESF used, and even if the assumption of almost complete information for every patient is not true. This can be seen on Tables 4.5 and 4.6.

Finally, we compare different allocation rules in $\mathcal{C}_a$:

- pairwise allocation, $\mathcal{A}_p$;
- allocation $\mathcal{A}_e$, defined by Rule 4.3.1;
- Rule 4.1.1 from Anscombe; and
- another rule, labelled $\mathcal{A}_b$, which assigns, at every step, all new patients to the currently best identified treatment.

We choose for all four situations $\zeta = 0.2$ and $\beta_1^2 = 0.42$ or $\beta_1^2 = 0.6$. Results obtained by simulation (2,500 trials) are displayed in Tables 4.9 and 4.10.

In Tables 4.9 and 4.10, we note that rules $\mathcal{A}_b$ and $\mathcal{A}_e$ leads to rather satisfying results. Indeed, the allocation rule $\mathcal{A}_b$ is much simpler than $\mathcal{A}_e$, defined by Rule 4.3.1.

Allocation	Steps	ESF	information time			time based on obs.		
			$\hat{\gamma}$	ASN	obs.	$\hat{\gamma}$	ASN	obs.
ns	2, 5, 19	α_P^*	5.2	240.0	96.1	5.1	240.0	98.4
$\mathcal{A}_p$	2, 5, 24	α_P^*	5.7	236.1	97.4	5.4	238.2	98.7
$\mathcal{A}_e$	2, 5, 24	α_P^*	5.4	236.2	97.5	4.8	238.7	99.1
ns	2, 5, 19	α_C^*, $s = 0.5$	5.4	240.0	95.8	5.4	240.0	97.3
$\mathcal{A}_p$	2, 5, 24	α_C^*, $s = 0.5$	5.5	234.2	96.4	5.7	236.0	97.4
$\mathcal{A}_e$	2, 5, 24	α_C^*, $s = 0.5$	5.6	234.4	96.5	5.2	236.1	97.5
ns	2, 5, 19	α_C^*, $s = 1$	5.4	240.0	96.4	5.0	240.0	97.5
$\mathcal{A}_p$	2, 5, 24	α_C^*, $s = 1$	5.8	237.1	98.0	5.4	239.0	97.7
$\mathcal{A}_e$	2, 5, 24	α_C^*, $s = 1$	5.1	237.7	98.5	4.9	239.2	98.1
ns	2, 5, 19	α_C^*, $s = 2$	5.2	240.0	97.8	4.9	240.0	99.8
$\mathcal{A}_p$	2, 5, 24	α_C^*, $s = 2$	5.3	239.5	99.6	5.3	239.8	99.9
$\mathcal{A}_e$	2, 5, 24	α_C^*, $s = 2$	5.1	239.4	99.6	5.0	239.9	99.9
ns	2, 5, 19	α_{OBF}^*	5.3	240.0	97.6	5.1	240.0	100.0
$\mathcal{A}_p$	2, 5, 24	α_{OBF}^*	5.2	239.9	99.9	5.2	239.9	99.9
$\mathcal{A}_e$	2, 5, 24	α_{OBF}^*	5.0	239.9	99.9	5.1	240.0	99.7

Table 4.5: *Example 4.4.1, with $\beta_1^2 = 1$; empirical Type I error rate, $\hat{\gamma}$, ASN, ITN, and percentage of observations used.*

Allocation	Steps	ESF	information time				time based on obs.			
			$\hat{\gamma}$	ASN	ITN	obs.	$\hat{\gamma}$	ASN	ITN	obs.
ns	2,5,19	α_P^*	75.5	240.0	120.0	37.4	65.0	240.0	120.0	48.3
$\mathcal{A}_p$	2,5,24	α_P^*	69.1	196.3	98.1	66.5	68.5	207.7	103.8	74.1
$\mathcal{A}_e$	2,5,24	α_P^*	48.4	210.3	51.8	78.6	46.1	219.4	51.4	84.3
ns	2,5,19	α_C^*, $s = 0.5$	76.1	240.0	120.0	36.6	71.4	240.0	120.0	41.8
$\mathcal{A}_p$	2,5,24	α_C^*, $s = 0.5$	67.6	188.5	94.3	62.0	69.3	195.2	97.6	65.7
$\mathcal{A}_e$	2,5,24	α_C^*, $s = 0.5$	50.8	202.4	53.4	74.1	49.0	209.3	51.8	78.0
ns	2,5,19	α_C^*, $s = 1$	74.1	240.0	120.0	38.9	61.1	240.0	120.0	52.6
$\mathcal{A}_p$	2,5,24	α_C^*, $s = 1$	69.2	201.6	100.8	70.0	68.0	212.7	106.3	77.9
$\mathcal{A}_e$	2,5,24	α_C^*, $s = 1$	47.4	214.6	51.3	81.3	44.5	223.4	51.5	87.2
ns	2,5,19	α_C^*, $s = 2$	69.0	240.0	120.0	44.4	40.8	240.0	120.0	72.4
$\mathcal{A}_p$	2,5,24	α_C^*, $s = 2$	66.6	219.9	110.0	83.5	62.2	220.4	110.2	86.1
$\mathcal{A}_e$	2,5,24	α_C^*, $s = 2$	42.1	229.1	51.9	91.5	36.2	230.6	47.9	93.7
ns	2,5,19	α_{OBF}^*	64.8	240.0	120.0	48.7	24.8	240.0	120.0	87.7
$\mathcal{A}_p$	2,5,24	α_{OBF}^*	63.8	232.4	116.2	93.8	62.9	236.0	118.0	97.0
$\mathcal{A}_e$	2,5,24	α_{OBF}^*	39.2	236.0	52.5	97.0	38.0	238.1	52.7	98.8

Table 4.6: *Example 4.4.1, with $\beta_1^2 = 1.42$; empirical power, $\hat{\gamma}$, ASN, ITN, and percentage of observations used.*

Allocation	Steps	ESF	information time			time based on obs.		
			$\hat{\gamma}$	ASN	obs.	$\hat{\gamma}$	ASN	obs.
ns	6,12,19	α_P^*	5.9	240.0	96.4	6.1	240.0	97.6
$\mathcal{A}_p$	6,12,31	α_P^*	5.3	234.9	97.0	5.5	236.8	98.1
$\mathcal{A}_e$	6,12,31	α_P^*	5.3	234.9	97.0	4.6	237.1	98.3
ns	6,12,19	α_C^*, $s = 0.5$	5.9	240.0	96.3	6.3	240.0	97.7
$\mathcal{A}_p$	6,12,31	α_C^*, $s = 0.5$	5.4	233.4	96.2	5.0	235.2	97.2
$\mathcal{A}_e$	6,12,31	α_C^*, $s = 0.5$	5.5	233.4	96.2	5.1	235.2	97.3
ns	6,12,19	α_C^*, $s = 1$	6.1	240.0	96.5	5.4	240.0	98.2
$\mathcal{A}_p$	6,12,31	α_C^*, $s = 1$	5.1	235.9	97.5	5.1	237.9	97.5
$\mathcal{A}_e$	6,12,31	α_C^*, $s = 1$	4.9	235.9	97.5	4.5	238.0	97.8
ns	6,12,19	α_C^*, $s = 2$	6.5	240.0	96.7	5.5	240.0	99.1
$\mathcal{A}_p$	6,12,31	α_C^*, $s = 2$	5.0	238.3	98.9	5.2	239.5	99.7
$\mathcal{A}_e$	6,12,31	α_C^*, $s = 2$	4.8	238.4	99.0	4.3	239.6	99.7
ns	6,12,19	α_{OBF}^*	6.4	240.0	96.8	5.4	240.0	99.3
$\mathcal{A}_p$	6,12,31	α_{OBF}^*	5.0	239.2	99.5	5.1	239.9	99.9
$\mathcal{A}_e$	6,12,31	α_{OBF}^*	4.3	239.4	99.5	4.0	239.9	100.0

Table 4.7: *Example 4.4.1, with $\beta_1^2 = 1$; empirical family-wise error rate, $\hat{\gamma}$, ASN, ITN, and percentage of observations used.*

Allocation	Steps	ESF	information time				time based on obs.			
			$\hat{\gamma}$	ASN	ITN	obs.	$\hat{\gamma}$	ASN	ITN	obs.
ns	6,12,19	α_P^*	87.1	240.0	120.0	43.4	82.4	240.0	120.0	47.0
$\mathcal{A}_p$	6,12,31	α_P^*	68.6	180.0	90.0	61.9	68.0	192.5	96.2	68.7
$\mathcal{A}_e$	6,12,31	α_P^*	51.8	194.8	51.5	72.9	45.6	208.0	47.7	80.3
ns	6,12,19	α_C^*, $s = 0.5$	87.4	240.0	120.0	43.2	84.1	240.0	120.0	45.6
$\mathcal{A}_p$	6,12,31	α_C^*, $s = 0.5$	66.7	174.4	87.2	59.7	68.6	180.8	90.4	62.8
$\mathcal{A}_e$	6,12,31	α_C^*, $s = 0.5$	53.0	188.7	55.4	70.4	49.3	196.0	49.5	74.2
ns	6,12,19	α_C^*, $s = 1$	86.8	240.0	120.0	43.7	79.4	240.0	120.0	49.1
$\mathcal{A}_p$	6,12,31	α_C^*, $s = 1$	69.1	185.7	92.8	64.6	67.0	198.8	99.4	72.3
$\mathcal{A}_e$	6,12,31	α_C^*, $s = 1$	49.2	201.2	49.5	76.2	42.6	214.4	47.4	83.9
ns	6,12,19	α_C^*, $s = 2$	85.8	240.0	120.0	44.5	67.8	240.0	120.0	57.3
$\mathcal{A}_p$	6,12,31	α_C^*, $s = 2$	66.4	203.4	101.7	74.5	62.2	220.4	110.2	86.1
$\mathcal{A}_e$	6,12,31	α_C^*, $s = 2$	41.6	218.9	47.7	86.0	36.2	230.6	47.9	93.7
ns	6,12,19	α_{OBF}^*	85.2	240.0	120.0	44.8	50.9	240.0	120.0	69.0
$\mathcal{A}_p$	6,12,31	α_{OBF}^*	65.3	211.1	105.5	79.0	57.4	233.4	116.7	95.7
$\mathcal{A}_e$	6,12,31	α_{OBF}^*	39.0	226.1	48.1	90.2	32.6	236.6	48.4	98.0

Table 4.8: *Example 4.4.1, with $\beta_1^2 = 1.42$; empirical power, $\hat{\gamma}$, ASN, ITN, and percentage of observations used.*

Sample sizes used by $\mathcal{A}_e$ and $\mathcal{A}_b$ are almost equal, but $\mathcal{A}_e$ is more powerful. On the contrary, when we use $\mathcal{A}_b$, $\hat{\sigma}_{ITN}$ — the standard deviation for the number of patient assigned to the inferior treatment — and often the ITN are smaller. This assertion must be counterbalanced, since in the situation $\beta_1^1 = \beta_1^2 = 1$, $\hat{\sigma}^2_{ITN}$ when we use $\mathcal{A}_b$ is between 50% and 100% much larger than when $\mathcal{A}_e$ is used.

Allocation	Steps	ESF	$\hat{\gamma}$	ASN	ITN	RD	$\hat{\sigma}_{ITN}$
$\mathcal{A}_p$	2,5,24	α_P^*	69.4	187.9	94.0		30.7
$\mathcal{A}_e$	2,5,24	α_P^*	62.2	197.6	66.5	32.7	30.1
$\mathcal{A}_b$	2,5,24	α_P^*	58.8	197.3	60.3	38.9	27.3
Anscombe	2,5,24	α_P^*	56.4	197.7	65.7	33.5	40.8
$\mathcal{A}_p$	2,5,24	α^*_{OBF}	68.0	214.6	107.3		22.4
$\mathcal{A}_e$	2,5,24	α^*_{OBF}	58.2	220.6	65.4	40.7	25.0
$\mathcal{A}_b$	2,5,24	α^*_{OBF}	58.1	220.6	65.3	40.8	25.0
Anscombe	2,5,24	α^*_{OBF}	55.1	220.6	71.3	35.3	40.7
$\mathcal{A}_p$	6,12,31	α_P^*	68.6	177.5	88.7		35.0
$\mathcal{A}_e$	6,12,31	α_P^*	60.8	186.4	64.5	30.8	32.3
$\mathcal{A}_b$	6,12,31	α_P^*	56.0	187.2	55.3	40.9	25.1
Anscombe	6,12,31	α_P^*	54.9	186.4	58.2	37.5	34.7
$\mathcal{A}_p$	6,12,31	α^*_{OBF}	66.7	205.1	102.5		24.9
$\mathcal{A}_e$	6,12,31	α^*_{OBF}	54.8	214.2	61.5	42.6	21.9
$\mathcal{A}_b$	6,12,31	α^*_{OBF}	54.4	214.1	60.8	43.2	21.7
Anscombe	6,12,31	α^*_{OBF}	52.7	214.2	64.3	40.0	33.2

Table 4.9: *Example 4.4.1, with* $\beta_1^2 = 1.42$ *and* $\zeta = 0.2$*; times (calendar) for intermediate tests, ESF used, empirical power,* $\hat{\gamma}$*, Average number of patients (ASN), average number of patients assigned to the inferior treatment (ITN), criterion RD defined in (4.1.7)* $\times 100$*, and standard deviation* $\hat{\sigma}_{ITN}$ *for the number of patients assigned to the inferior treatment.*

We also point out the influence of ζ is on Rule 4.3.1, by comparing Table 4.9 ($\zeta = 0.2$) with Tables 4.6 and 4.8 ($\zeta = 0.1$). It turns out that $\zeta = 0.2$ seems more suitable than $\zeta = 0.1$. In fact, the simulations that we have made with different values for ζ show that $\zeta = 0.2$ seems to be a satisfying value. Finally, when power is high, as in Table 4.10, α_P^* is undoubtly better than α^*_{OBF}.

Allocation	Steps	ESF	$\hat{\gamma}$	ASN	ITN	RD	$\hat{\sigma}_{ITN}$
$\mathcal{A}_p$	2,5,24	α_P^*	95.8	146.5	73.3		29.9
$\mathcal{A}_e$	2,5,24	α_P^*	91.1	158.6	50.0	37.0	24.1
$\mathcal{A}_b$	2,5,24	α_P^*	89.5	158.6	47.3	40.3	20.6
Anscombe	2,5,24	α_P^*	88.4	158.6	48.8	38.5	26.9
$\mathcal{A}_p$	2,5,24	α_{OBF}^*	94.8	180.3	90.1		26.1
$\mathcal{A}_e$	2,5,24	α_{OBF}^*	86.7	194.6	55.5	43.9	18.2
$\mathcal{A}_b$	2,5,24	α_{OBF}^*	86.7	194.6	54.5	43.9	18.2
Anscombe	2,5,24	α_{OBF}^*	85.9	194.6	56.6	41.8	27.0
$\mathcal{A}_p$	6,12,31	α_P^*	95.0	130.2	65.1		33.2
$\mathcal{A}_e$	6,12,31	α_P^*	90.2	140.8	45.3	35.7	24.6
$\mathcal{A}_b$	6,12,31	α_P^*	87.9	140.8	41.6	40.9	18.2
Anscombe	6,12,31	α_P^*	87.6	140.8	42.6	40.3	20.7
$\mathcal{A}_p$	6,12,31	α_{OBF}^*	94.8	169.6	84.8		26.1
$\mathcal{A}_e$	6,12,31	α_{OBF}^*	85.6	184.5	50.6	45.2	15.0
$\mathcal{A}_b$	6,12,31	α_{OBF}^*	85.2	184.5	50.4	45.4	14.8
Anscombe	6,12,31	α_{OBF}^*	85.0	184.5	51.0	44.7	18.8

Table 4.10: *Example 4.4.1, with $\beta_1^2 = 1.6$ and $\zeta = 0.2$; times (calendar) for intermediate tests, ESF used, empirical power, $\hat{\gamma}$, Average number of patients (ASN), average number of patients assigned to the inferior treatment (ITN), criterion RD defined in (4.1.7) $\times 100$, and standard deviation $\hat{\sigma}_{ITN}$ for the number of patients assigned to the inferior treatment.*

4.5 Generalization

In this section, we consider the three-treatment case. We suppose that two treatments are compared with a control, which is labelled 1.

4.5.1 Dealing with three treatments

We replace the expression (4.3.1) by

$$\tilde{S}(k)' = \left[\tilde{S}^{(2)}(k), \tilde{S}^{(3)}(k)\right]' = \left[\hat{\beta}_1^1(k) - \hat{\beta}_1^2(k), \hat{\beta}_1^1(k) - \hat{\beta}_1^3(k)\right]'. \tag{4.5.1}$$

Then, we can immediately generalize Lemma 4.3.1.

Lemma 4.5.1 *Distribution of $\tilde{S}(k)$.*

- *The random variable $\tilde{S}(k)$, given $\mathcal{F}_{k-1}$, is normally distributed;*

- *under* H_0: $\{\beta_1^1 = \beta_1^2\} \cap \{\beta_1^1 = \beta_1^3\}$, *we have*

$$\begin{aligned} \mathrm{E}\left(\tilde{S}(k) \mid \mathcal{F}_{k-1}\right) &= G(k)G(k-1)^{-1}\tilde{S}(k-1), \text{ and} \\ \mathrm{var}\left(\tilde{S}(k) \mid \mathcal{F}_{k-1}\right) &= G(k) - G(k)G(k-1)^{-1}G(k), \end{aligned}$$

where $G(k) = \mathrm{var}\left[\tilde{S}(k)\right]$, *and* $\mathcal{F}_k$ *is the sigma-algebra generated by* $\{\tilde{S}(l) : l = 1, \ldots, k\}$.

We choose allocation rules in $\mathcal{C}_a$, and our framework is defined by the following assumptions.

Assumptions 4.5.1 *We assume that the trial is such that*

- $m_0 = m^1(0) = m^2(0) = m^3(0)$ *patients are assigned to each treatment when the trial starts;*
- *at step* k, $m_+^1(k) + m_+^2(k) + m_+^3(k) = 3m$ *new patients join the trial,* $m_+^i(k) = 3mp_i(k)$ *are assigned to treatment* i, $i = 1, 2, 3$, *with* $p_1 + p_2 + p_3 = 1$; *and*
- $n_i^h(k)$, *the number of measures on patient* i *assigned to treatment* h *is written*

$$n_i^h(k) = g_i^h(k)n, \tag{4.5.2}$$

where the number of measurements n, *is the maximum number of measures that we can take on a patient, and* $0 \leq g_i^h(1) \leq \cdots \leq 1$.

Let $S^{*(i)}(k)$ be the standardized value of $\tilde{S}^{(i)}(k)$, $i = 1, 2$. When m and n are large, we have

$$S^{*(i)}(2)^2 = \frac{3\nu\tilde{S}^{(i)}(2)^2 p_1 p_i}{p_1 + p_i} m + o(1), \tag{4.5.3}$$

$$S^{*(i)}(k)^2 = 3\nu A m + o(1), \text{ for } k > 2, \text{ where} \tag{4.5.4}$$

$$A = \frac{\left[(k-1)(p_1 q_i - p_i q_1)\left(q_i \tilde{S}^{(i)}\{k\} + q_j \tilde{S}^{(j)}\{k\}\right) + p_i\left(\{k-1\}q_1 + p_1\right)\tilde{S}^{(i)}(k)\right]^2}{(k-1)\left[(q_1 p_i - q_i p - 1)^2 + (k-1)(p_1 q_i^2 + p_i q_1^2) + 2p_1 p_i (p1 + p_i)\right] + p_1 p_i (p_1 + p_i)},$$

$$\begin{aligned} i &\neq j, \ i, j \in \{1, 2\}, \\ p_i &= p_i(k), \\ q_i &= q_i(k-1) = \sum_{l=2}^{k-1} p_i(l)/(k-2), \text{ and} \\ \nu &= \sigma^{-2}\left(D_{[2,2]}\right)^{-1}. \end{aligned}$$

If we set $\delta_2 = \beta_1^1 - \beta_1^2$, $\delta_3 = \beta_1^1 - \beta_1^3$, and $|\delta| = |\delta_2| \vee |\delta_3|$, then Lemmas 4.3.2, 4.3.3, and 4.3.4 still hold.

4.5.2 Allocation

Because we face a problem of multiplicity when we deal with three treatments, we consider two simple practical situations. Firstly, we can use twice Rule 4.3.1:

- with treatment 1 and treatment 2, assigning $3m/2$ new patients per step; and
- with treatment 1 and treatment 3, assigning $3m/2$ new patients per step.

Another credible situation may be considered: consider a trial for which the proportion of patients assigned to the control is fixed in advance. Suppose we are at step k, and that $|\delta| > \epsilon$. Conditions (4.3.5) are replaced by:

$$\begin{aligned} p_2 + p_3 &= 1 - p_1, \\ |S^{*(2)}(k-1)| \vee |S^{*(3)}(k-1)| &\leq Q_{k-1}, \\ |S^{*(2)}(k)| \vee |S^{*(3)}(k)| &> Q_k. \end{aligned} \tag{4.5.5}$$

If we choose $|S^{*(2)}(k)| > Q_k$, respectively $|S^{*(3)}(k)| > Q_k$, instead of $|S^{*(2)}(k)| \vee |S^{*(3)}(k)| > Q_k$, we obtain the optimal values (p_2^*, p_3^*), respectively (p_2^{**}, p_3^{**}). Then we choose the pair which minimizes a cost function previously defined. This procedure does not lead to any theoretical difficulty. We would need one more term in the Taylor series (4.5.3) and (4.5.4), so as to improve the precision of expansions.

4.5.3 Prospects

Future simulations will show if the two conjectures defined above are worth considering, and asymptotic properties should be investigated.

In conclusion, we would like to stress that further steps should be considered in practice, when one uses any of the procedures proposed in Chapter 2 and 4. We can for instance counterbalance any decision by using a non-parametric model. Lee and DeMets (1991, 1992) chose twice the same example, which deals with the study of calcium supplement effects on bone density (Smith, Sempos, Smith, and Gilligan, 1989). They rejected the null hypothesis of no treatment difference between the placebo and the calcium supplement when they used a group sequential test based on (1.3.2). But they did not reject the null hypothesis when they used a group sequential rank test. The discrepancy was imputed to some outliers which had too much effect on the test based on the parametric model.

Finally, the possibility of bias should be always considered. Clinical trials which are stopped early due to evidence of benefit or toxicity among treatments are prone to exaggerate the magnitude of treatment difference (Hughes et al., 1992). We can estimate the bias either by numerical approximation, or by using parametric bootstrap (Pinheiro and DeMets, 1997).

List of symbols and abbreviations

$\xrightarrow{\text{a.s.}}$	convergence almost surely
$\xrightarrow{\mathcal{L}^p}$	convergence in mean of order p
$\xrightarrow{p}$	convergence in probability
$\vee$	maximum of
$\wedge$	minimum of
a.s.	almost surely
$A_{(i,j)}$	matrix A has i rows and j columns
$A_{[i,j]}$	element of the matrix A located in line i and column j
$B(\cdot)$	Brownian motion
$\mathrm{cov}(\cdot,\cdot)$	covariance
E	matrix which contains only 1s
$\mathrm{E}(\cdot)$	expectation
$F_{n,m}$	Fisher distribution with n and m degrees of freedom
Φ	cumulative distribution function of a zero-mean standardised normal random variable
I	identity matrix
$\mathcal{N}_p(\mu,\Sigma)$	normal distribution of dimension p (omitted if $p=1$) with expectation μ and variance Σ.
pr	probability
q_{z_α}	α-quantile of a zero-mean standardised normal random variable
$\mathrm{var}(\cdot)$	variance
χ^2_n	χ^2 distribution with n degrees of freedom

ASN	Average Sample Number
CI	Confidence Interval
ECMO	Evaluation of Extracorporeal Membrane Oxygenation
EM	Expectation Maximization (algorithm)
ESF	Error Spending Function
FDA	Food and Drug Administration

FWE	Family-wise Error Rate
ITN	Inferior Treatment Number
LANDEM	algorithm that generates sequential boundaries using the procedure of Lan and DeMets
LS	Least Squares
ML	Maximum Likelihood
MULNOR	Multivariate Normal algorithm
OSWALD	Object-oriented Software for the Analysis of Longitudinal Data in *S*
PEST	Planning and Evaluation of Sequential Trials
RCI	Repeated Confidence Interval
REML	Restricted Maximum Likelihood
SPRT	Sequential Probability Ratio Test

References

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association* **58**, 365–383.

Armitage, P. (1975). *Sequential Medical Trials* (Second ed.). Oxford: Blackwell.

Armitage, P., C. K. McPherson, and B. C. Rowe (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* **132**, 235–244.

Bauer, P. (1991). Multiple testing in clinical trials. *Statistics in Medicine* **10**, 871–890.

Bauer, P. and K. Köhne (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing . *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Berry, D. A. (1985). Interim analyses in clinical trials: Classical vs. Bayesian approaches. *Statistics in Medicine* **4**, 521–526.

Berry, D. A. (1987). Statistical inference, designing clinical trials, and pharmaceutical company decisions. *The Statistician* **36**, 181–189.

Betensky, R. A. (1997). Early stopping to accept H_0 based on conditional power: Approximations and comparisons. *Biometrics* **53**, 794–806.

Brunier, H. and J. Whitehead (1994). *PEST Version 3 Operating Manual.* University of Reading, Department of Applied Statistics.

Chow, Y. S. and H. Teicher (1988). *Probability Theory, Independence, Interchangeability, Martingales* (Second ed.). New York: Springer–Verlag.

Chèvre, C. (1992). La conception séquentielle des essais cliniques de comparaison de deux traitements. Travail de semestre, Swiss Federal Institute of Technology, DMA, Lausanne, Switzerland.

Contente Domingues, F. and I. Guerreiro (1988). *A vida a bordo na carreira de India.* Lisboa: Instituto de Investigaço Cientfica Tropical.

Crowder, M. J. and D. J. Hand (1990). *Analysis of Repeated Measures*. London: Chapman & Hall.

DeMets, D. L. (1984). Stopping guidelines vs stopping rules: A practitioner's point of view. *Communications in Statistics, A* **13**, 2395–2417.

DeMets, D. L. and K. K. G. Lan (1984). An overview of sequential methods and their application in clinical trials. *Communications in Statistics, A* **13**, 2315–2338.

Diggle, P. J., K. Y. Liang, and S. L. Zeger (1994). *Analysis of Longitudinal Data*, Volume 13 of *Statistical Science Series*. Oxford: Oxford University Press.

Eisele, J. R. (1994). The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference* **38**, 249–262.

Falissard, B. and J. Lellouch (1992). A new procedure for group sequential analysis in clinical trials. *Biometrics* **48**, 373–388.

Follmann, D. A., M. A. Proschan, and N. L. Geller (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* **50**, 325–336.

Friedman, L. M., C. D. Furberg, and D. L. DeMets (1985). *Fundamentals of clinical trials* (Second ed.). Littleton, MA: P.S.G.

Gange, S. J. and D. L. DeMets (1996). Sequential monitoring of clinical trials with correlated responses. *Biometrika* **83**, 157–167.

Geller, N. L. and S. J. Pocock (1987). Interim analyses in randomized clinical trials: Ramifications and guidelines for practitioners. *Biometrics* **43**, 213–223.

Ghosh, B. K. (1970). *Sequential Tests of Statistical Hypotheses*. Reading, Massachusetts: Addison-Wesley.

Ghosh, B. K. and P. K. Sen (1991). *Handbook of Sequential Analysis*. New York: Marcel Dekker.

Giesbrecht, F. G. and J. C. Burns (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. *Biometrics* **41**, 477–486.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385.

Harville, D. A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Annals of Statistics* **4**, 384–395.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* **46**, 1251–1271.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple Comparison Procedures*. New York: John Wiley.

Hughes, M. D. (1993a). Reporting Bayesian analyses of clinical trials. *Statistics in Medicine* **12**, 1651–1663.

Hughes, M. D. (1993b). Stopping guidelines for clinical trials with multiple treatments. *Statistics in Medicine* **12**, 901–915.

Hughes, M. D., L. S. Freedman, and S. J. Pocock (1992). The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* **48**, 41–53.

Jennison, C. and B. W. Turnbull (1989). Interim analyses: The repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society, Series B* **51**, 305–361.

Jennison, C. and B. W. Turnbull (1991). Exact calculations for sequential t, χ^2 and F tests. *Biometrika* **78**, 133–141.

Jennison, C. and B. W. Turnbull (1993). Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary responses. *Biometrics* **49**, 31–43.

Jennison, C. and B. W. Turnbull (1995). Distribution theory for group sequential analysis of general linear models. Technical Report **95:01**, University of Bath, Department of Mathematical Sciences, UK.

Jennison, C. and B. W. Turnbull (1997a). Distribution theory of group sequential t, χ^2 and F tests for general linear models. *Sequential Analysis* **16**, 295–317.

Jennison, C. and B. W. Turnbull (1997b). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92**, 1330–1341.

Jennrich, R. I. and M. D. Schluchter (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805–820.

Kackar, R. N. and D. A. Harville (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853–862.

Kenward, M. G. and J. H. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997.

Kim, K. and D. L. DeMets (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **74**, 149–154.

Kim, K. and D. L. DeMets (1992). Sample size determination for group sequential clinical trials with immediate response . *Statistics in Medicine* **11**, 1391–1399.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* **2**, 93–113.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lan, K. K. G. and D. L. DeMets (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

Lan, K. K. G. and D. L. DeMets (1989). Group sequential procedures: Calendar versus information time. *Statistics in Medicine* **8**, 1191-1198.

Lan, K. K. G., D. M. Reboussin, and D. L. DeMets (1994). Information and information fractions for design and sequential monitoring of clinical trial. *Communications in Statistics, A* **23**, 403–420.

Lange, R. L., R. J. A. Little, and M. G. Taylor (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881–896.

Lee, J. W. and D. L. DeMets (1991). Sequential comparison of changes with repeated measurements data. *Journal of the American Statistical Association* **86**, 757–762.

Lee, J. W. and D. L. DeMets (1992). Sequential rank tests with repeated measurements in clinical trials. *Journal of the American Statistical Association* **87**, 136–142.

Lee, J. W. and D. L. DeMets (1995). Group sequential comparison of changes: Adhoc versus more exact method. *Biometrics* **51**, 21–30.

Lee, S. J., K. Kim, and A. A. Tsiatis (1996). Repeated significance testing in longitudinal clinical trials. *Biometrika* **83**, 779–789.

Liang, K. Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lind, J. (1753). *A Treatise of the Scurvy in Three Parts. Containing an inquiry into the Nature, Causes and Cure of that Disease, together with a Critical and Chronological View of what has been published on the subject.* London: A. Millar.

Lindstrom, M. J. and D. M. Bates (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* **83**, 1014–1022.

Little, R. J. A. (1988). Analysis of data with missing values: Discussion. *Statistics in Medicine* **7**, 347-355.

Liu, W. (1995). A group sequential procedure for all-pairwise comparisons of k treatments based on the range statistic. *Biometrics* **51**, 946–955.

Liu, W. (1996). Somme group sequential procedures for comparing several treatments with a control. *Applied Statistics* **23**, 371–383.

Lohr, S. L. (1992). Multivariate normal probabilities of star-shaped regions. *Applied Statistics* **42**, 576–582.

Louis, T. A. (1975). Sequential allocation in clinical trials comparing the means of two gaussian populations. *Biometrika* **62**, 359–369.

Meinert, C. L. and S. Tonascia (1986). *Clinical trials. Design, conduct, and analysis.* Oxford: Oxford University Press.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* New York: John Wiley.

O'Brien, P. C. and T. R. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

O'Quigley, J., M. Pepe, and L. Fisher (1990). Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics* **46**, 33–48.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

Peace, K. E. (1992). *Biopharmaceutical Sequential Statistical Applications.* New York: Marcel Dekker.

Peto, R., M. C. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto, and P. Smith (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I: Introduction. *British Journal of Cancer* **34**, 585–612.

Pinheiro, J. C. and D. L. DeMets (1997). Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. *Biometrika* **84**, 831–845.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* **38**, 153–162.

Pocock, S. J. (1993). Statistical and ethical issues in monitoring clinical trials. *Statistics in Medicine* **12**, 1459–1469.

Pocock, S. J. (1995). Life as an academic medical statistician and how to survive it. *Statistics in Medicine* **14**, 209–222.

Pocock, S. J., N. L. Geller, and A. A. Tsiatis (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.

Proschan, M. A., D. A. Follmann, and N. L. Geller (1994). Monitoring multi-armed trials. *Statistics in Medicine* **13**, 1441–1452.

Reboussin, D. M. (1995). Flexible monitoring of multi-arm clinical trials with repeated measurements. In *American Statistical Association*, Proceedings of the Biopharmaceutical Section, pp. 111–114.

Reboussin, D. M., D. L. DeMets, K. Kim, and K. K. G. Lan (1992). Programs for computing group sequential bounds using the Lan–DeMets method. Technical Report **60**, University of Wisconsin, Department of Biostatistics, Madison, Wisconsin, USA.

Reboussin, D. M., K. K. G. Lan, and D. L. DeMets (1992). Group sequential testing of longitudinal data. Technical Report **72**, University of Wisconsin, Department of Biostatistics, Madison, Wisconsin, USA.

Robbins, H. E. and D. Siegmund (1974). Sequential tests involving two populations. *Journal of the American Statistical Association* **69**, 132–139.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Schervish, M. J. (1984). Multivariate normal probabilities with error bound. *Applied Statistics* **33**, 81–94.

Seber, G. A. (1984). *Multivariate Observations*. New York: John Wiley.

Senn, S. (1993). *Cross-over Trials in Clinical Research*. New York: John Wiley.

Senn, S. (1997). *Statistical Issues in Drug Development*. New York: John Wiley.

Siegmund, D. (1977). Repeated significance tests for a normal mean. *Biometrika* **64**, 177–189.

Siegmund, D. (1985). *Sequential Analysis*. New York: Springer–Verlag.

Smith, D. M., W. H. Robertson, and P. J. Diggle (1996). Oswald: Object-oriented Software for the Analysis of Longitudinal Data in *S*. Technical Report MA 96/192, Department of Mathematics and Statistics, University of Lancaster, United Kingdom.

Smith, E. L., C. T. Sempos, P. E. Smith, and C. Gilligan (1989). Calcium supplementation and bone loss in middle-aged women. *American Journal of Clinical Nutrition* **50**, 833–842.

Stallard, N. and K. M. Facey (1996). Comparison of the spending function method and the Christmas tree correction for group sequential trials. *Journal of Biopharmaceutical Statistics* **6**, 361–373.

Sutradhar, B. C. (1986). On the characteristic function of multivariate Student t distribution. *Canadian Journal of Statistics* **14**, 329–337.

Tanner, M. A. (1996). *Tools for Statistical Inference* (Third ed.). New York: Springer–Verlag.

Thall, P. F., R. Simon, and S. S. Ellenberg (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**, 303–310.

Wald, A. (1947). *Sequential Analysis*. New York: John Wiley.

Ware, J. H. (1989). Investing therapies of potentially great benefit: ECMO (with discussion). *Statistical Science* **4**, 298–340.

Wassertheil-Smoller, S. (1995). *Biostatistics and Epidemiology* (Second ed.). New York: Springer–Verlag.

Wei, L. J. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association* **73**, 559–563.

Wetherill, G. B. and K. D. Glazebrook (1986). *Sequential Methods in Statistics* (Third ed.). London: Chapman & Hall.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials* (Second ed.). New York: John Wiley.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.

Zoubeidi, T. (1989). Asymptotic efficiency of a sequential allocation rule. *Sequential Analysis* **8**, 205–231.

Zoubeidi, T. (1994). Optimal allocations in sequential tests involving two populations with covariates. *Communications in Statistics, A* **23**, 1215–1225.

Zoubeidi, T. (1996). Efficient allocations in group sequential tests. *Communications in Statistics, A* **25**, 1769–1781.

Curriculum vitae

Bernard Cerutti

Département de Mathématiques
Ecole Polytechnique Fédérale de Lausanne
CH - 1015 Lausanne
e-mail: bernard.cerutti@epfl.ch

Date of birth: December 23, 1969.
Nationality: EU French.

Occupation

Assistant at the Swiss Federal Institute of Technology in Lausanne, chair of Statistics, since November 1, 1994.

Education

1993 Post graduate certificate in statistics (DEA), Universités Paris I (Panthéon–Sorbonne) et Paris VII (Denis Diderot), Grade A.

1992 MA in applied mathematics (Maîtrise), Université de Franche-Comté, Grade A.

1991 Bachelor's degree in mathematics (Licence), Université de Franche-Comté, Grade A.

1987 Secondary school examination (mathematics and physics).

Languages

French: Mother tongue.

English: Cambridge Advanced Certificate in English (1997).

Italian: Good knowledge, numerous stays in Italy.