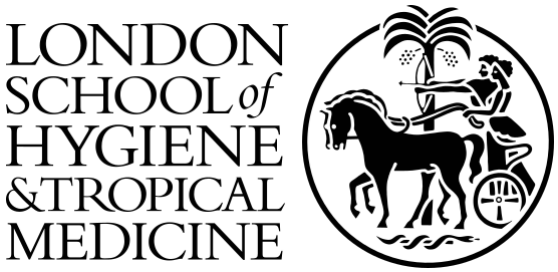


ANALYSIS OF REPEATED MEASURES IN CLINICAL TRIALS USING SUMMARY STATISTICS

THESIS

presented for the

DEGREE
of
DOCTOR OF PHILOSOPHY

in the faculty of medicine
(Field of Study:Statistics)

by

Lars Johan Frison

Medical Statistics Unit
London School of Hygiene and Tropical Medicine
University of London

ABSTRACT :

This thesis is concerned with statistical methodology for randomized clinical trials with repeated measurements over time, as regards both data analysis and the implications for study design. The inherent within-subject dependencies for repeated measurements necessitate analyses that take account of their covariance structure. There exists a whole battery of methods for analysing repeated measures designs, ranging from very simple (e.g. separate t-tests at each time-point) to very complicated (e.g. multi-level models with arbitrary error structures), but I will focus on "the summary statistic approach" which has recently become increasingly popular.

When interest centres around the average response to treatment over time, a logical choice of summary statistic is the mean of each subject's post-randomisation measurements, with appropriate adjustment for pre-treatment measurements. Among the class of "mean summary statistics" analysis of covariance (ANCOVA) is shown to be superior to its competitors. In particular, variance formulae are derived both under a general covariance structure and more specific cases (e.g. compound symmetry), allowing direct comparisons of efficiency among different summary statistics and repeated measures designs. The importance of precise estimates of the pre-entry levels and the consequences for sample size requirements are emphasized.

Some additional topics in relation to mean summary statistics, notably; the bias in estimation if pre-treatment means differ, the choice between additive or multiplicative models, and the summary statistic "area under the curve", are also investigated. For studies with restrictions on the range of baseline measurements the negative consequences incurred by "regression to the mean" are explored, especially regarding the variance for between-group comparisons.

For a more general class of true treatment effects over time, the optimal linear summary statistic under any covariance structure is derived. Special interest is devoted to the case of linearly diverging mean treatment curves, where the optimal alternative to the comparison of slopes is defined.

Asymptotic relative efficiencies are shown to be a useful tool when contrasting different designs and different summary statistics, both in the planning and reporting of repeated measures clinical trials. Finally, comparisons with other approaches are made, and recommendations given based on the need to balance theoretical considerations with practical matters.

ACKNOWLEDGEMENTS

I am grateful to my company Astra Hässle for generous financial support, and in particular to Mats Lörstad for giving me the opportunity to do this thesis, and to Sven Eriksson for learning me a lot about medical statistics. Finally, special thanks to Stuart Pocock for excellent guidance throughout the period of this work.

CONTENTS :

TABLES :

FIGURES :

1 INTRODUCTION: REPEATED MEASURES AND CLINICAL TRIALS

1.1 INTRODUCTION

In the realm of clinical trials it is more of a rule than an exception that each subject enrolled is assessed more than once with regard to the variable(s) comprising the primary objective of the investigation. These multiple recordings may relate either to baseline (run-in) visits or to measurements made during the treatment period, and there may be several visits performed both before as well as after the time of randomisation.

Heuristically, the longitudinal study allows each person to be used as his/her own control so that the ever-present heterogeneity among persons is reduced. Another advantage of performing a repeated measurements experiment is the possibility of considering a variety of research hypotheses in the same experiment. Indeed, a major difference between longitudinal and cross-sectional data is that the former provide information about the correlations between responses measured at different times, whereas the latter only provide information about the population marginal structure.

Increasing the number of measurements on each subject in a clinical trial will obviously increase the available information on treatment effects. The optimal way to allocate additional measurements over time at the design stage (e.g. before or after randomisation), and the best way to utilize the additional measurements at the analysis stage, is, however, not obvious. These considerations, pursued with emphasis on practical methodology rather than abstract theory, will form the main thread of this dissertation.

1.2 OBJECTIVES AND HYPOTHESIS FOR REPEATED MEASURES DESIGNS

Concentrating primarily on randomised clinical trials (RCTs), the possible main hypotheses in a typical repeated measures study can broadly be classified into three main categories. These consists of the main effect of treatment, the main effect of time, and the interaction of the two.

For a trial with one post-treatment evaluation, only the first of these hypotheses can be tested, and this test of an overall treatment effect is in most instances the one underlying the decisions concerning the primary objective in a trial. The medical question seeking an answer might for instance be; will our new treatment lower the average serum cholesterol level for a certain population of patients compared to standard treatment.

The main effect of time is usually of less interest. For example the finding that average levels of diastolic blood pressure across all subjects, ignoring treatment group, varies between different time-points will rarely be the answer to a main hypothesis. In some instances, however, there might be interest in detecting seasonal variations or diurnal variations.

The test for an interaction effect, that is for a treatment effect which depends upon the length of time in the trial, will often be of interest. There might, for instance, exist theories hypothesizing that the treatment effect will increase, attenuate or stabilize with time, or that the treatment effect is of a transient nature.

1.3 TYPES OF DESIGNS AND TYPES OF RESPONSES

In all that follows emphasis will be on randomised clinical trials, although some of the methods may be applicable also to laboratory experiments, uncontrolled experimental designs and sample survey designs. Within the context of randomised clinical trials there are two main types of design, the parallel group and the cross-over. My emphasis will be on the former.

There is also a need to decide on what type of responses we will concentrate on. Here the choice has been to investigate repeated observations of <u>quantitative</u> outcome measures on each subject. Thus, we will not be concerned with binary or categorical data, count data nor survival type data, though some of the ideas may extrapolate to such problems.

1.4 APPROACHES COMMONLY USED

This introductory section will be confined to a brief résumé of the various analyses strategies for repeated quantitative measures in clinical trials. More detailed descriptions of what these approaches do, and comparison with the summary statistic approach, will be saved for chapter 6.

Separate univariate analyses for each post-randomisation visit appear frequently in the medical literature and in clinical study reports. Matthews et al (1990), and Crowder and Hand (1990), provide informative discussions of the weaknesses of such an approach.

Repeated measures ANOVA, a modification of split-plot ANOVA, is also commonly used. This approach is well described in many standard text-books, like Fleiss (1986), and Milliken (1990). Relevant articles discussing several aspects of repeated measures ANOVA have been written by Rouanet and Lepine (1970), Wallenstein (1982), and Yates (1982).

Hotelling's T^2, a multivariate analogue of the univariate t-test, is sometimes used, even though it is not quite appropriate for this task. Descriptions of this approach appear in Chatfield and Collins (1980), and Crowder and Hand (1990).

Multivariate analysis of variance, MANOVA, can be used for many different designs, also for repeated measures studies. Descriptions with emphasis on repeated measures designs may be found in Crowder and Hand (1990), Fleiss (1986), Hand and Taylor (1987), and Rouanet and Lepine (1970).

Kenward (1987) has developed a refinement of MANOVA, labelled the "ante-dependence" approach, which is more economical in use of degrees of freedom to estimate the covariance structure among the repeated measurements. This is also discussed by Crowder and Hand (1990).

Useful references in relation to the summary statistic approach are; Matthews et al (1990), Dawson and Lagakos (1991), Rowell and Walters (1976), and Frison and Pocock (1992). As already hinted ANCOVA will often be the recommended summary statistic, good references for analysis of covariance being; Cochran (1957), Cox and McCullagh (1982), Fleiss (1986), and Senn (1989).

Many more complicated methods have been proposed. Most of them fall in the mixed model class (e.g. both fixed and random effects appear in the model) as for instance in multi-level models. Relevant references are; Crowder and Hand (1990), Jones (1993), Laird and Ware (1982), and Gumpertz and Pantula (1989).

1.5 THE COVARIANCE STRUCTURE

Interdependence between measurements on the same subject is the distinguishing factor between longitudinal and cross-sectional designs. For many of the approaches commonly used to analyse repeated measurements designs a correct specification of the covariance structure is essential for a valid and efficient analysis of the data. Hence, a parsimonious parametrization of the covariance structure is needed.

For the summary statistics approach, no knowledge about the covariance structure is needed for the validity of the analysis. As will be seen later on, however, the covariance structure has a great impact on the relative efficiencies between various summary statistics. It is also useful to be able to assume plausible covariance structures at the design stage, to ensure that appropriate repeated measurement design strategies and powerful summary statistics are chosen.

1.5.1 Some models for the covariance structure

We now move on to look at some specific classes of covariance structures. Assuming that the covariance structure is the same in both treatment groups, and that we have t repeated measurements, we want to impose some structure on a covariance matrix of the form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{21} & \cdot & \sigma_{t1} \\ \sigma_{12} & \sigma_2^2 & & \cdot \\ \cdot & & \cdot & \cdot \\ \sigma_{1t} & \cdot & \cdot & \sigma_t^2 \end{bmatrix}$$

In total, an otherwise unstructured matrix has t(t+1)/2 parameters, t variances and t(t-1)/2 covariances. It is helpful to rewrite Σ as $\Sigma = D_\sigma^T \cdot R \cdot D_\sigma$, where $D_\sigma^T = [\sigma_1 \quad \sigma_2 \quad \ldots \quad \sigma_t]$ is the vector of standard deviations and $R = \begin{bmatrix} 1 & \rho_{21} & \cdot & \rho_{t1} \\ \rho_{12} & 1 & & \cdot \\ \cdot & & \cdot & \cdot \\ \rho_{1t} & \cdot & \cdot & 1 \end{bmatrix}$ is the correlation matrix. Having thus removed the variances (but not assumed they are equal), we will concentrate on parameterizing the correlation structure.

The absolutely simplest structure is independence, when there are no random effects and all correlations are zero. This is not realistic for repeated measurements designs, since for RCT's there will always be within-subject variability present, which necessarily implies correlations different from zero.

The simplest generalization is compound symmetry. Which in spite of its simplicity is widely adopted as an underlying assumption for many of the approaches commonly used, explicitly or implicitly. This popular covariance model goes under many other names, like; random intercepts model, exchangeability model, and split-plot model. The correlation structure is given by $\rho_{ij} = \rho$ for all i and j (here we are also assuming that $\sigma_i = \sigma$ for all i), with ρ confined to lie in the interval [-1,1].

Another popular alternative is the first-order autoregressive

model, with $R = \begin{bmatrix} 1 & \rho & \rho^2 & . & \rho^t \\ \rho & 1 & & & . \\ \rho^2 & & 1 & & \rho^2 \\ . & & & . & \rho \\ \rho^t & . & \rho^2 & \rho & 1 \end{bmatrix}$. This structure originates

from an exponentially decreasing trend in the correlation pattern. A decreasing trend is plausible, however, in practice, exponentially decreasing is to "steep".

A further alternative is the first-order moving average model (which is frequently used in time-series analysis), with a

correlation structure determined by $\sigma_{ij} = \begin{cases} 1 & ,i = j \\ \rho & ,|i-j| = 1 \\ 0 & ,|i-j| > 1 \end{cases}$.

A banded or general autoregressive structure has one parameter for each diagonal in the matrix, specifically $\sigma_{ij} = \theta_k$, $k = |i-j|+1$, and there are t unknown parameters (including the variance, here assumed the same for all time points).

A flexible family of correlation structures, with only two parameters, was introduced by Muñoz et al (1992), which they called a damped exponential correlation structure. The correlation between two observations separated by s units of time is modelled as γ^{s^θ}, where γ is the correlation between elements separated by one s-unit, and θ is a damping parameter. Several of the one-parameter models are included as special cases in this family. For instance, with $\theta=0$ we have compound symmetry, with $\theta=1$ a first-order autoregressive model, and with $\theta \to \infty$ a first-order moving average process.

In addition, for $0<\theta<1$ we obtain a family of correlation structures with decay rate between those of compound symmetry and first-order autoregressive models, this is what is called attenuated exponential decay, and should offer plausible models in most circumstances. Having $\theta>1$ results in what might be termed accelerated exponential decay.

The feasibility of some of these models to explain the correlation structure on real data will be explored in section 2.3.

1.5.2 Examples of correlation structures from clinical trials

To give some objective evidence on how the correlation structures for repeated measurements in clinical trials actually turn out, a number of such examples are summarized in table 1.5.1. These examples represent the most recent experience of such trials that have been encountered in the Medical Statistics Unit at London School of Hygiene and Tropical Medicine and all have two randomised treatment groups. The aim is to obtain a reasonably representative sample of trials covering a variety of diseases and quantitative outcome measures.

For each trial table 1.5.1 lists the disease, the number of randomised patients, the numbers of pre- and post-treatment measurements and the mean time between post-treatment measurements, and then for each outcome measure three types of mean correlations. The mean of the pair-wise correlations among the pre-entry measurements is labelled "pre", the corresponding mean for the correlations among the post-treatment measurements is labelled "post", and "mix" refers to the mean of the correlations among all pre-entry post-treatment pairs of measurements. The final column in the table gives the estimated slope (decrease) in correlation with "time" between visits (where "time" denotes the number of visits apart). This allows a feeling to be gained for the degree of linear decay in correlation with time. The plausibility of a simple linear decrease is explored in section 2.3. In nearly all instances the post-treatment visits were at equally spaced intervals.

Table 1.5.1 : Summary of the correlations in repeated measurements from a sample of clinical trials.

Disease	Number of patients	Number of visits pre (p)	Number of visits post (r)	Mean time between post visits (mths)	Outcome measure	Mean correlation[1] pre	Mean correlation[1] mixed	Mean correlation[1] post	Estimated slope[2]
Coronary heart disease	152	3	8	1.5	CPK	.65	.62	.67	-.012
					ALAT	.69	.64	.67	-.017
					ASAT	.69	.70	.76	-.006
					Alkaline phosphatase	.79	.73	.75	.004
Coronary heart disease	219	2	4	3	HDL	.74	.74	.84	-.006
					Triglycerides	.68	.56	.56	-.066
					Total cholesterol	.65	.52	.65	-.011
Hypertension	55	3	12	1	Heart rate	.64	.56	.61	-.010
					Systolic blood pressure	.62	.56	.70	-.006
Hypertension	3450	1	7	2	SBP	-	.23	.44	-.029
					DBP	-	.44	.55	-.024
Intermittent claudication	504	2	2	6	Ankle/arm ratio of SBP	.74	.62	.65	-
Angina	251	1	3	4	Treadmill test distance	-	.53	.77	-.002
Childhood asthma	138	1	10	3	FEV_1	-	.70	.81	-.006
		1	5	6	PD_{20} (histamine resp.)	-	.47	.85	-.032
Multiple sclerosis	162	1	3	1	Muscle tone score	-	.70	.80	-.010
Low back pain	459	2	3	8	Back pain score	.85	.29	.75	.000
HIV infection	545	1	6	4	CD_4 cell count	-	.68	.77	-.021

1. pre is the mean of the correlations among the pre-treament visits, post similarly among the post-treatment visits mix is the mean of the correlations among pretreatment posttreatment pairs of measurements.

2. Estimated (by least squares) decrease in correlation per visit apart among the posttreatment visits.

Certain general characteristics emerge from these trials. The correlations between post-treatment visits mostly average between 0.6 and 0.8. A similar magnitude of correlation exists between pre-treatment visits, when $p \geq 2$. The average mixed pre-post correlation is mostly of similar magnitude, but with a tendency to be slightly lower. Most examples show a slight decline in correlation (amongst post-treatment visits) as the time interval between measurements increase. The extent and pattern of this decline is illustrated in figure 1.5.1, where the 11 first variables from table 1.5.1 are included. To reduce the mass of data (in total 531 distinct correlation coefficients) without imposing any specific structure (apart from smoothness) on the patterns over time, the correlation structure for each of the 11 variables has been approximated by a smoothed curve (using the function SM50 in SAS, SAS, 1992) through its correlation coefficients (in the figure, A=CPK, B=ALAT, and so on). There certainly appears to be a slow decrease in correlation with time for these variables, whether this decay is more complicated than a linear function is impossible to judge with the eye.

It is interesting to observe one or two exceptions, in the table, from the general pattern outlined above. The hypertension trial in elderly patients had somewhat lower correlations for blood pressure, and this can be attributed to the fact that treatment regimens were adjusted over time in each patient according to observed blood pressure; for example, a patient whose blood pressure stayed high received additional dosage or supplementary drugs. This is perhaps an unusual adaptive feature not commonly encountered in studies with repeated measurements. The low back pain study had a low mixed correlation, and this reflects the fact that a proportion of patients were cured (back pain score=0) and such prospect of cure was not closely associated with the original severity of disease.

Figure 1.5.1: Correlation coefficients versus time between measurements
Smoothed curves given for the 11 first variables in table 1.5.1

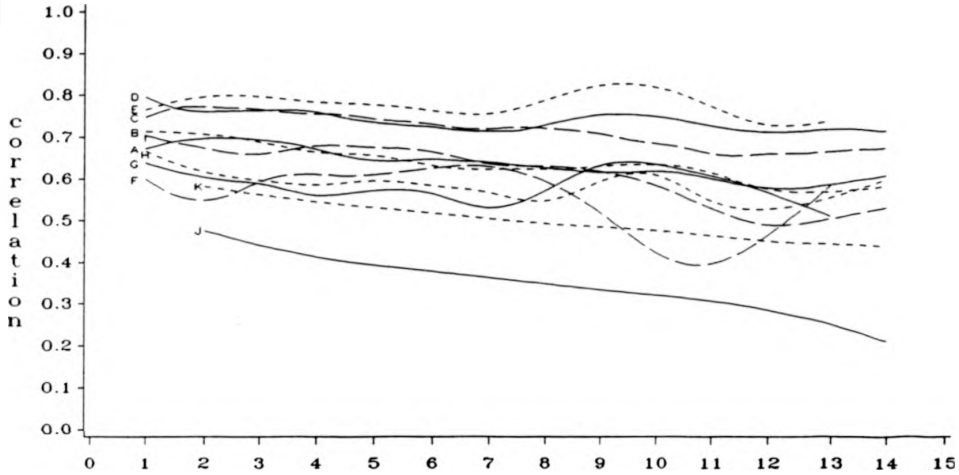
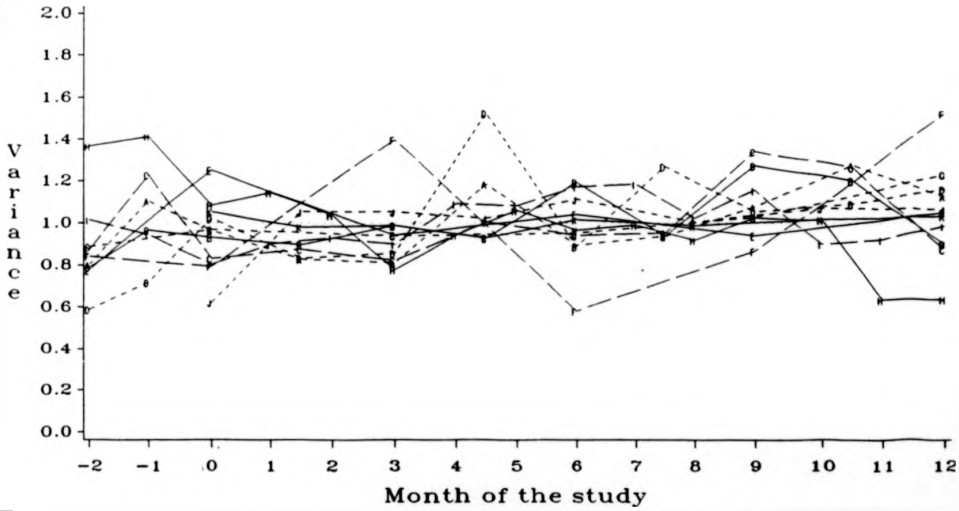


Figure 1.5.2: Variances over time for the 11 first variables in table 1.5.1
(For each variable the variances are scaled such that the overall mean equals 1)



We also need to consider the assumption of homogeneous variances over time. It is not possible to make any general conclusions for all possible untransformed biological variables, since variances often tend to increase with increasing mean values, and vice versa. However, to heuristically investigate how plausible an equi-variance assumption might be in practice, the variables underlying figure 1.5.1 are reused in figure 1.5.2 (using the same labelling of the curves) for illustrating how variances typically may change with time in clinical trials. For ease of comparison the variances have been scaled such that their average for each variable equals one. There is little evidence of a consistent pattern of change over time from this figure. However, calculating some summary statistics, there appears to be a small increase in variance with time. For instance, the mean of the pre-entry variances is .948, grouping the post-treatment variances into four-month periods, the averages are .985 (first 4 months), .996 (middle 4 months), and 1.069 (last 4 months). Calculating linear regression coefficients for the increase (decrease) in variance over time for the 11 variables, and testing whether the median of these is zero with Wilcoxon's signed-ranks test, results in rejection of the null hypothesis, $p=0.02$.

In summary, in most of these examples correlations tend to decline slightly over time and mixed correlations are somewhat lower. Also, variances might increase slightly with time. But there is in many of the examples no major departure from the compound symmetry assumption, which will often hold as an adequate approximation in practice.

1.6 THE SUMMARY STATISTIC APPROACH

1.6.1 Introduction

A profile (sequence of repeated measurements over time) usually consists of several observations from an underlying continuous process, and it is this process about which inferences are required. It may well be that the process itself is best represented by some summary statistics or derived variates calculated from the original measurements. This approach, which is termed the summary statistic approach, is particularly valuable when a direct comparison of mean profiles is inappropriate.

As described by Matthews et al (1990) this method considers the individual subject as the basic unit of analyses and uses the responses for each subject to construct a single number which summarizes some relevant aspect of that subject's response curve. Given the appropriate choice of summary statistics, the subsequent analysis is straightforward, since each statistic is treated like a conventional response and orthodox techniques can be applied. Very few assumptions are required to justify the validity of such an analysis. Estimates of error for the summary statistics are based solely on the randomisation in the experimental design, not on any assumptions about the covariance structure of the repeated measurements. If the statistics have a distribution that is far from normal then non-parametric methods can be used.

The simplicity and validity of the summary statistic technique are thus attractive features for the effective communication of clinical trial results. An appropriate choice of summary statistics enables the analysis to focus on relevant and clinically interpretable aspects of the response. What is not always clear is which summary statistic to use in a given situation. There exists many possible alternatives, primarily this choice is governed by the medical question underlying the trial. From an efficiency point of view, the way the outcome variable changes with time, and the covariance structure for the repeated measurements, will also have important consequences for this choice.

The summary statistic approach is not a new idea. Apart from the often obvious choice to analyse some kind of within-subject average value, the analysis of the individual regression coefficients resulting from orthogonal polynomial contrasts has sometimes been advocated. One of the earliest contributions in this respect being the classical paper 'Growth-rate determination in nutrition studies with the bacon pig, and their analysis' by Wishart (1938). Some other authors who have written on this topic are; Bradstreet (1993), Rowell and Walters (1976), and Leech and Healy (1959).

The terms "linear contrasts" and "orthogonal polynomials" indicate particular types of summary statistics. However, even when the underlying curves follow a polynomial, because the repeated measurements are intercorrelated, the use of least squares estimates is not optimal in any sense, but merely convenient (Potthoff and Roy, 1964). This will be elaborated on, and the summary statistics actually being optimal will be derived, in chapter 5.

It is important to bear in mind that it is the differences between the group time trends that determines the efficacy of a summary statistic, rather than the shape of the group trends themselves.

1.6.2 The General Linear Summary Statistic

The majority of the commonly used summary statistics are linearly weighted combinations of the outcomes. As a basis for much that follows, we now introduce a definition of a general linear summary statistic, and then give its expected value and variance under a general covariance structure.

It will be assumed that in a two treatment RCT with a continuous outcome variable, y, each subject has p measurements made before randomisation and r measurements after randomisation. The covariance matrix, consisting of the p+r variances and the $(p+r)(p+r-1)/2$ distinct covariances, is denoted by Σ. We assume that Σ is the same in both treatment groups.

Then, the general linear summary statistic, S_{ij}, where i indexes treatment group (usually A or B), and j indexes subject within treatment group (j=1,..,n_i), is given by

$$S_{ij} = \sum_{k=-(p-1)}^{r} c_k y_{ijk} = \mathbf{c}'\mathbf{y}_{ij}$$

where c_k denotes the weights for each measurement k (k=-(p-1),..,0,1,..r). Thus, S_{ij} is the summary statistic for subject j in treatment group i.

Denoting the true underlying mean vector for treatment group i by μ_i, the first two moments for the general linear summary statistic are given by;

$$E[S_{ij}] = \mathbf{c}'\mu_i \quad \text{and} \quad Var[S_{ij}] = \mathbf{c}'\Sigma\,\mathbf{c}$$

Assuming p=1 visit is made before and r=3 visits are made after the randomisation, a few straight forward examples of summary statistics in this general class are given by;

Post-randomisation mean: $\mathbf{c}' = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$

Change, last value-baseline: $\mathbf{c}' = \begin{bmatrix} -1 & 0 & 0 & 1 \end{bmatrix}$

Linear regression coefficient: $\mathbf{c}' \propto \begin{bmatrix} -3 & -1 & 1 & 3 \end{bmatrix}$

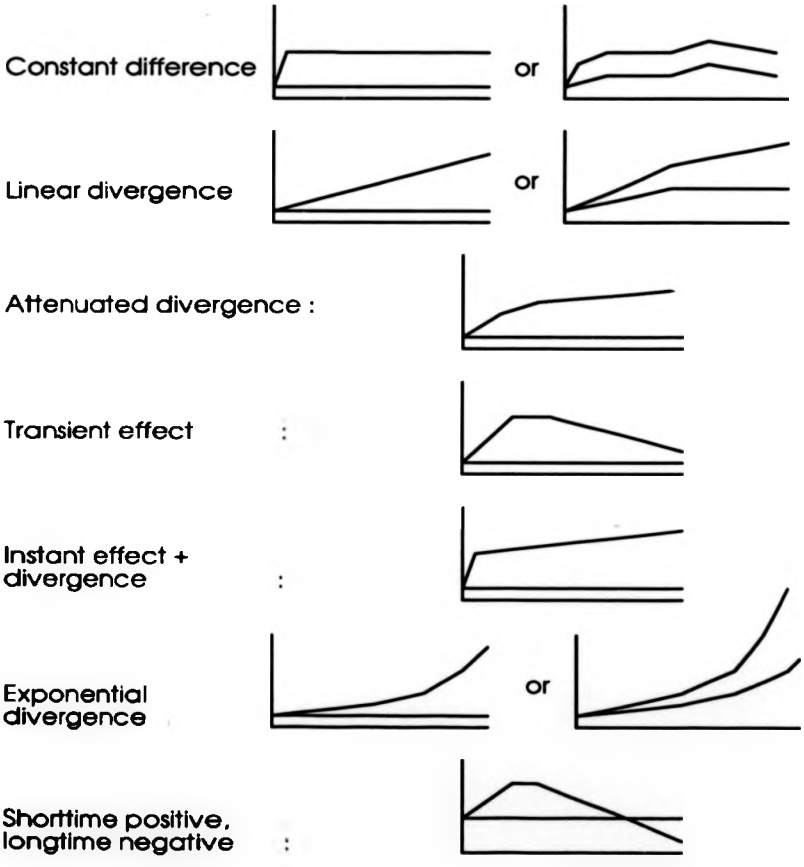
General formulae for the c'-vectors for the most common linear summary statistics will be given in section 5.4.

1.6.3 Categorization of response profiles

The best choice of a summary statistic, as far as efficiency and informativeness is concerned, depends on the clinical objectives of the study, the covariance structure, and the difference between the groups in the time trends (group means over time). The first thing to consider is to appropriately address the primary objective of the study.

This emphasizes one of the main attractiveness with the summary statistics, the possibility for a tailor-made approach to the analysis. For instance, is the clinical objective to maintain the subjects on a pH>6 during continuous pH measurement, then we might use the percentage of time each subject has spent above this threshold as a summary statistic. The covariance structure has already been considered in the preceding section, here some common classes of differences in mean response profiles over time for two treatment groups will be given.

Table 1.6.1: Examples of classes of differences in mean response profiles over time.



It is worth emphasizing that the actual shapes of the group mean profiles have no direct influence on the analysis, it is the difference between the mean profiles that matter. A constant difference might originate from an almost instant treatment effect, which remains stable over the time period under study. Alternatively, any differences in treatment effects over time which remain stable after a quick initial response falls in this category. This might be exemplified by the CPK-example (from a coronary heart disease study) described in section 2.5, and also by many studies on systolic and diastolic blood pressure lowering drugs.

Linear divergence is meant to mean a steadily increasing difference between the mean response profiles as time passes on. Examples of RCT's involving this type of divergence are often found in studies on pulmonary function data, e.g. PD_{20} (histamine response, see Van Essen-Zandvliet et al, 1992) and FEV_1 (see Diem and Liukkonen, 1988).

Attenuated divergence is something in between the two earlier mentioned categories. The difference between the mean curves increases over the whole study period, but the rate of divergence gets smaller and smaller. This model is often plausible, for instance, for CD4 cell counts in studies on HIV infection (see Dawson and Lagakos, 1991).

A transient effect might, for instance, be the result of a single-dose regimen, here, the mean curves diverges during the first phase of the study, until a maximum is achieved, after this the curves converge, and finally becomes identical again. An example of this kind might be found in Matthews et al (1990) in the context of aspirin concentration in the blood over time after a single dose at time zero.

The final three classes of differences in mean profiles in table 1.6.1 may be expected in certain applications. An instant effect, followed by some kind of divergence is suggested by the example on the concentration of steroisomers of a topical ophthalmic medication in the blood (Bradstreet, 1993).

A degree of divergence that increases with time appears to distinguish the groups in a study reported by Diggle (1988) concerned with the body-weights of rats. This might be modelled by, for example, a quadratic or an exponential divergence. Finally there is the possibility that a drug might show a short-time positive effect, which in the longer term turns out to be an adverse effect. This is sometimes found in cancer trials on tumour size data (see Chi, 1990).

1.6.4 Choice of summary statistics

When a constant difference in group time trends is anticipated, many plausible summary statistics are available. Often these are based on the average of each subjects post-treatment measurements, with or without some adjustment for the baseline level; e.g. post-treatment mean (POST), mean change (post-pre) (CHANGE), percentage change (from baseline to post mean), and analysis of covariance (using post mean as dependent variable, and baseline as covariate) (ANCOVA). Some alternative choices are; median value (of each subjects post-treatment measurements), and the area under the curve (the total area under a subjects response curve, formed by addition of the areas under the curve between each pair of consecutive observations, usually relying on a linear interpolation between the respective measurements) (AUC).

When group trends exhibit a linear divergence over time one might choose the linear regression coefficient for each subject (with or without baseline covariate adjustment) (SLOPE) or perhaps some other measure of rate of change (to be defined in chapter 5). Sometimes one of the summary statistics outlined in relation to a constant difference between mean curves might be useful, or a modification of one of these, like the mean of the last couple of measurements with a baseline covariate adjustment.

Peaked curves, such a plasma concentration curves (of some substance) over time, might be analysed using; maximum response (concentration) (c_{max}), time to reach maximum (t_{max}), and the area under the curve.

In some studies continuous 24 hour measurements are performed, for instance of the pH in the gastric juice in relation to anti-ulcer therapies. Useful summary measures for such studies might be; percentage time above some threshold (like pH 6), number of episodes below a certain ("at risk") level, and time to reach a predefined controlled level.

1.7 STRUCTURE OF THE REST OF THE THESIS

Chapters 2 and 3 cover the topic of "mean summary statistics". That is, summary statistics based on some kind of average of the post-treatment measurements for each subject, with or without some adjustment for baseline measurement(s). In section 2.1 a simple model is defined for RCT's with repeated measurements, and general formulae are given for the estimated difference in treatment effects and its variances, for the mean summary statistics; POST, CHANGE and ANCOVA. The statistical properties of these three commonly used approaches are explored, and the superiority of ANCOVA is documented. Sections 2.2 and 2.3 make more precise quantitative comparisons between the three approaches for two different classes of covariance structures, compound symmetry, and decaying correlations with time. While the three methods can be formulated as significance tests (two-sample t-tests and a covariance adjusted test of difference in mean respectively) emphasis is on estimating the magnitude of treatment difference.

There is little previous published information on statistical design considerations in repeated measures studies. Hence, section 2.4 is focused on the choice of the number of pre and post-treatment measurements, and the use of power calculations for determining the required number of subjects in repeated measures designs. Section 2.5 presents analyses of an example, and section 2.6 discusses the value and limitations of these relatively simple approaches.

The extent of bias in estimation if ANCOVA is not used, conditional on an observed mean pre-treatment difference, is described in section 3.1. Section 3.2 gives some guidance on the relative merits of increasing sample size or number of measurements for the efficiency of the analysis. Section 3.3 considers the issue of additive or multiplicative effects, and instances when the log-transformation is particularly useful are pointed out. A further summary statistic, the area under the curve, is explored in section 3.4. The final two sections on mean summary statistics, 3.5 and 3.6, are aimed at the recommended approach, ANCOVA. They investigate the optimal allocation of a fixed number of measurements before and after randomisation, and the issue of whether to use multiple baselines individually as separate covariates or as a single mean summary covariate in the ANCOVA model.

Chapter 4 is devoted to "regression to the mean", that is, the phenomenon that an individual with an extreme first measurement will tend to be closer to the centre of the distribution for a later measurement. Emphasis is on the effects of restrictive baseline values, as obtained from selection criteria. Sections 4.1 and 4.2 give some background for within-group comparisons with the necessary formulae for the effects of regression to the mean on means and variances for measurements taken both pre-entry and post-randomisation. Special interest is in the use of repeated pre-entry measurements to decrease the regression to the mean-effect. In section 4.3 some results for between-group comparisons are given. For studies where selection criteria are used for enrolling subjects, the value of allowing for an additional pre-entry visit, not underlying the selection, is evaluated, and results for the impacts on the variances of the three mean summary statistics are given.

Chapter 5 covers "optimal linear summary statistics", where the optimality refers to maximization of the between-group difference relative to its within-group standard deviation, under specified choices of the covariance structure and the mean group differences over time.

To enable comparisons between summary statistics and repeated measures designs, the notion of asymptotic relative efficiency (Pitman efficiency) for linear summary statistics is introduced in section 5.1. Based on Fisher's linear discriminant function, the optimal linear summary statistic is defined in section 5.2. In section 5.3 emphasis is on analysis of rate of change, and the optimal alternative to the slope as a summary statistic is given. Section 5.4 gives explicit formulae for the weights of the individual measurements to be used for some of the summary statistics. Section 5.5 gives optimal choices of, and relative efficiencies among, some summary statistics under specific classes of assumptions concerning the anticipated alternative hypothesis and the covariance matrix. In the last section the chapter's general relevance is reviewed.

The final chapter gives an overall perspective of the work done and the needs for some further research. In section 6.1 some of the approaches commonly used for repeated measures data are described, and their advantages and disadvantages relative to the summary statistic approach are discussed. Section 6.2 discusses the need for further methodology, e.g. allowance for missing values. Section 6.3 gives final conclusions on how the methods of this thesis should have an impact on the design and analysis of repeated measures clinical trials in everyday practice.

2 MEAN SUMMARY STATISTICS: THE FUNDAMENTAL ISSUES

In many clinical trials one's prime objective is to assess the average response to treatment over time, often (but not necessarily) in anticipation that treatment response is liable to occur quickly and to remain reasonably stable over time.

For a situation of this kind, the logical choice of summary statistic is some kind of mean of each subject's post-randomization measurements, possibly after adjusting in some way for pre-treatment measurements. This chapter will be concerned with this class of summary statistics, henceforth labeled "mean summary statistics".

2.1 GENERAL RESULTS

2.1.1 A simple model

In this section a simple model for randomized trials with repeated measures will be defined. Now we will restrict ourselves to the case of investigations encompassing two treatment groups. Most of the results, however, can in quite obvious ways be generalized to trials with more groups.

Going back to the model, suppose a randomized clinical trial has two treatment groups (i=A or B) with n_i patients per group, and suppose all patients have p pre-treatment visits k= -(p-1) ... 0, and r post-treatment visits, k=1 ... r. A quantitative measurement x is observed at every visit for every patient, and we adopt the simple model:

$x_{ijk} = \mu_{ik} + e_{ijk}$ for i=A or B, j=1,..,n_i and k=-(p-1),..,0,1,..,r

μ_{ik} is the true underlying mean response for treatment i at time k. As a result of randomization we can assume $\mu_{Ak} = \mu_{Bk}$ for the pre-treatment visits $k \leq 0$. e_{ijk} is the j^{th} patient "error" or residual variation around the underlying mean μ_{ik}, and these errors will not be independent within patients.

Hence, let $\Sigma = \{\sigma_{kl}\}$ be the covariance matrix for all pairs of measurement times k,l. For simplicity we assume this is the same for both treatments.

It is helpful to define 3 submatrices:

$\Sigma_{post} = \{\sigma_{kl}\}$ for k,l = 1 ... r ,

$\Sigma_{pre} = \{\sigma_{kl}\}$ for k,l = -(p-1) ... 0 , and

$\Sigma_{mix} = \{\sigma_{kl}\}$ for k = -(p-1) ... 0, and l = 1 ... r

so that we can display $\Sigma = \begin{bmatrix} \Sigma_{pre} & \Sigma'_{mix} \\ \Sigma_{mix} & \Sigma_{post} \end{bmatrix}$.

Also define $\sigma_{kl} = \rho_{kl} \cdot \sigma_k \cdot \sigma_l$ where ρ_{kl} is the within-treatment group pairwise correlation between a patient's measurements at visits k and l, and σ_k, σ_l are the standard deviations at visits k,l within each treatment group. We expect the correlations ρ_{kl} to be substantial (typically greater than 0.5 in most trials, see table 1.5.1 and the examples given there) since they reflect the consistency of patient effects over time, which are otherwise not explicitly included in this simple model.

2.1.2 The three approaches

Even for the subclass of clinical trials where interest centers around overall levels of response, the choice of summary statistics is wide. Possible candidates could for instance be; post-treatment mean, mean change relative to baseline, covariance analysis (with baseline value as covariate), end-value, end-value - baseline, area under the curve, median post-treatment, etc.

In this chapter we will be mainly concerned with the first three of these statistics. More precise definitions of the three approaches are as follows:

1) Post-treatment means (POST): a simple analysis using the mean of each patient's post-treatment measurements as summary measure.

2) Mean changes (CHANGE): a simple analysis of each patient's difference between mean of post-treatment measurements and mean of baseline measurements, the latter often consisting of just a single baseline value per patient.

3) Analysis of covariance (ANCOVA): between-patient variations in baseline measurements are taken into account, by using the mean baseline measurement for each patient as covariate in a linear model for treatment comparison of post-treatment means.

For brevity, these methods will henceforth be referred to as POST, CHANGE and ANCOVA respectively.

2.1.3 Estimates and variance formulae

Let $\bar{\Sigma}_{post}, \bar{\Sigma}_{pre}$ and $\bar{\Sigma}_{mix}$ be the respective means of the r^2, p^2 and rp components of the three submatrices $\Sigma_{post}, \Sigma_{pre}$ and Σ_{mix} defined above. Using this notation we can define the following variance formulae for the three analysis approaches under investigation.

1) Post-treatment means (POST):

For each individual the summary statistic is $\bar{x}_{ij.}^{post} = \frac{1}{r}\sum_{k=1}^{r} x_{ijk}$

Overall post-treatment mean difference =

$$\frac{1}{n_A}\sum_{j=1}^{n_A} \bar{x}_{Aj.}^{post} - \frac{1}{n_B}\sum_{j=1}^{n_B} \bar{x}_{Bj.}^{post} = \bar{\bar{x}}_{A..}^{post} - \bar{\bar{x}}_{B..}^{post}$$

which has expected value = $\bar{\mu}_{A.}^{post} - \bar{\mu}_{B.}^{post}$

For an individual patient, the variance for the summary statistic is:

$$Var\left[\frac{x_{ij1}+\ldots+x_{ijr}}{r}\right]=\frac{1}{r^2}\cdot\sum_{k=1}^{r}Var\left[x_{ijk}\right]+2\cdot\frac{1}{r^2}\cdot\sum_{g<h\le r}Cov\left[x_{ijg},x_{ijh}\right]=\overline{\Sigma}_{post}$$

The last equality is easily seen to hold from a term-by-term comparison with the covariance matrix.

Thus, $Var\left[\overline{\overline{x}}_{A..}^{post}-\overline{\overline{x}}_{B..}^{post}\right]=\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\overline{\Sigma}_{post}$

2) Mean changes (CHANGE):

For each individual the summary statistic is the mean change,

$$\frac{1}{r}\cdot\sum_{k=1}^{r}x_{ijk}-\frac{1}{p}\cdot\sum_{k=-(p-1)}^{0}x_{ijk}=\overline{x}_{ij.}^{post}-\overline{x}_{ij.}^{pre}$$

Then the overall treatment difference in these mean changes

$$=\frac{1}{n_A}\sum_{j=1}^{n_A}(\overline{x}_{Aj.}^{post}-\overline{x}_{Aj.}^{pre})-\frac{1}{n_B}\sum_{j=1}^{n_B}(\overline{x}_{Bj.}^{post}-\overline{x}_{Bj.}^{pre})=(\overline{\overline{x}}_{A..}^{post}-\overline{\overline{x}}_{A..}^{pre})-(\overline{\overline{x}}_{B..}^{post}-\overline{\overline{x}}_{B..}^{pre})$$

which has expected value again equal to $\overline{\mu}_{A.}^{post}-\overline{\mu}_{B.}^{post}$ since the pre-treatment expected values are the same for both treatments.

In the same fashion as for POST it is easily shown that

$$Var\left[(\overline{\overline{x}}_{A..}^{post}-\overline{\overline{x}}_{A..}^{pre})-(\overline{\overline{x}}_{B..}^{post}-\overline{\overline{x}}_{B..}^{pre})\right]=\left(\frac{1}{n_A}+\frac{1}{n_B}\right)(\overline{\Sigma}_{post}+\overline{\Sigma}_{pre}-2\cdot\overline{\Sigma}_{mix})$$

3) Analysis of covariance (ANCOVA):

The model for ANCOVA based on the individual's post-treatment mean $\bar{x}_{ij.}^{post}$, with the pre-treatment mean $\bar{x}_{ij.}^{pre}$ as a covariate, is as follows: $\bar{x}_{ij.}^{post} = \bar{\mu}_{i.}^{post} + \beta \cdot \left(\bar{x}_{ij.}^{pre} - \bar{\mu}_{..}^{pre}\right) + \varepsilon_{ij}$ where ε_{ij} are independent random errors with assumed constant variance. With estimate $\hat{\beta}$ obtained by least squares, we may define $\bar{x}_{ij.}^{cov} = \bar{x}_{ij.}^{post} - \hat{\beta} \cdot \left(\bar{x}_{ij.}^{pre} - \bar{\bar{x}}_{..}^{pre}\right)$, where $\bar{\bar{x}}_{..}^{pre}$ stands for the overall pre-treatment mean averaged over groups.

Then the estimated mean treatment difference =

$$\frac{1}{n_A} \cdot \sum_{j=1}^{n_A} \bar{x}_{Aj.}^{cov} - \frac{1}{n_B} \cdot \sum_{j=1}^{n_B} \bar{x}_{Bj.}^{cov} \quad = \bar{\bar{x}}_{A.}^{cov} - \bar{\bar{x}}_{B.}^{cov},$$

which again has expected value = $\bar{\mu}_{A.}^{post} - \bar{\mu}_{B.}^{post}$.

From the variance formula for ANCOVA (see Fleiss, 1986)

$$Var\left(\bar{\bar{x}}_{A.}^{cov} - \bar{\bar{x}}_{B.}^{cov}\right) = \frac{n_A + n_B - 2}{n_A + n_B - 3} \cdot \left[1 - corr^2\left(\bar{x}_{ij.}^{pre}, \bar{x}_{ij.}^{post}\right)\right] \cdot \mathrm{var}\left(\bar{x}_{ij.}^{post}\right) \left[\frac{1}{n_A} + \frac{1}{n_B} + \frac{\left(\bar{\bar{x}}_{A.}^{pre} - \bar{\bar{x}}_{B.}^{pre}\right)^2}{(n_A + n_B - 2) \cdot \mathrm{var}\left(\bar{x}_{ij.}^{pre}\right)}\right]$$

The first term corresponds to the loss of one degree of freedom, due to estimation of the slope. The additional correction factor in the last term allows for the fact that sampling error of the estimated slope leads to a correlation between $\bar{\bar{x}}_{A.}^{cov}$ and $\bar{\bar{x}}_{B.}^{cov}$.

Using the above notation for components of the pooled within-subject covariance matrix,

$$Var\left(\bar{\bar{x}}_{A.}^{cov} - \bar{\bar{x}}_{B.}^{cov}\right) = \frac{n_A + n_B - 2}{n_A + n_B - 3} \cdot \left(\bar{\Sigma}_{post} - \frac{\bar{\Sigma}_{mix}^2}{\bar{\Sigma}_{pre}}\right) \left[\frac{1}{n_A} + \frac{1}{n_B} + \frac{\left(\bar{\bar{x}}_{A.}^{pre} - \bar{\bar{x}}_{B.}^{pre}\right)^2}{(n_A + n_B - 2) \cdot \bar{\Sigma}_{pre}}\right]$$

As the sample size increases the first term approaches unity, and in randomized trials the last term becomes negligible. Hence, for any reasonable size of trial (say 50 ≥ subjects per group) we can use the simpler approximation:

$$Var\left(\bar{\bar{x}}_A^{cov} - \bar{\bar{x}}_B^{cov}\right) \approx \left(\frac{1}{n_A} + \frac{1}{n_B}\right) \cdot \left(\bar{\bar{\Sigma}}_{post} - \frac{\bar{\bar{\Sigma}}_{mix}^2}{\bar{\bar{\Sigma}}_{pre}}\right).$$

In summary, for a randomized clinical trial all three estimates of the mean treatment difference have the same expected value, $\bar{\mu}_{A.}^{post} - \bar{\mu}_{B.}^{post}$.

Given the common sample size adjustment $\left(\frac{1}{n_A} + \frac{1}{n_B}\right)$, the comparison of variance magnitudes may be expressed as:

POST, variance proportional to $\bar{\bar{\Sigma}}_{post}$

CHANGE, variance proportional to $\bar{\bar{\Sigma}}_{post} + \bar{\bar{\Sigma}}_{pre} - 2 \cdot \bar{\bar{\Sigma}}_{mix}$

ANCOVA, variance approximately proportional to $\bar{\bar{\Sigma}}_{post} - \frac{\bar{\bar{\Sigma}}_{mix}^2}{\bar{\bar{\Sigma}}_{pre}}$

It can be readily shown that ANCOVA always has a smaller variance than POST.

ANCOVA also produces a smaller population variance than CHANGE,

Proof:

$$Var[\text{ANCOVA}] \le Var[\text{CHANGE}] \Leftrightarrow \bar{\bar{\Sigma}}_{post} - \frac{\bar{\bar{\Sigma}}_{mix}^2}{\bar{\bar{\Sigma}}_{pre}} \le \bar{\bar{\Sigma}}_{post} + \bar{\bar{\Sigma}}_{pre} - 2\bar{\bar{\Sigma}}_{mix} \Leftrightarrow$$

$$-\frac{\bar{\bar{\Sigma}}_{mix}^2}{\bar{\bar{\Sigma}}_{pre}} \le \bar{\bar{\Sigma}}_{pre} - 2\bar{\bar{\Sigma}}_{mix} \Leftrightarrow \bar{\bar{\Sigma}}_{mix}\left(2 - \frac{\bar{\bar{\Sigma}}_{mix}}{\bar{\bar{\Sigma}}_{pre}}\right) \le \bar{\bar{\Sigma}}_{pre} \Leftrightarrow \frac{\bar{\bar{\Sigma}}_{mix}}{\bar{\bar{\Sigma}}_{pre}}\left(2 - \frac{\bar{\bar{\Sigma}}_{mix}}{\bar{\bar{\Sigma}}_{pre}}\right) \le 1$$

After making the substitution: $x = \frac{\bar{\Sigma}_{mix}}{\bar{\Sigma}_{pre}}$, we can express the left-hand side of the inequality as: $f(x) = x(2-x)$. It is easily shown that this function reaches its maximum at x=1 where the value of the function is 1, i.e.: $\bar{\Sigma}_{mix} = \bar{\Sigma}_{pre} \Rightarrow$ Var[ANCOVA] = Var[CHANGE], otherwise always: Var[ANCOVA] < Var[CHANGE]

For a design with two measurements, one pre-entry and one post-treatment, this superiority of ANCOVA may be shown in a more direct way. A general covariance matrix for this kind of design is given by:

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.$$

The corresponding variances for the three approaches are related as follows:

$$Var[POST] \quad \propto \sigma_y^2$$
$$Var[CHANGE] \propto \sigma_y^2 + \sigma_x^2 - 2\rho\sigma_x\sigma_y$$
$$Var[ANCOVA] \propto \sigma_y^2(1-\rho^2)$$

Submitting the data to an arbitrary scaling, whereby all the measurements are divided by σ_y $\left(x^* = x/\sigma_y, y^* = y/\sigma_y\right)$, we arrive at the following covariance matrix:

$\Sigma^* = \begin{bmatrix} \sigma_{x^*}^2 & \rho\sigma_{x^*} \\ \rho\sigma_{x^*} & 1 \end{bmatrix}$. The variances for the arbitrarily scaled data are now proportional to:

$$Var[POST] \quad \propto 1$$
$$Var[CHANGE] \propto 1 - \rho^2 + (\sigma_{x^*} - \rho)^2$$
$$Var[ANCOVA] \propto 1 - \rho^2$$

The superiority of ANCOVA relative to POST as well as to CHANGE is now evident. It may also be observed that Var[CHANGE] ≤ Var[POST] if $\rho \geq \sigma_{x^*}/2$, i.e., if $\beta \geq .5$ (since $\beta = Cov(x^*, y^*)/Var(x^*) = \rho/\sigma_{x^*}$).

Hence, for the mean summary statistic approach we have shown that (disregarding the minor correction factors for the variance of ANCOVA), ANCOVA is superior to a) ignoring pre-treatment readings and b) simply subtracting pre-treatment readings for each individual.

2.2 RESULTS WITH COMPOUND SYMMETRY

When trying to derive new and useful results, there are basically two different directions one might take. First one can go for general results which are valid in most circumstances. Then few assumptions are needed, the results hold globally, but usually little can be said about specific examples. The second option is to make more assumptions. Then the generalizability gets more restricted, but more specific and useful results (in an applied sense) will be achieved for these examples in line with the assumptions chosen.

In the preceding section, the more general road was followed, with variance formulae valid for any variance/covariance matrix. In this section more assumptions will be imposed, allowing us to produce more specific results.

More specifically, compound symmetry will be assumed. That is, we will assume equal variances for all time-points and both treatments and also equal correlations between all pairs of time-points. Thus,

$$Var(x_{ijk}) = \sigma^2 \text{ and } Corr(x_{ijk}, x_{ijl}) = \rho \text{ for } k \neq l.$$

Admittingly these are quite restrictive assumptions, nevertheless they are used quite frequently, both in the literature, and for some standard statistical techniques. For example repeated measurements ANOVA (without correcting the degrees of freedom for the F-statistic) uses an only slightly more general assumption for the covariance matrix, labeled the Huynh-Feldt type H-structure (Huynh and Feldt, 1970), whereby all normalized contrasts among all repeated measurements have to have the same variance.

From the real-world examples of covariance matrices presented in table 1.5.1, it was seen that compound symmetry often is a quite realistic model for the joint variability in a data set. Also, as is shown in section 2.2.3, the results derived below are quite insensitive to modest departures from these assumptions.

2.2.1 Comparison of variances with a single baseline

Often there is just a single pre-treatment measurement and several (r) measurements after randomization for each patient and we now focus on this simple case.

Under the assumption of compound symmetry, the variances for the three approaches currently under investigation can be rewritten as :

$$\text{POST, variance} = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\frac{\sigma^2}{r}\left[1+(r-1)\rho\right]$$

$$\text{CHANGE, variance} = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\frac{\sigma^2}{r}\left[1+(r-1)\rho+r(1-2\rho)\right]$$

$$\text{ANCOVA, variance} \approx \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\frac{\sigma^2}{r}\left[1+(r-1)\rho-r\rho^2\right]$$

Figure 2.2.1 compares the resulting variances (arbitrarily scaled, σ^2 is factored out), when assuming a correlation of 0.6 (which is often found in practice, as was observed from the examples in table 1.5.1). With this degree of intra-subject correlation, and with one pre-entry measurement, the ranking order between the three approaches is quite clear, with POST performing least favourably and ANCOVA most favourably. Also evident from this figure is the increase in precision gained by increasing the number of post-treatment visits, which is of an identical magnitude for all the three approaches.

Figures 2.2.2 to 2.2.4 makes pairwise comparisons between the approaches by plotting ratios of variances for various values of ρ and r. First comparing POST with CHANGE, with a correlation of 0.5 the variances are identical irrespective of the number of post-treatment visits. With lower correlations POST is more favourable, with higher correlations CHANGE is more advantageous. For any given correlation, the approach (POST or CHANGE) which is more efficient with a single post-treatment measurement, will be increasingly more favourable as the number of visits post-randomisation increase.

Comparing POST with ANCOVA (figure 2.2.3), the former becomes more inferior the larger the correlation ρ. This inferiority is somewhat more marked if the number of post-treatment visits is substantial. If ρ=0, then the pre-treatment measurements are of no value, that is β=0 in ANCOVA in which case the two approaches are in principle equivalent. We may plausibly expect ρ in the range .5 to .7, in which case the variance for ANCOVA will be around 40% to 60% less than for POST. This reflects the serious loss of statistical efficiency incurred by failing to take account of pre-treatment measurements.

CHANGE becomes less inferior to ANCOVA as the correlation ρ increases (see figure 2.2.4). Again, for any value of ρ the inferiority of CHANGE becomes somewhat more accentuated as the number of post-treatment measurements increases. For the plausible values of ρ in the range .5 to .7, the variance for ANCOVA will be around 20% to 40% less than for CHANGE.

Figure 2.2.1 :

Variances for POST, CHANGE and ANCOVA depending on r, assuming equicorrelation with ρ=0.6 and one pretreatment measure

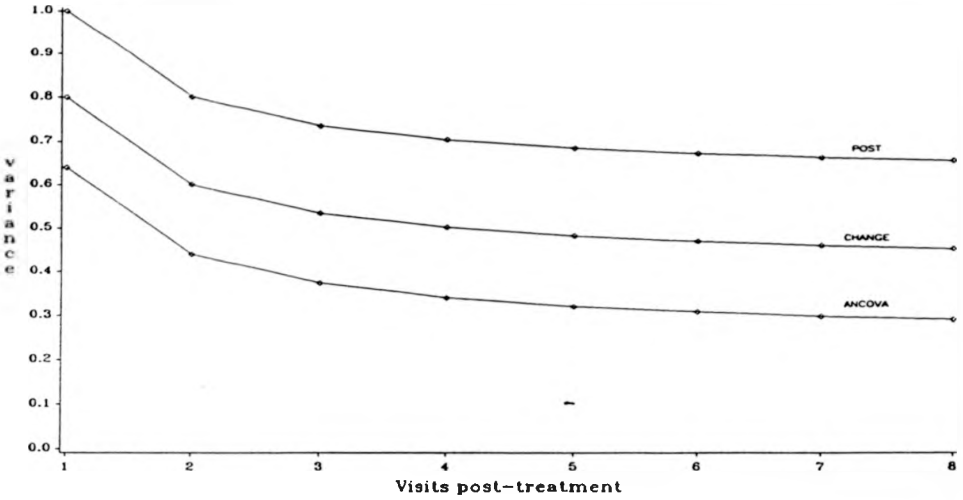


Figure 2.2.2 :

Dependence of Var(Change)/Var(Post) on r and ρ, assuming equicorrelation ρ and one pretreatment measure.

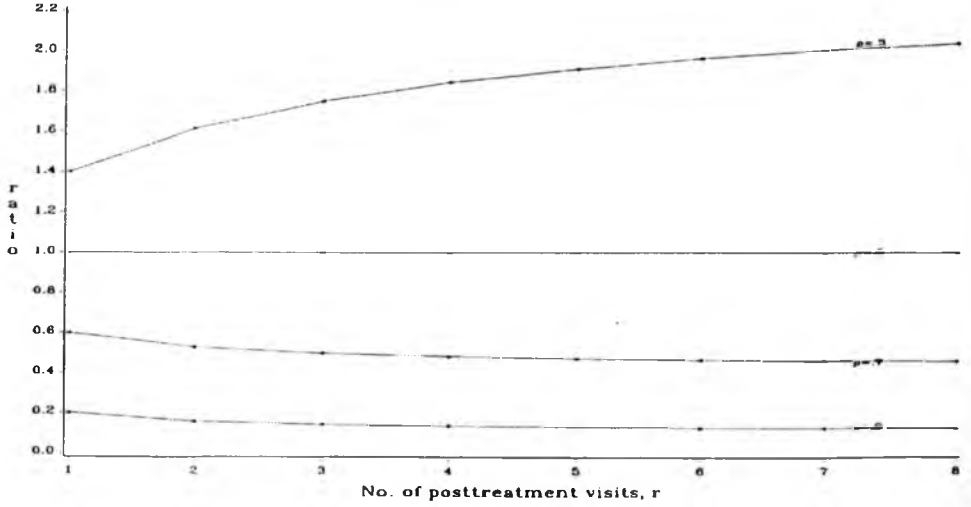


Figure 2.2.3 :

Dependence of Var(Ancova)/Var(Post) on r and ρ, assuming equicorrelation ρ and one pretreatment measure.

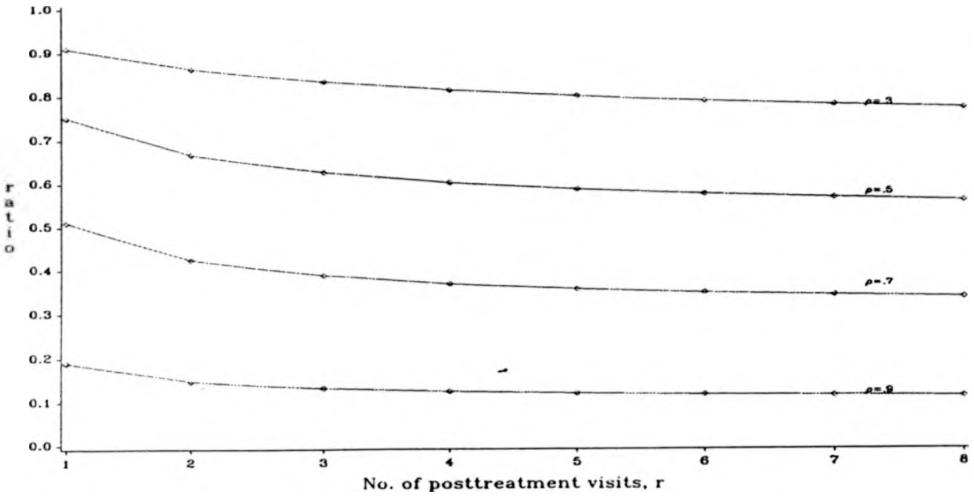
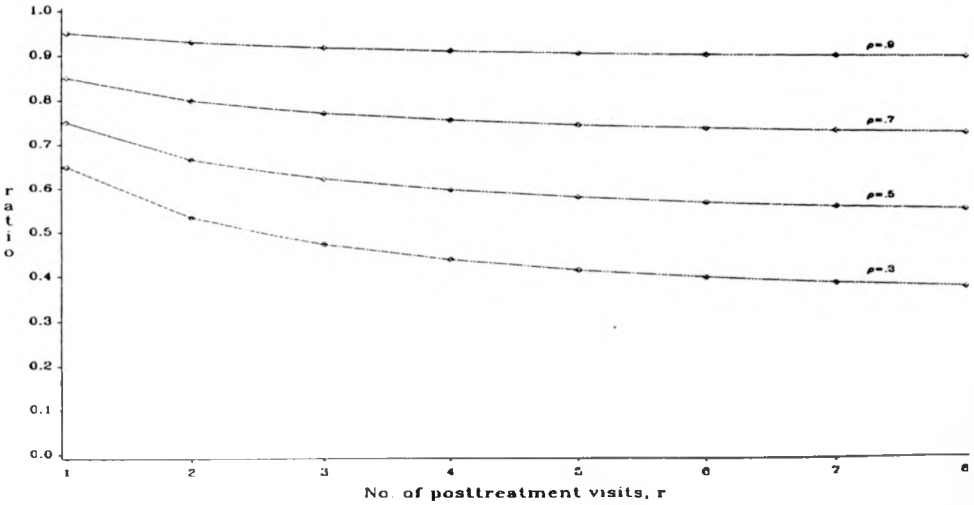


Figure 2.2.4 :

Dependence of Var(Ancova)/Var(Change) on r and ρ, assuming equicorrelation ρ and one pretreatment measure.



Note again that for ρ=0.5 POST and CHANGE have identical variances. Our examples (see section 1.5) suggest ρ will commonly be somewhat higher, so that CHANGE will be better than POST. However, with just a single pre-treatment measurement it seems likely that both analyses will be substantially inferior to ANCOVA in most practical circumstances.

2.2.2 Consequences of having more pre-treatment measurements

It is often possible to have more than one pre-treatment visit in a repeated measures design (all pre-treatment visits occurring before randomization), and here we consider the improved efficiency for both ANCOVA and CHANGE. Of course the time lapses between pre-treatment measurements may effect the correlation structure, but for simplicity we continue to explore the statistical properties under compound symmetry.

With r post-treatment measurements and p pre-treatment measurements we have

$$\bar{\Sigma}_{post} = \sigma^2\left[\frac{1+(r-1)\rho}{r}\right], \quad \bar{\Sigma}_{pre} = \sigma^2\left[\frac{1+(p-1)\rho}{p}\right] \text{ and } \bar{\Sigma}_{mix} = \sigma^2\rho$$

$$\text{For CHANGE, variance} = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma^2\left[\frac{1+(r-1)\rho}{r} - \frac{(p+1)\rho-1}{p}\right]$$

$$\text{For ANCOVA, variance} = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma^2\left[\frac{1+(r-1)\rho}{r} - \frac{p\rho^2}{1+(p-1)\rho}\right]$$

First, consider the advantage of extra pre-treatment visits while keeping the number of post-treatment visits fixed. Then, CHANGE becomes superior to POST provided ρ>1/(p+1). This means that provision of two or more pre-treatment measurements will make CHANGE the better option unless correlations are small, which appears unlikely in practice.

More important is the extent to which extra pre-treatment measurements make CHANGE closer in statistical efficiency to ANCOVA. From the above formulae it is easy to show that if $\rho=0.5$ then ANCOVA with p pre-treatment measurements has the same variance as CHANGE with p+1 pre-treatment measurements. For $\rho>0.5$, which is quite likely in practice, this gap between the two methods is narrowed more rapidly.

Table 2.2.1 compares ANCOVA and CHANGE for r=10 post-treatment visits and p=1,..,5 pre-treatment visits, all variances being expressed as a proportion of the ANCOVA variance for p=1. For instance, for p=5 and $\rho=0.7$ the variance reduction for ANCOVA relative to CHANGE is only 5%. This is because the observed pre-treatment mean more closely estimates the true pre-treatment level for each patient. Consequently the "regression to the mean" problem (the tendency for variables that are extreme on its first measurement to be closer to the center of the distribution for a later measurement) in a mean changes analysis is reduced and the estimated slope $\hat{\beta}$ in ANCOVA becomes closer to unity. It is worth pointing out at this stage that CHANGE is not only inferior to ANCOVA in terms of variances, for any true $\beta<1$ CHANGE (by always assuming $\beta=1$) will overcorrect for any existing mean pre-treatment differences, and thus give (conditionally) biased results.

For ANCOVA, addition of more pre-treatment visits is always helpful, but especially so if ρ is large. For instance, if $\rho=0.7$, then having a second pre-treatment visit reduces the variance by 36%. Further somewhat less substantial gains are made by adding a third pre-treatment visit, and so on. This proportionate gain for ANCOVA, as shown in Table 2.2.1 for r=10, is reduced slightly for a smaller number of post-treatment visits.

Table 2.2.1 : The dependence of the variances for ANCOVA and CHANGE on the number of pre-treatment measurements p and the equi-correlation ρ between time-points assuming r=10 post-treatment visits. For each ρ, variances are divided by the variance for ANCOVA with p=1.

ρ	Analysis	Number of pre-treatment measurements, p				
		1	2	3	4	5
.3	Ancova	1.000	0.827	0.719	0.645	0.591
	Change	2.750	1.500	1.083	0.875	0.750
.5	Ancova	1.000	0.722	0.583	0.500	0.444
	Change	1.833	1.000	0.722	0.583	0.500
.7	Ancova	1.000	0.640	0.490	0.407	0.355
	Change	1.375	0.750	0.542	0.438	0.375
.9	Ancova	1.000	0.574	0.421	0.343	0.296
	Change	1.100	0.600	0.433	0.350	0.300

In some repeated measures designs there may be a fixed total number of visits p+r=t, and we can therefore only increase the number of pre-treatment visits p at the expense of the number of post-treatment visits r.

Then, for CHANGE, variance = $\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma^2\frac{t(1-\rho)}{p(t-p)}$

and for ANCOVA, variance = $\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma^2\frac{t(1-\rho)}{p(t-p)}\cdot\left[1-\frac{(t-p)(1-\rho)}{t\left[1+(p-1)\rho\right]}\right]$

For CHANGE, the variance is minimized for p=r, that is equal numbers of pre and post readings when t is even and p=(t-1)/2 or (t+1)/2 when t is odd.

However it is more important to consider the choice of p for ANCOVA given a fixed total number of visits p+r=t. In general, the "minimum variance" choice of p for any given t becomes larger as ρ increases, because the pre-treatment readings are of greater use, that is $\hat{\beta}$ becomes larger. More specifically, we can show that for any choice of integer p' then if $\rho=1/(t-2p')$, the variances of ANCOVA for p=p' and p=p'+1 are the same. If $\rho<1/(t-2p')$ then p=p' produces a smaller variance, and if $\rho>1/(t-2p')$ then p=p'+1 produces a smaller variance. Thus, the optimal choice is p=p' when ρ lies between $1/[t-2(p'-1)]$ and $1/(t-2p')$ for $p'>0$. Also p=0 if $\rho<1/t$.

For example, if t=10 measurements in total, to achieve minimum variance for ANCOVA we would set p'=5 and divide them equally between pre- and post-treatment readings if $\rho>1/2$ and set p'=4 if $1/2>\rho>1/4$. Smaller values of ρ are unlikely to occur in practice. Hence, if the aim is to minimize the ANCOVA variance, p should be not much smaller than t/2, since precision of the individual's pre-treatment mean level is almost as important as precision of the post-treatment mean level. For more considerations of the optimal choice of p for a given t, in particular for more general covariance structures, see section 3.5.

Of course, reduction in variance is not the only criterion affecting the choice of p. We usually wish to concentrate on the post-treatment readings to describe the shape of mean change over time (for example, is the treatment difference constant, increasing or peaked?) and post-treatment measurements may be required at certain intervals for patient monitoring. Departure from the equi-correlation assumption is also relevant. For instance, if the average correlation between pre and post readings was considerably lower than the average pairwise correlation between pairs of post-treatment readings then the "minimum variance" p would be further from t/2. Nevertheless, the above results appropriately reflect the merit of having multiple pre-treatment readings if practicable.

However, it may sometimes be unfeasible or unethical to obtain multiple pre-treatment measurements at adequate intervals. For instance, if randomization must occur soon after the first visit, there may be no opportunity for repeat pre-treatment visits or their spacing may have to be so close in time that they do not provide sufficiently independent measurements to improve estimation of the subject's "true baseline". In most applications, it may be difficult to define what is an adequate minimum spacing, though having p>1 can only do good!

It should be noted that the greatest gain in efficiency is by having p=2 rather than p=1. For instance, with t=10 readings in all and ρ=0.7, the reduction in ANCOVA variance for p=2 versus p=1 is 34% while for p=5 versus p=1 the reduction is 53%. In practice, some compromise is needed between precision of overall treatment effect estimation (p sufficiently large) and adequate description of the time pattern of treatment response (r sufficiently large).

The statistical consequences of increasing the number of post-treatment readings r is the same for all three methods of analysis. Under equi-correlation assumption the reduction in each variance by having r+1 rather than r post-treatment readings is equal to

$$\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\frac{\sigma^2(1-\rho)}{r(r+1)}.$$

The practical consequence of this reduction in variance for increasing r might best be viewed in the context of power calculation, as described in section 2.4.

2.2.3 Sensitivity analysis for the compound symmetry assumption

Since many of the comparisons and recommendations in this chapter are based on the assumption of compound symmetry, it is important to consider the impact that departure from this assumption have on the results presented so far.

We suspect non-equal correlations is a more serious problem than inequality of variances (see figures 1.5.1 and 1.5.2 and the accompanying comments), and will focus on the alterations to the variances of the mean summary statistics as a means of illustrating the implications of unequal correlations.

Let $\bar{\rho}_{post}, \bar{\rho}_{mix}$ and $\bar{\rho}_{pre}$ be the mean pairwise correlations (excluding the "self-correlations" of 1) in the post-post, pre-post and pre-pre covariance submatrices $\Sigma_{post}, \Sigma_{mix}$ and Σ_{pre} respectively. Then, based on the general variance formulae for the summary statistics given in section 2.1, and by substituting the general means for the submatrices of Σ for the means we get when using the mean pairwise correlations given above, it is easily shown that the variances of treatment differences are proportional to the following:

POST: $$\frac{1+(r-1)\bar{\rho}_{post}}{r}$$

CHANGE: $$\frac{1+(r-1)\bar{\rho}_{post}}{r}+\frac{1+(p-1)\bar{\rho}_{pre}}{p}-2\bar{\rho}_{mix}$$

ANCOVA: $$\frac{1+(r-1)\bar{\rho}_{post}}{r}-\frac{\bar{\rho}_{mix}^2}{\left(\frac{1+(p-1)\bar{\rho}_{pre}}{p}\right)}.$$

Therefore, determination of variances and its dependence on r, p and the method of analysis can all be documented if one knows the values of the three parameters $\bar{\rho}_{post}, \bar{\rho}_{mix}$ and $\bar{\rho}_{pre}$. With these formulae it is possible to take the time-spacing of pre-entry measurements into account. Often pre-randomisation visits are performed with shorter time-intervals in between, than are post-randomisation visits. If this is considered to produce higher correlations among the pre-entry measurements, we can adjust the assumed value for $\bar{\rho}_{pre}$ accordingly.

The practical consequences of such departures from compound symmetry are usefully explored in the context of sample size calculations. We return to this in subsection 2.4.3.

An alternative method of exploring the consequences of assuming equi-correlation, when correlations are not stable over time, is as follows. If p=1 baseline measurement, suppose $\bar{\rho}_{mix}$ and $\bar{\rho}_{post}$ differ, but the overall mean correlation (which under these circumstances equals $\frac{2\cdot\bar{\rho}_{mix}+(r-1)\cdot\bar{\rho}_{post}}{r+1}$) is kept fixed.

Having $\bar{\rho}_{mix} > \bar{\rho}_{post}$ seems illogical in practice. Hence, we anticipate underestimation of the variances for our three mean summary statistics under the simplifying assumption of compound symmetry. Specifically, the absolute magnitude of the underestimation [all times $\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\cdot\sigma^2$] is given by;

$$POST \quad : \quad \frac{2(r-1)}{r(r+1)}\cdot\left(\bar{\rho}_{post}-\bar{\rho}_{mix}\right)$$

$$CHANGE: \quad \frac{2(r-1)}{r}\cdot\left(\bar{\rho}_{post}-\bar{\rho}_{mix}\right)$$

$$ANCOVA: \quad \frac{2(r-1)}{r(r+1)}\cdot\left(\bar{\rho}_{post}-\bar{\rho}_{mix}\right)-\bar{\rho}_{mix}^2+\frac{1}{(r+1)^2}\left[(r-1)\bar{\rho}_{post}+2\bar{\rho}_{mix}\right]^2$$

The underestimation for CHANGE will always be (r+1) times as big as for POST, the underestimation for ANCOVA will always lie somewhere in between. An example of how the variances for the three approaches are affected by different mean correlations in the different submatrices of Σ is given in figure 2.2.5. Here a study design encompassing 1 pre and 3 post-treatment visits is assumed, and the dependence of the variances on the difference $\bar{\rho}_{post}-\bar{\rho}_{mix}$ is visualized when the overall mean correlation is 0.6 .

Figure 2.2.5 :

Variances for the three approaches depending on the difference in mean correlations, post−mix. Assuming 1+3 visits, equivariance over time, and an overall mean correlation of 0.6 . P=POST, C=CHANGE, A=ANCOVA

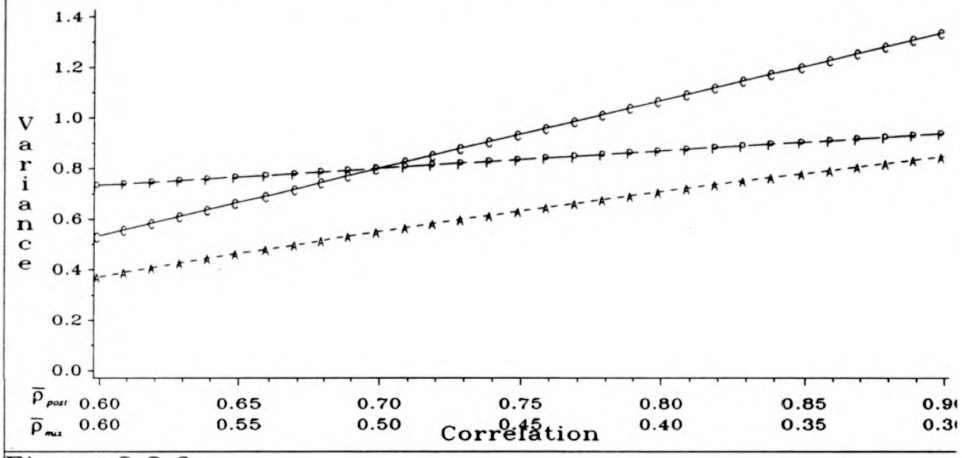
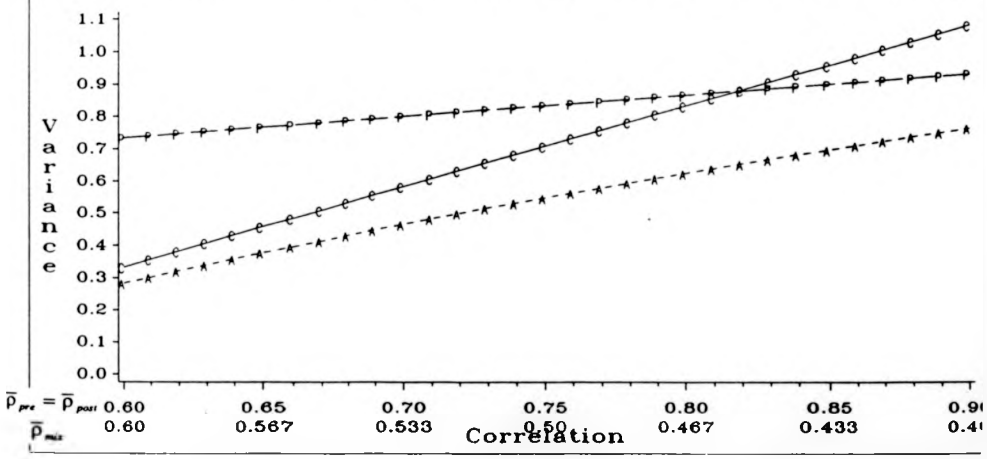


Figure 2.2.6 :

Variances for the three approaches depending on the difference in mean correlations, post−mix. Assuming 2+3 visits, equivariance over time, and an overall mean correlation of 0.6 . P=POST, C=CHANGE, A=ANCOVA



For p>1 pre-treatment visits, and for any specific values of $\bar{\rho}_{post}, \bar{\rho}_{pre}, \bar{\rho}_{mix}$, explicit formulae for the degree of underestimation caused by assuming compound symmetry are easily defined, but it gets more complex to discern any clear pattern. We simply compare the variances for the summary statistics from the formulae given on page 51, with the corresponding formulae under compound symmetry given in subsection 2.2.2.

As an illustration figure 2.2.6 compares variances for the mean summary statistics for a design with 2 pre and 3 post-treatment visits. The assumption is imposed that $\bar{\rho}_{post} = \bar{\rho}_{pre}$, and that $\bar{\rho}_{mix}$ differs from these by such an amount that the overall correlation remains 0.6.

2.3 RESULTS WHEN CORRELATIONS DECAY WITH INCREASING TIME INTERVALS

2.3.1 Modelling correlations for some real examples

Many of the results presented so far have relied on the assumption of compound symmetry for the covariance structure. Even if, as was shown in section 1.5, this often is a quite realistic approximation of the truth, there exists many biological variables where it is known that correlations decay with increasing time intervals between visits. To take account of this when comparing different methods for the analysis of, and design for, repeated measurements studies, we need to find a simple but adequate model for the correlation matrix when compound symmetry is known not to be flexible enough in approximating the true covariance structure.

As a first step we will investigate some of the examples displayed in table 1.5.1. In doing so we will be comparing the ability of five different models for the underlying correlation structure. These are as follows;

Compound symmetry :

$\rho_{ij} = \rho$, for all i≠j.

First-order autoregressive model :

$\rho_{ij} = \gamma^{s}$, where γ is the correlation between adjacent visits (separated by one "unit") and s is the time-interval between visits i and j in such "units".

Damped exponential model :

$\rho_{ij} = \gamma^{s^{\theta}}$, where γ and s are as above, and θ is a parameter controlling the degree of "damping" of the exponential decrease (See Muñoz et al, 1992).

Linearly decreasing correlations with time :

$\rho_{ij} = \gamma - b \cdot s$, with γ and s as above, and b is the estimated linear (least squares) regression coefficient of correlations on time (ignoring non-independence of pairs).

Quadratically decreasing correlations with time :

$\rho_{ij} = \gamma - b_1 \cdot s - b_2 \cdot s^2$, with parameters as above and with addition of a quadratic term.

To give some feeling for the flexibility of the damped exponential correlation structure, figure 2.3.1 is given. For this example it has been assumed that the correlation between adjacent visits is 0.8, and different curves are shown for some of the possible values of the second parameter, θ, in the model, $\rho = \gamma^{d^{\theta}}$.

The curve for θ=0 is the special case of compound symmetry, the one with θ=1 is the special case of a first-order autoregressive curve, those in between have a damped exponential decrease, and the bottom two belong to the class having an accelerated exponential decreasing correlation structure.

Moving on to real examples, we start with the ALAT data referred to in table 1.5.1. In this study 11 visits were performed, 3 before and 8 after randomisation. The maximal time interval between visits was 14 months, with all successive visits being separated by either 1 or 1.5 months. There are 55 estimated correlation coefficients, ranging from .52 to .79, and with an overall mean of .65 . The correlation structure may be seen in figure 2.3.2 along with the five curves resulting from least squares estimation under the five models outlined above. Clearly the auto-regressive model does not fit the data, compound symmetry appears slightly to simplistic, while the remaining three models give very similar results, and they all seem to represent the data quite adequately. However, the linear curve has the advantage of relying on one less parameter.

The estimates for the parameters in the five models, along with the error sums of squares around the estimated curves, are given in table 2.3.1 below.

Figure 2.3.1 : Examples of exponentially decreasing correlation structures

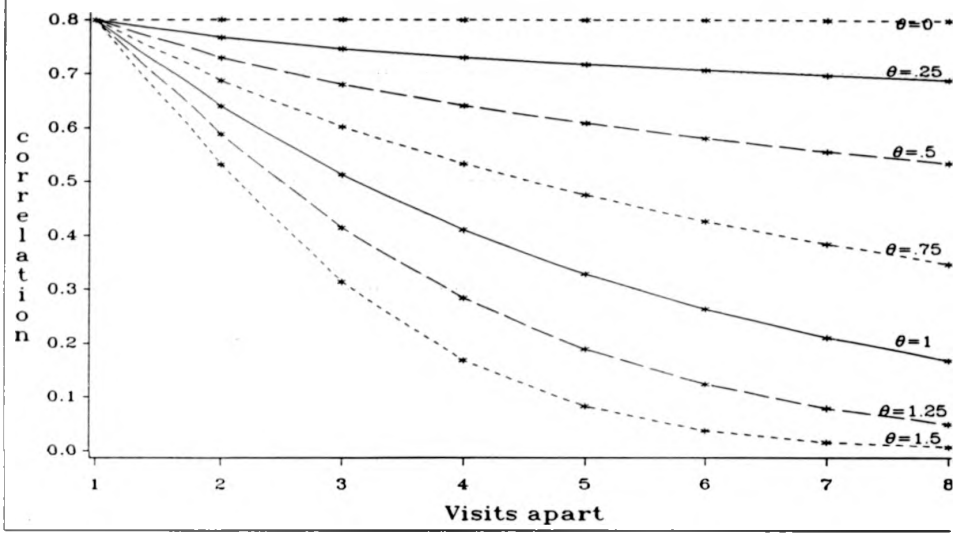


Table 2.3.1 : Estimated correlation structures for the five models with sums of squared deviations for the observed correlation coefficients around the estimated curves (SS_{error}). ALAT data.

Correlation structure	Estimated model for ρ	SS_{error}
Autoregressive	$.936^{t}$	.982
Compound symmetry	.654	.222
Linear decrease	.718-.011t	.138
Damped exponential	$.73^{t^{.19}}$	.136
Quadratic decrease	$.735-.018t+.00054t^{2}$	.135

Figure 2.3.2 : Modelling of the correlation structure, the ALAT–example

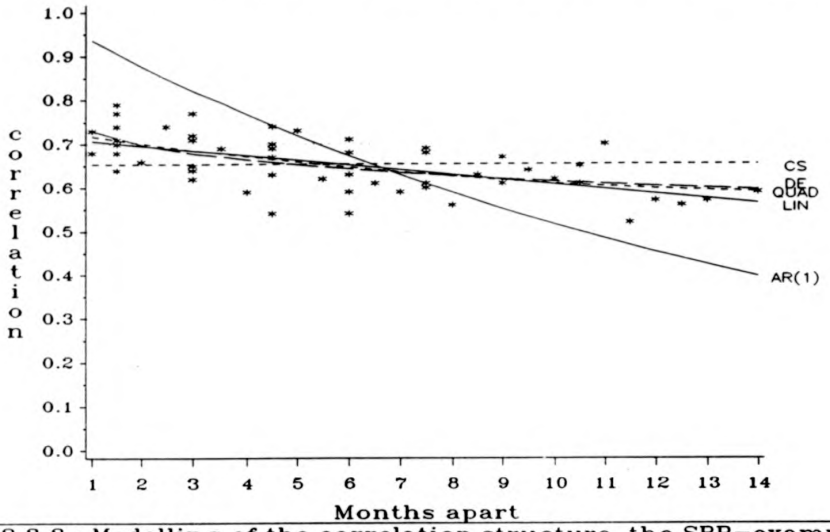
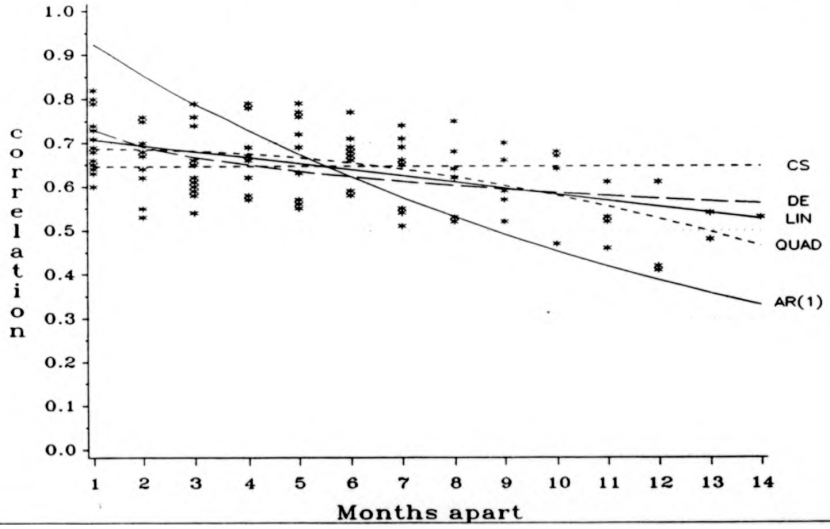


Figure 2.3.3 : Modelling of the correlation structure, the SBP–example



CS=Compound symmetry, DE=Damped exponential, LIN=Linear decrease, QUAD=Quadratic decrease, AR(1)=Autoregressive.

The next example concerns CPK in the same study, the figure (not shown) is almost identical to what was seen for ALAT. The descriptive statistics for the correlation coefficients are consequently very similar to the earlier example with a range from .51 to .78 and a mean of .65 . From table 2.3.2 we can draw about the same conclusions as we did above relating to the appropriateness of the respective models.

Table 2.3.2 : Estimated correlation structures for the five models with error sums of squares. CPK data.

Correlation structure	Estimated model for ρ	SS_{error}
Autoregressive	$.936^{t}$	1.154
Compound symmetry	.651	.272
Linear decrease	.705-.0095t	.210
Damped exponential	$.72^{t''}$	.217
Quadratic decrease	$.709-.011t+.00013t^{2}$	.210

The final example is from the smaller (n=55) of the two hypertension trials included in table 1.5.1, and the outcome measure chosen is systolic blood pressure (SBP). This design encompassed 15 visits, 3 of which were performed before randomisation, and all successive time intervals between visits were 1 month. For this example, there is as expected, due to the smaller N, more variability among the correlation coefficients, with a range from .41 to .82, once again the mean is equal to .65 . The correlation structure for this example, shown in figure 2.3.3, is somewhat different from the two earlier.

There is a decrease in correlation with increasing time intervals, but the observed curvature goes in the opposite direction, as evidenced by the negative estimate for the quadratic term in the quadratic regression in table 2.3.3 below. This slight negative curvature may of course be due to chance, but it is only the quadratic model (of the five models under consideration) that is able to capture such correlation structures.

Table 2.3.3 : Estimated correlation structures for the five models with error sums of squares. SBP data.

Correlation structure	Estimated model for ρ	SS_{error}
Autoregressive	$.924^{t}$	1.928
Compound symmetry	.646	.785
Linear decrease	.722-.014t	.618
Damped exponential	$.73^{t^{.25}}$	.602
Quadratic decrease	$.686+.0024t-.0013t^{2}$	.593

Thus, from these three real examples we conclude that a first-order autoregressive model is best forgotten since it generates too steep a trend. If a simple one-parameter model is desired compound symmetry is not grossly unreasonable. When compound symmetry is too restrictive, a model based on a linear regression of correlation coefficients on the time intervals between visits appears appropriate and there seems little gain in incorporating a quadratic term or in using the damped exponential correlation structure.

Henceforth this section will be concerned with models for the correlation structure based on a linear decrease with time.

2.3.2 Linearly decreasing correlations

Under compound symmetry, comparing variances for the mean summary statistics under different designs is very convenient since the degree of correlation is assumed not to depend on time between measurements. As soon as one moves away from this simple structure one has to consider the impact of time intervals on the variances for the different mean summary statistics.

What also matters here is the shape of the alternative hypothesis, $\mu_{A_t} - \mu_{B_t} = \delta_t$ over time. We will assume δ_t to be constant over time, and return to the issue of a non-constant δ_t in section 5.5.

For now, emphasis is on the influence of linearly decreasing correlations over time on the variances of our treatment comparisons. When investigating the advantage incurred by adding a visit to a certain design, one can either keep the time intervals between successive visits intact and prolong the total study duration, or keep the study duration intact and "squeeze" in the visits by shortening the time intervals between the visits. This subsection will be divided into two parts looking at these two alternative strategies. Furthermore, some results will be given reflecting the influence on the variances for the mean summary statistics caused by simply changing the time intervals between the visits while keeping the number of visits constant.

2.3.2.1 Comparison of variances when time intervals between visits remain fixed

In this sub-subsection we are assuming that all time intervals between successive visits are equal, and that the addition of visits implies an increased study duration. Specifically, we will assume that the correlation coefficient between two given time-points drop by a constant amount c for each further visit there are between them.

For a general design with p pre and r post-randomisation visits, a correlation structure of this type may be depicted as follows:

$$\Sigma = \begin{bmatrix} 1 & & & & \\ \gamma & 1 & & & \\ \gamma - c & & 1 & & \\ \gamma - 2c & & & 1 & \\ \vdots & & & & \vdots \\ \gamma - (p+r-2)c & \cdots & \gamma - c & \gamma & 1 \end{bmatrix},$$

where γ is the correlation between adjacent visits. Using some straightforward algebra, the following equalities may be shown to hold:

$$\overline{\Sigma}_{pre} = \frac{1}{p}\left[1+(p-1)\left(\gamma - \frac{(p-2)}{3}\cdot c\right)\right]$$

$$\overline{\Sigma}_{post} = \frac{1}{r}\left[1+(r-1)\left(\gamma - \frac{(r-2)}{3}\cdot c\right)\right]$$

$$\overline{\Sigma}_{mix} = \gamma - \frac{(p+r-2)}{2}\cdot c$$

It is the changes in the means of the variances and covariances in each of the three submatrices, Σ_{pre}, Σ_{mix} and Σ_{post}, that jointly will decide how the variances for our mean summary statistics (see section 2.1 for the general variance formulae) are affected when we increase the number of visits pre and/or post-randomisation.

In particular, the decreases for the means of the entities in the three distinct submatrices for an increasing number of visits are given by:

$$\overline{\Sigma}_{pre}^{(p)} - \overline{\Sigma}_{pre}^{(p+1)} = \frac{1-\gamma+\frac{c}{3}\left(p^2+p-2\right)}{p(p+1)}$$

$$\overline{\Sigma}_{post}^{(r)} - \overline{\Sigma}_{post}^{(r+1)} = \frac{1-\gamma+\frac{c}{3}\left(r^2+r-2\right)}{r(r+1)}$$

$$\overline{\Sigma}_{mix}^{(p+r)} - \overline{\Sigma}_{mix}^{(p+r+1)} = \frac{c}{2}$$

Assuming, for simplicity, that we have p=1 visit pre-treatment, the variance formulas for our mean summary statistics are as follows:

$$Var[POST] = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\cdot\frac{1}{r}\left[1+(r-1)\left(\gamma-\frac{(r-2)}{3}\cdot c\right)\right]$$

$$Var[CHANGE] = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\cdot\left\{\frac{1}{r}\left[1+(r-1)\left(\gamma-\frac{(r-2)}{3}\cdot c\right)\right]+1-2\gamma+(r-1)\cdot c\right\}$$

$$Var[ANCOVA] = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\cdot\left\{\frac{1}{r}\left[1+(r-1)\left(\gamma-\frac{(r-2)}{3}\cdot c\right)\right]-\left(\gamma-\frac{(r-1)}{2}\cdot c\right)^2\right\}$$

Moving on to the resulting change in variance for the preferred mean summary statistic, ANCOVA, incurred by adding an extra post-treatment visit when there is p=1 pre-treatment, we arrive at the following change in variance:

$$\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\cdot\left[\frac{1-\gamma+\frac{c}{3}\left(r^2+r-2\right)}{r(r+1)}-\gamma\cdot c+\frac{c^2}{4}(2r-1)\right]$$

This formula may be compared with the corresponding formula derived under a compound symmetry model (which is identical for all the three approaches).

Change in variance under compound symmetry: $\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\cdot\frac{1-\rho}{r(r+1)}$.

We are now in a position where we can draw some inferences on the value of increasing the number of visits under a model of linearly decreasing correlations with increasing time intervals.

For a fixed number of measurements pre-treatment, increasing the number of measurements post-randomisation will not always decrease the ANCOVA variance. Even if the precision increases for the dependent variable (the post-treatment mean), this will under certain circumstances be offset by the decrease in $\overline{\Sigma}_{mix}$.

How γ and c interrelates to determine when an upturn in variance (with increasing r) occurs may be judged analytically by the sign of the change in ANCOVA variance for increasing r given in the formula above. Table 2.3.4 gives the smallest c for which the variance starts increasing as a function of γ and r.

Table 2.3.4 : The smallest size of the decrease in correlation per further visits apart (c), for which the ANCOVA variance starts to increase when further post-treatment visits are added, as a function of the original number of post visits (r), and the starting correlation (γ). Assuming a linear decrease in correlation with time and one pre-entry visit.

	Number of post-treatment visits				
γ	1	2	3	5	10
.9	.057	.026	.014	.006	.002
.8	.131	.063	.035	.015	.005
.7	.234	.133	.077	.032	.009
.6	.400	-	-	-	.022*
.5	-	-	-	-	-

* For fixed γ and r, the relationship of the ANCOVA variance depending on c is a quadratic function. When c increases from zero the ANCOVA variance also increases, but for each γ and r there is a worst possible c, after which the ANCOVA variance starts decreasing again with successively larger c. When this occurs it is mostly for impossible combinations of the parameters, when γ=.6 and r=10, however, the ANCOVA variance starts decreasing again when c is larger than .037 .

The above table indicates for which combinations of the parameters γ, c and r, that the ANCOVA variance would actually increase when adding a further post visit to the design. To get some feeling for the relative changes in the variance for different values of r, for fixed plausible choices of γ (=.7) and c (=.02), figure 2.3.4 is given.

Figure 2.3.4 :
Variances for ANCOVA, CHANGE and POST, for linearly decreasing correlations Depending on number of post visits for 1 pre, assuming a correlation of 0.7 for adjacent visits and a drop of 0.02 for each visit further apart. (The assumptions imply that each visit added increases the study duration).

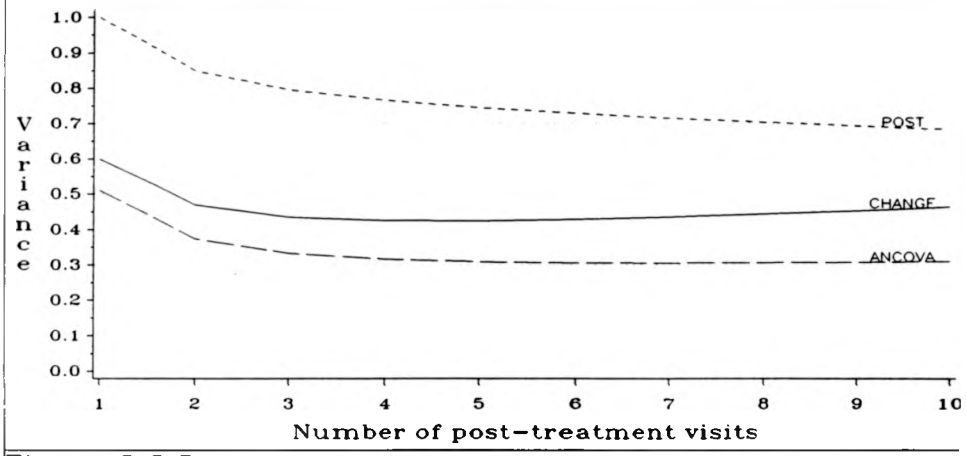
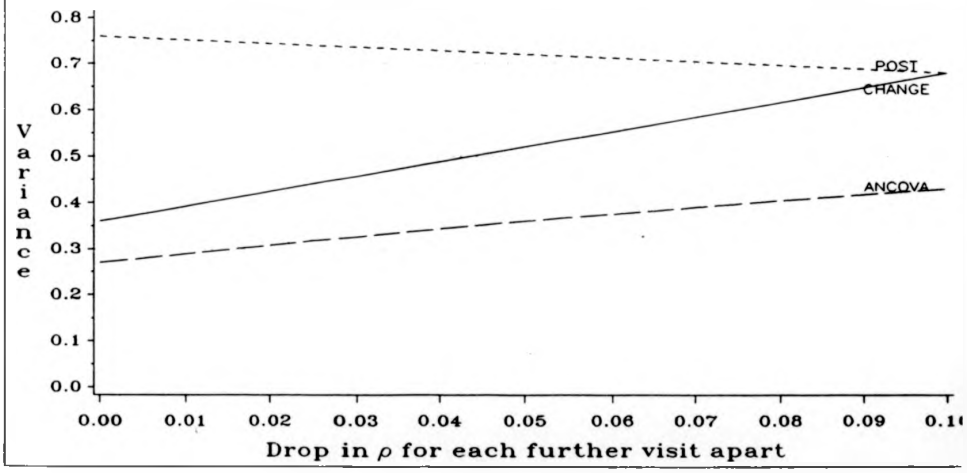


Figure 2.3.5 :
Variances for ANCOVA, CHANGE and POST, for linearly decreasing correlations For a fixed number of visits 1 pre and 5 post, but depending on the degree of decay in correlation assuming a correlation of 0.7 for adjacent visits.



We see that when there is only one pre-entry measurement, and there is a slight decline in correlation with time, provision of more than a handful (in this example 4) post visit will not improve the efficiency of our analysis (when based on ANCOVA, for other, in these circumstances optimal choices of summary statistics, see section 5.5).

When both the number of visits before and after randomisation are fixed, we may illustrate the effect of the degree of linear decrease in correlations over time on the respective variances for POST, CHANGE and ANCOVA. This has been done under some plausible assumptions in figure 2.3.5 (the POST and CHANGE variances are linearly related to c, for ANCOVA there is a slight curvature).

We will now consider the consequences of increasing the number of pre-treatment measurements. As a first step we will give a reworked version of table 2.2.1, but instead of assuming compound symmetry we have assumed linearly decreasing correlations with time with a drop of 0.02 for each further visit separating two time-points. From a comparison of the two tables we can conclude that the advantage of increasing the number of pre-treatment evaluations is much smaller for a model based on linearly decaying correlations with time. Obviously the pre-treatment mean will be estimated with better precision when the number of baselines increase, but this is counteracted by the decrease in $\overline{\Sigma}_{mix}$ with its consequent lower dependence between post-treatment and pre-treatment means.

We also need to consider whether it is plausible to assume the same values for γ and c in all the three submatrices $\overline{\Sigma}_{pre}, \overline{\Sigma}_{mix}$ and $\overline{\Sigma}_{post}$. As observed in table 1.5.1 it is often the case that correlations tend to be slightly lower in $\overline{\Sigma}_{mix}$.

Table 2.3.5 : The dependence of the variance for ANCOVA and CHANGE on the number of pre-treatment measurements p and the correlation ρ for adjacent visits, assuming linearly decreasing correlations with a decay of 0.02 for each further visit apart. We are further assuming r=10 post-treatment visits. For each ρ, variances are divided by the variance for ANCOVA with p=1.

ρ	Analysis	p=1	p=2	p=3	p=4	p=5
0.3	Ancova	1.000	0.937	0.913	0.908	0.913
0.3	Change	3.246	2.058	1.694	1.537	1.461
0.5	Ancova	1.000	0.865	0.816	0.800	0.801
0.5	Change	2.043	1.354	1.151	1.070	1.036
0.7	Ancova	1.000	0.834	0.789	0.782	0.792
0.7	Change	1.491	1.071	0.960	0.926	0.923
0.9	Ancova	1.000	0.915	0.923	0.957	1.001
0.9	Change	1.175	1.030	1.024	1.054	1.098

The value of adding further pre or post-treatment visits to a design, under the model for the correlation structure under investigation, is strongly dependent on the degree of decline for the correlations with increasing time intervals between evaluations. As a further illustration of this, figures 2.3.6 and 2.3.7 are given, where the proportional decrease (increase) in variance for ANCOVA is shown for increasing number of pre (or post)-treatment visits, under some plausible assumptions.

Figure 2.3.6 :
Proportional decrease in variance for ANCOVA when adding further pretreatmen visits, when there are 5 visits posttreatment. Depending on the degree of linear decrease for the correlations over time when assuming a correlation o 0.7 for adjacent visits. All variances are divided by the variance for p=1.

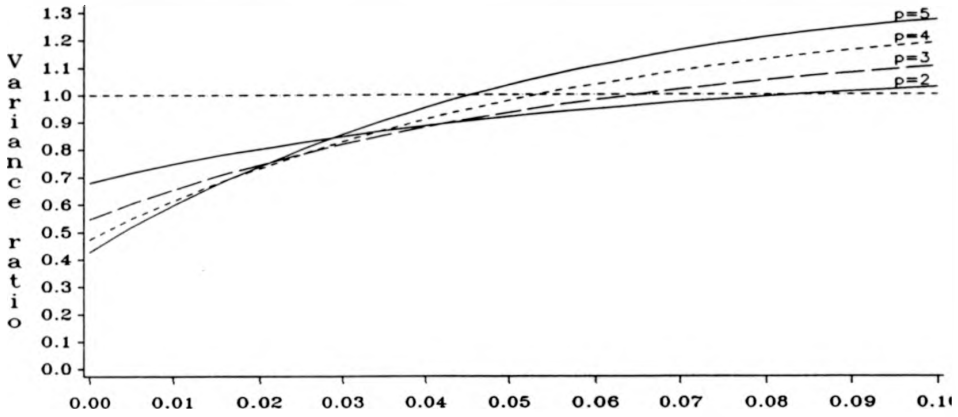
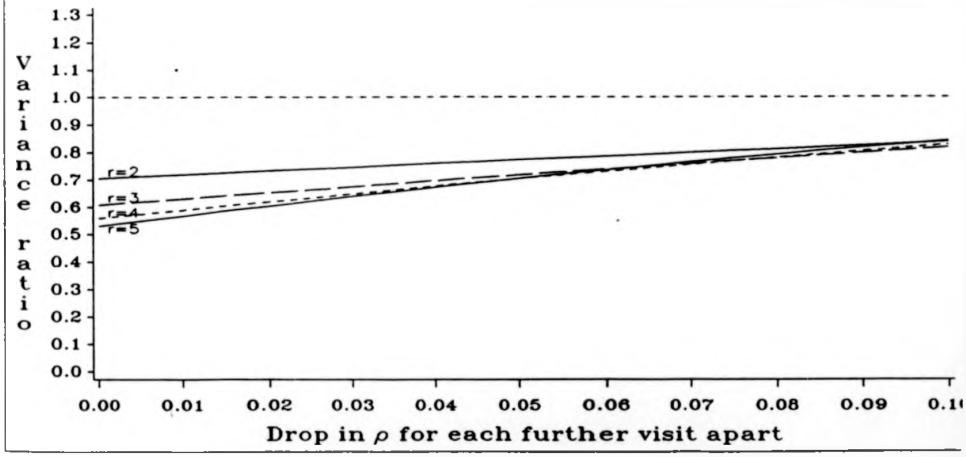


Figure 2.3.7 :
Proportional decrease in variance for ANCOVA when adding further posttreatment visits, when there are 1 visits pretreatment. Depending on the degree of linear decrease for the correlations over time when assuming a correlation of 0.7 for adjacent visits. All variances are divided by the variance for r=1.



2.3.2.2 Comparison of variances when the total study duration is fixed

We will now be concerned with a different strategy, keeping the total study period fixed, and letting the between-visits time interval change as a function of the number of visits incorporated in the design. Under compound symmetry, the variances for our mean summary statistics are unaffected by the length of time intervals, this is not the case when correlations decay with time. As before, the correlation between adjacent visits will be denoted by γ. The total decay in correlation will be denoted by b, this is the anticipated difference between γ and the correlation between the very first and the very last visits.

The correlation matrix for this model has the following structure:

$$\Sigma = \begin{bmatrix} 1 & & & & \\ \gamma & 1 & & & \\ \gamma - \frac{b}{p+r-2} & & 1 & & \\ \gamma - \frac{2b}{p+r-2} & & & 1 & \\ \vdots & & & & \ddots \\ \gamma - \frac{(p+r-2)b}{p+r-2} & \cdots & \gamma - \frac{b}{p+r-2} & \gamma & 1 \end{bmatrix}$$

The means of the variances and covariances in the three submatrices; pre, mix and post, are equal to:

$$\bar{\Sigma}_{pre} = \frac{1}{p}\left[1+(p-1)\left(\gamma - \frac{(p-2)}{3(p+r-2)}\cdot b\right)\right]$$

$$\bar{\Sigma}_{post} = \frac{1}{r}\left[1+(r-1)\left(\gamma - \frac{(r-2)}{3(p+r-2)}\cdot b\right)\right]$$

$$\bar{\Sigma}_{mix} = \gamma - \frac{b}{2}$$

As $\bar{\Sigma}_{mix}$ is independent of the number of visits, many of the relationships between the mean summary statistics will be more straightforward. The changes for $\bar{\Sigma}_{pre}, \bar{\Sigma}_{mix}$ and $\bar{\Sigma}_{post}$ incurred by increasing the number of visits is given by:

$$\bar{\Sigma}_{pre}^{(p)} - \bar{\Sigma}_{pre}^{(p+1)} = \frac{1}{p(p+1)}\cdot\left[1-\gamma - \frac{b\{p(3-p-r-pr)+2(r-1)\}}{3(p+r-2)(p+r-1)}\right]$$

$$\bar{\Sigma}_{post}^{(r)} - \bar{\Sigma}_{post}^{(r+1)} = \frac{1}{r(r+1)}\cdot\left[1-\gamma - \frac{b\{r(3-p-r-pr)+2(p-1)\}}{3(p+r-2)(p+r-1)}\right]$$

$$\bar{\Sigma}_{mix}^{(p+r)} - \bar{\Sigma}_{mix}^{(p+r+1)} = 0$$

For p=1 measurement pre-treatment the reduction for $\bar{\Sigma}_{post}$ simplifies to: $\bar{\Sigma}_{post}^{(r)} - \bar{\Sigma}_{post}^{(r+1)} = \frac{1}{r(r+1)}\cdot\left[1-\gamma+\frac{2}{3}\cdot b\right]$

Continuing with the investigation of results when one baseline measurement is available, we get:

$$Var[ANCOVA] = \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\cdot\left\{\frac{1}{r}\left[1+(r-1)\left(\gamma - \frac{(r-2)}{3(p+r-2)}\cdot b\right)\right] - \left(\gamma - \frac{b}{2}\right)^2\right\}$$

Under the current assumptions, the reduction in variance by having r+1 rather than r post-treatment readings, for p=1 pre-treatment, may be calculated from: $\frac{1-\gamma+\frac{2}{3}\cdot b}{r(r+1)}$. For the current model, with a fixed study duration, the variance for ANCOVA (as well as for POST and CHANGE) will always decrease with increasing number of post-treatment recordings. Further, because $\bar{\Sigma}_{mix}$ always remain unchanged, the reduction will be the same for all the three approaches (this is for p=1 measurement pre-treatment).

2.3.2.3 Comparison of variances with increasing study duration but with a fixed number of visits

Finally, with respect to linearly decreasing correlations, we will consider an alternative way to improve the precision in our estimates of treatment effects, other than increasing the number of repeated measurements.

When correlations decay with time, it is possible to decrease the variance for the pre-treatment and post-treatment means simply by prolonging the study duration. The reason for this is that the measurements get increasingly less dependent the further apart they are, and thus provide more individual information.

However, when extending the study duration one has to consider whether the assumption of a constant difference between mean response curves over time remains realistic. Also, practicalities often dictate the study duration. The results given below relies on $\mu_{Ai} - \mu_{Bi} = \delta_i$ being constant (=δ) over the time intervals under consideration.

We will denote the originally intended study duration by T, and investigate how the variances change when this duration is increased to fT, where f ≥ 1. Starting with the means in the three submatrices of the total within-subject covariance matrix, we find that by moving from a duration of T to fT:

$\bar{\Sigma}_{post}$ decreases by $\frac{(r-1)(r-2)}{3 \cdot r} \cdot c \cdot (f-1)$

$\bar{\Sigma}_{pre}$ decreases by $\frac{(p-1)(p-2)}{3 \cdot p} \cdot c \cdot (f-1)$

$\bar{\Sigma}_{mix}$ decreases by $\frac{(p+r-2)}{2} \cdot c \cdot (f-1)$

The consequent impacts on the variances for our mean summary statistics, when assuming p=1 measurement pre-entry, is:

POST , variance changes by $\frac{(r-1)(r-2)}{3 \cdot r} \cdot c \cdot (f-1)$ as we move from T to fT

CHANGE, variance changes by $-\frac{2}{3} \cdot c \cdot (f-1)\left(r-\frac{1}{r}\right)$ as we move from T to fT

ANCOVA, variance changes by

$(r-1) \cdot c \cdot (f-1)\left[\frac{(r-2)}{3r} - \gamma + \frac{(r-1) \cdot c \cdot (f+1)}{4}\right]$ as we move from T to fT.

It is easy to confirm that, for f≥1, POST will always gain in precision, while CHANGE will always lose precision. For ANCOVA the variance may change in either direction depending on the degree of the correlations. With high correlations (when ANCOVA gets closer to CHANGE) ANCOVA tends to lose precision, with lower correlations (when ANCOVA gets closer to POST) ANCOVA tends to gain precision. For given γ and r, the larger c is, the more likely will it be that the ANCOVA variance decreases with increasing study duration.

2.4 SAMPLE SIZE DETERMINATION

2.4.1 A general covariance structure

As in the conventional approach to power calculation we define α and β as the type I and type II errors for the test of our hypothesis. It is covenient to assume that sample sizes are large enough that the normal approximation to the t-distribution can be applied. In that case, for two equal sized treatment groups of size n, for a general summary statistic ($S_{ij} = \sum_{k=-(p-1)}^{r} c_k y_{ijk} = \mathbf{c}'\mathbf{y}_{ij}$), under a general shape for the alternative hypothesis ($\mu_A - \mu_B = \delta$) and with a general Σ, we require that; $n = \frac{2 \cdot \mathbf{c}'\Sigma\mathbf{c}}{(\mathbf{c}'\delta)^2} \cdot f(\alpha,\beta)$, where $f(\alpha,\beta) = \left[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\right]^2$, Φ being the cumulative distribution of a standardized normal deviate.

Correspondingly, given the sample sizes, the approximate power may under this general scenario be calculated from;

$$1-\beta = \Phi\left(\frac{\sqrt{n/2}\cdot \mathbf{c}'\delta}{\sqrt{\mathbf{c}'\Sigma\mathbf{c}}} + \Phi^{-1}\left(\alpha/2\right)\right)$$

These formulae for the determination of sample sizes and power may be used in conjunction with any linear summary statistic, under any assumed vector of mean treatment differences over time, and under any plausible covariance structure.

This section will only be concerned with constant treatment effects, to make similar comparisons and evaluations under any circumstances (for instance, linearly diverging mean treatment curves) is straightforward. An immediate observation is that conditional on a given treatment effect, the required sample sizes are directly proportional to the variance of the summary statistic.

Among the mean summary statistics ANCOVA has been shown to be consistently more efficient than its competitors. Returning to the notation of section 2.1, we consider the alternative hypothesis $\bar{\mu}_A^{post} - \bar{\mu}_B^{post} = \delta$. For ANCOVA we than require, for a general Σ, that;

$$n = \frac{2\left(\bar{\Sigma}_{post} - \frac{\bar{\Sigma}_{mix}^2}{\bar{\Sigma}_{pre}}\right)}{\delta^2} f(\alpha, \beta).$$

In the following sections some ways to utilize this formula at the design stage, under some plausible models for the covariance structure, will be indicated. Also, the actual gains expected by using ANCOVA instead of the simpler approaches, POST and CHANGE, in reducing sample sizes and/or increasing the power, will be illustrated.

2.4.2 Compound symmetry

Under compound symmetry, for two equal sized groups of size n, we require for ANCOVA that;

$$n = \frac{2\sigma^2}{\delta^2}\left[\frac{1+(r-1)\rho}{r} - \frac{p\rho^2}{1+(p-1)\rho}\right] f(\alpha, \beta).$$

For the other two methods of analysis, POST and CHANGE, we have respectively

$$n = \frac{2\sigma^2}{\delta^2}\left[\frac{1+(r-1)\rho}{r}\right] f(\alpha, \beta), \quad n = \frac{2\sigma^2}{\delta^2}\left[\frac{1+(r-1)\rho}{r} - \frac{(p+1)\rho - 1}{p}\right] f(\alpha, \beta).$$

The corresponding formulae for calculation of power are obtained in a straightforward manner by a direct substitution of the variances of the summary statistics.

For illustration, consider the alternative hypothesis $\delta = 0.4 \cdot \sigma$, and let $\rho = 0.7$, often a realistic value for practical use.

Figure 2.4.1a shows the required sample size n in each group for a variety of study designs and analysis approaches: for r=1,...,8 post-treatment measurements, for p=1 or 3 pre-treatment measurements and for POST, CHANGE and ANCOVA.

The simplest possible design has r=1 and p=0. The POST analysis (a two-sample t-test) requires around n=100 patients per group. Increasing the number of post-treatment readings has some effect on decreasing n, but with no use of pre-treatment readings n remains at around 75 even with r=8.

The CHANGE analysis with p=1 pre-treatment measurements (a two-sample t-test comparing mean changes) leads to a required n around 60 for r=1 post-treatment measurements, which can be reduced to n<40 if r is increased to 4 or more post-treatment measurements. The superiority of ANCOVA is illustrated by a further fall in sample size. For instance, with p=1 and r≥4 we can reduce n to below 30 if ANCOVA is used.

The advantage of increasing the number of pre-treatment measurements in substantial. For instance, with p=3 and r≥4 ANCOVA requires n<20 patients per group. For p=3, CHANGE is similar to ANCOVA.

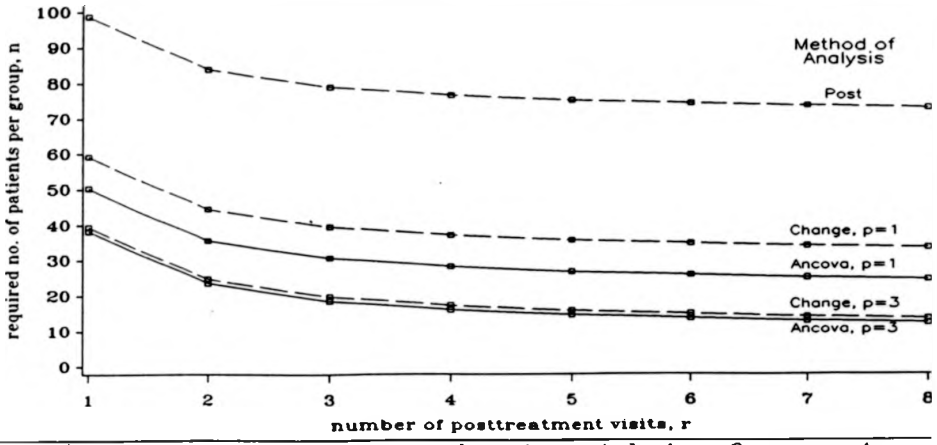
Figure 2.4.1b is given for comparative purposes, and will be explored in the next subsection.

Many simplifying assumptions must by necessity be made at the design stage for a study when making sample size determinations. The size of the treatment effect one wishes to detect (δ), often quantified in terms of a proportion of the standard deviation of a single measurement, will strongly affect the sample sizes called for.

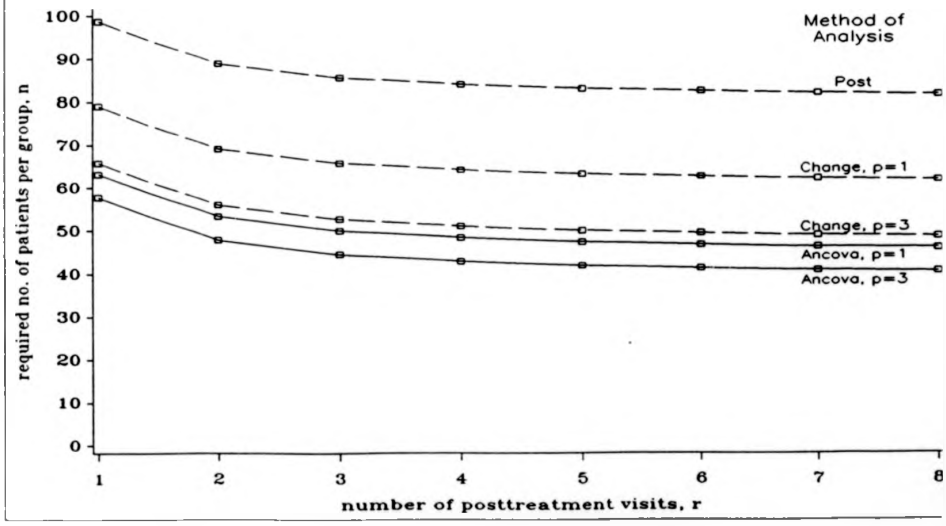
The number of patients needed is inversely proportional to the square of δ. This relationship is exemplified in figure 2.4.2 where a study with 1 pre and 4 post-treatment measurements is under planning. Common choices of α (=0.05) and β (=0.20) have been made, and compound symmetry with ρ=0.6 is anticipated.

Figure 2.4.1
Example of Power calculations for a repeated measures design. Alternative hypothesis $\delta=0.4\sigma$, $\alpha=.05$, $\beta=.2$

a) assuming $\rho=.7$ (pre–pre, pre–post and post–post)



b) assuming $\rho=.8$ pre–pre and post–post, but $\rho=.6$ pre–post



Desiring to detect a treatment effect of $0.5\cdot\sigma$ a total of 90 patients is called for. Would a δ as large as $0.6\cdot\sigma$ be realistic 64 patients would be enough, while settling for $\delta=0.4\cdot\sigma$ would increase the necessary sample size to 138. By this little example we see how quite small alterations in the desired δ have rather large implications on the required sample size. From the figure we may also note how much less efficient POST and CHANGE would be, and also how much smaller treatment effects ANCOVA would be able to detect for any given sample size.

The impact of an increase in sample size on power is far from linear. An illustration of how that relationship might look is given in figure 2.4.3. We are again looking at an intended design with 1 pre and 4 post-treatment visits, it is desired to detect a δ of $0.5\cdot\sigma$ and compound symmetry with $\rho=0.7$ is assumed. Settling on 40 patients in total, ANCOVA has a power of .82, CHANGE has .73 and POST only .43 . Doubling the sample sizes, ANCOVA reaches a power of .98, CHANGE has .95 and POST .72. Not even with 100 patients in total will POST reach the power ANCOVA obtains with 40 patients. Looking at power curves of this type is important, both to reach an acceptable power, but also to avoid over-sized clinical trials.

The two preceding examples have both investigated one specific type of design. Given that we know that we should use ANCOVA, and assuming that, for practical reasons, the sample size is limited to 60 subjects, what can we do to reach an acceptable power, given also that $\delta=0.5\cdot\sigma$ and compound symmetry with $\rho=0.6$ is anticipated? With only 1 pre and 1 post evaluation the power is .66 (see figure 2.4.4), which is felt to low. Adding a second post visit raises the power to .82 . If it is required to reach a power of .90 we are still not satisfied. Adding, also, a second pre visit increases our power to .89, which is further improved to .94 with a total of 2 pre and 3 post-treatment measurements, which, in this case, would be our selected design.

Returning to the issue of the assumptions we have to make at the design stage, one of the advantages with ANCOVA is its robustness. In relation to the degree of correlation (under compound symmetry) this is illustrated in figure 2.4.5.

Figure 2.4.2 :
Number of patients needed depending on the ratio (std/delta)
Assumptions : α=0.05, 1−β=0.80, ρ=0.6, visits=1+4

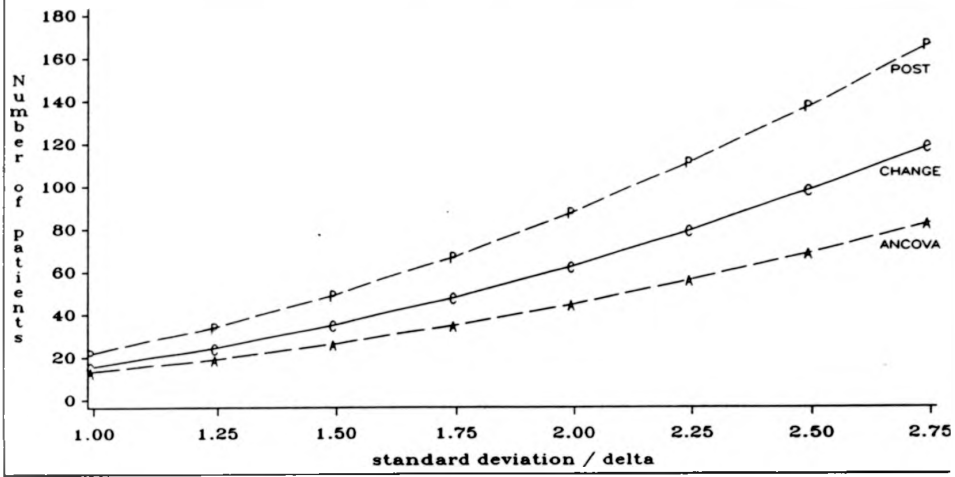


Figure 2.4.3 :
Power achieved depending on number of patients per group
Assumptions : (std/δ)=2, ρ=0.7, α=0.05, 1+4 visits

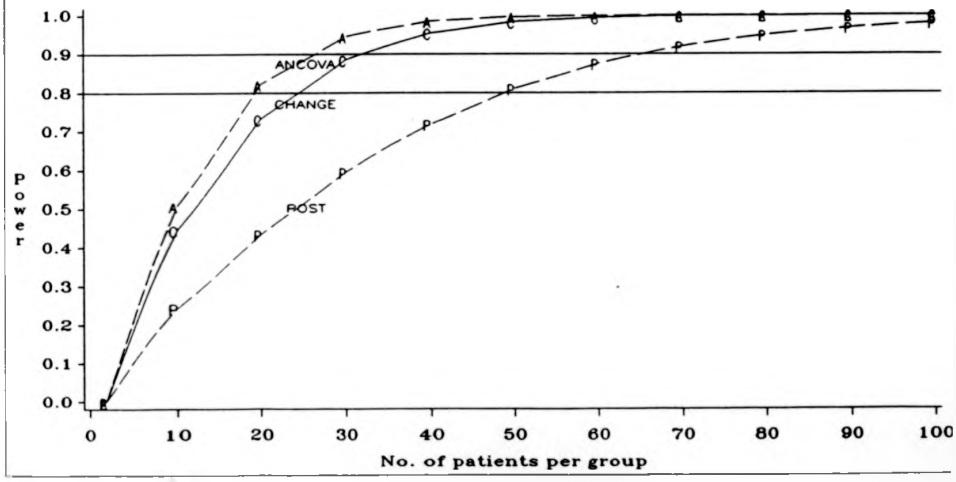


Figure 2.4.4 :
Power achieved depending on number of visits pre+post
Assumptions : (std/delta)=2, ρ=0.6, α=0.05, patients=30+30

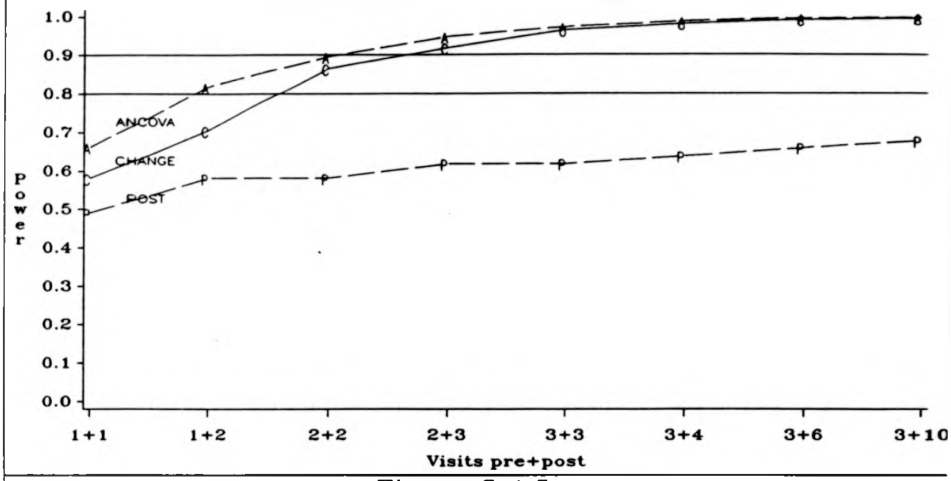
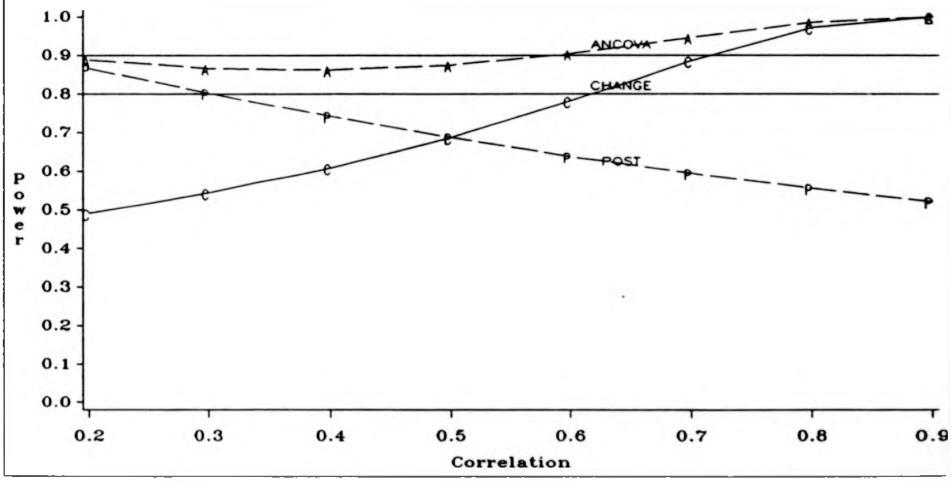


Figure 2.4.5 :
Power achieved depending on the correlation
Assumptions : (std/delta)=2, α=0.05 patients=30+30, 1+4 visits



We are making the same assumptions as in the preceding example, except that a design of 1 pre and 4 post-treatment measurements is desired, and that we are very uncertain on the level of ρ (perhaps due to difficulties in assessing the degree of measurement error to be anticipated). However, as long as the other assumptions are met, the degree of equicorrelation present matters not very much for this example, the power will never drop below .86 . This particular feature of robustness is evidently not shared by POST or CHANGE.

2.4.3 Sensitivity of the compound symmetry assumption

Utilizing the notation and results of subsection 2.2.3 we may now investigate the impact that unequal correlations have on sample size determinations under a compound symmetry assumption.

As indicated in subsection 2.2.3, determination of trial size and its dependence on r, p and the method of analysis can all be documented if one knows the values of the three parameters $\bar{\rho}_{post}$, $\bar{\rho}_{mix}$ and $\bar{\rho}_{pre}$. Given the theoretically infinite variety of correlation structures that could exist, one cannot reach completely generalizable quantitative conclusions on these design issues. However, we will attempt to elucidate some practical suggestions based on certain realistic departures from compound symmetry.

First, consider p=1 pre-treatment reading and the consequence of having $\bar{\rho}_{mix}$ different from $\bar{\rho}_{post}$ ($\bar{\rho}_{pre}$ is non-existent if p=1). Suppose non-equality of correlations can be represented by a decline in ρ of magnitude b per visit apart, all visits being equally spaced. Judged by the examples displayed in figure 1.5.1 this simple structure is likely to be an adequate approximation of the true covariance structure in many situations. If data exist from a previous trial, this slope b can be estimated from the full correlation matrix. Then it can be shown that $\bar{\rho}_{post} - \bar{\rho}_{mix} = b(r+1)/6$.

If power calculations take account of $\bar{\rho}_{mix}$ being less than $\bar{\rho}_{post}$, as in figure 2.4.1b where it has been assumed that $\bar{\rho}_{mix} = 0.6$ while $\bar{\rho}_{pre}$ (when it exists) and $\bar{\rho}_{post} = 0.8$, the relationships between the summary statistics change. Again, the five curves are parallel, but the sample size reductions for CHANGE and ANCOVA compared with POST become less marked. The difference between the elevation of the ANCOVA and POST curves is now 35.5 compared to 48.4 in figure 2.4.1a. The advantage of CHANGE over POST has decreased even more, from 39.5 to 19.8.

Let us next consider the decline in sample size with increasing r and how this could be affected by unequal correlations. Initially we assume p=1. Suppose correlations get weaker the further apart visits are, as often is the case. For a fixed total follow-up time T it can be shown that for r equally spaced visits the mean of all pairwise distances is (r+1)T/3r. This declines with r (by a maximum of one-third for r=∞ compared with r=2) so that $\bar{\rho}_{post}$ increases with r.

Concentrating on ANCOVA, we note that increasing r is liable to increase slightly $\bar{\rho}_{post}$, while under the current model for the decline in correlations, $\bar{\rho}_{mix}$ will remain unchanged (see subsection 2.3.2.2 for more details) so that the resulting effects on the trends in sample size with r will tend to level off slightly more quickly than under a compound symmetry assumption.

When considering the merit of p>1 baseline readings, the extent to which $\bar{\rho}_{pre}$ and $\bar{\rho}_{mix}$ differ from $\bar{\rho}_{post}$ has some bearing on the power calculations. If the repeat baselines are close together $\bar{\rho}_{pre}$ might be increased, whereas having baselines further back in time might reduce $\bar{\rho}_{mix}$, either of these possibilities leading to an increase in the required sample size for ANCOVA. For instance, for p=3 baselines in figure 4.2.1b ANCOVA has the required n decreased by 5.5 (for any value of r) relative to ANCOVA with p=1 baseline. The corresponding drop in required sample size in figure 2.4.1a was 12.1 (for any value of r).

Overall, substantial improvement in statistical efficiency with repeat baselines are possible provided $\bar{\rho}_{mix}$ is not radically reduced and $\bar{\rho}_{pre}$ is not too large. The magnitude of benefit is dependent on $\bar{\rho}_{mix}$ and $\bar{\rho}_{pre}$, but like other parameters in power calculation their values may not be known in advance. Thus, while the recommendation to have more than one baseline if possible is of general relevance to repeated measures trials, the precise extent of statistical improvement cannot be reliably predetermined unless one has some prior knowledge (for example from a previous trial) regarding the correlation structure.

2.4.4 Linearly decreasing correlations

When the compound symmetry assumption is felt to restrictive, and if it is known that correlations will decline with time, the model with linearly decreasing correlations put forward in section 2.3 is a simple but often realistic and robust alternative.

As already noted, the required sample sizes are directly proportional to the variances of the respective summary statistics, and there is no need to repeat the formulae here. Instead, emphasis will be on illustrating the dependence of the power for ANCOVA, and also of the number of patients needed, on the covariance structure as decided by γ and b (using the notation from section 2.3).

Figure 2.4.6 gives contours for four levels of power (A=.95, B=.90, C=.80 and D=.70) as a function of γ and b, assuming: $(std/\delta)=4$, $\alpha=0.05$, 200 patients in total, and 1+4 visits pre respectively post-treatment. For each given point on any of the contours we may read off what b at most can be to give the specified power for a certain γ, or correspondingly, what γ at least has to be to achieve a certain power for a given b. The contour for b=.3 needs some extra clarification. As long as the total decline in correlation (b, the difference between the ρ for adjacent visits, and the ρ between the very first and the very last visit) over the study period is less than .18, the power will exceed .70 for all γ.

Figure 2.4.6 :
γ necessary between adjacent visits to achieve a certain power for ANCOVA depending on total decline, b, in correlation
Assumptions : (δ/std)=.25, α=0.05, 100+100 patients, 1+4 visits

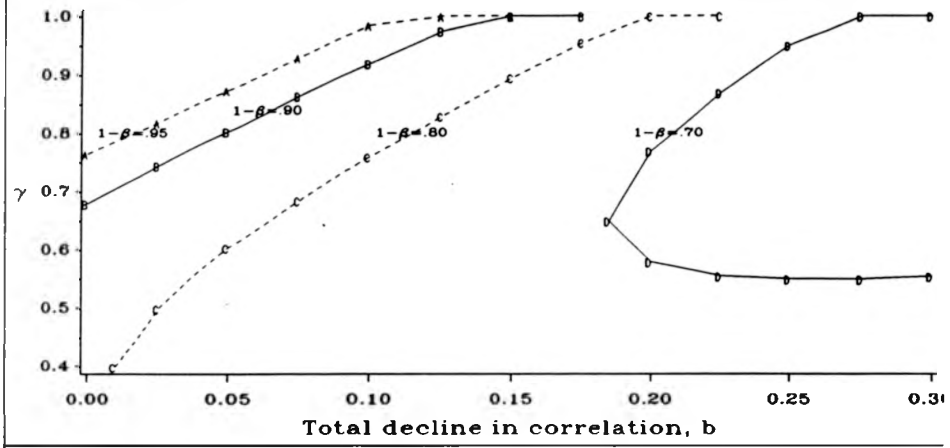
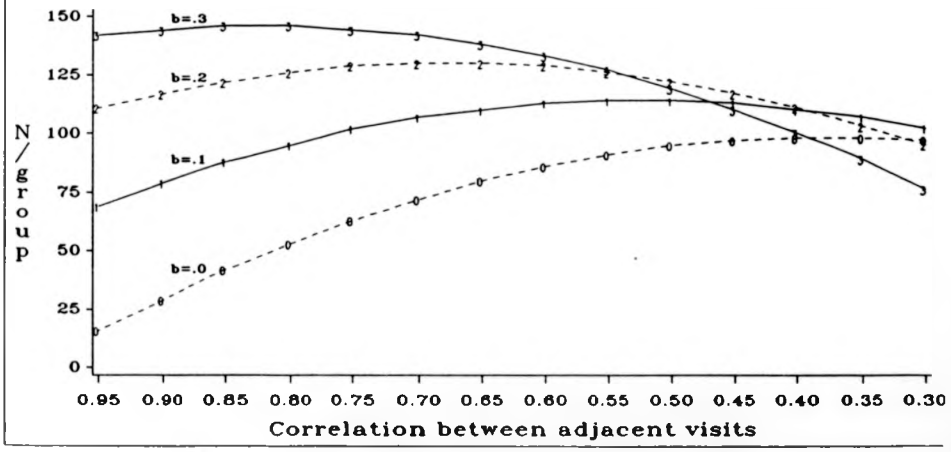


Figure 2.4.7 :
Number of patients needed (per group) when using ANCOVA, depending on correlation between adjacent visits and its linear decline with time
Assumptions : (δ/std)=.25, α=0.05, 1−β=.80, 1+4 visits



With larger b, γ has either to be above the upper arm or below the lower arm, of the contour, to satisfy the specified power requirement. Generally, we see that when the decline in correlation is substantial, the loss in power has to be taken into consideration.

The required sample sizes will not necessarily decrease as γ increase, this is exemplified with figure 2.4.7. This example illustrates the sample sizes needed for a design with 1 pre and 4 post-treatment visits, where it is desired to detect a δ of $0.25 \cdot \sigma$. The four curves are labeled with the level of b anticipated (0=.0, 1=.1, etc.), and γ is given on the x-axis. With γ in the most likely range, say .5 to .8, we see that declining correlations affect the required sample sizes substantially. When b is large, say .2, the required number of subjects may actually increase when γ increases. Generally, when correlations increase we lose precision in $\bar{y}^{pre}$ and $\bar{y}^{post}$, this is because the effective sample size is getting smaller, each individual measurement is giving us less new information. However, the dependency between $\bar{y}^{pre}$ and $\bar{y}^{post}$ also increases, and as is obvious from the usual ANCOVA variance, this effect is very important for increasing the overall efficiency. Normally, increasing correlations imply less ANCOVA variance, but when correlations decrease with time, and $\bar{\rho}_{mix}$ is constrained to be smaller than $\bar{\rho}_{post}$, this balance may shift, as was noted in figure 2.4.7.

2.4.5 Use of a specific predefined covariance matrix

When a certain drug has reached phase III, the research personnel at the pharmaceutical company concerned with its development usually have quite good knowledge of the effects of their treatment on the outcome measures of primary interest. Normally, they are also well aware of at which time points and for which time intervals they are interested in detecting treatment effects.

Specifically, from all the trials performed earlier on this drug, and from the literature concerned with the same type of treatments and diseases, fairly accurate knowledge usually exists (or could exist, if investigated properly) relating to the type of treatment effects anticipated over time, and for the type of covariance structure to be expected.

In this subsection an illustrative example will be given on how such prior information might be used at the design stage of a clinical trial to increase the power to detect an effect of a new treatment regimen. To begin with, a purely hypothetical scenario will be outlined. We will anticipate that we are dealing with a disease for which there is a well defined primary outcome measure, which is continuous and fairly normally distributed. The treatment duration typically lasts for four weeks, and provision of more than one baseline visit is not considered feasible. A rather quick response to treatment which remains reasonably stable over time is expected. Hence, use of one of the mean summary statistics seems appropriate.

Assessments of efficacy in this disease are typically performed weekly, and from the joint information available from earlier investigations it is known that the covariance structure for one pre-entry and four post-treatment evaluations, for this kind of design, will be well approximated by (for notation, see page 18) :

$\Sigma = D_\sigma^T \cdot R \cdot D_\sigma$, with $D_\sigma^T = \left[\sqrt{10} \quad \sqrt{7} \quad \sqrt{7} \quad \sqrt{7} \quad \sqrt{7}\right]$, and

$$R = \begin{bmatrix} 1 & & & & \\ .6 & 1 & & & \\ .5 & .8 & 1 & & \\ .4 & .7 & .8 & 1 & \\ .4 & .6 & .7 & .8 & 1 \end{bmatrix}$$

A few remarks regarding this structure are in order here. The decreasing variances after randomisation reflect the dependence of the variability on the overall mean response, which drops quite substantially after the initiation of treatment (implying that we have to be wary about the risk of unequal covariance matrices between groups). The effects of this treatment effect explains also the relatively lower correlations between pre and post-treatment evaluations as compared to the post-post correlations. In addition to this difference there is also a decrease in correlation with increasing time-intervals between evaluations. In summary, a constant treatment effect after randomisation is anticipated, with a covariance structure as specified above.

Assessing the efficiencies of our mean summary statistics the following results emerge after having substituted the appropriate values of the means for the three submatrices of Σ into the respective variance formulae of the three analysis approaches:

Var[POST] $\propto$ 5.60
Var[CHANGE] $\propto$ 7.65
Var[ANCOVA] $\propto$ 4.02

Appreciating that the investigators for practical reasons are unable to include more than 150 patients in total in this study, and calculating the resulting power, when having decided on a type I error of .05 and having been told by the investigators that they want to be able to detect a difference in mean treatment effect of one unit ($\delta=\sigma/\sqrt{7}$) , we end up with the following;

Power for; ANCOVA = 0.86
POST = 0.73
CHANGE = 0.60

The superiority of ANCOVA is obvious. The relative inefficiency of CHANGE, under these circumstances with relatively high correlations, is perhaps unexpected, but it is explained by the low resulting regression coefficient for the post-treatment mean on the pre-treatment reading for this kind of covariance structure, which has $\beta=0.397$.

It may also be noted that, had it been possible to include multiple baselines, great gains in efficiency might have been possible. Provision of additional post-treatment measurements offer a limited advantage.

The recommendation that should have been done at the design stage for this hypothetical example is straightforward, use an analysis of covariance with each subjects mean post-treatment as dependent variable.

2.5 ANALYSIS OF AN EXAMPLE

We now illustrate the value of some of the issues discussed above with a practical example involving real data. A randomized trial of 152 patients with coronary heart disease compared an active drug with a placebo during a 12 month follow-up period. The liver enzyme CPK in serum was measured to study a possible adverse drug effect on the liver. Each patient had three pre-treatment measurements, taken 2 months before, 1 month before and at randomization, and eight post-treatment measurements, taken at every 1.5 months after randomization.

Figure 2.5.1 shows the results as commonly displayed in a medical journal, with means by treatment group for every time point. While there is a consistent pattern of higher post-treatment means on the active drug, the standard errors are substantial. The common but misguided practice of separate significance testing for each post-treatment time point reveals a varied collection of t-statistics, whether we use means, mean changes or ANCOVA. The t-values range from 0.35 (ANCOVA for visit 12 with visit 0 as covariate) to 3.34 (ANCOVA for visit 4.5 with means of visit -2, -1 and 0 as covariate) with around half the time-point-specific significance tests having $P<0.05$ whichever method of analysis is used.

However, this plethora of significance tests is based on the false premise that each time point is of separate interest in its own right. In reality, the primary hypothesis is more global (across all post-treatment measurements, is there a tendency for an elevation in CPK on the active drug?).

In exploring the correlation structure in these data, each pairwise correlation ρ_{kl} has been estimated by $\hat{\rho}_{kl}$, the observed correlations obtained from a weighted average of the two treatment groups' covariance matrices, weights being proportional to sample size. Figure 2.5.2 plots $\hat{\rho}_{kl}$ by the time between visit k and l; pre-pre, pre-post and post-post pairs are denoted by different symbols. There is a general consistency in the correlations, all being in the range 0.5 to 0.8. Also, the three types of pairs show similar magnitude.

Figure 2.5.1 :

Mean level of CPK over time for drug A (n=76) and drug B (n=76)
(standard error of mean shown only for 3 month visit, others are of similar magnitude).

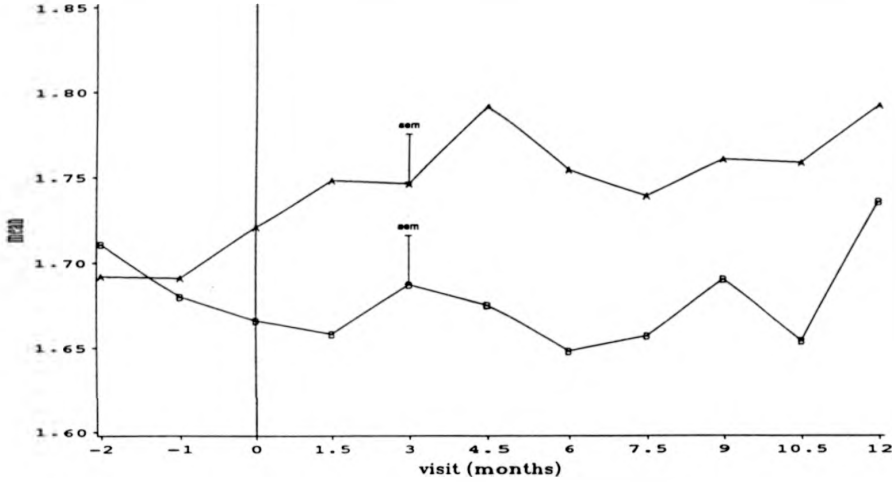
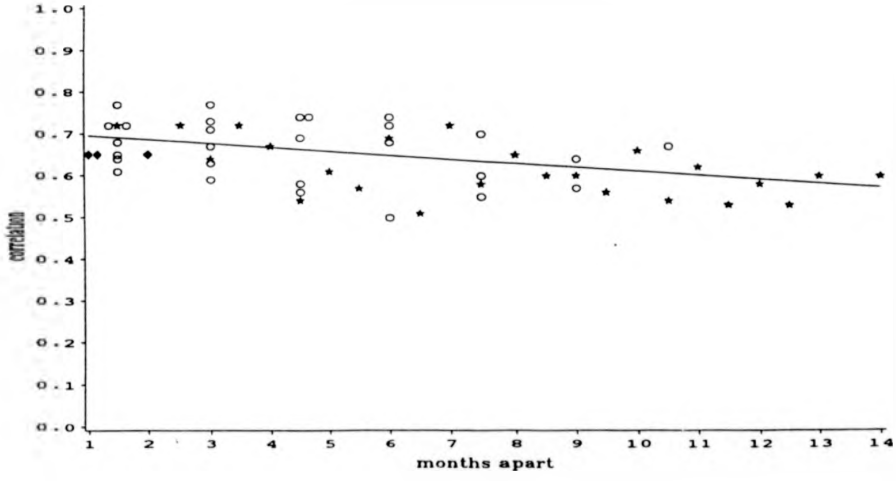


Figure 2.5.2 :

CPK, n=152, correlation coefficents versus time between visits,,
◆=pre,pre, ★=pre,post, ○=post,post.



There is a slight decline in correlation amongst more distant pairs of time points, the estimated slope being -0.009 per month apart. This indicates only slight departure from the assumption that ρ_u is constant for any k≠1. Also, the variance σ_i^2 varied little between visits.

From the discussion in earlier sections, the most appropriate method of analysis for the data in figure 2.5.1 is ANCOVA based on each patient's mean of the eight post-treatment measurements as dependent variable with the mean of the three pre-treatment measurements as covariate. Table 2.5.1 shows CHANGE and POST for comparison, and also includes for illustration ANCOVA and CHANGE as if only a single pre-treatment measurement (visit 0) had been available.

Table 2.5.1 : ANCOVA, CHANGE and POST analyses for the CPK data, n=76 patients in each treatment group, r=8 post-treatment measurements, p=1 or 3 pre-treatment measurements; $\hat{\beta}$ is estimated regression coefficient.

	Number of pre-treatment measurements	Estimated mean diff. in CPK (IU/l)	Standard error	t-statistic	P-value
ANCOVA ($\hat{\beta}$ =0.83)	3	-.066	.021	3.24	.001
ANCOVA ($\hat{\beta}$ =0.63)	1	-.043	.025	1.72	.09
CHANGE	3	-.062	.022	2.89	.004
CHANGE	1	-.023	.030	0.77	.44
POST		-.085	.037	2.31	.02

ANCOVA is seen to produce a smaller standard error and hence stronger evidence of a treatment difference, especially if the mean of all three baseline readings is used as covariate. Since $\hat{\beta}$ is close to 1 in this case the CHANGE analysis is only marginally inferior. POST suffers from two problems: the standard error is much larger, and also failure to take account of the slightly higher average pre-treatment mean level on active drug leaves an upward bias in the estimated treatment effect.

With just a single pre-treatment reading (visit 0) rather than the mean of three, the standard errors for ANCOVA and CHANGE are substantially increased. Given the more pronounced pre-treatment imbalance at visit 0, the CHANGE analysis is prone to a downward bias, this being related to the smaller $\hat{\beta}$ for ANCOVA when p=1.

2.6 SUMMARY AND DISCUSSION

The emphasis of this chapter has been on studies where the main interest is in an overall (mean) response during treatment, the aim has been to explore the statistical properties of some simple approaches to repeated measures using summary statistics. While there are many possible summary statistics, we have focused on the mean post-treatment response of each subject as being a logical choice in many such trials. Consequently, ANCOVA using the mean pre-treatment level as a covariate is the preferred method of analysis. In practice, we suspect ANCOVA is not used nearly enough, so that too many trial reports of quantitative outcome variables, with or without repeated measurements, rely on inferior analyses using just post-treatment values or post-pre differences.

Further, we feel that little attention has been given to the statistical design of clinical trials with repeated measurements. The methods presented for determining sample size and the number of pre- and post-treatment measurements should be of practical use in the planning of such trials. Specifically, the importance of obtaining precise estimates of the subjects pre-treatment levels should be acknowledged, and more than one pre-treatment measure be obtained whenever feasible. For the statistical efficiency of the treatment comparison, this is almost as important as obtaining precise estimates of the post-treatment levels. We feel that the examples presented support the use of the compound symmetry assumption as a realistic guide to the quantitative planning of clinical trials with repeated measurements. However, it should again be emphasized that no such assumption is made when it comes to data analysis.

As an alternative to compound symmetry, the assumption of a slight linear decrease in correlations with time could be made, using methods given in this chapter. This latter approach should provide a safeguard against the sometimes slightly optimistic results (in terms of statistical power) which are suggested under a compound symmetry assumption.

In conclusion, this chapter has presented methods and results for the choice of approach to analysis, and for appropriate statistical designs, which when used sensibly in conjunction with repeated measures clinical trials may greatly improve the efficiency of the statistical analysis.

3 MEAN SUMMARY STATISTICS: SOME ADDITIONAL TOPICS

3.1 BIAS IN ESTIMATION IF PRE-TREAMENT MEANS DIFFER

As described earlier we will adopt the simple model: $x_{ijk} = \mu_{ik} + e_{ijk}$ where i is the index for treatment (A or B), j indexes subject and k visit (ranging from -(p-1),..,0,1,..,r). For a randomised clinical trial, $\mu_A^{pre} = \mu_B^{pre}$, so the expected value of $\bar{\bar{x}}_{A..}^{pre} - \bar{\bar{x}}_{B..}^{pre}$ is zero. Accordingly, at the design stage (before $\bar{\bar{x}}_{A..}^{pre}$ and $\bar{\bar{x}}_{B..}^{pre}$ are observed), all three methods of analysis produce (on average) unbiased estimates of $\bar{\mu}_{A.}^{post} - \bar{\mu}_{B.}^{post}$. However, conditional on any particular observed pre-treatment difference in means $\bar{\bar{x}}_{A..}^{pre} - \bar{\bar{x}}_{B..}^{pre} = d^{pre} \neq 0$ there exists scope for bias.

One rationale behind ANCOVA is that the covariance adjustment removes that component of the observed difference in post-treatment means that is predicted on purely statistical grounds from the observed difference in pre-treatment means. For non-randomised designs (i.e. observational or non-equivalent groups studies) this removal of bias due to inequality of pre-treatment means is only true if there is no measurement error in the pre-treatment readings (see Snedecor and Cochran, 1989). For alternative approaches to analysis in these situations, see also Caroll (1989) and Huitema (1980).

Moving back to the randomised clinical trial, the presence of measurement errors in the pre-treatment recordings will, as observed earlier, result in an attenuation of the slope β in ANCOVA, compared with regression on the true underlying (but unknown) pre-treatment means for each subject.

If we define a variance σ_e^2 for measurement error, and suppose that this is a sub-component of the overall variance σ^2, and that both σ^2 and σ_e^2 are the same for all time-points.

Then, for p=1 pre-treatment reading, the expected value of the observed slope $\beta_{obs} = \left[\frac{\sigma^2 - \sigma_e^2}{\sigma^2}\right] \cdot \beta_{true}$.

For p>1 pre-treatment measurements the attenuation in slope becomes less marked, and specifically $\beta_{obs} = \left[\frac{\overline{\Sigma}_{pre} - \frac{\sigma_e^2}{p}}{\overline{\Sigma}_{pre}}\right] \cdot \beta_{true}$

One might now argue that conditional on a certain difference in pre-treatment means, the use of an attenuated slope, with a certain consequent less degree of adjustment, would imply that also ANCOVA is affected by some bias. This is, however, not the case (see Senn, 1990). The adjustment used by ANCOVA is $\beta \cdot d^{pre}$. We have already seen that the effect of the measurement error on the slope is to decrease β from $\frac{\overline{\Sigma}_{mix}}{\overline{\Sigma}_{pre}}$ to $\left[\frac{\overline{\Sigma}_{pre} - \frac{\sigma_e^2}{p}}{\overline{\Sigma}_{pre}}\right] \cdot \frac{\overline{\Sigma}_{mix}}{\overline{\Sigma}_{pre}}$, or in the case of one pre-treatment reading from ρ to $\left(\frac{\sigma^2 - \sigma_e^2}{\sigma^2}\right) \cdot \rho$. At the same time the expected value of d^{pre} is affected. If we label our observed difference in pre-treatment means allowing for measurement error d_{obs}^{pre}, then given a particular value for this entity, d_{obs}^{pre} is a biased estimator of d_{true}^{pre} (the average over all randomisations, of course, in both cases is equal to zero).

In a randomised clinical trial we have the relationship $E\left[d_{true}^{pre} \middle| d_{obs}^{pre}\right] = \gamma \cdot d_{obs}^{pre}$, where γ is the regression of true values on observed values and satisfies the relationship $\gamma = \frac{\beta_{obs}}{\beta_{true}}$.

Hence, we can write down our covariate adjustment when working on the observed values as; $\beta_{obs} \cdot d_{obs}^{pre} = \beta_{obs} \cdot \frac{\beta_{true}}{\beta_{obs}} \cdot d_{true}^{pre} = \beta_{true} \cdot d_{true}^{pre}$.

I.e. our expected degree of adjustment is the same whether pre-entry measurements are affected by measurement errors or not, and ANCOVA is unbiased. The impact of the measurement errors is only a loss in precision.

This conclusion reinforces the general message that ANCOVA is the best of the three methods considered and the only one which produces unbiased estimators in the presence of chance observed imbalance, irrespective of whether baseline recordings are subject to measurement error.

Of course no technique can hope to adjust for unobserved imbalance, but where we have randomised we are justified in regarding the variances of our estimators as appropriately expressing our uncertainty.

The bias for POST and CHANGE conditional on an observed difference in pre-treatment means, d_{obs}^{pre}, are as follows;

POST, bias is $\beta_{obs} \cdot d_{obs}^{pre} = \frac{\overline{\Sigma}_{mix}}{\overline{\Sigma}_{pre}} \cdot d_{obs}^{pre}$

CHANGE, bias is $-(1-\beta_{obs}) \cdot d_{obs}^{pre} = -\left(1-\frac{\overline{\Sigma}_{mix}}{\overline{\Sigma}_{pre}}\right) \cdot d_{obs}^{pre}$

It is worth noting that the POST bias is in favour of the group being (by chance) better of at baseline, while CHANGE overcorrects for any chance baseline imbalance and has a bias in favour of the group being worse off.

Having more than one pre-treatment measurement will reduce this bias. If we adopt the compound symmetry assumption, then,

for POST, the bias = $\left\{\frac{p\cdot\rho}{[1+(p-1)\rho]}\right\}\cdot d_{obs}^{pre}$,

for CHANGE, the bias = $\left\{\frac{-(1-\rho)}{[1+(p-1)\rho]}\right\}\cdot d_{obs}^{pre}$.

For p=1 this means that the POST bias = $\rho\cdot d_{obs}^{pre}$ and the CHANGE bias = $-(1-\rho)\cdot d_{obs}^{pre}$. For $\rho>0.5$ (which is usually the case), POST contains more bias than CHANGE, Furthermore, for p>1 pre-treatment measurements, this aspect of inferiority for POST becomes more marked. For instance, if p=3 and $\rho=0.7$ (say), then the POST bias = $.875\cdot d_{obs}^{pre}$ while the CHANGE bias = $.125\cdot d_{obs}^{pre}$. However, with more pre-treatment readings we can expect d_{obs}^{pre} to become smaller.

Overall, if there exists a pre-treatment difference, then POST may be seriously biased. CHANGE may also contain a certain degree of bias, especially if the correlations between pre and post measurements are relatively small, but this bias will be reduced considerably if the number of pre-treatment measurements is increased. For some Monte Carlo-simulations on the bias introduced by chance baseline differences, confirming the results given here, especially the unbiasedness of ANCOVA, see Overall and Magee (1992).

Strongly related to the question of biased estimates in the presence of baseline imbalance, is the question of type I error rates for different approaches to analysis conditional on baseline imbalance. Senn (1989) gives analytical results proving that only ANCOVA maintains the proper type I error rate for RCT's, POST and CHANGE may often be far off conditional on a given mean pre-treatment difference.

3.1.1 Effects on variances when pre-treatment means differ

The variances for the estimated treatment effect when using POST or CHANGE are not directly affected by chance observed imbalance between groups. It is different for ANCOVA, as is obvious from its variance formula;

$$\text{Var(ANCOVA)} = \left(\frac{1}{n_A}+\frac{1}{n_B}+\frac{(d^{pre}_{obs})^2}{(n_A+n_B-2)\cdot\overline{\Sigma}_{pre}}\right)\cdot\left(\overline{\Sigma}_{post}-\frac{\overline{\Sigma}^2_{mix}}{\overline{\Sigma}_{pre}}\right)\cdot\left(\frac{n_A+n_B-2}{n_A+n_B-3}\right),$$

which depends on the difference in pre-treatment means. This slight increase in the ANCOVA variance is a price we have to pay for non-orthogonality between treatment groups and pre-entry measurements. Thus, when using ANCOVA for RCT's, baseline balance has nothing to do with validity, only with efficiency.

In the following table it can be seen how this variance increases with the difference in pre-treatment means.

Table 3.1.1 : Proportional increase in variance for ANCOVA caused by chance observed mean pre-treatment differences. (This increase is independent of the correlation and the number of post-treatment measurements). SEM stands for standard error of the mean.

d^{pre}_{obs}/SEM	10+10 pat.	50+50 pat.	250+250 pat.
0	1.000	1.000	1.000
0.5	1.014	1.003	1.001
1	1.056	1.010	1.002
1.5	1.125	1.023	1.005
2	1.222	1.041	1.008

For large trials (say, hundreds of subjects) there is nothing to worry about. For medium-sized trials (say, 50 to 100 subjects per group) we might lose some precision if we are unlucky with the randomisation (typically an increase in variance of a few per cent).

With small sample sizes baseline imbalance might be of a real concern, and should perhaps be accounted for in the power calculations. For instance with only 10+10 patients and a standardized mean pre-treatment difference of 2, the ANCOVA variance would increase with 22 per cent.

One further table is given, showing the relationship between the variances for CHANGE and ANCOVA depending on mean pre-treatment differences, number of subjects, and the degree of correlation. Here we are assuming compound symmetry for the derivation of the results.

Table 3.1.2 : Proportional increase in Var(CHANGE) compared to Var(ANCOVA) depending on standardized baseline imbalance, sample size and correlation. Assuming compound symmetry and 1+4 visits.

		Number of subjects per group :				
d_{obs}^{π}/SEM	ρ	10+10	25+25	50+50	100+100	250+250
0	.4	1.816	1.883	1.903	1.913	1.919
1	.4	1.721	1.845	1.884	1.904	1.915
2	.4	1.486	1.738	1.829	1.875	1.904
0	.6	1.389	1.440	1.456	1.463	1.468
1	.6	1.316	1.411	1.441	1.456	1.465
2	.6	1.136	1.329	1.399	1.434	1.456
0	.8	1.124	1.166	1.178	1.184	1.188
1	.8	1.065	1.142	1.166	1.179	1.186
2	.8	0.920	1.076	1.132	1.161	1.179

We see from this table that the superiority of ANCOVA relative to CHANGE decreases when pre-treatment means differ and sample sizes are small. In extreme cases the CHANGE variance may actually be smaller. However, when this happens, for large standardized baseline differences, CHANGE is likely to give biased results (unless β is close to 1), taking validity into consideration, ANCOVA should always be chosen before CHANGE.

3.2 INCREASING SAMPLE SIZE OR NUMBER OF VISITS

Consider the design of a repeated measures clinical trial, and suppose the calculated power for the intended sample size is too low to be acceptable. We may assume that plausible values have been chosen for the difference in treatment effect (assumed constant after randomisation) and for the covariance structure. What can be done under these circumstances to raise the power to a desired level? We assume further that an efficient approach to analysis has been specified, i.e. ANCOVA. Then two options to improve the situation remains. Either one has to increase the sample size, or one has to increase the number of repeated measurements taken on each subject. (A third alternative, when compound symmetry do not apply, might be to change the timing of the measurements, see subsection 2.2.3).

In comparing the relative merits of these two options we will not directly consider the issue of cost, and the natural extension of evaluating cost-effectiveness. The comparisons will be made solely in terms of precision. However, it would not be difficult to have a costings model, involving costs both per patient and per visit. To keep the exposition simple compound symmetry will be assumed for the covariance structure, though extensions to other structures are relatively straightforward.

We will make these comparisons with emphasis on ANCOVA, and assess the usefulness of increasing either the number of pre-treatment visits or the number of post-treatment visits, relative to increasing the total number of subjects, for a two-group RCT with equal sample sizes. The way we go about doing this is by equating the variance formula for ANCOVA when there are p pre- and r post-treatment visits and n+x subjects per group, to the corresponding variance formula with an additional measurement (pre- or post-treatment) but with n subjects per group.

Assessing the value of a further post-treatment visit for ANCOVA when there are p measurements pre-treatment, we solve the following equation (based on the variance formula for ANCOVA given on page 46);

$$\frac{2\sigma^2}{n+x}\cdot\left(\frac{1+(r-1)\rho}{r}-\frac{p\rho^2}{1+(p-1)\rho}\right)=\frac{2\sigma^2}{n}\cdot\left(\frac{1+r\rho}{r+1}-\frac{p\rho^2}{1+(p-1)\rho}\right)$$

resulting in $x=\dfrac{n(p\rho-\rho+1)}{r(p\rho+r\rho+1)}$.

This is the additional number of subjects needed per group to raise the power by the same amount as the addition of a further visit would do. In the simplified case with p=1 we get;

$$x=\frac{n}{r(1+\rho+r\rho)}.$$

The corresponding general formulae for POST and CHANGE are;

$$POST\ :\ x=\frac{n(1-\rho)}{r(1+r\rho)},\qquad CHANGE\ :\ x=\frac{n\cdot p}{r(1+p+r)}.$$

Similarly, contrasting an increase in the number of pre-treatment readings, for a fixed number post-randomisation, relative to an increase in sample size, we arrive at the following equality

for ANCOVA; $x=\dfrac{n\cdot r\cdot\rho^2}{(1+r\rho+p\rho)(1+(p-1)\rho)}$ (for CHANGE we get,

$x=\dfrac{n\cdot r}{p(p+r+1)}$).

To give some feeling for the relative increases in sample size that are needed to compensate for not providing for a further visit in the study design, a few examples will be summarized in the two tables given below. These examples may usefully be compared with the sample size figure 2.4.1 on page 76.

Table 3.2.1 : Percentage of increase in sample size needed to increase the power by the same amount as provision of an additional post-treatment visit would. Assuming p=1, analysis will be based on ANCOVA, compound symmetry, and a constant treatment effect.

ρ	Number of post-treatment visits (before addition) 1	2	3	5
.4	55.6	22.7	12.8	5.9
.6	45.5	17.9	9.8	4.3
.8	38.5	14.7	7.9	3.4

From the above table we see that when there is only one post-treatment measurement to start with, provision of an additional visit after randomisation is likely to increase the power by the same amount as an increase in sample size in the order 40 to 50 per cent (for ρ in the plausible range around .6). When the originally intended design has more post-treatment measurements, increasing the number of subjects might be a better option (for instance, increasing the sample size by 10% is likely to be more efficient than increasing the number of post-treatment visits from 3 to 4). This is because we already have quite precise estimates of the subjects post-randomisation levels, increasing the number of pre-entry evaluations might be a better option in this case.

Table 3.2.2 : Percentage of increase in sample size needed to increase the power by the same amount as provision of an additional pre-treatment visit would. Assuming r=4, analysis will be based on ANCOVA, compound symmetry, and a constant treatment effect.

ρ	Number of pre-treatment visits (before addition) 1	2	3
.4	21.3	13.4	9.4
.6	36.0	19.6	12.6
.8	51.2	24.5	14.9

From table 3.2.2 we see that provision of 2 pre-entry measurements rather than 1 is likely to be as efficient as increasing the sample size with somewhere between 30 to 50 per cent, when there are 4 measurements after randomisation. With fewer post-treatment measurements the value of adding a second pre-entry measurement would be somewhat less impressive relative to increasing the number of subjects.

These tables should not be taken to suggest that it might be more efficient to add further post measurements rather than pre, because it is usually not. Such comparisons should be based on the variance formulae given in the preceding chapter, rather than on indirect comparisons from different designs here.

General practical conclusions relating to the usefulness of adding further measurements, relative to having more subjects, is difficult to give. This will depend on costs as well as other practical matters, like availability of subjects and time. However provision of two measurements, rather than one, both pre and post-randomisation is likely to decrease the required number of subjects substantially in most applications.

3.3 ADDITIVE OR MULTIPLICATIVE EFFECTS

One of the most frequent assumptions made when searching for an appropriate statistical model is that of additive effects. However, many variables measured in clinical experiments have at least one, often several, of the following characteristics:

1. The treatment effect depends on the initial value for a given subject, that is (substituting "covariate" for the pre-entry measurement) we have a treatment-by-covariate interaction.

2. The standard deviation of the dependent variable increases when the mean level of the variable increases (a covariate-by-residual interaction).

3. The residual variance around the fitted (separately for the groups) regression lines (of dependent variable on covariate) are different, a treatment-by-residual interaction.

4. The responses have a log-normal distribution.

These four characteristics are in many ways related (actually number 4 implies the first three), and for variables with one or more of these properties the log-transformation will often succeed both in reducing the heteroscedasticity, the treatment-by-covariate interaction, and in producing distributions that are more nearly normal.

Several other types of transformations could be considered under these circumstances, for instance, the class of power transformations (Draper and Smith, 1981), but we will restrict ourselves to evaluation of the logarithmic transformation.

An alternative to a transformation when the treatment effect is assumed to be multiplicatively related to the pre-entry value is to analyse either the ratio, Y/X (henceforth labelled RATIO), or the percentage change, 100*(Y-X)/X (henceforth labelled %CHANGE). These two summary statistics are mathematically equivalent, i.e.;

$$\%CHANGE = 100*(Y-X)/X = 100*(Y/X) - 100 = 100*RATIO - 100 .$$

Sometimes %CHANGE may be more clinically meaningful, even if there is no difference in the fit of the models.

RATIO will be considered further, and it will be shown that by using this summary statistic, a specific model, consisting of both additive and multiplicative effects, is implicitly assumed. Under certain special circumstances RATIO will be shown to be the optimal summary statistic.

The rest of this section will evaluate different underlying data-generating models, and show what kind of observed response by covariate relationship they are likely to produce. Hence, some guidance will be given in choosing which model is correct, and thereby in deciding on whether a transformation might be needed. More formal goodness-of-fit comparisons goes outside the scope of this thesis. For comparisons between non-nested models see Royston and Thompson (to appear in Biometrics) and the references therein. Further, based on analytical results, the transformations needed to achieve complete additivity, under some different models, will be given. Finally, some of the proposed methods will be illustrated in an example selected from table 1.5.1, where multiplicative effects appears to be present.

3.3.1 Some simple data-generating models

In comparing models with additive and/or multiplicative effects, for simplicity we consider a simple design with one pre-entry measurement (X), and one post-randomisation measurement (Y). Main interest is in the comparison of the following two models;

1. $Y_{ij} = \alpha_i + \beta \cdot X_{ij} + \varepsilon_{ij}$
2. $\log(Y_{ij}) = \gamma_i + \delta \cdot \log(X_{ij}) + \eta_{ij}$ (i.e. $Y_{ij} = e^{\gamma_i} \cdot X_{ij}^{\delta} \cdot e^{\eta_{ij}}$)

In both models the response is allowed to depend on three effects; treatment, covariate, and residual. The coefficients have of course different interpretations in the two models.

The residuals, ε_{ij} and η_{ij} will in be assumed to follow normal distributions.

As a starting point for our comparisons we will look at the kinds of data structures that are likely to be observed under the different data-generating models.

Under model 1 all effects are additive, we have parallel regression lines, homoscedasticity between groups, and within-group variances which are independent of the covariate level. For each group the X and Y variables jointly have a bivariate normal distribution, the only between group difference is the additive treatment effect.

Under the second model all effects are multiplicatively interrelated. The treatment-by-covariate interaction gives rise to non-parallel regression lines. The covariate-by-residual interaction will make the variances increase with increasing covariate values. The treatment-by-residual interaction causes the residual variance to differ around the two (separately fitted) regression lines. After a log-transform for both X and Y variables everything becomes additive, and the transformed variables will follow a bivariate normal distribution.

Motivated by our interest in when RATIO is a sensible summary statistic to use, a third model, consisting of a mixture of additive and multiplicative effects, will be given some consideration. This is similar to model 2, the difference being that the residual variance now is equal around the two (separately fitted) regression lines (i.e. there is no treatment-by-residual interaction). However, there is still a treatment-by-covariate interaction, and a covariate-by-residual interaction. For non-transformed data the model is; $Y_{ij} = X_{ij}^{\theta} \cdot (\mu_i + \tau_{ij})$. Here μ_i represents the overall treatment mean in group i, θ is the regression coefficient for Y on X, and τ_{ij} is the residual.

3.3.2 Transformations necessary to achieve additivity

We shall now look for transformations of our two variables that will change multiplicative relationships to additive, specifically, we want to have a completely additive statistical model and to correct the dependent variable for differences in the covariate level by an amount that is predicted on purely statistical grounds (e.g. analysis of covariance adjustment).

Model 1 has $Y_{ij} = \alpha_i + \beta \cdot X_{ij} + \varepsilon_{ij}$, by a simple subtraction we accomplish are goal; $Y_{ij} - \beta \cdot X_{ij} = \alpha_i + \varepsilon_{ij}$. If this is the underlying data-generating model the recommended approach is to analyse the summary statistic $Y_{ij} - \beta \cdot X_{ij}$ for each subject, this is, of course, ANCOVA.

Our main contender as the true underlying model is number 2, the completely multiplicative model. Here, $Y_{ij} = e^{\gamma_i} \cdot X_{ij}^{\delta} \cdot e^{\eta_{ij}}$, taking logarithms this changes to, $\log(Y_{ij}) = \gamma_i + \delta \cdot \log(X_{ij}) + \eta_{ij}$. This may be rewritten as, $\log(Y_{ij}) - \delta \cdot \log(X_{ij}) = \gamma_i + \eta_{ij}$, or alternatively as, $\log\left(\frac{Y_{ij}}{X_{ij}^{\delta}}\right) = \gamma_i + \eta_{ij}$. All we need to do if this is the correct model is to log-transform both Y and X and then use ANCOVA, or equivalently analyse the summary statistic $\log\left(\frac{Y_{ij}}{X_{ij}^{\delta}}\right)$.

Model 3 has; $Y_{ij} = X_{ij}^{\theta} \cdot (\mu_i + \tau_{ij})$, and we can readily see that $\frac{Y_{ij}}{X_{ij}^{\theta}} = \mu_i + \tau_{ij}$ achieves the desired aims. We thus analyse the same ratio as for the preceding model , but without a log-transformation. We may also note that when the regression coefficient for a regression of Y on X is one, the optimal summary statistic is Y/X. This fact has also been noted by Cochran (1957, p.263), who observed that analysing the percentage change is optimal when Y/X is independent of X and has constant variance.

The appropriate transformations to use under these three models are summarized in table 3.3.1.

Table 3.3.1: Recommended summary statistics under three different models for making an appropriate covariate adjustment, and for converting multiplicative relationships to additive.

Model	Optimal summary statistic	Special case: β=0	Special case: β=1
$Y_{ij} = \alpha_i + \beta \cdot X_{ij} + \varepsilon_{ij}$	$Y_{ij} - \beta \cdot X_{ij}$	Y_{ij}	$Y_{ij} - X_{ij}$
$Y_{ij} = e^{\gamma_i} \cdot X_{ij}^{\delta} \cdot e^{\eta_{ij}}$	$\log\left(\frac{Y_{ij}}{X_{ij}^{\delta}}\right)$	$\log(Y_{ij})$	$\log\left(\frac{Y_{ij}}{X_{ij}}\right)$
$Y_{ij} = X_{ij}^{\theta} \cdot (\mu_i + \tau_{ij})$	$\frac{Y_{ij}}{X_{ij}^{\theta}}$	Y_{ij}	$\frac{Y_{ij}}{X_{ij}}$

3.3.3 The triglycerides example

From the second coronary heart disease study referred to in table 1.5.1 the outcome measure triglycerides has been chosen to illustrate the methods of this subsection. The third of the four post-treatment measurements, the 6 months visit, was chosen as dependent variable, with the first of the two pre-entry measurements as covariate. The descriptive statistics for the two treatment groups were as follows:

	N	Pre-entry Mean	Pre-entry Std	At 6 months Mean	At 6 months Std
Drug A	109	1.815	(.97)	1.610	(.92)
Drug B	110	1.829	(1.02)	1.843	(1.10)

Observed covariance structures, within-groups, respectively pooled:

$$\hat{\Sigma}_A = \begin{bmatrix} .945 & \\ .551 & .839 \end{bmatrix} \qquad \hat{\Sigma}_B = \begin{bmatrix} 1.039 & \\ .750 & 1.231 \end{bmatrix} \qquad \hat{\Sigma}_{pooled} = \begin{bmatrix} .992 & \\ .651 & 1.036 \end{bmatrix}$$

$$\hat{\rho}_A = .619, \quad \hat{\beta}_A = .583 \qquad \hat{\rho}_B = .663, \quad \hat{\beta}_B = .722 \qquad \hat{\rho}_{pooled} = .642, \quad \hat{\beta}_{pooled} = .656$$

Figure 3.3.1, together with these descriptive statistics, give us clear indications that we may not have additive effects. The regression coefficients from the two groups are quite different, .583 (standard error .072) in group A, versus .722 (standard error .078) in group B. Even though this suggests a treatment-by-covariate interaction, the test for such an interaction (which has low power) is non-significant (F=1.69, p=.19).

The variance for the dependent variable is also seen to increase with increasing baseline values, indicating a covariate-by-residual interaction. Grouping the pre-entry measurements into quartiles (<1.14, 1.14-1.63, 1.64-2.18, >2.18), the variance of the residuals (with residuals calculated from separately fitted linear regression lines) equals, respectively, .18, .55, .67, and 1.03. Thus confirming the visual impression.

Whether there also appears to be a treatment-by-residual interaction, i.e. different overall variances around the two separately fitted regression lines, is more difficult to judge by the eye. This residual variance equals .52 for group A, and .69 for group B, with the test for equality of variances giving F(109,108)=1.33, and p=.14 . Certainly, however, a log-transformation seems well motivated.

A second set of descriptive statistics, now on a <u>log-scale</u> are given below:

	N	Pre-entry Mean	Pre-entry Std	At 6 months Mean	At 6 months Std
Drug A	109	.472	(.498)	.343	(.506)
Drug B	110	.488	(.472)	.484	(.492)

Figure 3.3.1 : Triglycerides, pre-entry vs 6 months, for drug A (n=109, dotted line, stars) and drug B (n=110, dashed line, diamonds). Separate linear regression lines fitted for each group

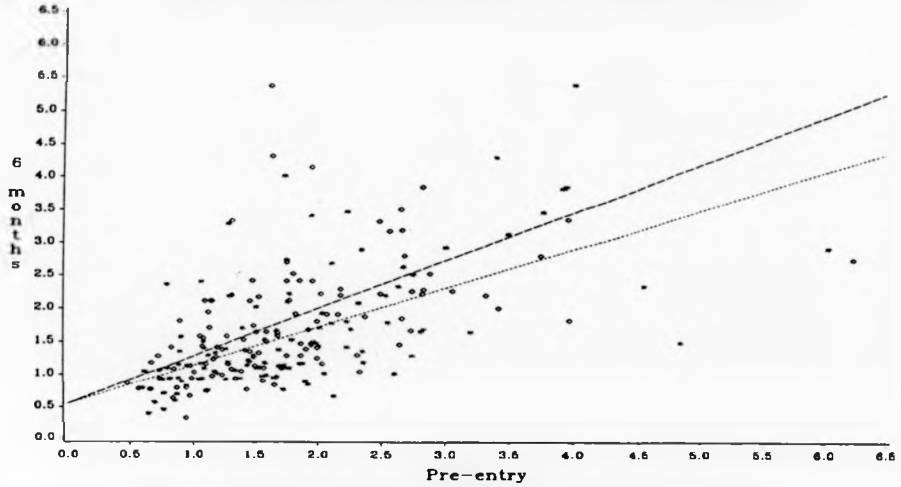
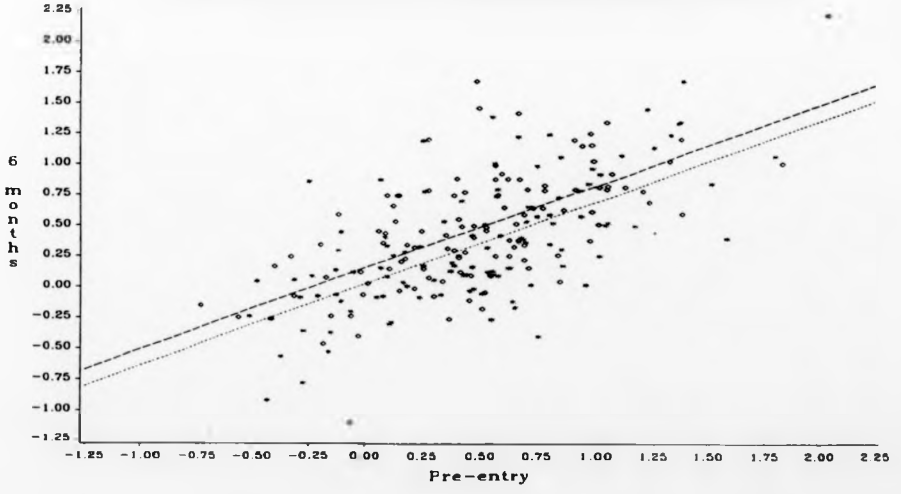


Figure 3.3.2 : Log(Triglycerides), pre-entry vs 6 months, for drug A (n=109, dotted line, stars) and drug B (n=110, dashed line, diamonds). Separate linear regression lines fitted for each group



Observed covariance structures, within-groups, respectively pooled:

$$\hat{\Sigma}_A = \begin{bmatrix} .248 & \\ .165 & .256 \end{bmatrix} \qquad \hat{\Sigma}_B = \begin{bmatrix} .223 & \\ .148 & .242 \end{bmatrix} \qquad \hat{\Sigma}_{pooled} = \begin{bmatrix} .235 & \\ .156 & .249 \end{bmatrix}$$

$$\hat{\rho}_A = .654, \quad \hat{\beta}_A = .665 \qquad \hat{\rho}_B = .638, \quad \hat{\beta}_B = .665 \qquad \hat{\rho}_{pooled} = .646, \quad \hat{\beta}_{pooled} = .665$$

Figure 3.3.2 looks very different from figure 3.3.1, here the within-group regression lines are parallel, and the data-scatter looks much more reminiscent of a bivariate normal distribution. Let us now see what impact a log-transformation has on the ANCOVA analysis.

Table 3.3.2 : Analysis using ANCOVA of the triglycerides data on original and log-scale. For comparative purposes the remaining summary statistics from table 3.3.1 are also included.

Summary statistic	Estimated treatment effect	Standard error	t-statistic	p-value
$Y_{ij} - \beta \cdot X_{ij}$	.225	.106	2.126	.035
$\log(Y_{ij}/X_{ij}^{\delta})$	.131	.052	2.544	.012
Y_{ij}/X_{ij}^{θ}	.152	.068	2.236	.026
$Y_{ij} - X_{ij}$	.220	.115	1.910	.057
$\log(Y_{ij}/X_{ij})$	.126	.056	2.249	.026
Y_{ij}/X_{ij}	.125	.063	1.972	.050
Y_{ij}	.234	.138	1.700	.091
$\log(Y_{ij})$	.142	.067	2.102	.037

Note: the estimated treatment effects and their standard errors are not directly comparable since different scales are being used.

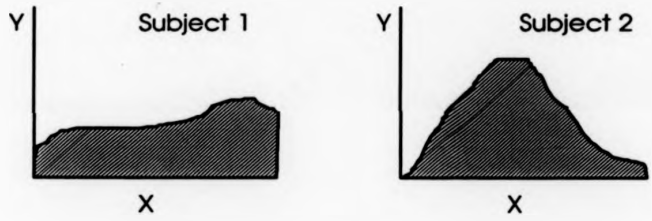
For this data set a logarithmic transformation appears well justified. Analysis of covariance on the log-transformed data gave the strongest evidence of a treatment difference, and is also, based on the covariate adjustment and the observed distributions, the most reliable approach in terms of validity.

The completely additive and the completely multiplicative models are just two possibilities among a vast number of choices. Even for the very simple model consisting of only treatment effect and residual, depending on the relationship between these effects, there are an infinite number of possible true underlying models. The purpose of this section was to give some advice in choosing between models, and also to emphasize the importance of checking the plausibility of the chosen model in explaining the variability in the data and in fulfilling the assumptions made for the analysis.

3.4 THE AREA UNDER THE CURVE

One of the more popular summary statistics which has not been discussed so far is "the area under the (response) curve", AUC. This summary statistic is calculated by adding the areas under the curve between each pair of consecutive observations for a given subject. In that way we obtain the total area between that subjects response curve and the x-axis for the full study period. This total area then constitute the summary statistic. Examples of hypothetical AUC's for two subjects are given in the figure below.

Figure 3.4.1 : AUC's for two hypothetical subjects when response was recorded continuously over time.



Several alternative ways for the calculation of AUC's are in use. Two decisions have to be made in choosing between them; that of the potential use of baseline(s), and that of interpolation. These two issues will be discussed in the following.

Regarding baselines, two different strategies are available when calculating the AUC for a given subject, subtracting or not subtracting that recording from all the measurements made under treatment. To distinguish between the two, the former will be referred to as AUC_{change}, and the latter as AUC_{post}. As will be shown there is a strong relationship between AUC_{post} and POST, and between AUC_{change} and CHANGE.

Suppose one wishes to use AUC's for describing and making inferences of the responses to treatment. Then the best option, in terms of both validity and efficiency, is clearly to do this with an analysis of covariance adjustment for the pre-entry level.

Also, in the same way as we get equivalent results when analysing POST or CHANGE with a covariate adjustment for the pre-entry level, we will get equivalent results whether we choose AUC_{post} or AUC_{change} as dependent variable in an ANCOVA model.

In the remaining parts of this section we will not explicitly mention covariate adjustments. Instead emphasis will be on evaluation of the usefulness of AUC_{post} relative to POST, and of AUC_{change} relative to CHANGE as dependent variables.

Apart from the choice of subtracting or not subtracting an existing baseline from all the measurements for a given subject, we also have to decide on the choice of interpolation. By definition, the true individual's area under the curve calls for a continuous recording of the variable of interest, otherwise we do not have access to a proper response curve, we only have information on the position of the curve at certain time points. An exception to this rule is studies involving continuous ambulatory 24-hour measurements (e.g. of blood pressures). Normally, though, we have to use some kind of interpolation between the successive measurements. We will base our results on the most widely used method, the trapezoidal rule, which is based on linear interpolation between the repeated measurements. More sophisticated techniques are available, see Crowder and Hand (1990).

Having decided to use the trapezoidal rule, the general formula without baseline subtraction, but with one measurement before and r measurements after randomisation, and with the r+1 recordings taken at the time points; $t_0, t_1, \ldots, t_r$, is (see Matthews et al, 1990):

$$AUC_{post} = \frac{1}{2}\sum_{i=0}^{r-1}(t_{i+1} - t_i)(y_i + y_{i+1})$$

Without access to a pre-entry measurement the formula is:

$$AUC_{post} = \frac{1}{2}\sum_{i=1}^{r-1}(t_{i+1} - t_i)(y_i + y_{i+1})$$

With a pre-entry evaluation, and using the baseline subtraction, we get:

$$AUC_{change} = \frac{1}{2}\sum_{i=0}^{r-1}(t_{i+1} - t_i)\left[(y_i - y_0) + (y_{i+1} - y_0)\right]$$

When there are more than one recording pre-randomisation, one simply substitutes the average of these, $\bar{y}^{pre} = \frac{1}{p}\sum_{i=-(p-1)}^{0} y_i$, for y_0 in the preceding formula.

To be able to make direct comparisons of the variances between AUC_{post} and POST, and between AUC_{change} and CHANGE, we need to change the units of measurement for AUC_{post} and AUC_{change} to give them the same expected value as POST and CHANGE have. This will not affect the efficiency of any of the analyses, only the units of the measurements. This feature of scale-invariance for linear summary statistics is discussed in more detail in section 5.2.

For simplicity, we will begin to make our comparisons under the assumption of equal distances between any two adjacent time points. Making the expected values for the summary statistics equal is accomplished, under a model of constant treatment effects, by dividing the resulting AUC's with the total time period for the study. After this scaling, the total study period is equal to one. AUC_{post} may now be calculated, in the absence of baselines, from:

$$AUC_{post} = \sum_{i=1}^{r} c_i y_i \text{ , where } c_i = \begin{cases} \frac{1}{2(r-1)} & \text{, i = 1 and r} \\ \frac{1}{(r-1)} & \text{,i = 2,...,r-1} \end{cases}$$

These weights are seen to be very similar to what we use with POST, where each measurement receives the weight 1/r. The difference being that AUC_{post} gives only half the weight to the first and the last recordings relative to what it gives to the intermediate ones. As a result AUC_{post} gives $\frac{r}{(r-1)}$ times the weight to the intermediate measurements relative to what POST does.

Under the current assumptions of equal time intervals between successive measurements, and further, under compound symmetry, the variance for AUC_{post} with our transformed time-scale is:

$$Var[AUC_{POST}] = \frac{\sigma^2}{4(r-1)^2}\{(2r-3)[2+(2r-3)\rho]+\rho\}$$

This may be compared with the variance for POST:

$Var[POST] = \frac{\sigma^2}{r}[1+(r-1)\rho]$. The difference between the two being:

$Var[AUC_{POST}] - Var[POST] = \frac{\sigma^2(1-\rho)(r-2)}{2r(r-1)^2}$, which is strictly non-negative, implying that we are bound to lose precision by using the AUC approach as opposed to POST. To illustrate the magnitude of this inferiority table 3.4.1 is given.

Table 3.4.1 : Dependence of Var[POST]/Var[AUC_{post}] on the number of post-treatment measurements r and the correlation ρ, Assuming compound symmetry, equidistance between consecutive visits, and no pre-entry evaluations.

Number of post-treatment measurements, r	Correlation, ρ .3	.5	.7	.9
3	.948	.970	.985	.996
4	.961	.978	.989	.997
6	.978	.989	.995	.999
10	.991	.996	.998	.999

For many repeated measures designs it will not be the case that all time intervals between successive visits are equal, e.g. visits may be more frequent early on. If treatment effects are constant, and under compound symmetry, it is easy to show that the more irregular the time intervals are the more inferior will AUC_{post} be relative to POST. The reason for this is the successively more unequal weights used by AUC_{post}.

We will now compare variances when we have access to one or more pre-entry evaluations (for simplicity assuming equal time intervals between visits). Firstly, without subtracting the baseline(s), AUC_{post} may be calculated from:

$$AUC_{post} = \sum_{i=0}^{r} c_i y_i \quad , \text{ with } c_i = \begin{cases} \frac{1}{2(r-1/2)} & ,i = 0 \text{ and } r \\ \frac{1}{(r-1/2)} & ,i = 1,\ldots,r-1 \end{cases}$$

This corresponds to the formula for the AUC as given by Matthews et al (1990), for the case when all time intervals between successive measurements are equal, and when the total time period has been scaled such that the expected value for the AUC is the same as the expected values we have for POST and CHANGE.

One peculiarity with this particular summary statistic should be noted. While POST ignores existing mean pre-treatment differences between groups, and CHANGE usually overcorrects for them (see subsection 3.1), AUC_{post} actually inflates this imbalance (when it exists) by having a positive weight for the baseline measurement.

Investigating the efficiency of AUC_{change} instead, and contrasting this with CHANGE, we will start off by giving the formula for its calculation:

$$AUC_{change} = \sum_{i=0}^{r} c_i y_i \quad , \text{ with } c_i = \begin{cases} -1 & ,i = 0 \\ \frac{1}{(r-1/2)} & ,i = 1,\ldots,r-1 \\ \frac{1}{2(r-1/2)} & ,i = r \end{cases}$$

When multiple baselines are available $\bar{y}^{pre}$ should be substituted for y_0. The variances for AUC_{change} are, as may be seen in table 3.4.2, quite similar to the corresponding variances for CHANGE, there is, however, a small degree of loss in efficiency incurred by not having equal weights for all the post-treatment measurements. It may further be observed that the relative efficiency between AUC_{change} and CHANGE is independent of the degree of correlation, ρ.

Table 3.4.2: Dependence of Var[CHANGE]/Var[AUC_{change}] on the number of post-treatment measurements r and the correlation ρ, Assuming compound symmetry, equidistance between consecutive visits, and one pre-entry evaluation.

Number of pre and post-treatment measurements, p+r	Correlation, ρ .3	.5	.7	.9
1+2	.964	.964	.964	.964
1+3	.980	.980	.980	.980
1+5	.992	.992	.992	.992
1+10	.998	.998	.998	.998

The AUC, as a summary measure, is often analysed as an understandable feature in relation to an individual response curve, usually without having any direct physical interpretation. One exception may be noted, however, as discussed by Crowder and Hand (1990). In so called first-order kinetics the instantaneous rate of exchange between compartments of a substance is in direct proportion to the difference in concentrations at the interface. For such solutions, and for compartments in series, it can be shown that the total area under the curve is inversely proportional to the elimination rate constant of the substance. Thus when primary interest centres on the elimination rate, the AUC should be analysed.

In conclusion, in relation to the area under the curve approach, AUC_{post} behaves very similar compared to POST, and the same holds for the comparison between AUC_{change} and CHANGE. The respective pairs of summary statistics are highly correlated. There is, however, under the assumptions of a constant treatment effect and under compound symmetry, a slight loss in efficiency incurred by choosing an AUC-approach. Also, one should beware of giving positive weights for pre-randomisation visits, and of giving less weight to final visits.

3.5 OPTIMAL ALLOCATION OF VISITS FOR ANCOVA

Having decided to use ANCOVA at the design stage for a repeated measures study, there might be a wish to go one step further and to consider the allocation of measurements before and after the randomisation to optimise the precision for a given total number of visits.

This topic was touched upon in subsection 2.2.2 in the context of a constant treatment effect after randomisation and under the assumption of compound symmetry. Under these circumstances it was shown that, conditional on the total number of measurements, t, and the equicorrelation, ρ, the optimal choice for the number of pre-entry evaluations was given by;

p=p' when ρ lies between $\begin{cases} \frac{1}{t-2(p'-1)} \text{ and } \frac{1}{t-2p'} & \text{, for } p'>0 \\ 0 \text{ and } \frac{1}{t} & \text{, for } p'=0 \end{cases}$

A more direct way of deriving the optimal p is given by the formula; $\hat{p}=\begin{cases} \frac{1}{2}\left(t+1-\frac{1}{|\rho|}\right) & \text{for } |\rho|>\frac{1}{t+1} \\ 0 & \text{for } |\rho|\le\frac{1}{t+1} \end{cases}$

However, here p is treated as continuous, the optimal choice will be either the "smallest larger" or the "largest smaller" integer. In most instances a simple rounding off procedure will give the optimal choice.

As an illustration, figure 3.5.1 gives the optimal number of pre-entry visits for three different choices of t; 10, 7 and 4, and depending on the degree of correlation. One general result is that with t even, as soon as ρ≥.5, the optimal choice has equal number of visits before and after randomisation. We may also note that only when correlations are really small will the optimal choice call for substantially more measurements post than pre-randomisation, this situation is unlikely to occur for a repeated measures design.

Figure 3.5.1 : Optimal number of pre−entry visits for ANCOVA, assuming compound symmetry and a fixed total number of visits; 10, 7 or 4

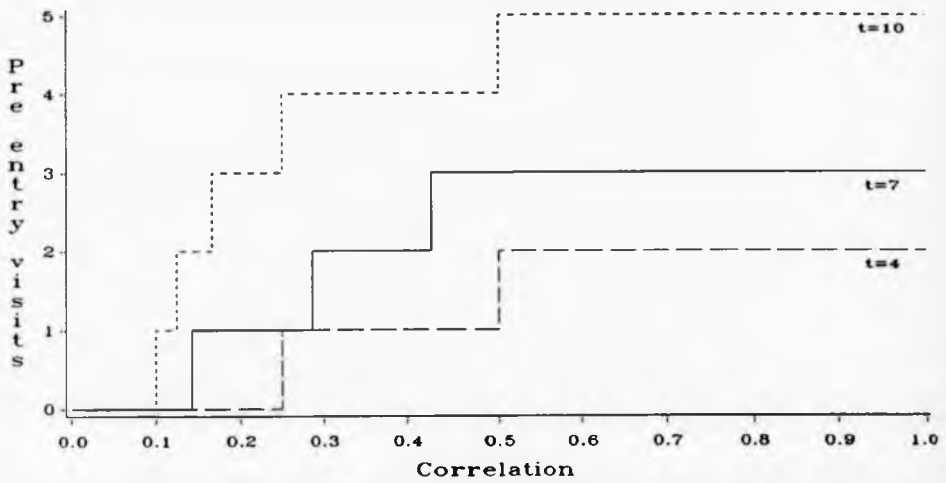
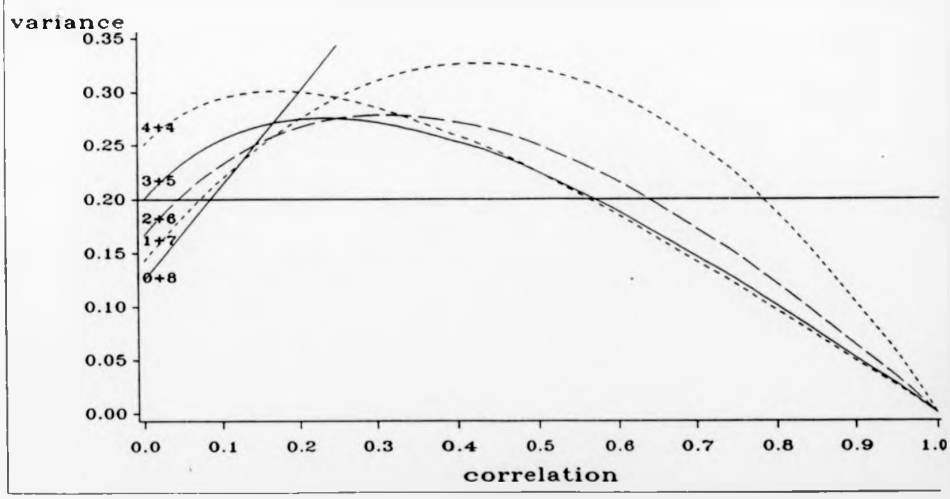


Figure 3.5.2 : Variances for ANCOVA assuming a total of 8 visits, but different numbers pre and post, by degree of correlation



To illustrate not only the optimal choice of p and r, given t and ρ, but also the relative differences in precision expected to occur depending on the choice of p and r, figure 3.5.2 is given. Here, a repeated measures design encompassing a total of 8 measurements is evaluated, and the resulting variances for five alternative ANCOVA's are shown as a function of ρ, under the assumption of compound symmetry and a constant treatment effect.

The five ANCOVA's are based on 4,3,2,1 and 0 (the degenerate case of POST) pre-entry measurements, with correspondingly 4,5,6,7 and 8 measurements post-treatment. We see that, for ρ in the plausible range .5 to .8, having p equal to 3 or 4 are about equally effective, using p=2 is slightly less efficient, allowing for only one baseline would imply a quite substantial loss in efficiency, while the choice p=0 is literally speaking out of the picture.

We will now look beyond the simplifying assumption of compound symmetry, but we will stick to a constant treatment effect. The reason is that we otherwise move outside the direct scope of ANCOVA. With non-constant treatment effects over time, using equal weights for all post-randomisation measurements will not be optimal, and other summary statistics might be called for. Considerations of this kind will be pursued in chapter 5.

Obtaining general results for the choice of the number of pre and post-treatment measurements under a completely general covariance structure is not algebraically tractable. We want to minimize the ANCOVA variance, $\bar{\Sigma}_{post} - \bar{\Sigma}^2_{mix}/\bar{\Sigma}_{pre}$, depending on the choice of p, given t. To be able to do this we have to condition on the covariance structure. One plausible alternative for repeated measures designs (as was found in section 2.3) is to assume a linear decrease in correlation with increasing time intervals between assessments.

Using the notation of section 2.3, the averages for the three submatrices of Σ are given by;

$$\bar{\Sigma}_{pre} = \frac{1}{p}\left[1+(p-1)\left(\gamma - \frac{b(p-2)}{3(t-2)}\right)\right]$$

$$\bar{\Sigma}_{mix} = \gamma - b/2$$

$$\bar{\Sigma}_{post} = \frac{1}{r}\left[1+(r-1)\left(\gamma - \frac{b(r-2)}{3(t-2)}\right)\right]$$

Given t,γ (the correlation between adjacent visits) and b (the total decay in correlation over the study period), we can now readily compute the variance for ANCOVA for any design choice of p. Since $\bar{\Sigma}_{mix}$ is independent of p we only have to consider the impact of the choice of the number of pre-entry evaluations on $\bar{\Sigma}_{pre}$ and $\bar{\Sigma}_{post}$. We want both $\bar{\Sigma}_{pre}$ and $\bar{\Sigma}_{post}$ to be small in order to minimize the expression $\bar{\Sigma}_{post} - (\gamma - b/2)^2 / \bar{\Sigma}_{pre}$.

The relationships with changing p's, however, goes in opposite directions. As p increases $\bar{\Sigma}_{pre}$ decreases and $\bar{\Sigma}_{post}$ increases, and vice versa when p decreases. The end result on the ANCOVA variance will depend on both γ and b as well as on t, and again, general results are intractable. However, conditional on γ, b and t, computation of ANCOVA variances for any p are straightforward, and this has been done in producing table 3.5.1.

Table 3.5.1: Optimal number of pre-entry visits (p) for minimizing the ANCOVA variance depending on the correlation between adjacent visits (γ), and the total decline (assumed linear) in correlation over the study duration (b). We are assuming a design consisting of t=8 visits in total, a constant treatment effect, as well as equal variances for all time-points.

γ	b=.0	b=.1	b=.2	b=.3	b=.4
.9	4	4	3	2	2
.7	4	4	3	3	2
.5	3=4	3	3	2	1

For a given starting correlation, γ, the optimal p decreases with increasing b. This is expected since an increasing b will decrease $\overline{\Sigma}_{mix}$, and thereby the regression coefficient of $\overline{y}^{post}$ on $\overline{y}^{pre}$, which implies that the value of a covariate adjustment diminishes.

The function curve determining the optimal p given t, γ and b, is quite flat around its minimum. This may be exemplified by the resulting ANCOVA variances (arbitrarily scaled) for the case when t=8, γ=.7 and b=.2, which are as follows :

p	r	Var[ANCOVA]
0	8	243.3
1	7	120.1
2	6	103.7
3	5	100.0
4	4	101.6
5	3	108.1
6	2	123.6
7	1	172.7

As a general rule, for repeated measures designs with an anticipated stable treatment effect after randomisation, and with a total number of visits not exceeding 10, choosing p between 2 and t/2 will in most circumstances result in an analysis close to the optimal efficiency. When t>10 having p>2 might well be worth wile.

3.6 SEPARATE BASELINES OR THEIR MEAN

The simplest form of ANCOVA uses multiple pre-entry evaluations as a single mean, and for each subject the summary statistic is $\bar{x}_{ij.}^{post} - \hat{\beta} \cdot \left(\bar{x}_{ij.}^{pre} - \bar{\bar{x}}_{..}^{pre} \right)$. Even if, as shown in chapter 5, this is optimal when the covariance structure adheres to compound symmetry, it may be far from optimal for other covariance structures. Thus, an evaluation of the possible advantages of including all pre-entry measurements separately in the ANCOVA model, and using the summary statistic $\bar{x}_{ij.}^{post} - \hat{\gamma}_1 \cdot \left(x_{ij1}^{pre} - \bar{\bar{x}}_{..1}^{pre} \right) - \ldots - \hat{\gamma}_p \cdot \left(x_{ijp}^{pre} - \bar{\bar{x}}_{..p}^{pre} \right)$ is motivated. The choice between these two summary statistics is the main objective of this section.

3.6.1 Adjustment for multiple covariates

Apart from having to decide whether multiple baselines should be used separately or as a single mean, there will often be a desire to include other prognostic factors in the statistical model to further increase efficiency, and to enhance understanding of the underlying model. In most clinical trials a whole battery of prognostic variables are recorded on all patients before treatment commences. Surely the investigators would not have wasted their time and money on collecting all this data if they thought it would be of no relevance to the primary outcome measures. In a recent paper Tukey (1993) expresses the view that "we have a scientific method obligation to do what we can to milk our covariates as thoroughly as is reasonable". On the other hand, with selective use of significant covariates from a large choice of covariates, this can lead to "overprediction".

In his section on "adjustment for many covariates", Tukey motivates the use of compound covariates (i.e. linear combinations) as opposed to multivariate analyses. He suggests that it is often desirable to first construct one or two compound covariates and then work with them.

What determines an individual covariate's value in contributing to a given model is not its univariate correlation with the outcome, but rather its influence on the multiple correlation coefficient R between the compound covariate and the dependent variable.

The formula for the multiple correlation (see Flury, 1989) corresponds very nicely to the formula for a univariate correlation. This may be seen from the following equalities, where y is the dependent variable and x the (vector of) covariate(s):

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} = \left\{ \sigma_{xy} \left(\sigma_x^2\right)^{-1} \sigma_{xy} / \sigma_y^2 \right\}^{1/2}$$

$$R_{xy} = \quad \ldots \quad = \left\{ \Sigma_{mix}^{T} \Sigma_{pre}^{-1} \Sigma_{mix} / \sigma_y^2 \right\}^{1/2}$$

For a given dependent variable, when multiple covariates are available, one will typically want to increase R^2 as much as possible without incorporating too many covariates, see Rencher (1993). We will concentrate on the special case of when to use pre-entry measurements separately or averaged for ANCOVA.

3.6.2 Pre-entry measurements separately or averaged for ANCOVA

We are now interested in a scenario where we have one dependent variable y, which may or may not be the average of r post-treatment measurements, and p pre-entry measurements, $x_1,..,x_p$ (or more generally prognostic variables).

The following two models will be contrasted:

$$y_{ij} = \mu_i + \beta\left(\bar{x}_{ij.} - \bar{\bar{\bar{x}}}_{...}\right) + \varepsilon_{ij}$$

$$y_{ij} = \mu_i + \gamma_1\left(x_{ij1} - \bar{\bar{x}}_{.1}\right) + \ldots + \gamma_p\left(x_{ijp} - \bar{\bar{x}}_{.p}\right) + \eta_{ij}$$

The question is; are we better off using $\bar{x}_{ij.}$ as covariate or using $x_1,..,x_p$ individually in a multiple analysis of covariance?

The answer to this will depend on the covariance structure, as well as on the sample sizes. With small samples the advice will usually be to use a single mean, since the regression coefficients for the separate covariates would be to unreliably estimated. With larger samples, and when the covariance structure differs from compound symmetry, we may gain precision by using separate covariates. In this latter situation, the decision will depend on a balancing between the possible gain in efficiency relative to the increase in complexity of the model.

We will begin by having a look at the simplest situation, when we have two pre-entry measurements.

3.6.2.1 Two pre-entry measurements (covariates)

With two covariates and one dependent variable, the covariance structure is given by;

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{12} & \sigma_2^2 & \\ \sigma_{1y} & \sigma_{2y} & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} \Sigma_{pre} & \Sigma_{mix} \\ \Sigma_{mix}^T & \sigma_y^2 \end{bmatrix}$$

We will be contrasting the following two summary statistics:

$$\text{ANCOVA}_1 = y_{ij} - \beta\left(\bar{x}_{ij} - \bar{\bar{\bar{x}}}_{..}\right)$$

$$\text{ANCOVA}_2 = y_{ij} - \gamma_1\left(x_{ij1} - \bar{\bar{\bar{x}}}_{..1}\right) - \gamma_2\left(x_{ij2} - \bar{\bar{\bar{x}}}_{..2}\right)$$

Assuming known covariance matrices the ANCOVA variances are given by;

$$\text{Var}[\text{ANCOVA}_1] \propto \sigma_y^2 - \bar{\bar{\Sigma}}_{mix}^2 / \bar{\bar{\Sigma}}_{pre} = \sigma_y^2 - \beta \cdot \bar{\bar{\Sigma}}_{mix}$$

$$\text{Var}[\text{ANCOVA}_2] \propto \sigma_y^2 - \gamma_1 \cdot \sigma_{1y} - \gamma_2 \cdot \sigma_{2y}$$

Where; $\beta = \bar{\bar{\Sigma}}_{mix} / \bar{\bar{\Sigma}}_{pre}$, $\gamma_1 = \dfrac{\sigma_{1y}\sigma_2^2 - \sigma_{2y}\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$, $\gamma_2 = \dfrac{\sigma_{2y}\sigma_1^2 - \sigma_{1y}\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$

To facilitate the interpretations of the relationships we normalize all three variables (divide them by their respective standard deviations). For $ANCOVA_2$ this has no effect at all on the efficiency of the analysis. However, for $ANCOVA_1$ it will affect the relative weights the two pre-entry measurements get (if the two pre-entry variances differ). The pre-entry measurement with the higher variance will contribute slightly less to the pre-entry mean in the normalized case (since it has been divided by a larger standard deviation). Normally this will make very little difference for the choice between the two ANCOVA approaches.

For normalized variables the covariance matrix simplifies to;

$$\Sigma = \begin{bmatrix} 1 & & \\ \rho_{12} & 1 & \\ \rho_{1y} & \rho_{2y} & 1 \end{bmatrix}$$, and correspondingly the variance formulae

changes to;

$$Var[ANCOVA_1] \propto 1 - \frac{(\rho_{1y} + \rho_{2y})^2}{2(1+\rho_{12})}$$

$$Var[ANCOVA_2] \propto 1 - \frac{\rho_{1y}^2 + \rho_{2y}^2 - 2\rho_{12}\rho_{1y}\rho_{2y}}{1-\rho_{12}^2}$$

Thus, $Var(ANCOVA_1)$ exceeds $Var(ANCOVA_2)$ by $\frac{(\rho_{1y} - \rho_{2y})^2}{2(1-\rho_{12})}$

Hence, what primarily matters is whether the two repeat baselines are equally correlated with the dependent variable or not. If they are we should use $ANCOVA_1$, otherwise, it might be worth-while to use $ANCOVA_2$. We also see that, for any given difference in correlation between the two baselines and the outcome, the relative importance of this difference will be magnified when the correlation between the baselines is substantial, and hence the motivation to opt for $ANCOVA_2$ will be larger.

So much for the covariance structure, we will now look at the impact of sample sizes, and the price we have to pay for estimating two regression coefficients instead of one (since we are hardly expected to know the true underlying covariance structure).

The variances (for non-normalized variables) are now given by;

$$\text{Var[ANCOVA}_1\text{]} = \left(\frac{1}{n_A}+\frac{1}{n_B}+\zeta_1\right)\left(\sigma_y^2-\hat{\beta}\cdot\overline{\Sigma}_{mix}\right)\frac{n_A+n_B-2}{n_A+n_B-3}$$

$$\text{Var[ANCOVA}_2\text{]} = \left(\frac{1}{n_A}+\frac{1}{n_B}+\zeta_2\right)\left(\sigma_y^2-\hat{\gamma}_1\sigma_{1y}-\hat{\gamma}_2\sigma_{2y}\right)\frac{n_A+n_B-2}{n_A+n_B-4}$$

where; $\zeta_1 = \dfrac{(\overline{\overline{x}}_{A..}-\overline{\overline{x}}_{B..})^2}{(n_A+n_B-2)\overline{\Sigma}_{pre}}$ and

$$\zeta_2 = \frac{\sigma_2^2(\overline{x}_{A.1}-\overline{x}_{B.1})^2+\sigma_1^2(\overline{x}_{A.2}-\overline{x}_{B.2})^2-2\sigma_{12}(\overline{x}_{A.1}-\overline{x}_{B.1})(\overline{x}_{A.2}-\overline{x}_{B.2})}{(n_A+n_B-2)(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}$$

For some background to these formulae, see Snedecor and Cochran, (1989, pp 386 and 441).

Comparing the correction factors for the estimated variances we see, firstly, that we lose one additional degree of freedom due to the estimation of a second regression coefficient. What is less clear is the relation between the ζ's, the correction factors making allowance for non-orthogonality between covariates and treatment groups.

A formal comparison based on the sampling distributions of the two ζ's will not be performed here, but it may be noted that ζ_1 is distributed as 1 plus $\left(F_{1,n-2}\right)/(n-2)$, see Laird and Wang (1990, p.410) for details.

Taking a simpler route and assuming knowledge of the underlying covariance structure, the expected values of the ζ's are given by;

$$E[\zeta_1] = \frac{n_A + n_B}{n_A \cdot n_B \cdot (n_A + n_B - 2)} \approx \frac{1}{n_A \cdot n_B}$$

$$E[\zeta_2] = \frac{2 \cdot (n_A + n_B)}{n_A \cdot n_B \cdot (n_A + n_B - 2)} \cdot \frac{\sigma_1^2 \sigma_2^2}{(\sigma_1^2 \sigma_2^2 - \sigma_{12}^2)} \approx \frac{2}{n_A \cdot n_B} \cdot \frac{\sigma_1^2 \sigma_2^2}{(\sigma_1^2 \sigma_2^2 - \sigma_{12}^2)}$$

I.e. ζ_2 is at least twice as large as ζ_1. Compared to $\left(\frac{1}{n_A} + \frac{1}{n_B}\right)$ both these correction factors are usually very small.

A hypothetical example might help in clarifying the relevance of the differences in performance between $ANCOVA_1$ and $ANCOVA_2$ depending on covariance structure and sample sizes.

Assuming an underlying covariance structure of $\Sigma = \begin{bmatrix} 1 & & \\ .6 & 1 & \\ .5 & .7 & 1 \end{bmatrix}$

the variances, calculated using the formulae given above for standardized variables, will be proportional to;

$Var[ANCOVA_1] \propto .55$
$Var[ANCOVA_2] \propto .50$

Having to estimate the regression coefficients, and using the expected values for the correction factors based on knowledge of the covariance structure, we can compare the expected variances for the two ANCOVA's depending also on sample size. Using the covariance structure from the example above, along with the last variance formulae given for the two ANCOVA's, it may be found that having 16 subjects or less per group will give a lower expected variance for $ANCOVA_1$, while providing for 17 or more subjects per group would give a better expected precision using $ANCOVA_2$.

These results are highly dependent on the difference $\rho_{1y} - \rho_{2y}$, assuming these correlations to be .58 and .62, instead of .5 and .7, the number of patients needed per group for making the expected variance of $ANCOVA_2$ the lower would increase to 430.

To further illustrate the differences between $ANCOVA_1$ and $ANCOVA_2$, two figures are gives. Figure 3.6.1 displays the variances based on knowledge of the covariance structure. We see to which extent the use of $ANCOVA_2$ gets more advantageous as the difference $|\rho_{1y} - \rho_{2y}|$ increases. This relationship is given for some different values of ρ_{12}.

Figure 3.6.2 gives the variance ratio between $ANCOVA_2$ and $ANCOVA_1$ for the normal case where we have to estimate the regression coefficients (with sizes of the correction factors as expected based on the specified covariance structure). This variance ratio is given depending on sample sizes, with different curves displaying the relationships for some possible choices of $|\rho_{1y} - \rho_{2y}|$.

In conclusion, in a repeated measures setting, when the two covariates are measuring the same variable as the outcome measure, there are no reasons to expect substantial differences between ρ_{1y} and ρ_{2y}, unless the correlations are strongly time dependent. The use of $ANCOVA_1$ would normally be recommended.

3.6.2.2 Three or more pre-entry measurements (covariates)

Disregarding the sample size correction and the other correction factors, an ANCOVA with p separate covariates (here pre-entry measurements) has the following variance;

$$\text{Var}[\text{ANCOVA}_p] = \sigma_y^2 - \gamma_1 \cdot \sigma_{1y} - \ldots - \gamma_p \cdot \sigma_{py} = \sigma_y^2 - \left(\Sigma_{mix}^T\right)^2 \cdot \Sigma_{pre}^{-1}$$

The vector of regression coefficients may be calculated from;

$$\gamma' = \Sigma_{mix}^T \cdot \Sigma_{pre}^{-1}$$

Figure 3.6.1 : Variances for ANCOVA1 and ANCOVA2 when assuming knowledge of the true covariance structure. Depending on the difference in "mixed" corrrelations. Different pair of curves for different "pre" correlations

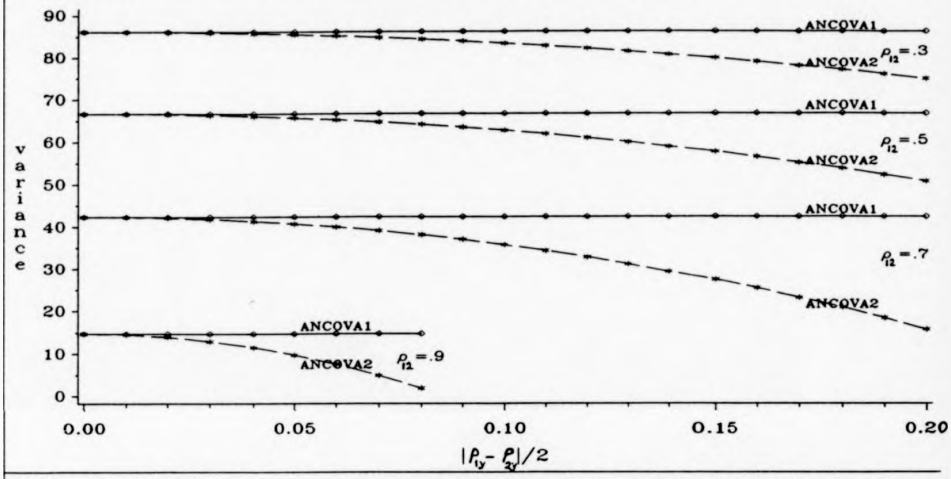
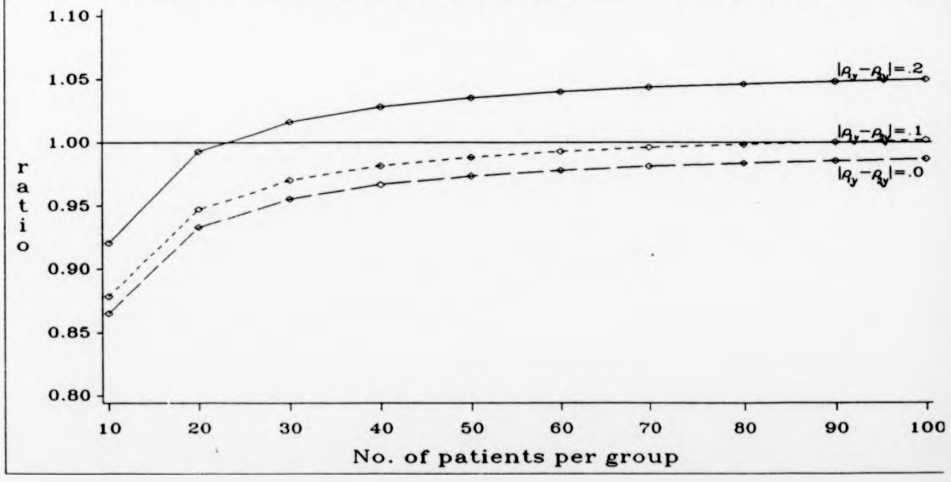


Figure 3.6.2 : Variance ratio, Var(ANCOVA1) / Var(ANCOVA2), based on expected values for the correction factors. Depending on sample sizes and differences between "mixed" correlations



For the special case with p=3 pre-entry measurements, and for normalized variables, we have the following covariance matrix;

$$\Sigma = \begin{bmatrix} 1 & & & \\ \rho_{12} & 1 & & \\ \rho_{13} & \rho_{23} & 1 & \\ \rho_{1y} & \rho_{2y} & \rho_{3y} & 1 \end{bmatrix} .$$

Our ANCOVA variance is; $\mathrm{Var}[\mathrm{ANCOVA}_3] \propto 1-\gamma_1\rho_{1y}-\gamma_2\rho_{2y}-\gamma_{31}\rho_{3y}$, where the three regression coefficients may be found from ("det" stands for determinant);

$$\gamma_1 = \frac{(1-\rho_{23}^2)\rho_{1y}-(\rho_{12}-\rho_{13}\rho_{23})\rho_{2y}-(\rho_{13}-\rho_{12}\rho_{23})\rho_{3y}}{\det[\Sigma_{pre}]}$$

$$\gamma_2 = \frac{(1-\rho_{13}^2)\rho_{2y}-(\rho_{12}-\rho_{13}\rho_{23})\rho_{1y}-(\rho_{23}-\rho_{12}\rho_{13})\rho_{3y}}{\det[\Sigma_{pre}]}$$

$$\gamma_3 = \frac{(1-\rho_{12}^2)\rho_{3y}-(\rho_{13}-\rho_{12}\rho_{23})\rho_{1y}-(\rho_{23}-\rho_{12}\rho_{13})\rho_{2y}}{\det[\Sigma_{pre}]}$$

For a given (assumed known) covariance structure it is now possible to derive the amount of decrease in ANCOVA variance attainable by using separate covariates instead of their mean.

A small example of this follows, assume the covariance structure is given by; $\Sigma = \begin{bmatrix} 1 & & & \\ .8 & 1 & & \\ .7 & .8 & 1 & \\ .6 & .7 & .8 & 1 \end{bmatrix}$.

Then $\mathrm{Var}[\mathrm{ANCOVA}_1]=.42$, while $\mathrm{Var}[\mathrm{ANCOVA}_3]=.35$. Thus, when compound symmetry does not apply it is possible to gain some efficiency by taking account of the different dependencies between the pre-entry measurements and the dependent variable.

It may be of interest to see how the two different ANCOVA models turn out for this example.

ANCOVA_1: $y_{ij} = \mu_i + .829 \cdot \left(\bar{x}_{ij} - \bar{\bar{\bar{x}}}_{..}\right) + \varepsilon_{ij}$

ANCOVA_3: $y_{ij} = \mu_i + .00 \cdot \left(x_{ij1} - \bar{\bar{\bar{x}}}_{.1}\right) + .167 \cdot \left(x_{ij2} - \bar{\bar{\bar{x}}}_{.2}\right) + .667 \cdot \left(x_{ij3} - \bar{\bar{\bar{x}}}_{.3}\right) + \eta_{ij}$

We see that, in the presence of the two latter pre-entry measurements, the first is of no value for decreasing the ANCOVA_3 variance. With a steeper decrease in correlation with time, the first regression coefficient would be negative.

The issue of the expected sizes of the necessary correction factors for the variance formulae, when we have to estimate the regression coefficients from the data, for the case with more than two covariates, has not been investigated. As before. however, the general recommendation is to use a single mean of the pre-entry measurements for most repeated measures studies, unless sample sizes are large and correlations clearly unequal.

4 REGRESSION TO THE MEAN

4.1 INTRODUCTION

This concept was introduced by Galton in an 1877 paper. In a later paper (1885) he exemplified the term in the following illuminating way: "Each peculiarity in a man is shared by his kinsmen, but on the average to a less degree".

In the current statistical literature regression to the mean is used to identify the following phenomenon: a variable that is extreme on its first measurement will tend to be closer to the centre of the distribution for a later measurement.

The topic of regression to the mean is well covered in the literature, general overviews have been give by Davis (1976), Cutter (1976), Ederer (1972), and Johnson and George (1991). Many articles have dealt with the problem of relating change to initial value, like Blomqvist (1977), Oldham (1962), MacGregor et al (1985), and Hayes (1988). The implications that regression to the mean has for screening (e.g. of cholesterol levels) have been addressed by Thompson and Pocock (1990), Roeback et al (1993), and Chen and Cox (1992). Almost all results are given under a normal-theory framework, some exceptions are Das and Mulder (1983), Davis (1986), and Senn (1990). Finally, papers dealing with the consequences for between-group comparisons are scarce, one recent reference is Chambless and Roeback (1993).

Regression to the mean has several implications for comparative clinical trials. Most evident are the effects for within-group comparisons, especially in the presence of selection criteria (e.g. when only subjects with a diastolic BP $\geq$ 95 are randomised). Then, any mean change during treatment for a given group will be partly due to regression to the mean (RTM)-effects. This has led to many misleading conclusions in the literature regarding treatment effects, effects on sub-groups, and dependencies between pre-entry levels and change during treatment.

The best way to get unbiased results is to have access to a control group, and to make between-group comparisons. This is where our main interest is, however, regression to the mean is still of concern. For instance, when we have a mean pre-randomisation difference between groups, in the absence of treatment effects, this mean difference is expected to decrease at a subsequent measurement. This effect of regression to the mean was noted in section 3.1, where it was shown that this implies biased estimates of treatment effects for POST and CHANGE, while ANCOVA remains valid.

Further, when studies involve selection criteria, the variances of our summary statistics will be affected. This will be explored below, and it will be shown that the variances for CHANGE and ANCOVA may increase quite substantially. However, there are remedies in terms of additional pre-entry evaluations not underlying the selection.

The underlying reason causing regression to the mean is within-subject variability, and this is present in any clinical trial. This class of variation may potentially consist of many different types of variance components. To facilitate the exposition to follow, the term within-subject variability will be divided into two distinct sub-components, called extraneous respectively intrinsic within-subject variability, to be explained below.

Let us assume that a quantitative observation over time in each subject has a true underlying mean level (assumed constant over the time-period of interest, apart from a possible treatment effect), and that the true (momentarily) level in the subject varies over time around this mean depending on sources of variation, labelled intrinsic within-subject variability, like; time of the day, time of the year, food intake, mood, concurrent diseases, amount of sleep, and so on.

In principle one could subdivide this further into systematic and non-systematic intrinsic within-subject variability depending on whether the variation related to a specific source is caused by a deliberate change of its level (like time of the day, morning versus afternoon) or whether it is due to an unintended change. However, we will not pursue this subdivision.

So far we have conceptualized a subject's true measurement at the time of recording. Additional variability in measurement, however, will almost always be present, caused by, for instance; imprecise measuring devices, errors when reading off, typing errors, inter-rater variability, and laboratory technique. These sources of variation will henceforth be labelled extraneous within-subject variability, or measurement error (though measurement error sometimes is given a wider definition).

Evidently, it is desirable to decrease both these sources of variation as much as possible, by for instance; more standardized recording devices, double entry of values into a computer, having several evaluators for each patient, taking multiple recordings, and so on.

4.2 EFFECTS ON WITHIN-GROUP COMPARISONS

4.2.1 Comparisons for normal distributions

One way to investigate the effects of regression to the mean for within-group comparisons is to consider a sample of subjects with values exceeding a pre-set cut-off point. Then, consider the distribution of a second measurement in the absence of treatment interventions.

Let us assume a bivariate normal distribution for the two measurements (at screening and post-randomisation), as for example in Gardner and Heady (1973), and Hayes (1988). Initially we will assume that the variances for the two measurements are equal, and when further repeated measurements are considered, that compound symmetry applies.

For a more general treatment allowing for correlated within-subject variation (by decomposing the RTM-effects into parts attributable to correlated within-subject variability respectively measurement error) see Johnson and George (1991).

The mean and variance for our variable of interest, x, in the absence of treatment effects, and before selection criteria are used, are denoted by μ and ω^2, respectively. The variance consists of two main components the between-subject part, σ^2, and the within-subject part, δ^2. The long-term true value of a subject (assumed constant over the time-periods of interest) is denoted by X. Let x_0 be the first measurement, and x_1 a second measurement, taken on the subjects with $x_0 \geq$ a cut-off point k at the first measurement occasion.

Further, if no constraints on x_0 and x_1:

$\rho_{x_0,x_1} = \sigma^2/\omega^2$ is the intra-class correlation coefficient.

$\rho_{X,x_i} = \sigma/\omega$ is the correlation between the long-term true value and the observed value obtained.

$z = \dfrac{k-\mu}{\omega}$, where k is the cut-off point for inclusion.

$\theta = \dfrac{\phi(z)}{1-\Phi(z)}$, where $\phi(z)$ is the pdf of the normal distribution, and $\Phi(z)$ is the corresponding cdf. Finally,

$\lambda = \theta(\theta - z)$.

Under the assumption of a bivariate normal distribution for the two repeated measurements, x_0 and x_1, the following equalities can be shown to hold (James, 1973):

$E[x_0 | x_0 \geq k] = \mu + \omega \cdot \theta$ and

$E[x_1 | x_0 \geq k] = \mu + \rho_{x_0,x_1} \cdot \omega \cdot \theta$. Hence, the expected RTM-effect is

$E[x_0 - x_1 | x_0 \geq k] = (1 - \rho_{x_0,x_1}) \cdot \omega \cdot \theta$.

This last formula reveals that, for a given cut-point, the regression to the mean depends on the size of the within-subject variance relative to the between-subject variance.

Correspondingly, for the variances we get:

$Var[x_0|x_0 \geq k] = \omega^2(1-\lambda)$ and

$Var[x_1|x_0 \geq k] = \omega^2(1-\rho^2_{x_0,x_1} \cdot \lambda)$. Hence the expected increase in variance for x_1 relative to x_0 would be $\omega^2\lambda(1-\rho^2_{x_0,x_1})$. Also

$Cov[x_0, x_1|x_0 \geq k] = \sigma^2(1-\lambda)$.

To illustrate the use of these formulae, consider screening a blood pressure lowering study with a cut-point chosen of 95 mmHg for inclusion (i.e. k=95 mmHg). Suppose the distribution of the screened population is N(90,36+16), where the total variance is given as a sum of the between and within components of variance, 36 and 16, respectively.

Then, the conditional expected value for a patient included in a study following a screening visit will be

$$E[x_0|x_0 \geq 95] = 90 + \sqrt{52} \cdot \phi\left(\frac{5}{\sqrt{52}}\right) \Big/ \left(1 - \Phi\left(\frac{5}{\sqrt{52}}\right)\right) = 90 + 9.27 = 99.27 \text{ mmHg}.$$

In the absence of treatment effects, the conditional expected value at a second measurement occasion will be $E[x_1|x_0 \geq 95] =$ $90 + \rho_{x_0,x_1} \cdot 9.27 = 96.42$ mmHg. Thus, the expected regression to the mean is in this case $E[x_0 - x_1|x_0 \geq 95] = 2.85$mmHg.

The conditional expected variance for a subject fulfilling the entry criteria is 52(1-.7611)=12.42 mmHg, whereas we would expect a variance of 33.03 when re-measuring our sample.

Apart from the direct regression to the mean, resulting in exaggerated treatment effects, other phenomena result. The correlation coefficient between x_0 and x_1 (and the regression coefficient, for a regression of x_1 on x_0) will be attenuated. To see this, suppose two pre-entry measurements have been performed, one for classification purposes (x_{01}), and an additional baseline not underlying the selection (x_{02}). Introducing the notation ρ_{ac} for the correlation between x_{01} and x_1, and ρ_{acw} for the corresponding correlation, between x_{02} and x_1, then the following two equalities can be shown to hold;

$$\rho_{acv} = \rho_{norm} \cdot \sqrt{\frac{1-\lambda}{1-\rho_{norm}^2 \cdot \lambda}}, \qquad \rho_{norm} = \frac{\rho_{acv}}{\sqrt{1-\lambda(1-\rho_{acv}^2)}}$$

Using these formulae it may be seen to which extent the correlation between x_{01} and x_1 is expected to be decreased due to regression to the mean, and hence how much less useful a covariate adjustment is likely to be. For instance, for the example above, we would expect a correlation of 36/52=0.69 between x_{02} and x_1, and a correlation of 0.42 between x_{01} and x_1 .

The approach with double baselines, with different purposes (one for classification, and one to be used in the analysis), was proposed by Ederer (1972). This will be helpful, but it may not totally guarantee avoidance of regression to the mean. In fact, assuming homoscedasticity, $E[x_{02} - x_1 | x_{01} \geq k] = \theta \cdot \omega \cdot (\rho_{x_{01},x_{02}} - \rho_{x_{01},x_1})$, see Davis (1976). Under compound symmetry this expression will equal zero, but when correlations decline with time it is likely to be positive, and then there will still be some regression to the mean around (but usually very small).

However, for practical reasons it will often not be possible to include such an additional baseline in the design. Also, when it is possible, it may be difficult to justify the inclusion of subjects where the measurement for this second baseline disagrees to much with the selection criteria.

A further drawback, caused by selection criteria, is that in spite of the underlying normal distribution for the population, a typical sample arrived at through an inclusion criterion will not be normal (for x_1). It will typically follow some skewed distribution, thus, invalidating the use of procedures based on the normal distribution, for small samples.

A further understanding of the reasons underlying the regression to the mean effect, and the role of the within-subject variation, can be achieved as follows. Assume that the true value (long-term true), and the observed value obtained (as influenced by intrinsic variation and measurement error) at a screening visit, jointly have a bivariate normal distribution (before making use of a selection criterion).

For instance, using the numerical values of the example above,
$N(\mu_X, \mu_{X_1}, \sigma_X^2, \sigma_{X_1}^2, \rho_{X,X_1}) = N(90,90,36,52,0.832)$.

This can be illustrated as in figure 4.2.1, which displays data for 500 hypothetical subjects simulated from the above bivariate normal distribution. The marginal distribution on the horizontal axis follows a N(90,36)-distribution, and all variation is due to between-subject variation. The corresponding marginal distribution for the vertical axis is N(90,36+16). Without the extra within-subject variation, all subjects would fall on the diagonal line, the correlation would be one, and all patients would be correctly included/excluded.

The horizontal and vertical reference lines, positioned at 95 mmHg, divide the subjects into four sub collections. Those falling to the right of the vertical reference line are the subjects we in principle are aiming to randomise, the one's having true dbp's of 95 or above.

The subjects above the horizontal reference line are the one's actually included, since they scored 95 or above at the screening visit. The subjects in the lower-left corner have both true and measured dbp's below 95, and are correctly excluded. These in the lower-right quadrant are unnecessarily excluded, they have true underlying values exceeding 95, but owing to measurement error and/or intrinsic within-subject variability, they scored below 95 on this specific occasion. In the upper-right position we find the correctly included subjects, and finally in the upper-left quadrant the undesiredly included patients. This last mentioned group is the one primarily causing the RTM-effect, at a subsequent visit these subjects are expected to have, on the average, lower values.

A further understanding of the reasons underlying the regression to the mean effect, and the role of the within-subject variation, can be achieved as follows. Assume that the true value (long-term true), and the observed value obtained (as influenced by intrinsic variation and measurement error) at a screening visit, jointly have a bivariate normal distribution (before making use of a selection criterion).

For instance, using the numerical values of the example above,
$N(\mu_X, \mu_{x_i}, \sigma_X^2, \sigma_{x_i}^2, \rho_{X,x_i}) = N(90,90,36,52,0.832)$.

This can be illustrated as in figure 4.2.1, which displays data for 500 hypothetical subjects simulated from the above bivariate normal distribution. The marginal distribution on the horizontal axis follows a N(90,36)-distribution, and all variation is due to between-subject variation. The corresponding marginal distribution for the vertical axis is N(90,36+16). Without the extra within-subject variation, all subjects would fall on the diagonal line, the correlation would be one, and all patients would be correctly included/excluded.

The horizontal and vertical reference lines, positioned at 95 mmHg, divide the subjects into four sub collections. Those falling to the right of the vertical reference line are the subjects we in principle are aiming to randomise, the one's having true dbp's of 95 or above.

The subjects above the horizontal reference line are the one's actually included, since they scored 95 or above at the screening visit. The subjects in the lower-left corner have both true and measured dbp's below 95, and are correctly excluded. These in the lower-right quadrant are unnecessarily excluded, they have true underlying values exceeding 95, but owing to measurement error and/or intrinsic within-subject variability, they scored below 95 on this specific occasion. In the upper-right position we find the correctly included subjects, and finally in the upper-left quadrant the undesiredly included patients. This last mentioned group is the one primarily causing the RTM-effect, at a subsequent visit these subjects are expected to have, on the average, lower values.

Table 4.2.1: Exact probabilities, from the bivariate normal distribution, for the outcome of selected subjects at a screening visit and a subsequent repeated measurement occasion (without treatment effects). Assuming screening from a $N(90, 6^2+4^2)$-distribution, with an entry criteria of 95mmHg.

	Re-measure, untreated										
	070-075	075-080	080-085	085-090	090-095	095-100	100-105	105-110	110-115	115-120	Total
At screening											
095-100	0.01	0.15	1.80	9.48	21.67	21.62	9.43	1.79	0.15	0.01	66.10
100-105	.	0.01	0.15	1.36	5.44	9.43	7.13	2.35	0.34	0.02	26.23
105-110	.	.	0.01	0.08	0.59	1.79	2.35	1.35	0.34	0.04	6.55
110-115	.	.	.	.	0.03	0.15	0.34	0.34	0.15	0.03	1.03
115-120	.	.	.	.	.	0.01	0.02	0.04	0.03	0.01	0.10
Total	0.01	0.16	1.96	10.92	27.73	33.00	19.27	5.87	1.01	0.11	100.0

Figure 4.2.1 : Fivehundred random data points from a bivariate normal distribution, N(90,90,36,52,.832).

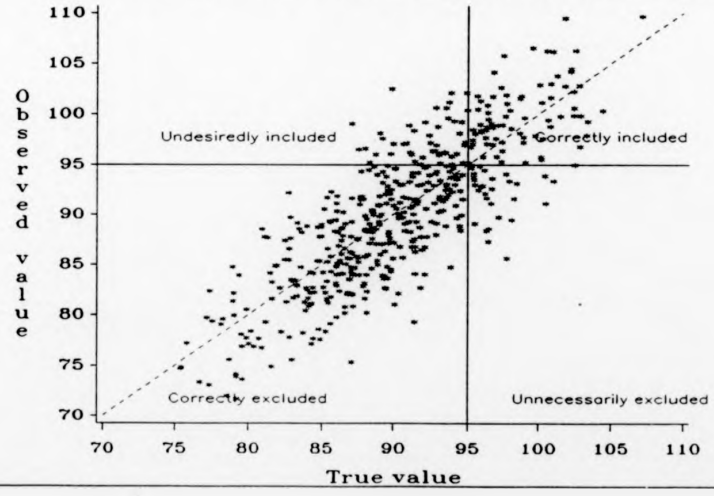


Table 4.2.1 is based on the same underlying distribution as figure 4.2.1, but now the results are analytical, based on the bivariate normal distribution. The table contains only the part of this distribution over the horizontal reference line (i.e. with $x_0 \geq 95$). Further, it gives the exact probabilities (in per cent, adding to 100 for the table) of falling into a grid of squares formed by categorizing the vertical and horizontal axes of figure 4.2.1 into 5 mmHg intervals. For instance, we see from the first row of the table, that we expect 66% of the selected subjects to have observed values (x_0) in the range 95 to 100, of these half are expected to have observed values below 95 when re-measured (x_1), in the absence of treatment effects. From this table we get a feeling for how the distribution of measurements are expected to change from screening to re-measure (for this example) simply because of regression to the mean.

For any given example it is possible to calculate the expected proportions of subjects falling into the four categories. To achieve this, we have to refer to the formula for the bivariate normal. To get the proportion (from the total bivariate normal distribution) of undesiredly included patients above, we have to calculate the following integral;

$$\int_{95}^{\infty}\int_{-\infty}^{95} \frac{1}{2\pi\sigma_X\sigma_{x_0}\sqrt{1-\rho_{X,x_0}^2}} \cdot \exp\left\{-\frac{1}{2\left(1-\rho_{X,x_0}^2\right)}\left[\frac{(X-\mu_X)^2}{\sigma_X^2} - 2\cdot\rho_{X,x_0}\cdot\frac{(X-\mu_X)(x_0-\mu_{x_0})}{\sigma_X\sigma_{x_0}} + \frac{(x_0-\mu_{x_0})^2}{\sigma_{x_0}^2}\right]\right\} dX\,dx_0$$

The probabilities for the remaining three categories are obtained by obvious changes in the ranges of integration above. This can be accomplished using the formulae given in the Handbook of Mathematical Functions (1970). More easily, though, the function PROBBNRM in SAS (1992, available from versions 6.07 and onwards) can be used.

Continuing with our example, the expected proportions are as follows;

Correctly included	15.2%
Correctly excluded	70.6%
Unnecessarily excluded	5.0%
Undesiredly included	9.2%

I.e., over a third of our selected sample are from outside the target population, and about a fourth of the screened patients actually in our target group were unnecessarily excluded.

There are some general rules governing the regression to the mean phenomenon, which can be deduced from the formula for the RTM-effect $E[x_0 - x_1 | x_0 \geq k] = (1 - \rho_{x_0,x_1}) \cdot \omega \cdot \theta$. They can be categorized as follows;

* The larger the within-subject variability is compared to the between-subject variability, i.e. the smaller the correlation, ρ_{x_0,x_1}, the worse the RTM-effect.

* The more extreme the cut-point, the larger the θ (which goes from 0 when k=-∞ to ∞ when k=∞), the worse the RTM-effect.

Also, There is a strong dependence between the RTM-effect, and the proportion of undesiredly included patients.

What can be done to decrease these undesired effects ? Primarily, we should try to decrease the within-subject variance. This may be done by a more precise measuring technique and/or by taking repeated pre-entry measurements.

We now investigate the gains that can be made by using repeated measurements at the screening visit. Following Gardner and Heady (1973), we assume independent random variation around the true underlying value for each subject (i.e. equicorrelation). For a more general treatment, see Johnson and George (1991). Then, with p pre-entry measurements, using the mean of these when classifying the subjects, the summary statistic used will follow a $N\left(\mu, \sigma^2 + \frac{\delta^2}{p}\right)$-distribution. Returning once more to our hypothetical example, the joint bivariate normal distribution of X (the true underlying mean) and $\bar{x}_{ij}^{pre}$ (the observed pre-entry mean) will be $N\left(90, 90, 36, 36 + \frac{16}{p}, \frac{36}{36 + 16/p}\right)$.

For instance, having 3 recordings pre-entry, instead of one, this would decrease our expected RTM-effect from 2.85 to 1.12.

Also, relative to basing the selection on one pre-entry evaluation, the ratio (correctly included)/(correctly included + undesiredly included) increases from .62 to .72, and the ratio (correctly included)/(correctly included + unnecessarily excluded) increases from .75 to .87. Further, the misclassified subjects will be nearer, on average, to the cut-off point if we use multiple pre-treatment recordings.

A further way to decrease the RTM-effect, is to indirectly change the position of the cut-point relative to the underlying distribution. This might be achieved by a more careful definition of the population one is screening from. Thus, avoid investigating subjects not ill enough to merit their inclusion in the study.

Figures 4.2.2 and 4.2.3 illustrate, respectively, the dependencies of RTM-effects on the position of the cut-off point relative to the underlying distribution, and the remedies possible by the use of multiple pre-entry measurements.

Figure 4.2.2 : Expected % of correctly included, undesiredly included, unnecessarily excluded, and correctly excluded subjects. Depending on the position of the cut-point (k) relative to the underlying distribution, here assumed N(90,36+16)

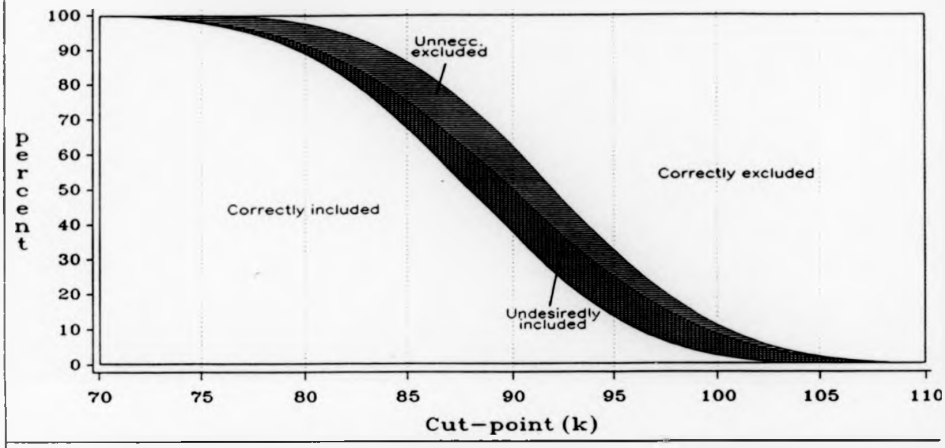
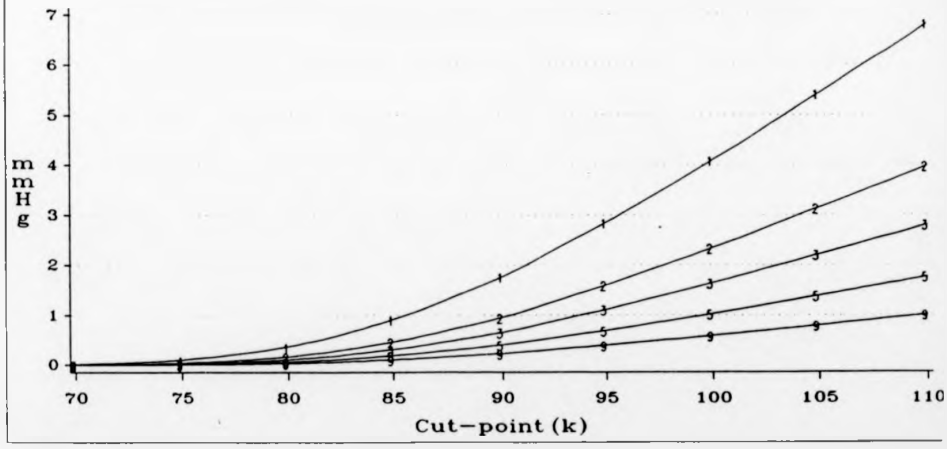


Figure 4.2.3 : RTM-effect (in mmHg) as a function of the position of the cut-point relative to the underlying distribution, assumed N(90,36+16). Different curves depending on the number of pre-entry measurements (1,2,3,5 or 9). Assuming compound symmetry



4.2.2 Regression or digression

In the preceding subsection all results given were based on the assumption of equal variances at the screening and post-treatment visits (and on assuming sampling from a bivariate normal distribution). Under these circumstances we found that the RTM-effect was expected to be $\omega \cdot \theta\left(1-\rho_{x_0,x_1}\right)$. All the three factors involved in this expression are strictly non-negative, and apart from degenerate null-cases (e.g. k=-∞), regression to the mean will indeed take place. However, this is not necessarily the case when the variances for x_0 and x_1 are allowed to differ.

Relaxing the assumption of homoscedasticity, and using the same notation as earlier in this section, the main results for the effects caused by the use of a selection criterion, generalize to:

$$E\left[x_0 \middle| x_0 \geq k\right] = \mu + \omega_{x_0} \cdot \theta$$

$$E\left[x_1 \middle| x_0 \geq k\right] = \mu + \beta_{x_0,x_1} \cdot \omega_{x_0} \cdot \theta, \qquad \text{where } \beta_{x_0,x_1} = \rho_{x_0,x_1} \cdot \frac{\sigma_{x_1}}{\sigma_{x_0}}$$

$$E\left[x_0 - x_1 \middle| x_0 \geq k\right] = \omega_{x_0} \cdot \theta \cdot \left(1-\beta_{x_0,x_1}\right)$$

$$Var\left[x_0 \middle| x_0 \geq k\right] = \omega_{x_0}^2 \left(1-\lambda\right)$$

$$Var\left[x_1 \middle| x_0 \geq k\right] = \omega_{x_1}^2 \left(1-\rho_{x_0,x_1}^2 \cdot \lambda\right)$$

Also, when conditioning on a specific value:

$$E\left[x_1 \middle| x_0 = k\right] = \mu + \beta_{x_0,x_1}\left(k-\mu\right)$$

$$E\left[x_0 - x_1 \middle| x_0 = k\right] = \beta_{x_0,x_1}\left(k-\mu\right)$$

$$Var\left[x_1 \middle| x_0 = k\right] = \omega_{x_1}^2 \left(1-\rho_{x_0,x_1}^2\right)$$

It may be seen that, whether using a selection criterion or conditioning on a specific value, the RTM-effect is no longer restricted to be non-negative. Specifically:

$\rho_{x_0 \cdot x_1} \cdot \omega_{x_1} < \omega_{x_0} \Rightarrow$ Regression to the mean

$\rho_{x_0 \cdot x_1} \cdot \omega_{x_1} > \omega_{x_0} \Rightarrow$ Digression from the mean

The interesting implication is that regression to the mean will not necessarily take place. A dual phenomena, henceforth termed "digression from the mean" might actually take place. The specific factor deciding whether regression or digression will be expected is whether the regression coefficient, $\beta_{x_0 \cdot x_1}$, is smaller or larger than one.

In most but not all applications the regression coefficient (of x_1 on x_0) will be smaller than one. In Egger et al (1985), the estimated regression coefficients on seven variables in four different arthritis studies from a total of ten different treatment arms were reported. Out of the seventy regression coefficients two exceeded unity, being 1.21 (based on 61 subjects) and 1.09 (based on 68 subjects), both reported from the variable "grip strength".

In summary, when $\omega_{x_1} > \omega_{x_0}$, regression to the mean will not necessarily take place, if $\rho_{x_0 \cdot x_1} \cdot \omega_{x_1} > \omega_{x_0}$ we will actually experience digression from the mean. When $\rho_{x_0 \cdot x_1} \cdot \omega_{x_1} = \omega_{x_0}$ the RTM-effect will be zero. In practice, digression from the mean is unlikely to occur, but it is worth being aware of the possibility.

4.2.3 Some results for general distributions

Most published research relating to regression to the mean assume normal distributions for the derivation of results. There are a few exceptions, however. For instance, Das and Mulder (1983), suggested a change in terminology, from 'regression to the mean' to 'regression to the mode'.

This suggestion was based on a general formula for the regression effect for a given measurement x_0 when a subsequent measurement x_1 is observed. They used the following assumptions; the 'true' values are arbitrarily distributed, but the measurement errors are normally distributed with constant variance, written to simplify the below formulae as $(1-\rho_{x_0,x_1})\sigma^2_{x_0}$. The regression effect conditional on a given value, x_0=k, is defined as $E[x_0 - x_1 | x_0 = k]$. and Das and Mulder showed that their model yields the formula;

$$E[x_0 - x_1 | x_0 = k] = -(1-\rho_{x_0,x_1}) \cdot \sigma^2_{x_0} \cdot \frac{d}{dx_0} \ln[g(x_0)],$$

where $g(x_0)$ is the density function of the true measurements, $\sigma^2_{x_0}$ is their variance, and ρ_{x_0,x_1} is the correlation between x_0 and x_1.

As they pointed out, it is obvious from this formula, that for unimodal distributions, the regression is to the mode and not to the mean (since $\frac{d}{dx_0}\ln[g(x_0)]$ is zero at the mode of the distribution).

However, these results are based on conditioning on x_0=k, not on $x_0 \geq k$, which is the usual situation when selecting for inclusion.

In this case the regression effect may be defined as $E[x_0 - x_1 | x_0 \geq k]$, and Das and Mulder found that;

$E[x_0 - x_1 | x_0 \geq k] = \frac{(1-\rho_{x_0,x_1}) \cdot \sigma^2_{x_0} \cdot g(k)}{1-G(k)}$, where G(k) is the distribution function corresponding to g(k).

As pointed out by Senn (1990), this expression is always positive and thus justifies the term 'regression to the mean'. Senn also showed that, if the measurement errors have some other type of distribution than the normal, even when one is conditioning on a specific value, the regression might be to the mean and not to the mode.

Moving now to the variance, it has been shown earlier that, for the normal distribution, the variance decreases for all types of truncation. That is, $Var[x_0] \geq Var[x_0 | x_0 \geq k]$. To explore whether this finding carries over to some other common distributions, a simulation study has been performed. The results of this study are summarized in table 4.2.2 below.

For each distribution investigated, table 4.2.2 lists the choice of parameters, the cut-point (k) for selection, the number of subjects screened (10000 in each instance), the mean and variance before selection, the number of selected subjects, and the mean and variance after selection. These simulations indicate that variances decrease after truncation for most reasonably symmetric unimodal distributions. Only for one of the reported distributions, the chi-squared with one degree of freedom, did there appear to be an increased variance for the truncated case. For the exponential distribution it is well known that the variance is unaffected when it is truncated (the memory-less property).

It should be emphasized that these results are of a preliminary nature. If feasible, more general analytical results would be more valuable.

Table 4.2.2 : Observed means and variances for all subjects screened and for all patients included, for some distributions with varying degrees of truncation.

Distribution	Cut-point (k)	Subjects screened	Subjects selected	Mean (untruncated)	Mean (truncated)	Variance (untruncated)	Variance (truncated)
Normal(0,1)	0	10000	5021	.004	.798	.993	.358
	0.5	10000	3099	.002	1.140	.998	.266
	1.645	10000	491	.017	2.094	.995	.162
Exp(1)	1	10000	3715	1.01	2.00	1.02	1.02
	2	10000	1378	1.01	3.00	1.02	1.06
Gamma(3,3)	1	10000	4170	0.99	1.53	0.33	0.22
(8,2)	15	10000	4358	15.8	25.6	123	96
Poisson(25)	25	10000	4384	24.9	29.4	24.7	9.8
	30	10000	1388	25.1	33.3	24.6	5.7
Weibull(10,.1)	10	10000	3707	9.5	10.6	1.30	0.21
(2,1/75)	50	10000	6431	66.2	85.3	1178	724
Chi-squared$_1$	0.5	10000	4807	1.00	1.92	2.02	2.56
	1	10000	3105	0.99	2.53	1.99	2.79
	2	10000	1521	0.99	3.62	1.95	3.09
Chi-squared$_2$	2	10000	3695	2.01	4.03	4.09	4.04
3	3	10000	3907	2.99	5.38	5.98	4.91

4.3 EFFECTS ON BETWEEN-GROUP COMPARISONS

4.3.1 Effects on variances caused by inclusion criteria

The effect that the regression to the mean has on the precision for the estimated difference in treatment effects, when basing the analysis on either; ANCOVA, CHANGE or POST, will be investigated in this subsection. For simplicity, assuming one post-treatment measurement only.

In particular, the variance achieved for the three different methods of analysis when using the screening visit as baseline will be compared with what we achieve if we use the measurements from another pre-entry visit for the subjects included in the study.

Utilizing the notation and assumptions introduced earlier in this chapter, and assuming $\rho_{x_{01},x_1} = \rho_{x_{02},x_1}$, the covariance matrix for one pre-entry and one post-treatment measurement, when using the screening visit as baseline (i.e. based on x_{01}), is;

$$\Sigma_{screen} = \begin{bmatrix} \omega^2(1-\lambda) & \\ \sigma^2(1-\lambda) & \omega^2(1-\rho^2\lambda) \end{bmatrix}$$

With new baseline measurements (x_{02}) for selected subjects, we obtain;

$$\Sigma_{new} = \begin{bmatrix} \omega^2(1-\rho^2\lambda) & \\ \sigma^2(1-\rho^2\lambda) & \omega^2(1-\rho^2\lambda) \end{bmatrix}$$

In most instances there will not be a second pre-entry measurement, so Σ_{screen} is more relevant. However, we now show that if one can incorporate an additional pre-entry visit in the design, much is gained as regards precision.

As observed earlier, the expected regression to the mean is $\omega \cdot \theta \cdot (1-\rho)$ for both treatment groups, and the decrease in the variance for the measurement at the screening visit is $\omega^2 \cdot \lambda \cdot (1-\rho^2)$ as compared to the variance resulting from a new measurement occasion.

Moving on to the three different methods of analysis, and substituting the relevant components of the covariance matrices given above into the general variance formulae for ANCOVA, CHANGE and POST, the following equalities result.

Variances when using the screening visit as baseline:

$$Var_{x_{01}}[POST] \quad \propto \omega^2(1-\rho^2\lambda)$$

$$Var_{x_{01}}[CHANGE] \propto \omega^2(1-\rho^2\lambda) - (2\sigma^2 - \omega^2)\cdot(1-\lambda)$$

$$Var_{x_{01}}[ANCOVA] \propto \omega^2(1-\rho^2\lambda) - \frac{\sigma^4}{\omega^2}\cdot(1-\lambda)$$

Variances when using measurements from a new pre-entry visit as baseline:

$$Var_{x_{02}}[POST] \quad \propto \omega^2(1-\rho^2\lambda)$$

$$Var_{x_{02}}[CHANGE] \propto \omega^2(1-\rho^2\lambda)\cdot 2\cdot\left(1-\frac{\sigma^2}{\omega^2}\right)$$

$$Var_{x_{02}}[ANCOVA] \propto \omega^2(1-\rho^2\lambda)\cdot\left(1-\frac{\sigma^4}{\omega^4}\right)$$

Differences in the above variances; using the screening visit minus using a new pre-entry visit.

$$Var_{x_{01}}[POST] - Var_{x_{02}}[POST] \qquad \propto 0$$

$$Var_{x_{01}}[CHANGE] - Var_{x_{02}}[CHANGE] \propto \lambda \cdot (1-\rho^2) \cdot (2\sigma^2 - \omega^2)$$

$$Var_{x_{01}}[ANCOVA] - Var_{x_{02}}[ANCOVA] \propto \lambda \cdot (1-\rho^2) \cdot \frac{\sigma^4}{\omega^2}$$

λ, $(\lambda = \theta(\theta - z)$, where $\theta = \frac{\phi(z)}{1-\Phi(z)}$, and $z = \frac{k-\mu}{\omega}$), is always positive, going from zero when $\frac{k - \mu}{\sigma} = -\infty$ to one when $\frac{k - \mu}{\sigma} = +\infty$.

Hence, for ANCOVA the variance will always be smaller if one is using a new pre-entry visit as baseline instead of the screening visit.

For CHANGE, this depends on the sign of $(2\sigma^2 - \omega^2) = (\sigma^2 - \delta^2)$, i.e. when $\rho_{x_0, x_1} = \sigma^2/\omega^2 \geq .5$ (for untruncated variables) the variance will be larger if we use the screening visit as baseline.

Now, the question is, in general, is the variance reduction for ANCOVA sufficient to justify the effort of an extra baseline measure? Would we gain just as much by an extra post-treatment measure? Using the parameterization of the covariance matrix above, and assuming compound symmetry, the ANCOVA variance with two post measures and one pre (the screening visit), is proportional to $\frac{1}{2}(\sigma^2 + \omega^2)(1 - \rho^2\lambda) - \frac{\sigma^4}{\omega^2}(1-\lambda)$. This variance will almost always exceed the variance we get with an extra baseline, as seen in the example below.

Example :

Again assume data is from a N(90,36+16) distribution, and that we have 95 as our lower limit for including subjects into our study. We will, thus, have the following numerical values for the parameters of interest;

$$\sigma^2 = 36,\ \delta^2 = 16,\ \omega^2 = 52,\ \rho = .693,\ z = .693,\ \theta = 1.286,\ \lambda = .761$$

$$\Sigma_{screen} = \begin{bmatrix} 12.4 & \\ 8.6 & 33.0 \end{bmatrix}, \quad \Sigma_{new} = \begin{bmatrix} 33.0 & \\ 22.9 & 33.0 \end{bmatrix}$$

Variances when using the screening visit as baseline:

Var[POST] ∝33.0
Var[CHANGE]∝28.2
Var[ANCOVA]∝27.1

Variances when using a new pre-entry visit as baseline:

Var[POST] ∝33.0
Var[CHANGE]∝20.3
Var[ANCOVA]∝17.2

The drop in ANCOVA variance when an extra baseline is provided for represents a major benefit. Adding instead an extra post measure gives Var[ANCOVA]∝22.0 . Thus, when feasible, provision of an extra baseline, giving unrestricted knowledge of the subjects pre-entry level, may be well worth-while.

4.4 SUMMARY AND DISCUSSION

Regression to the mean (RTM) causes undesired effects in many clinical trials. The negative consequences for within-group comparisons include biased estimates of treatment effects, exaggerated claims of effects on sub-groups (e.g. the subjects being worst off at baseline), and spurious correlations between initial values and changes. These effects are well known in theory, though, not always recognized in practice.

The effects of regression to the mean on between-group comparisons are less well appreciated. When pre-entry means differ between groups, and inferior approaches to analysis are used (i.e. not ANCOVA), biased estimates of treatment effects result. Also, when selection criteria are used, the ANCOVA variance may increase substantially, since the baseline will be of reduced value for the purpose of covariate adjustment.

In this chapter we have reviewed the existing results for within-group comparisons, and practical suggestions have been given for reducing the RTM-effects in terms of repeated baselines. Most published research have been based on normal distribution theory. We have included some more general results regarding the effects of RTM, both on means for within-group changes and their variances.

Also, some results on between-group comparisons were given, especially relating to the consequences of selection criteria on the variances of our mean summary statistics. One helpful solution is to add an extra unrestricted baseline (i.e. not under the selection criteria). Major benefits in variance reduction may be gained with such an additional baseline. However practicalities may often prevent this possibility.

So far it has been assumed that the selection criterion is based only on the level of the intended dependent variable at one or, as an average, at several pre-entry visits. In practice, the selection can be more complicated involving a combination of several prognostic factors, which are correlated with each other, and with the dependent variable.

An example in hypertension trials is to have selection criteria related to both diastolic and systolic blood pressure. Another example might be that the dependent variable, for instance time until end of an exercise test (e.g. on a treadmill), is not allowed to vary more than something like 15% between two repeated pre-entry measurements for a given subject.

The practical consequences of any selection criteria involving the variable of main interest for the subsequent analysis is to limit the value of the pre-entry measurement as a covariate in a statistical model. When feasible, it is recommended to provide for a second baseline, or to base the selection on other grounds than the pre-entry level of the intended dependent variable. However, the latter option is heavily dependent on clinical/practical circumstances.

5 OPTIMAL LINEAR SUMMARY STATISTICS

So far attention has been confined to well known mean summary statistics, such as ANCOVA, CHANGE and POST. Depending on the expected response profiles (mean treatment curves) over time, and the hypothesis of interest (e.g. is the objective to assess differences in rate of change, overall effects over specific time intervals, etc.), there may often be some other logical choice of summary statistic for each subject. I.e. for steadily diverging mean response curves we might choose to analyse the data using each patient's regression coefficient (SLOPE) as the summary statistic in anticipation that this will increase statistical power compared with mean summary statistics. However, is this necessarily the case ? For instance, the mean curves for two treatment groups might separate quite quickly in the beginning of the study, and continue to separate, but more slowly during the later phases. Such considerations raise a number of important issues:

* Will ANCOVA perform better than SLOPE ?

* Is there an optimal summary statistic, and in that case what will it look like ?

* How much more efficient is one summary statistic compared to another in a given setting ?

These are the questions that will be tackled in this section, and hopefully resolved, through the derivation of "the optimal linear summary statistic", and through the use of the concept of asymptotic relative efficiency (Pitman efficiency).

5.1 ASYMPTOTIC RELATIVE EFFICIENCY FOR LINEAR SUMMARY STATISTICS

A linear summary statistic is any linear combination of a subject's measurements, i.e. $S=\mathbf{c}'\mathbf{Y}$, where $\mathbf{c}$ is a vector of weights, and $\mathbf{Y}$ is the vector of responses (see Dawson and Lagakos, 1991). This class of summary statistics incorporates almost all of the one's considered in this dissertation, for instance, SLOPE, POST, CHANGE, AUC and ANCOVA (when assuming a known covariance matrix).

Examples of summary statistics outside this subclass includes t_{max} (the time to reach maximum) and c_{max} (the maximal concentration/level).

If μ_i denotes the true mean vector of $\mathbf{Y}$ in group i, and Σ is the covariance matrix of $\mathbf{Y}$, assumed common to both treatment groups, then the asymptotic relative efficiency (ARE) of a test based on $S_1=\mathbf{c}_1'\mathbf{Y}$ versus one based on $S_2=\mathbf{c}_2'\mathbf{Y}$ is equal to the ratio of their squared non-centrality parameters (Cox and Hinkley, 1974), and can be written as;

$$ARE(S_1:S_2)=\frac{(c_1'[\mu_1-\mu_2])^2}{c_1'\Sigma c_1}\Big/\frac{(c_2'[\mu_1-\mu_2])^2}{c_2'\Sigma c_2}.$$

The appropriateness of this equality follows from the asymptotic normality of linear summary statistics, and is an immediate generalization of a result concerning large-sample power functions for maximum likelihood-ratio tests given on page 337 of Cox and Hinkley (1974), see also Lehmann (1993). To evaluate this expression numerically, we must specify $\mathbf{c}_1$ and $\mathbf{c}_2$, and assume knowledge of μ_i and Σ.

The ARE is a number that reflects the relative power of one statistical test to another. For example, if the ARE of S_2 to S_1 in a particular setting is 0.75, this means that using S_1 would be more efficient than S_2, in that, asymptotically, only about 75% as many subjects would be needed for it to have the same power. Thus, ARE's that are close to one imply that the two summary statistics have similar power against the treatment effect being considered, while values far from one imply the opposite.

There is another way of utilizing the ARE formula. Having decided on which summary statistic to use, we may contrast two different designs, we interpret the ARE as the relative design efficiency under the given test. For example, having decided to use ANCOVA, and having specified Σ and $[\mu_1-\mu_2]$ (not necessarily constant over time), we may compare ARE's between designs consisting of different number of pre and post-treatment measurements.

The variance formulae given earlier are only (strictly) valid to use for this purpose when $\left[\mu_1 - \mu_2\right]$ is constant over time.

Although the concept of ARE is derived from asymptotic theory, it is also closely related to variance ratios between the summary statistics being compared. Thus, the ARE's indicates the relative precision of the summary statistics in estimating model parameters.

For example, the ARE of SLOPE to POST is simply the ratio of the variance of POST to SLOPE, after the former has been standardized to have the same expected value as the latter (i.e. the measurements have been scaled such that $E\left[\bar{\bar{x}}_{A.}^{post} - \bar{\bar{x}}_{B.}^{post}\right] = E\left[\overline{SLOPE}_{A.} - \overline{SLOPE}_{B.}\right]$), see Dawson and Lagakos (1991).

In ensuing sections the ARE's between various summary statistics have been computed, under different assumptions regarding the differences between the mean treatment vectors and the underlying covariance structure. However, first we derive "the optimal linear summary statistic".

5.2 THE OPTIMAL LINEAR SUMMARY STATISTIC

For any choice of; repeated measures design, vector of assumed mean treatment differences over time, and assumed underlying covariance structure, it is possible, extending the results of O'Brien (1984) and of Pocock, Geller and Tsiatis (1987), to derive an "optimal linear summary statistic".

The "optimal linear summary statistic", defined below, will maximize the power to detect a treatment effect under the assumptions chosen. I.e., no other linearly weighted combination of the outcomes for each subject will give a more powerful test statistic.

Theorem:

Letting δ' ($=[\mu_A-\mu_B]'$) be the known vector of true mean treatment differences, and Σ be the covariance matrix (assumed known, and identical between treatments), the optimal weights, c', for a linear summary statistic are obtained from $\delta'\Sigma^{-1}$, as c' proportional to $\delta'\Sigma^{-1}$.

This result follows from the extended Cauchy-Schwarz inequality, which states: Let β and δ be any two p+r (where p and r are the number of pre and post-treatment measurements, respectively, see chapter 2) vectors and let Σ be a positive definite (p+r)(p+r) matrix. Then $(\beta'\delta)^2 \leq (\beta'\Sigma\beta)(\delta'\Sigma^{-1}\delta)$ with equality if and only if $\beta=k\Sigma^{-1}\delta$ (or $\delta=k\Sigma\beta$) for some constant k. A proof is given on page 64 in Johnson and Wichern (1988).

Then, for an arbitrary non zero p+r vector c,

$\max_{c\neq 0} \frac{(c'\delta)^2}{c'\Sigma c} = \delta'\Sigma^{-1}\delta$ with the maximum attained when $c'=k\delta'\Sigma^{-1}$ for any constant $k\neq 0$.

This optimality theorem may be viewed either as an extension of O'Brien's generalized least squares procedure, or as an application of results from discriminant analysis, and the use of Fisher's linear discriminant function (see, for example, Chatfield and Collins, 1980), adapted to repeated measures designs.

Indeed, these coefficients were derived by Fisher (1936) when searching for a linear combination of a set of variables which had maximum between-group difference relative to its within-group standard deviation.

The weights, $c'=\delta'\Sigma^{-1}$, are scale-invariant but not shift-invariant, i.e., multiplying any vector of weights, c_i, by a constant will not affect the ARE derivations, but adding a constant to all the weights will change the performance of the summary statistic (unless the summary statistic has equal weights, i.e. POST in the absence of baselines).

To exemplify the scale-invariance, consider a design with 1 pre and 3 post-treatment visits. In this situation the weights for CHANGE would normally be written as $\mathbf{c}' = \left[-1 \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}\right]$. Pre-multiplying this vector with, for example, 3, would change the vector to $\mathbf{c}' = [-3 \quad 1 \quad 1 \quad 1]$, but it would not change the outcome of our analysis, only the units for our estimate of the difference in treatment effect.

That CHANGE is not shift-invariant is obvious from the appearance of the vector $\mathbf{c}' = \left[2 \quad 3\frac{1}{3} \quad 3\frac{1}{3} \quad 3\frac{1}{3}\right]$, resulting from the addition of 3 to all the weights in the vector. We would then be estimating something very different from a mean change.

To allow for this arbitrary scaling, it is convenient to scale all the summary statistics in a consistent fashion: we henceforth set the sum of the weights for the post-treatment visits equal to one. For example, with p=1 and r=5 visits pre and post-randomisation, and using ANCOVA, we will use the weights

$$\mathbf{c}' = \left[-\beta \quad \frac{1}{5} \quad \frac{1}{5} \quad \frac{1}{5} \quad \frac{1}{5} \quad \frac{1}{5}\right].$$

An alternative approach for a consistent scaling would be to present the vector $\mathbf{c}'$ in its orthonormalized form, i.e. scaled such that $\mathbf{c}'\mathbf{c}=1$.

For any given set of data, if we substitute $\mathbf{d}'$ $(=[\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B]')$ for $\boldsymbol{\delta}'$ and S^{-1} for Σ^{-1}, and use $\mathbf{c}' = \mathbf{d}'S^{-1}$ as the weights for the summary statistic. Than, this will form the basis for a test that maximizes the discrimination between two treatment groups in that no larger t-statistic can be achieved using any other linear summary statistic.

Letting the observed data decide the weights is not a valid technique if we want to make a confirmatory analysis, this method would produce much inflated type I error rates.

It is valid, though, to let data from one study indicate what kind of summary statistics will be likely to be most powerful in a forthcoming study. Further, if a prior hypothesis is hopelessly wrong, to be realistic, one may have to alter the approach to analysis. For instance changing from CHANGE to SLOPE if an hypothesized stable treatment difference over time turned out to be a linear divergence between the treatment curves over the course of the study. If so, however, one should be cautious with the conclusions, and a further confirmatory study would usually be recommended.

Furthermore, it is interesting to see what kind of summary statistics will be most powerful under various specific types of mean treatment difference profiles, and under some plausible choices of covariance structures.

Also of interest is the versatility of some common linear summary statistics to be powerful under various models, and the differences in ARE's between the summary statistics under plausible assumptions for a study at the design stage.

5.3 ANALYSIS OF RATE OF CHANGE

Mean summary statistics aside, summary statistics dealing with the analysis of rate of change are probably the most common. In particular analysis using each individual subjects linear regression coefficient (SLOPE) as a summary statistic entertains a wide usage, dating back, at least, to Wishart (1938).

The use of SLOPE when mean treatment curves diverge approximately linearly with time has several appealing features. These include ease of calculation, and ease of interpretation and communication. One might also think it has comparatively high power. However, since observations on a given subject are intercorrelated, least squares is not optimal, but merely convenient (Potthoff and Roy, 1964). The degree of sub-optimality that the within-subject dependencies impose on SLOPE is clarified below.

In this section we will review some of the more relevant results, firstly, relating to the precision of SLOPE depending on the frequency and timing of measurements, and, secondly, relating to some alternatives to SLOPE, one of them indeed achieving optimality in terms of statistical power for a linearly diverging alternative hypothesis.

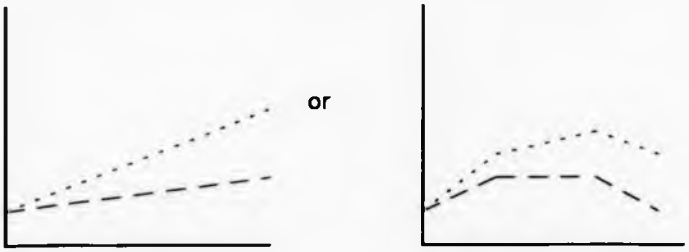
5.3.1 Analysis using SLOPE

Based on the model for repeated measurements given in section 2.1, and assuming that the true responses from each subject increases (decreases) linearly with time, we will adopt the simple model:

$$x_{ijk} = \mu + \beta_i(t_k - t_0) + \varepsilon_{ijk} ,$$

where i indexes treatment group (i=A or B), j indexes subject within group ($j=1,..,n_i$), and k indexes the repeated measurements ($k=0,..,r$). The overall baseline mean is denoted by μ, β_i is the (assumed common) slope for treatment group i, and ε_{ijk} is the residual variation around the slope, which is interdependent within subject.

However, for between-group comparisons we need not restrict ourselves to situations where we have linearity <u>within</u> groups. All that matters for the power of the subsequent statistical test is that we have linear divergence <u>between</u> the two mean treatment-group curves. Thus, we might substitute μ_k for μ in the model given above, where μ_k represents the underlying true mean response at time k disregarding treatment effects. That is, this section is concerned both with trials with linearity within groups, as represented by the left-hand figure below, and more generally in trials where we have linearity only for the difference between the two curves, as for the right-hand figure below.



If, for the latter type of trial, we were to have knowledge of the true μ_k's, and substituted these from the observed measurements, we would be back to linearity also within groups. Hence, the treatment effect gives rise to a continual linear increase (decrease) in mean response over time relative to the true underlying μ_k's.

Based on the above model (with or without fixed μ), for a repeated measures design where one baseline and r post-treatment visits are performed at the time-points; $t_0, t_1, \ldots, t_r$, the summary statistic for subject j in group i, SLOPE_{ij}, may be calculated as;

$$\text{SLOPE}_{ij} = \frac{\sum_{k=0}^{r}(t_k - \bar{t}_\cdot)\cdot y_{ijk}}{\sum_{k=0}^{r}(t_k - \bar{t}_\cdot)^2}$$

Then, using this summary statistic as dependent variable for each subject, standard two-sample t-tests may be performed between the two groups based on the overall treatment difference in mean SLOPE's, $\overline{SLOPE_A} - \overline{SLOPE_B}$.

With access to more than one pre-entry measurement one should preferably take the average of these and use this as Y_0, rather than including them as separate measurements. Needless to say, SLOPE_{ij} estimates the mean rate of change in the dependent variable per time unit for subject j in group i.

An explicit variance formula for SLOPE, for a general covariance structure, and a general choice of time-intervals between measurements, will be complicated. However, given the

weights used in the derivation of each $SLOPE_{ij}$, e.g. $\frac{(t_k - \bar{t}_.)}{\sum_{i=0}^{r}(t_i - \bar{t}_.)^2} = c_k$

for measurement k, and denoting the vector of weights $[c_0, c_1, .., c_r] = \mathbf{c}'$, the general variance formula for summary statistics given in section 1.6 may be used, i.e.

$$\text{Var}[\overline{SLOPE_A} - \overline{SLOPE_B}] = \left(\frac{1}{n_A} + \frac{1}{n_B}\right)\mathbf{c}'\Sigma\mathbf{c}$$

A common approach is to adopt a model where both intercept and SLOPE are treated as between-subject random effects, and assuming equal variances around the lines at all time-points. Then the variance for SLOPE is given by (see Roe and Korn, 1993) :

$$Var\left[\overline{SLOPE_A} - \overline{SLOPE_B}\right] = \left(\frac{1}{n_A} + \frac{1}{n_B}\right) \cdot \left(\frac{\tau^2}{\sum_{k=0}^{r}(t_k - \bar{t}_.)^2} + \eta^2 \right)$$, where η^2 is

the between-subject variance in true underlying SLOPE's, and τ^2 is the within-subject variance around the regression line. This approach will not be pursued here.

In the remaining parts of this subsection, we will for simplicity assume compound symmetry (whereby we implicitly assume intercepts to be random and slopes to be fixed within groups). More general results, and comparisons with other approaches, will appear later in this chapter.

Given that we believe in linear relationships with time, and that we intend to use SLOPE for the analysis, three options for increasing the precision of the individual SLOPE's will be described:

A) Increasing r when time-intervals between measurements are fixed.

B) Increasing the study duration for a fixed r.

C) Changing the distribution over time for a given number of measurements over a fixed time period.

Assuming compound symmetry and a variance of τ^2 around each separate regression line (this is the within-subject part of the total variance σ^2), and further, equally spaced time-intervals of unit length, the variance formula for an individual $SLOPE_{ij}$ estimate, with r measurements in total, simplifies to (see Schlesselman, 1973):

$$Var\left[\hat{SLOPE}_{ij}\right] = \frac{12\tau^2}{(r^2-1)\cdot r}$$

Then, considering the merits of option A, the ratio in variance having r+1 measurements divided by having r, becomes:

$$\frac{Var\left[\hat{SLOPE}_{ij}^{(r+1)}\right]}{Var\left[\hat{SLOPE}_{ij}^{(r)}\right]} = \frac{r-1}{r+2}$$, with the variance reduction being $100\cdot\frac{3}{r+2}\%$

More generally, moving from r to r+s measurements (s≥1):

$$\frac{Var\left[\hat{SLOPE}_{ij}^{(r+s)}\right]}{Var\left[\hat{SLOPE}_{ij}^{(r)}\right]} = \frac{r^3-r}{(r+s)^3-(r+s)}$$, with the variance reduction being

$$100\cdot\frac{(r+s)^3-r^3-s}{(r+s)^3-r-s}\%$$

We now look at option B. Having, as before, r measurements made at equi-distant time-points, the study duration, d, equals r-1. Under compound symmetry we may rewrite our $\hat{SLOPE}_{ij}$ variance to (see Schlesselman, 1973); $Var\left[\hat{SLOPE}_{ij}\right] = \frac{12\tau^2(r-1)}{d^2\cdot r\cdot(r+1)}$

We immediately see that the variance is inversely proportional to the square of the study duration. Doubling the study duration while keeping the number of visits fixed will decrease the variance by 75%.

Hence, substantial gains in precision for the estimates of the individual SLOPE's may be gained either by increasing r for a fixed d, or by increasing d for a fixed r. However, we have to be careful not to overstate the relevance of these results, since strong assumptions have been made: that the variance for the dependent variable stay constant over time, while the actual values are increasing linearly with time. This will often not be the case. Also, the compound symmetry assumptions is likely to be violated when time-intervals between visits are changed.

Before moving to option C, a combination of options A and B will be considered. Given that, at the design stage, a certain precision is desired for the estimates of the individual SLOPE's. Then different combinations of study duration and number of measurements, achieving this variance, may be evaluated.

An illustration of such an approach is given in table 5.3.1. This example, based on compound symmetry and equi-distant time-intervals between visits, evaluates a range of measurements from r=2 to 6, and study durations of 1, 1.5 and 2 (arbitrary) units. Choosing the design with r=6 and d=2 as baseline, the relative increases in variance for the $\hat{SLOPE}_i$'s under various other designs are given in the body of the table (these relationships are independent of the degree of equicorrelation). Note, however, that d often is determined by practicalities.

Table 5.3.1: Relative increases in variance for the $\hat{SLOPE}_{ij}$'s under various designs relative to what is obtained with 6 measurements and a study duration of 2 units.

Number of measurements	Total study duration (arbitrary units) 1	1.5	2
2	5.60	2.49	1.40
3	5.60	2.49	1.40
4	5.05	2.24	1.26
5	4.48	1.99	1.12
6	4.00	1.78	1.00

The equality of the entries in the first and second rows were to be expected, since, for three equidistant measurements, SLOPE gives the weight zero to the middle one. However, while being of no use for decreasing the variance, the third measurement may be needed to confirm (or refute) linearity.

Let us now consider the third design option, C. That is, having decided on the number of measurements to be taken and on the study duration, we still have to decide on the distribution of our measurements during the study period. The fact that the equidistance strategy is sub-optimal should be clear simply by consideration of the weights for the SLOPE's when there are three measurements. These will be proportional to; [-1, 0, 1], i.e. the second visit was, as far as estimation of each $SLOPE_{ij}$ was concerned, wasted. In the sense of precision, substituting the middle measurement for a second pre-entry measurement (or a second final measurement), would be a better alternative.

When all measurements are independent, we know from linear regression (see Draper and Smith, 1981) that the optimal distribution of the measurements (in the sense of SLOPE precision) is to have half at the very first time-point and half at the very last. This result may be shown to carry over unchanged to the compound symmetry situation. Using this strategy, of course, calls for a complete faith in the underlying linear relationship with time, and hence is usually unrealistic.

However, it is also likely that compound symmetry will then not hold. We can not expect measurements taken at, or very close to, the same point in time to be as equally correlated as measurements taken far apart. This would only happen if all within-subject variability was due to measurement error, and none to true intrinsic within-subject variability.

Thus, to give any explicit advice on the optimal distribution of measurements we would have to separate the within-subject variability into two components; measurement error and intrinsic within-subject variability (see Johnson and George, 1991), and go ahead with the modelling of the covariance structure from there.

This will not be explored further here. As a guiding principle, though, under a model with linearly diverging mean curves, having relatively more measurements later on in the study will be more powerful than using an equally-spaced distribution for the measurements.

5.3.2 The optimal alternative to SLOPE

Given the non-optimality of SLOPE for correlated within-subject measurements (see Potthoff and Roy, 1964), a search for better alternatives is suggested. As a help in deriving the optimal linear summary statistic under compound symmetry, when mean treatment curves diverge in a linear fashion, it is convenient to, firstly, derive another summary statistic, "Regression through the origin" (RTO), before arriving at the optimal one, SLANC (a "SLOPE-based ANCOVA").

In the process, it will be shown that under a model of linear divergence, the three summary statistics; RTO, SLOPE and SLANC, are analogous to POST, CHANGE and ANCOVA, under a model with a constant treatment effect.

As outlined in the preceding section linear summary statistics are scale-invariant. Returning to the consistent rules for scaling introduced there (i.e. such that the weights for the post-treatment measurements add to one), $SLOPE_{ij}$ may be redefined as;

$SLOPE_{ij} = \dfrac{\sum_{k=0}^{r}(t_k - \bar{t}_\cdot)\cdot y_{ijk}}{\sum_{k=1}^{r}(t_k - \bar{t}_\cdot)}$, only the scaling in the denominator has changed.

Whatever scaling is used it is a fact that SLOPE will give negative weights to the earlier post-treatment visits, where we actually are anticipating positive treatment differences. Some investigators feel uneasy about this, as a consequence the summary statistic RTO (regression through the origin) has been put forward as an alternative to SLOPE (Senn, 1993). Starting with a zero weight for the baseline, this summary statistic gives weights to the measurements that are increasing linearly with a linear increase in time.

RTO_{ij} may be calculated as; $RTO_{ij} = \dfrac{\sum_{k=0}^{r}(t_k - t_0)\cdot y_{ijk}}{\sum_{k=1}^{r}(t_k - t_0)}$

For a design with four visits, RTO_{ij} will be proportional to;

$$0\cdot y_0 + 1\cdot y_1 + 2\cdot y_2 + 3\cdot y_3,$$

correspondingly, $SLOPE_{ij}$ will be proportional to;

$$-3\cdot y_0 - 1\cdot y_1 + 1\cdot y_2 + 3\cdot y_3.$$

We see that RTO makes no baseline adjustment (we could, however, use RTO as dependent variable in an analysis of covariance, and incorporate the baseline as a covariate, as suggested by Senn (1993)), while SLOPE assign as much weight to the baseline as to the post-treatment measurements. One might anticipate that the optimum would be somewhere in between, and, as will be shown, it usually is.

For our mean summary statistics we have the relationship;

ANCOVA = $(1-\beta)\cdot$POST + $\beta\cdot$CHANGE.

Under linear divergence we have a corresponding relationship;

SLANC = $(1-\beta)\cdot$RTO + $\beta\cdot$SLOPE.

Under the assumptions of compound symmetry and linear divergence between mean response curves (but with no assumption needed about linearity within groups), this last equality defines the optimal linear summary statistic. A proof of this, for the special case of equi-distant time-intervals between measurements, will be given below. For the validity of the equality above to hold, we have to scale RTO and SLOPE in a consistent manner, as we are doing by constraining the weights for the post-treatment measurements to sum to unity (i.e. using the formulae given earlier for this purpose).

Generally, $SLANC_{ij}$ may be calculated from;

$$SLANC_{ij} = \frac{(1-\beta)\cdot\sum_{k=0}^{r}(t_k - t_0)\cdot y_{ijk}}{\sum_{k=1}^{r}(t_k - t_0)} + \frac{\beta\cdot\sum_{k=0}^{r}(t_k - \bar{t}_.)\cdot y_{ijk}}{\sum_{k=1}^{r}(t_k - \bar{t}_.)}$$

A standard two-sample t-test may then be performed between the two groups based on the overall treatment difference in mean SLANC's, $\overline{SLANC_A} - \overline{SLANC_B}$.

Under compound symmetry, the optimal linear summary statistic for linear divergence between mean treatment curves, and, for comparative purposes, the same under a model with a constant treatment effect, may be summarized as in table 5.3.2.

Table 5.3.2: Optimal linear summary statistics, assuming compound symmetry, under linear divergence, respectively, under a constant difference, between mean treatment curves.

Treatment effect	β=0	General β	β=1
Constant	POST	ANCOVA	CHANGE
Linear divergence	RTO	SLANC	SLOPE

With equidistant unit time-intervals between measurements the three linear divergence summary statistics may be expressed as follows:

$$\text{SLOPE}_{ij} = \sum_{k=0}^{r} \frac{(2k-r)\cdot y_{ijk}}{r}$$

$$\text{RTO}_{ij} = \sum_{k=0}^{r} \frac{2\cdot k\cdot y_{ijk}}{r(r+1)}$$

$$\text{SLANC}_{ij} = \sum_{k=0}^{r}\left[\frac{2\cdot k}{r+1}\left(\frac{1}{r}+\beta\right)-\beta\right]\cdot y_{ijk}$$

Direct comparisons of variances are not meaningful, since the expected mean treatment differences varies between RTO, SLOPE and SLANC. Instead, comparisons based on the asymptotic relative efficiencies among these summary statistics will be given in section 5.5.

We will now give a proof of the optimality of SLANC under linear divergence and compound symmetry, for the special case of equi-distant unit time-intervals between measurements and one pre-entry evaluation.

From the optimal linear summary statistic theorem we know that, in general, the optimal choice is $\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}$. Firstly, we rewrite this optimal choice in the form of the general linear summary statistic, as given in subsection 1.6.2, then, $\text{OPTI}=\sum_{i=0}^{r} c_i y_i$.

Utilizing results from Rao (1973), dealing with Fisher's linear discriminant function, the weights, the c_i's, may be written as;

$c_i = \sum_{j=0}^{r} (\mu_{Aj} - \mu_{Bj}) \sigma^{ij}$, where the σ^{ij} are the elements of the inverse Σ^{-1}. Dividing all measurements by σ^2 (arbitrary scaling), the covariance matrix Σ has elements 1 on the main diagonal and ρ off the main diagonal. The inverse matrix Σ^{-1} has elements

$\sigma^{ii} = \frac{1+(r-1)\rho}{1+(r-1)\rho-r\rho^2}$ on the main diagonal, and $\sigma^{ij} = \frac{-\rho}{1+(r-1)\rho-r\rho^2}$ off the main diagonal.

Then, the optimal summary statistic has weights (proportional to) $c_i = [1+(r-1)\rho-r\rho^2] \sum_{j=0}^{r} (\mu_{Aj} - \mu_{Bj}) \sigma^{ij}$.

With equi-distant unit time-intervals this simplifies to

$$c_i = [1+(r-1)\rho-r\rho^2] \sum_{j=0}^{r} j\,\sigma^{ij} =$$

$$= [1+(r-1)\rho-r\rho^2] \cdot \left(i \cdot \left[\frac{1+(r-1)\rho}{1+(r-1)\rho-r\rho^2} \right] + \left(\frac{r(r+1)}{2} - i \right) \cdot \left[\frac{-\rho}{1+(r-1)\rho-r\rho^2} \right] \right) =$$

$$= i \cdot [1+(r-1)\rho] - \rho \cdot \left[\frac{r(r+1)}{2} - i \right]$$

After multiplication by $\frac{2}{r(r+1)}$ (to conform to our consistent scaling) this may be rewritten as; $c_i = \frac{2i}{r+1} \left[\frac{1}{r} + \rho \right] - \rho$, which for one pre-entry measure and under compound symmetry (so that $\rho=\beta$) is identical to the summary statistic labelled SLANC above.

Before concluding this section, some previous alternative approaches for improving upon SLOPE will be mentioned.

An obvious improvement would be to use an analysis of covariance with SLOPE as dependent variable and with the (mean of the) pre-treatment measurement(s) as covariate, or with the estimated intercept for each subject as covariate. Even if the first approach makes a proper covariate adjustment, the relative differences between the weights for the post-treatment measurements will remain fixed, only the weight for the baseline will change. Thus, the increments in the weights (the c_i's) will not be the same between c_0 and c_1 as it is between the remaining c_i's. As a consequence, a covariance adjusted SLOPE will differ from SLANC. The latter approach was suggested by Laird and Wang (1990). Both of these approaches improve on SLOPE, but neither reaches quite as far as SLANC.

It has also been suggested to use the fitted higher order polynomials of response against time for each subject, i.e. quadratic, cubic, quartic etc., as covariates, when having SLOPE as dependent variable (see Leech and Healy, 1959, and Kenward, 1985). To increase the power Leech and Healy further considered forming a linear combination of two summary statistics (scaled to) having the same expected value (in particular of SLOPE and POST), by forming the minimum-variance combination of the two based on the covariance matrix for the two summary statistics. They considered this type of combined summary statistics for both linear and quadratic divergence.

When quadratic divergence is thought plausible, analysis of quadratic regression coefficients might, if interest resides in the rate of rate of change, be recommended. For a useful reference, see Snedecor and Cochran (1989).

Acknowledging the fact that the variations around the regression lines for different subjects often are far from equal (e.g. when different subjects have different number of measurements), Matthews (1993) contrasted different schemes for arriving at weighted analyses of rates of change, as measured by linear regression coefficients.

That is, using SLOPE as a summary statistic for each subject, but weighting the subjects, in some way, according to the relative precision of their estimated regression coefficient. An extension to this approach would be to use a "weighted SLANC".

As a final reference, an approach suggested by C.R. Rao (1959) will be mentioned. Emphasizing the examination of whether differences exists between groups in mean rate of change during treatment, and realizing that the rate is rarely constant, but rather a complicated function of time, he suggested a method for transforming the time scale making the rate of change linear in a new time metameter. The linear regression coefficients in this new time scale would then be used as the summary statistics. This approach has a certain appeal, but we suspect that one will usually lack enough data to make reliable transformations of the time dimension. A natural extension to Rao's approach would be to use his transformation of the time scale, but then to use SLANC at the analysis stage.

5.4 WEIGHTINGS FOR LINEAR SUMMARY STATISTICS

In this section, the weightings used for various possible linear summary statistics will be explained. This will form a basis for more specific comparisons among these summary statistics, under different assumptions regarding treatment effects and covariance structures, in the remainder of this chapter. In table 5.4.1, outlined below, the weightings used for some of the summary statistics, when there are p pre and r post-treatment visits, are given. There are no constraints on the covariance matrix, other than it has to be identical between treatment groups. Further, as has been discussed above, all the $\mathbf{c}'$-vectors have been scaled so that the weights sum to one for the post-treatment visits.

One new summary statistic is introduced in this table, OPTI_CS, which is the optimal linear summary statistic under compound symmetry. The optimality holds if the true δ'-vector is as specified in the derivation of the weights, and under compound symmetry with knowledge of the true ρ. The weights for OPTI, the optimal linear summary statistic under a general covariance structure, are given by $\mathbf{c}' = \delta'\Sigma^{-1}$.

Three of the summary statistics in the table; POST, CHANGE and SLOPE, are unconditional, they are not based on any information from the covariance matrix. The remaining summary statistics are conditional, either on β or on information from δ.

ANCOVA, SLANC and OPTI_CS are all based on the principle of analysis of covariance. The three approaches use different dependent variables (weighted combinations of the post-treatment measurements) in the analysis, as described below, but they are all using the same covariate adjustment, $-\beta$ times the pre-treatment value (or the mean of several pre-treatment values). Hence, they are all based on analysis of covariance.

Table 5.4.1 : Weightings for some linear summary statistics. Design; p visits pre-treatment, r post-treatment. For OPTI_CS d_i is the assumed mean treatment difference for the i'th measurement. $\beta = \bar{\Sigma}_{mix} / \bar{\Sigma}_{pre}$, under compound symmetry and when p=1, $\beta=\rho$.

Summary statistic	Weightings	Special case 1 pre + 3 post (β=2/3)
POST	$\begin{cases} c_i = 0 & , i = -(p-1),\ldots,0 \\ c_i = \frac{1}{r} & , i = 1,\ldots,r \end{cases}$	$\left[0,\vdots\frac{1}{3},\frac{1}{3},\frac{1}{3}\right]$
CHANGE	$\begin{cases} c_i = -\frac{1}{p} & , i = -(p-1),\ldots,0 \\ c_i = \frac{1}{r} & , i = 1,\ldots,r \end{cases}$	$\left[-1,\vdots\frac{1}{3},\frac{1}{3},\frac{1}{3}\right]$
ANCOVA	$\begin{cases} c_i = -\frac{\beta}{p} & , i = -(p-1),\ldots,0 \\ c_i = \frac{1}{r} & , i = 1,\ldots,r \end{cases}$	$\left[-\frac{2}{3},\vdots\frac{1}{3},\frac{1}{3},\frac{1}{3}\right]$
SLOPE*	$\begin{cases} c_i = -\frac{1}{p} & , i = -(p-1),\ldots,0 \\ c_i = \frac{-r+2\cdot i}{r} & , i = 1,\ldots,r \end{cases}$	$\left[-1,\vdots-\frac{1}{3},\frac{1}{3},1\right]$
SLANC*	$\begin{cases} c_i = -\frac{\beta}{p} & , i = -(p-1),\ldots,0 \\ c_i = \frac{2\cdot i}{r+1}\left(\frac{1}{r}+\frac{\beta}{p}\right)-\frac{\beta}{p} & , i = 1,\ldots,r \end{cases}$	$\left[-\frac{4}{6},\vdots-\frac{1}{6},\frac{2}{6},\frac{5}{6}\right]$
OPTI_CS	$\begin{cases} c_i = -\frac{\beta}{p} & , i = -(p-1),\ldots,0 \\ c_i = \frac{1}{r\cdot\bar{d}^{post}}\left[d_i - \frac{r}{p}\left(\bar{d}^{post}-d_i\right)\cdot\beta\right] & , i = 1,\ldots,r \end{cases}$	$\left[-\frac{2}{3},\vdots c_1,c_2,c_3\right]$

* These are the weights for SLOPE and SLANC when all time-intervals between adjacent visits are equal, for the general case see below.

ANCOVA uses the mean of the post-treatment recordings as dependent variable. SLANC assumes a linear divergence between the mean response curves, reflected in linearly increasing weights for linearly increasing time-intervals since randomisation. OPTI_CS puts weights proportional to the assumed differences in mean response profiles for each visit, when composing the dependent variable. These weights are shifted downwards such that the pre-treatment visit (or mean of pre-treatment visits) receives the weight $-\beta$ $(= -\left(\overline{\Sigma}_{mix} / \overline{\Sigma}_{pre}\right))$ when the sum of the post-treatment weights has been scaled to unity. That is, each of the three approaches assumes the shape but not the magnitude of H_A is known in advance.

Continuing with the table above, for SLOPE and SLANC for the weights given, it is assumed, for illustration, that visits are performed with equidistant time-intervals. Relaxing this equidistance assumption, and assuming that the measurements for one baseline and r post-treatment visits are performed at the time-points $t_0, t_1, \ldots, t_r$, the weights for SLOPE are given by;

$$c_i = \frac{t_i - \bar{t}}{\bar{t} - t_0} \quad \text{for } i=0,\ldots,r.$$

Having more than one baseline, one simply takes c_0 (i.e. -1) times the average of these respective pre-entry measurements.

The weights to be used for RTO (regression through the origin) are not given in the table. With equidistant time-intervals scaled to unity the weights are found from;

RTO; $c_i = \dfrac{2 \cdot i}{r \cdot (r+1)}$, i=0,1,..,r , and with a general time-scale;

RTO; $c_i = \dfrac{t_i - t_0}{(r+1)(\bar{t} - t_0)}$

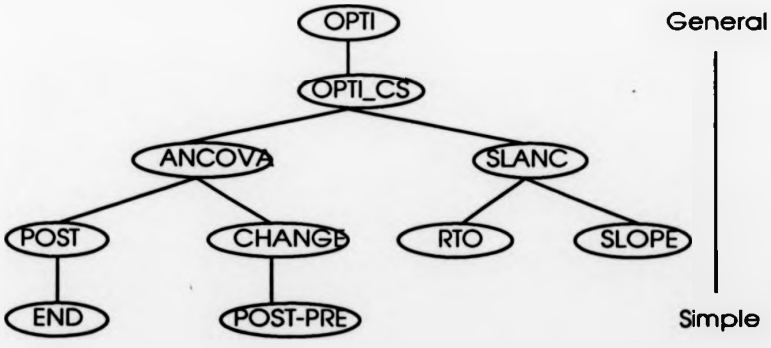
Likewise, for SLANC, when the measurements are taken at the time-points $t_0, t_1, \ldots, t_r$:

$$\text{SLANC:}\quad c_i = \frac{\beta(t_i - t_0)}{(r+1)(\bar{t} - t_0)} + \frac{(1-\beta)(t_i - \bar{t})}{(\bar{t} - t_0)}$$

As a numerical illustration, say, for p=1 and r=4 measurements pre and post-randomisation, and under compound symmetry with ρ=.7, SLANC obtains the weights [-.70, -.32, .06, .44, .82].

In figure 5.4.1 below a categorization according to the degree of generalizability of some of the more useful linear summary statistics is given. The bottom layer consists of simple univariate (time-point specific) statistics. In the next layer (in principle) all measurements are utilized, but the summary statistics are unconditional, thus, optimal use of baselines is not accomplished. The third layer consists of analyses of covariance where a "simple" (i.e. constant mean treatment difference over time, or linear divergence) alternative hypothesis is assumed. In the second highest layer, a general alternative hypothesis is allowed for, but compound symmetry is needed for strict optimality. At the very top we find the optimal choice under any covariance structure and any alternative hypothesis over time.

Figure 5.4.1: Linear summary statistics, hierarchical structure.



Finally, the weightings arrived at when deriving the optimal linear summary statistic under compound symmetry when no pre-entry measurements are available, will be given.

OPTI for compound symmetry with no baselines;

$$c_i=\frac{1}{\bar{d}_.}\cdot\left[\frac{d_i}{r}+\frac{\rho(d_i-\bar{d}_.)}{1-\rho}\right] \qquad , i=1,\ldots,r$$

For the special case with linearly diverging mean response curves the weights may be found from; $c_i=\frac{2}{r+1}\cdot\left[\frac{i}{r}+\frac{\rho\left(i-\frac{r+1}{2}\right)}{1-\rho}\right]$

, i=1,...,r

This is different from SLOPE (because of the non-optimality of least squares for inter-correlated measurements), and might be referred to as SLANC for the case where there is no baseline.

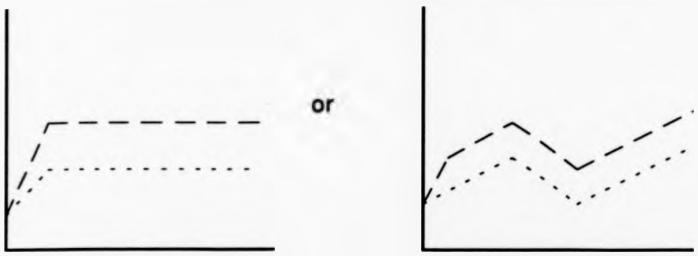
5.5 CHOICES OF SUMMARY STATISTICS UNDER SPECIFIC CLASSES OF ASSUMPTIONS

As described in section 5.2, given a clear idea of what the shape of the vector of mean treatment differences will look like, and what the covariance structure is likely to be, one can derive the optimal linear summary statistic in the sense of maximization of the expected t-statistic. More specifically, the optimal linear summary statistic maximizes the estimated treatment difference, **c'd**, relative to its standard error.

We will now look at some realistic classes of assumptions for repeated measures designs, and investigate how these assumptions affects the asymptotic relative efficiencies among various linear summary statistics, and also how the weightings for the optimal linear summary statistic changes with changes in these assumptions.

5.5.1 A constant difference in mean response profiles

This category of differences in treatment effects cover studies for which the mean curves might look like, for instance,



The only assumption is that the difference between the mean curves remain constant after an initial treatment effect, i.e. the lines are parallel.

5.5.1.1 Constant difference under compound symmetry

Under compound symmetry, and without baselines, one can do no better than POST. With any number of baselines and any number of post-treatment evaluations, ANCOVA is asymptotically the optimal choice (but ANCOVA has **c** estimated from the data, i.e. $\hat{\beta}$ not β).

Table 5.5.1: Optimal linear summary statistics for constant treatment effects (δ is proportional, not necessarily equal, to unity) and under compound symmetry. Asymptotic relative efficiencies compared to other summary statistics.

δ'	ρ	OPTI	POST	CHANGE	ANCOVA	SLANC	SLOPE
0,1,1,1	.1	ANCOVA	.98	.32	1	.82	.20
	.3	ANCOVA	.83	.48	1	.76	.28
	.5	ANCOVA	.62	.62	1	.71	.38
	.7	ANCOVA	.39	.78	1	.66	.46
	.9	ANCOVA	.13	.92	1	.62	.56
0,1,1,1	.7	ANCOVA	.39	.78	1	.66	.46
0,0,1,1,1	.7	ANCOVA	.28	.89	1	.73	.43
0,0,0,1,1,1	.7	ANCOVA	.23	.94	1	.76	.40
0,1	.7	ANCOVA	.51	.85	1	1	.85
0,1,1	.7	ANCOVA	.42	.80	1	.79	.60
0,1,1,1	.7	ANCOVA	.39	.78	1	.66	.46
0,1,1,1,1,1	.7	ANCOVA	.36	.75	1	.50	.32
0,1,1,1,1,1,1,1,1,1,1		ANCOVA	.33	.73	1	.31	.18

As already stated, for a constant difference between mean treatment curves, and under compound symmetry, ANCOVA (assuming a known covariance matrix) is always the optimal choice among all linear summary statistics. Under these circumstances SLANC and SLOPE would hardly be considered as approaches to use at the analysis stage, but they are included in the table for comparative purposes.

POST is a good choice only when there are no baselines, or when correlations are very low. CHANGE is an efficient approach when correlations are high and when one has more than one pre-entry measurement. When correlations are low and the number of post-treatment measurements are few, SLANC is a quite powerful choice even under these assumptions. ANCOVA is, of course, the method of choice, the only exception from this rule is when the number of subjects is very low, and when, thus, the regression coefficient β becomes unreliably estimated.

5.5.1.2 Constant difference and other covariance structures

In the following table, a constant difference between mean response curves is once more assumed. The difference being that we now investigate the impact of departures from compound symmetry on the relative merits of the different summary statistics under investigation. In particular, a banded covariance structure is assumed. That is, the correlation is assumed constant on each diagonal of Σ. For instance, for the second example of table 5.5.2, the variance is proportional to one, the correlation between adjacent visits is .733, the correlation between two visits having one visit in between is .683, finally, the correlation between the very first and the very last visit is .633 .

Table 5.5.2: Optimal linear summary statistics for a constant treatment effect and under covariance structures different from compound symmetry. Asymptotic relative efficiencies compared to other summary statistics. (In all instances $\delta'=[0,1,1,1]$, i.e. p=1 and r=3).

Σ (banded)	OPTI				POST CHANGE	ANCOVA	SLANC	SLOPE
σ^2: 1 1 1 1								
ρ : 1 .7 .7 .7	-.70	.33	.33	.33	.39	.78	1	.66	.46
σ^2: 1 1 1 1								
ρ : 1 .733 .683 .633	-.69	.47	.29	.25	.41	.76	.98	.57	.39
σ^2: 1 1 1 1								
ρ : 1 .767 .667 .567	-.71	.60	.22	.18	.42	.71	.92	.51	.33
σ^2: 1 1 1 1								
ρ : 1 .833 .633 .433	-.81	1.00	-.11	.11	.36	.52	.68	.35	.21
σ^2: 1 .9 .8 .7								
ρ : 1 .7 .7 .7	-.62	.25	.33	.42	.38	.63	.99	.76	.41
σ^2: 1 1.1 1.2 1.3								
ρ : 1 .7 .7 .7	-.76	.39	.33	.28	.39	.87	.99	.58	.47
σ^2: 1 1.33 1.67 2								
ρ : 1 .7 .7 .7	-.88	.48	.31	.21	.37	.95	.96	.46	.43
σ^2: 1 2 3 4								
ρ : 1 .7 .7 .7	-1.11	.59	.28	.13	.35	.86	.90	.30	.33
σ^2: 1 1.33 1.67 2								
ρ : 1 .767 .667 .567	-.88	.74	.19	.07	.37	.76	.79	.30	.34

In the top half of rows in table 5.5.2, the effect of assuming declining correlations over time, as opposed to equicorrelations, is illustrated for a design with 1 pre and 3 post-randomisation visits. Here, a banded correlation structure is assumed with correlation coefficients decided in such a way that the overall average correlation, for the within-subjects covariance matrix, in all instances remain at 0.7 (which was found to be a plausible choice in practice from the examples incorporated in table 1.5.1). The degree of decline in correlation for each further visit apart is .00, .05, .10 respectively .20 for the four rows reported here.

The effect of declining correlations over time, for the optimal linear summary statistic, is that more weight should be put on the first post-randomisation visit, relative to the later, since this is the one most highly correlated with the baseline. As long as the degree of decline over time for the correlation coefficient remains reasonably small, say below 0.10 per visit, ANCOVA is close in relative efficiency to the optimal choice (with more than three post-treatment visits the ARE for ANCOVA relative to OPTI will tend to drop for this type of covariance structure).

As for ANCOVA; CHANGE, SLOPE and SLANC all lose in efficiency the more the correlations tend to decline over time. This is not true for POST, but with this degree of within-subject dependency it is never a method to be recommended.

In the bottom half of table 5.5.2, equicorrelation with a ρ of 0.7 is assumed, now the consequences of departure from homoscedasticity is illustrated. Once again four visits are assumed, and four different scenarios are illustrated, with variances for the four time-points assumed as; 1,.9,.8,.7 , 1,1.1,1.2,1.3 , 1,1.33,1.67,2 , and finally 1,2,3,4. Here we can see that ANCOVA is very robust against heteroscedasticity, not even when variances increase fourfold over the study period does the ARE relative to the optimal choice drop below 0.90.

As could be expected, when variances increase, more weight should be put on the first visit, the one with best precision post-randomisation. The only realistic alternative to ANCOVA in this setting is CHANGE, with moderately increasing variances over time its ARE increases and approaches unity, this is because the regression coefficient β gets closer to one, which is implicitly what is assumed when one is analysing data with a mean change approach.

In the final row of the table, the joint effect of increasing variances and declining correlations over time is illustrated. For the specific example outlined, an analysis of covariance with the first post-randomisation visit as dependent variable (i.e. "throwing away" the later data), and with the pre-entry measurement as covariate, would actually be more powerful than ANCOVA. The ARE for this simple approach is 0.92 relative to the optimal choice.

5.5.2 A linear divergence between mean response profiles

What normally springs to mind in relation to linearly divergent treatment effects, is the situation where both mean treatment curves follow straight lines, but with different regression coefficients, as exemplified in the left figure on page 163. However, for linearly divergent treatment effects more complex situations like in the figure at the right on the same page, are equally well covered (i.e. we may have a μ_k which is non-constant over time in the model on page 162). All we assume is that the mean treatment difference increase linearly with time.

5.5.2.1 Linear divergence under compound symmetry

As previously explained in section 5.3, SLANC is the optimal linear summary statistic in this situation. This has equal increments between weights, but with the weights shifted downwards in such a way as to correspond to a covariance analysis. More specifically, when the $\mathbf{c}'$-vector has been scaled such that the weights for the post-treatment visits sum to one, the weight for the pre-entry measurement will be $-\beta$.

To get some idea on how the relative efficiencies among the summary statistics under investigation change depending on the number of measurements and the degree of correlation, the following table is presented.

Table 5.5.3: Optimal linear summary statistics for a linear divergence between mean response curves, and under compound symmetry. Asymptotic relative efficiencies compared to other summary statistics.

δ'	ρ	OPTI	POST	CHANGE	ANCOVA	SLANC	SLOPE
0,1,2,3	.1	SLANC	.80	.27	.82	1	.45
	.3	SLANC	.63	.36	.76	1	.60
	.5	SLANC	.44	.44	.71	1	.74
	.7	SLANC	.26	.51	.66	1	.85
	.9	SLANC	.08	.57	.62	1	.95
0,1,2,3	.7	SLANC	.26	.51	.66	1	.85
0,0,1,2,3	.7	SLANC	.20	.65	.73	1	.88
0,0,0,1,2,3	.7	SLANC	.18	.71	.76	1	.85
0,1	.7	SLANC	.51	.85	1	1	.85
0,1,2	.7	SLANC	.33	.63	.79	1	.84
0,1,2,3	.7	SLANC	.26	.51	.66	1	.85
0,1,2,3,4,5	.7	SLANC	.18	.38	.50	1	.88
0,1,2,3,4,5,6,7,8,9,10		SLANC	.10	.23	.31	1	.91

The claim that SLANC does for linear divergence what ANCOVA does for constant differences, may be verified by comparing the ARE's for ANCOVA in this table, with the ARE's for SLANC in the corresponding table (5.5.1) on constant differences. All the figures are identical.

A relevant issue in connection with linear divergence under compound symmetry is how much more powerful the optimal choice, SLANC, is relative to SLOPE and ANCOVA. This depends on the number of post-treatment measurements as well as on the degree of equi-correlation.

These relationships, contrasting ANCOVA with SLANC, and SLOPE with SLANC, are illustrated in figures 5.5.1 and 5.5.2, respectively. From figure 5.5.1 we see that, with few post-treatment measurements, ANCOVA is relatively powerful, especially when correlations are low. However, the more r increases the more superior will SLANC be relative to ANCOVA.

Figure 5.5.1 :
ARE's for ANCOVA relative to SLANC under linear divergence and compound symmetry, as a function of the number of post-treatment meaurements and the degree of equicorrelation (.3, .5, .7 or .9). Assuming 1 baseline

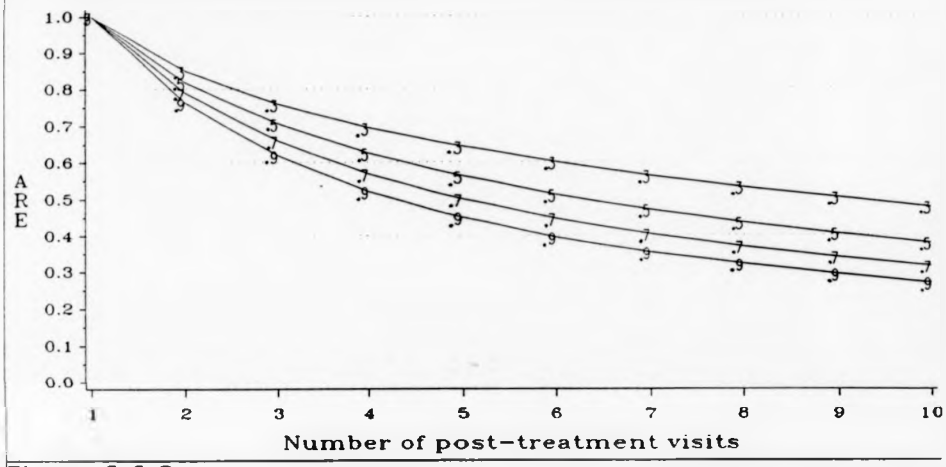
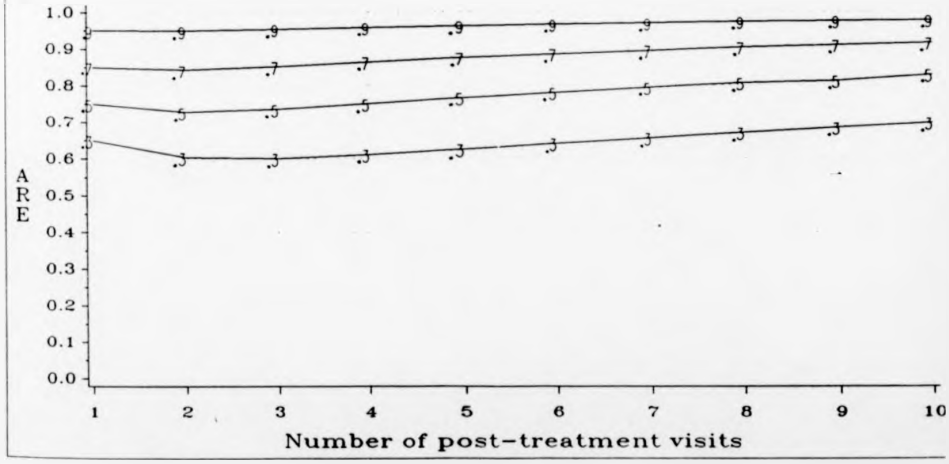


Figure 5.5.2 :
ARE's for SLOPE relative to SLANC under linear divergence and compound symmetry, as a function of the number of post-treatment meaurements and the degree of equicorrelation (.3, .5, .7 or .9). Assuming 1 baseline



From figure 5.5.2 it is revealed that the degree of inferiority for SLOPE relative to SLANC depends primarily on the correlation. With ρ in the plausible range .5 to .7 (as suggested by the examples in table 1.5.1) the ARE stays around .80. This implies that a SLOPE analysis would require about 25% more subjects to obtain the same power as a SLANC analysis. With increasing r SLOPE gets somewhat closer to SLANC in relative efficiency, this is because $\beta = \overline{\Sigma}_{mix}/\overline{\Sigma}_{pre}$ gets closer to unity. The increasing inferiority for SLOPE when r goes from 1 to 2 is because SLOPE in the latter case gives the weight zero to the middle measurement.

Returning to table 5.5.3, we see, from the middle third of the table, that an increase in the number of pre-treatment measurements improves the situation for ANCOVA, while SLOPE remains at about the same efficiency relative to SLANC.

5.5.2.2 Linear divergence and other covariance structures

The table below is composed in exactly the same way as the corresponding table in subsection 5.5.1.2, except that a linear divergence, instead of a constant difference, between mean response curves is assumed.

Table 5.5.4: Optimal linear summary statistics for linearly divergent mean response curves and under covariance structures different from compound symmetry. Asymptotic relative efficiencies compared to other summary statistics. (In all instances $\delta'=[0,1,2,3]$).

Σ (banded)	OPTI	POST CHANGE	ANCOVA	SLANC	SLOPE
σ^2: 1 1 1 1					
ρ : 1 .7 .7 .7	-.70 -.18 .33 .85	.26	.51	.66	1	.85
σ^2: 1 1 1 1					
ρ : 1 .733 .683 .633	-.63 -.14 .29 .85	.31	.57	.73	.99	.82
σ^2: 1 1 1 1					
ρ : 1 .767 .667 .567	-.57 -.10 .22 .88	.36	.60	.78	.98	.78
σ^2: 1 1 1 1					
ρ : 1 .833 .633 .433	-.43 .05 -.11 1.06	.43	.62	.81	.92	.70
σ^2: 1 .9 .8 .7					
ρ : 1 .7 .7 .7	-.58 -.20 .29 .91	.23	.37	.58	.99	.68
σ^2: 1 1.1 1.2 1.3					
ρ : 1 .7 .7 .7	-.80 -.17 .36 .81	.27	.62	.71	.99	.93
σ^2: 1 1.33 1.67 2					
ρ : 1 .7 .7 .7	-.98 -.13 .39 .74	.30	.76	.78	.97	.97
σ^2: 1 2 3 4					
ρ : 1 .7 .7 .7	-1.35 -.07 .41 .66	.33	.82	.85	.89	.86
σ^2: 1 1.33 1.67 2					
ρ : 1 .767 .667 .567	-.82 .02 .29 .69	.42	.86	.90	.97	.94

For all departures from compound symmetry investigated, SLANC stays close to OPTI in relative efficiency. Only with an extreme increase in variance will the ARE drop below 0.9. ANCOVA gets relatively more powerful as correlations decline over time and/or variances increase with time, and will often be almost as powerful as SLOPE.

SLOPE is usually a quite efficient approach to analysis under these circumstances, unless correlations and/or variances decrease with increasing time-intervals since randomisation. General results for increasing r are difficult to convey, these will depend on whether we assume that an increasing number of post-treatment measurements will imply a prolonged study period or shorter time-intervals between visits. Typically both ANCOVA and SLOPE will lose in efficiency relative to OPTI when r gets larger.

5.5.3 Other types of divergence in mean response profiles

Under a general covariance structure it is difficult to give any general advice relating to the relative merits of different possible summary statistics under various classes of differences in mean response profiles over time. However, we can explicitly evaluate results for any given choice of Σ and δ.

Under compound symmetry the optimal summary statistic will always be a covariance analysis, under any δ'-vector assumed. This may be seen from the formula for the derivation of the weights for OPTI_CS given in section 5.4. The dependent variable will be a weighted sum of the post-treatment measurements, and when this sum is scaled to one, the (sum of the) weight(s) for the pre-treatment measurement(s) will always be $-\beta$. Thus, we end up with an analysis of covariance.

In fact, under compound symmetry, the weights for the different measurements, for the optimal linear summary statistic, will always be proportional to the vector of mean treatment differences.

To give just a flavour of the calculations one can do at the design stage when comparing various approaches to the analysis of a forthcoming study, another table involving ARE calculations between some summary statistics is given. Here, some of the possible classes of differences in mean treatment effects outlined in section 1.7 are exemplified, for two specific examples of covariance structures, both with homoscedasticity and a mean correlation of 0.7, but with and without declining correlations with increasing time-intervals between measurements.

Table 5.5.5: Optimal linear summary statistics for some different classes of vectors of mean treatment differences, and for two different correlation structures. ARE's compared to some other summary statistics.

Σ (banded)	OPTI				POST CHANGE	ANCOVA	SLANC	SLOPE
Attenuated divergence : **(δ'=[0,1,1.5,1.75])**								
ρ : 1 .7 .7 .7	-.70	.03	.39	.58	.34	.67	.87	.93	.74
ρ : 1 .767 .667 .567	-.64	.21	.34	.45	.45	.76	.98	.87	.64
Transient effect : **(δ'=[0,2,1,.5])**								
ρ : 1 .7 .7 .7	-.70	1.07	.19	-.26	.21	.41	.53	.04	.01
ρ : 1 .767 .667 .567	-.82	1.24	.03	-.28	.16	.27	.34	.02	.00
Exponential divergence : **(δ'=[0,.5,2,5])**								
ρ : 1 .7 .7 .7	-.70	-.49	.13	1.37	.14	.28	.37	.90	.82
ρ : 1 .767 .667 .567	-.40	-.66	-.31	1.97	.18	.31	.40	.80	.69

With an attenuated divergence, either ANCOVA or SLANC could be chosen, as regards efficiency. Of course, other aspects than efficiency are involved in this choice of summary statistic, especially as relates to how one wishes to estimate and report a difference in treatment effects. One important aspect which has not been discussed explicitly so far, relating to the choice of summary statistic, is to decide on what we really are trying to estimate. That is, what will the "bottom line" answer be from our analysis. Specifically, do our summary statistic estimate between-group differences in outcomes, within-group outcomes, or within-subject outcomes? In principle, the results in this chapter are developed for between-group comparisons. However, often the summary statistics will be equally useful in estimating within-group and within-subject outcomes. This may be exemplified by the two figures given on page 163. When the underlying model is such that all subjects follow linear curves with time (as illustrated by the left-hand side figure), but with different rates of change between groups, a summary statistic like SLANC (or SLOPE) will be relevant in estimating all the three different types of outcomes outlined above.

However, when a model as illustrated by the right-hand side figure holds true, SLANC (or SLOPE) estimates are only directly meaningful for between-group comparisons.

With a transient difference in mean treatment effects, none of the five summary statistics for which the ARE's are reported is appropriate. For a study of this type, one should preferably drop the last measurement, perhaps also the next last, and than use ANCOVA. A simple analysis of covariance with the first post-treatment measurement as dependent variable obtains an ARE of .95 for the compound symmetry example, and .92 for the decreasing correlation example. The corresponding two ARE's when dropping just the final measurement are .75 and .58

With an exponential divergence, most of the post-treatment weights should be put on the last measurement, either by using SLANC or SLOPE, or by using an analysis of covariance with the last measurement as dependent variable. This piece of advice remains applicable as long as the variances do not increase too much for the later phases of the study. When the variances are proportional to δ' we get nearer to equal weights (for the post-treatment measurements) again. In fact, when the standard deviations are proportional to δ', the optimal summary statistic will put most of the post-treatment weight on the first measurement after randomisation.

5.6 SUMMARY AND DISCUSSION

A sound piece of statistical advice is that "the more you put into an analysis, the more you can expect out". This can be applied to the choice of summary statistics for the analysis of repeated measures designs. The more information that is available from past experience and medical knowledge about expected treatment effects and within-subject dependencies, the more sensible the choice can be made for a summary statistic to increase both validity and sensitivity. The extent to which such information exists (and if it exists) will vary depending on therapeutic area, type of variable being measured, and the clinical phase the treatment is in. It is safe to say that there will never be perfect knowledge, but some information will usually be available.

Unless more refined knowledge is available, the general advice; to use ANCOVA in all situations where a reasonably stable mean treatment difference over time is expected, and to use SLANC when a gradual divergence between mean curves seems plausible, will almost always result in valid and efficient inferences.

Not to use an analysis of covariance, when a pre-entry measurement is available, will always be a mistake. Apart from a highly probable loss of efficiency, there is also a risk that the results will not be strictly valid. For instance, CHANGE will overcorrect for a chance mean pre-treatment difference, and so will SLOPE. POST, on the other hand, simply ignores any such imbalance.

It is worth considering allowance of a cautious flexibility in the choice of summary statistic if the pre-defined δ' is clearly wrong. An example where this occurred is given in the childhood asthma trial reported by Van-Essen Zandvliet et al (1992), where in fact a linear divergence between mean curves over the study period was observed for FEV_1, rather than the a priori assumption of a stable treatment effect. Restricting oneself to always stick rigidly to pre-specified summary statistics under all circumstances would be a poor scientific approach.

One underlying assumption for the optimal linear summary statistic theorem is the equality of covariance matrices between treatment groups. An alternative route to follow, if the underlying distributions are normal but the covariance matrices appear unequal, would be to substitute the quadratic discriminant function (see Lachenbruch and Goldstein, 1979), for the linear discriminant function, above.

In common with the methods for multiple endpoints, referred to above, we may easily run into the problem of obtaining negative weights for post-treatment measurements, when we are anticipating positive treatment effects. This will, for instance, often occur when mean curves diverge linearly with time (for SLOPE this occurs by definition). A possible way of overcoming this problem might be to choose the maximin direction (Abelson and Tukey, 1963).

This choice would maximize the minimum power over the orthant (Tang et al, 1993). In principle this means finding $\max_{c \neq 0} \frac{(c'\delta)^2}{c'\Sigma c} = \delta'\Sigma^{-1}\delta$ subject to the constraint that all c_i's for the post-treatment measurements should be non-negative.

Finally, in this chapter methods have been given that will allow efficient summary statistic approaches to be chosen for almost any pattern of differences in mean responses between treatments over time, and for almost any covariance structure. When treatment effects are not stable over time it is frequently the case that inferior and sometimes misleading analyses are being performed. By a sensible use during planning of the optimal linear summary statistic theorem, and comparisons of asymptotic relative efficiencies, powerful and valid summary statistics may be chosen for most repeated measures studies with continuous outcome measures.

6 FURTHER PERSPECTIVES

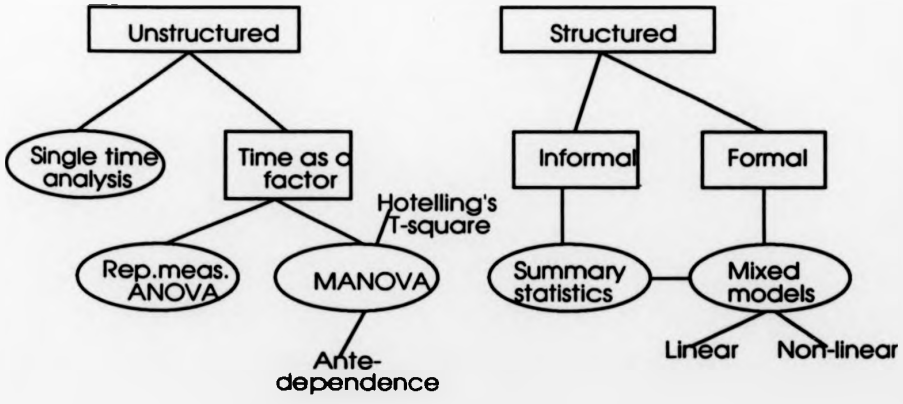
6.1 COMPARISON WITH OTHER APPROACHES

6.1.1 Introduction

The available approaches to analysis of continuous repeated measures data from comparative clinical trials can broadly be classified into two classes, unstructured and structured. In the former, no assumptions at all are made about the form or shape of the response profiles, and time is usually regarded as a factor, with the repeated measurement occasions as levels of the factor. The latter class incorporates some feature(s) or structure of the response profiles, formally or informally, in the analysis.

A schematic overview of the main classes of approaches is given in figure 6.1.1.

Figure 6.1.1: Schematic overview of the main classes of approaches for the analysis of continuous repeated measures data in comparative clinical trials.



In this thesis the usefulness of the summary statistic approach has been emphasized. Specific attention has been given to the choice of summary statistic and the choice of repeated measures design, with the latter choice being in terms of the number of pre (p) and post-treatment (r) measurements, as well as for the spacing in time of these evaluations. The objectives of these considerations have been to increase efficiency and enhance validity (e.g. removal of bias, incurred, for instance, by chance baseline imbalance between groups) under any plausible covariance structure and for any pre-declared difference in true mean responses over time between treatments. In this section, some of the more common alternative approaches will be described and their advantages and disadvantages relative to the summary statistics approach will be evaluated.

6.1.2 Some other approaches and the CPK example

Some of the more common approaches for the analysis of continuous repeated measures data will now be described. To illustrate their usefulness, and to contrast them with the summary statistic approach, the CPK example from section 2.5 will be re-analysed.

6.1.2.1 Repeated univariate, time-point specific, tests

Many reports of clinical trials rely on repeated significance testing for all time-points using univariate t-tests or Wilcoxon's rank-sum test. In the CPK example eight post-treatment measurements were obtained. Submitting these time-points, performed at 1.5, 3, 4.5, 6, 7.5, 9, 10.5 and 12 months, to univariate t-tests, based both on unadjusted post-treatment measurements and changes from baseline (mean of the three pre-entry evaluations), as well as covariance adjusted post-treatment measurements (with the pre-treatment mean as covariate), results in the following table.

Table 6.1.1: Univariate, time-point specific, analyses of the data from the CPK-example, t-tests based on post-treatment measurements, post-pre changes, and covariance adjusted post-treatment measurements.

Time-point (mths)	Estimated difference in the means of the: Post meas. (SEM)	Post-pre meas. (SEM)	Covariance adj. post meas. (SEM)	p-values for test of Post meas.	Post-pre meas.	Cov. adj. meas.
1.5	.090 (.040)	.067 (.023)	.070 (.023)	.025	.004	.002
3	.059 (.039)	.036 (.028)	.042 (.027)	.13	.20	.12
4.5	.117 (.048)	.094 (.038)	.098 (.037)	.015	.015	.009
6	.106 (.041)	.083 (.026)	.086 (.026)	.012	.002	.001
7.5	.082 (.042)	.059 (.033)	.065 (.032)	.055	.077	.043
9	.070 (.045)	.047 (.032)	.050 (.032)	.13	.15	.12
10.5	.104 (.049)	.081 (.039)	.085 (.039)	.037	.041	.030
12	.056 (.046)	.033 (.034)	.036 (.034)	.23	.33	.28

Adopting, for instance, a (test-wise) significance level of .05, we see that, whichever of the three approaches to analysis we choose, we would be able to declare a statistically significant treatment effect at the time-points 1.5, 4.5, 6 and 10.5 months. At 7.5 months only the test utilizing a covariance adjustment would reject the hypothesis of equal treatment means. While for the remaining time-points, at 3, 9 and 12 months, we would not be able to declare a significant treatment difference. Controlling the type I error by a Bonferroni correction, and multiplying all p-values for each approach to analysis by 8, only the outcomes at 1.5 and 6 months, when basing the analysis on changes or covariance adjusted means, would remain statistically significant.

Several criticisms can be made relating to this approach. To begin with, no account in the analysis is taken of the fact that measurements at different time points are from the same subjects, i.e. within-subject correlations are ignored. Also, it is inherent in the design of a repeated measures study that one will be interested in the effect of time, both on mean responses within groups, and on differences in mean responses between groups, these issues are ignored when each time point is analysed separately.

Further, dividing the results into "significant" and "non-significant" introduces an artificial dichotomy into serial data, which, for most biological variables change over time in a smooth and continuous manner. Generally, this is an inappropriate method, unless there are few time points, each of which are of interest in their own right.

6.1.2.2 Hotelling's T^2

Hotelling's T^2 is sometimes used, but this approach addresses the wrong question by not taking the directions of the mean differences at the various time-points into account. Whereas we, in principle, nearly always expect uni-directional departures from the null hypothesis. A further drawback is that baselines are ignored, unless the vector of post-treatment measurements for each subject is substituted for the vector of changes from pre-entry. Hotelling's T^2 is the multivariate analogue of the (square of the) univariate t-statistic. It is also a special case of the multivariate analysis of variance (MANOVA), to be described below. The definition of the test statistic is :

$$T^2 = \left(n_A^{-1} + n_B^{-1}\right)^{-1} (\bar{y}_{A.} - \bar{y}_{B.})^T \hat{\Sigma}^{-1} (\bar{y}_{A.} - \bar{y}_{B.}),$$

where n_A and n_B are the two sample sizes, $\bar{y}_{A.}$ and $\bar{y}_{B.}$ are the mean vectors of the repeated measurements for the two groups, and $\hat{\Sigma}$ is the pooled sample covariance matrix. It can be shown that, with t repeated measurements, $\frac{(n_A + n_B - t - 1)}{t(n_A + n_B - 2)} \cdot T^2 \sim F(t, n_A + n_B - t - 1)$.

For the CPK example we get; T^2=12.58, and hence, $F_{3,143}$=1.50 (p=.16). This example fails to achieve significance while previous methods (e.g. ANCOVA had p=.001, see section 2.5) have indicated a highly significant treatment difference. This confirms our disbelief in the usefulness of Hotelling's T^2 for repeated measures data.

6.1.2.3 Repeated measures ANOVA

Repeated measures data are often submitted to a repeated measures ANOVA, also called a split-plot-in-time ANOVA. This approach is well described in many standard textbooks, see for instance; Fleiss (1986), Milliken (1990), and Crowder and Hand (1990). It is based on a comparison of mean profiles treating time as a factor (in a formal ANOVA sense). This method is based on the model; $y_{ijk} = \alpha + \beta_i + \gamma_j + \delta_{ij} + \eta_{ij} + \varepsilon_{ijk}$,

where α is the overall mean, β_i a fixed group-effect, γ_j a fixed time-effect, δ_{ij} a fixed group-by-time interaction effect, η_{ij} a random between-subjects error, and ε_{ijk} a random within-subjects error. This model should preferably be extended with a covariate adjustment for the pre-entry level, e.g. by addition of the term $\theta \cdot y_{ij0}$. This will be discussed further below, but for simplicity it will not be included in table 6.1.2.

For the analysis, the available information is partitioned into between and within-subject variation. This leads us to an "orthodox" split-plot ANOVA (see, Fleiss, 1986), as illustrated in the table below.

Table 6.1.2: Analysis of variance table for data from a repeated measurements study.

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F-ratio
Groups	g-1	$t\sum n_i\left(\bar{\bar{y}}_{i\cdot} - \bar{\bar{\bar{y}}}_{\cdot\cdot}\right)^2$	MS_G	$F_G = MS_G/MS_S$
Subjects	n.-g	$t\sum\sum\left(\bar{y}_{ij\cdot} - \bar{\bar{y}}_{i\cdot}\right)^2$	MS_S	
Times	t-1	$n_{\cdot}\sum\left(\bar{\bar{y}}_{\cdot j} - \bar{\bar{\bar{y}}}_{\cdot\cdot}\right)^2$	MS_T	$F_T = MS_T/MS_R$
Interaction	(g-1)(t-1)	$\sum\sum n_i\left(\bar{y}_{ik} - \bar{\bar{y}}_{i\cdot} - \bar{\bar{y}}_{\cdot k} + \bar{\bar{\bar{y}}}_{\cdot\cdot}\right)^2$	MS_I	$F_I = MS_I/MS_R$
Residual	(n.-g)(t-1)	$\sum\sum\sum\left(y_{ijk} - \bar{y}_{ij\cdot} - \bar{y}_{i\cdot k} + \bar{\bar{y}}_{i\cdot\cdot}\right)^2$	MS_R	
Total	n.t-1	$\sum\sum\sum\left(y_{ijk} - \bar{\bar{\bar{y}}}_{\cdot\cdot\cdot}\right)^2$		

The first two lines in the table correspond to the between-subjects differences, and the following three to the within-subjects differences. For convenience, the indices for the summations have been left out, n. is short for the total sample size. F_G is the test for a group main effect, it is identical to a test of the equality between groups for the means (or the sums) over time for the subjects (i.e. identical to POST). It is valid irrespective of the nature of the covariance structure.

F_I is the test for a group-by-time interaction. The absence of an interaction implies that the group mean profiles are of the same shape, but they may be at different overall levels. For this test to be strictly valid we would have to be able to randomise the order of the time-points, which is clearly impossible. However, if the covariance structure is such that all normalised contrasts among the repeated measurements have the same variance (sphericity, also termed the Huynh-Feldt type H-structure, after Huynh and Feldt, 1970), the test is still valid (Compound symmetry is a special case of sphericity). If this does not hold an approximate procedure based on "ε-adjusted" degrees of freedom for the F-test has to be used to allow for this departure from the assumed covariance structure (see, Box, 1954). Then, both numerator and denominator degrees of freedom are multiplied by this correction factor, which is confined to lie in the range $(1/(t-1) \leq \varepsilon \leq 1)$. For instance, the test statistic F_I is, under the null hypothesis, assumed to follow a $F_{\varepsilon(g-1)(t-1),\varepsilon(n-1)}$ distribution.

There are two estimates of ε in common use, the Greenhouse-Geisser estimate (Geisser and Greenhouse, 1958, Greenhouse and Geisser, 1959) and the Huynh-Feldt estimate (Huynh and Feldt, 1976). A description of, and contrasting of, these correction factors goes beyond the scope of this thesis. Enough is to say that when the test for sphericity (which is given by most statistical packages) is rejected, adjusted degrees of freedom should be used.

F_T is the test of an overall time effect, averaged over treatment groups. This test is usually of little interest, and for its validity to hold the same assumptions about sphericity as above must hold, otherwise the approximate tests must be used.

Important between-subject covariates (e.g. pre-entry measurements) should, whenever possible, be included in the model. However, these will only enter the between-subjects part of the analysis. I.e. F_G would reproduce an ANCOVA analysis if the pre-entry mean was included in the model, but F_T and F_I would not be affected. Thus, the between-subjects part is identical to a mean summary statistic approach.

Generally, using standard statistical packages, subjects with missing values are excluded from repeated measures ANOVA. To prevent this, when there are only a few missing measurements, some kind of interpolated or estimated values might be substituted for the missing values. Alternatively, non-orthogonal applications of analysis of variance are available. However, a proper use of correction factors for the degrees of freedom is very complicated under these circumstances.

Repeated measures ANOVA has several drawbacks. Firstly, it is restricted to the comparison of mean profiles. The curve joining the means over the time points for a treatment group may not be a good descriptor of a typical curve for an individual. Important variation in the shapes and locations of curves for different subjects may be hidden. Using it, for instance, for peaked data is of dubious value. Secondly, the overall F-statistics ignores totally the time ordering of the data. Permuting the time-points in the same manner for all patients will not affect any of these statistics. Upon finding a significant group-by-time or time effect one is usually recommended, in a typical explorative data-analysis manner, to investigate a set of orthogonal (i.e. independent of each other) contrasts. Normally, one uses polynomial contrasts, and starts by looking at the contrast of (t-1)th degree, if this is non-significant one continues with the (t-2)th, and so on until finding an individual contrast making a significant contribution. When this is found, for say, the (t-k)th degree contrast, one declares that a polynomial function of degree (t-k) is needed to explain the overall changes in response over time (time effect), or the differences in mean changes in response over time between groups (group-by-time interaction).

In textbooks it is "fortunately" mostly the case that one ends up with a linear (or perhaps quadratic) polynomial, which makes interpretation feasible. The topics of multiple testing, power and sensitivity in this process are hardly ever touched upon.

For the CPK example, including the pre-entry mean as covariate, F_G=10.5 which on (1,149) degrees of freedom has p=.001, identical to ANCOVA as already described. F_I is also of interest, this statistic equals 0.78, so there are no indications of a treatment-by-time interaction. F_T (which normally is of little interest) equals 3.01 which on (t-1,(t-1)(n.-g)) = (7,1050) degrees of freedom has a p-value of 0.004. However, the rejection of the sphericity assumption (p<0.0001) necessitates adjustment of the degrees of freedom. The Huynh-Feldt ε equals 0.85 and the Greenhouse-Geisser ε = 0.81, use of the former changes the p-value to 0.007, still highly significant. Looking at figure 2.5.1 one might guess that there is a slow linear increase over time for the CPK-levels, averaged over groups. Let us proceed to test this using polynomial contrasts for the time dimension. The relevant significance tests are summarised in table 6.1.3 below.

Table 6.1.3: Tests of significance for polynomial contrasts over the time dimension (averaged over treatment groups) for the CPK example.

Contrast	F(1,150)	p-value
Linear	2.96	.09
Quadratic	5.36	.02
Cubic	6.34	.01
Quartic	0.00	.99
Quintic	0.68	.41
6th degree	5.74	.02
7th degree	0.01	.93

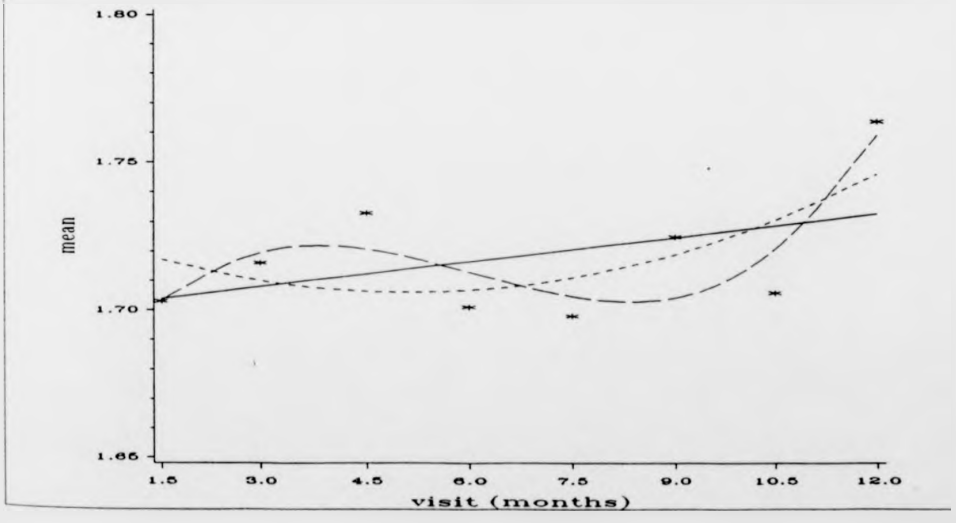
If we choose a conventional 5% level we end up needing a 6th-degree polynomial to describe the pattern of changes over time. Suggesting that such a model is necessary to describe the overall time dependencies is clearly not meaningful. However, to get a feeling for how well polynomials up to the third degree fit this example, figure 6.1.2 is given. We see, from the observed pattern of overall means over time, why the cubic contrast was significant.

Whether a model involving a cubic polynomial is medically meaningful, or simply represents an overreaction to random noise, is a different question.

For this example the group-by-time interaction was not significant, had it been, however, we would have needed a 5th-degree polynomial (the quintic contrast had p=0.04) to describe the differences in pattern of changes over time between the two groups.

In summary, this method provides overall comparisons between mean profiles for treatment groups, which might be useful if one knows little about the structure of the profiles, as a kind of hypothesis generating method. Upon finding a significant overall F-statistic various linear contrasts have to be investigated in order to gain an understanding of what kind of effects are involved. However, when allowing for a proper covariate adjustment for pre-entry levels, to reproduce an ANCOVA analysis for the between-subjects part, and with pre-specified hypotheses regarding contrasts to be tested for the time-dimension, to properly address questions relating to the treatment-by-time interaction, it is a conceivable alternative.

Figure 8.1.2 :
Overall means over time (*) for the CPK-example (n=152), with fitted polynomial curves (by least squares) up to third degree (i.e. linear, quadratic and cubic)



6.1.2.4 Repeated measures MANOVA

A repeated measures MANOVA is a multivariate analysis of variance applied to a repeated measures design. Relative to a repeated measures ANOVA, a different viewpoint is taken in that the t measurements $y_{ij1},\ldots,y_{ijt}$ on individual j in treatment group i are regarded as a single vector observation. With the multivariate approach, the overall comparisons are made in terms of sums of squares and cross-product matrices instead of with sums of squares. F-ratios are replaced by ratios of determinants or other functions of eigenvalues (e.g. Wilks' lambda, Hotelling-Lawley's trace, Roy's largest root or Pillai-Bartlett's trace). These test-statistics are obtained from $H \cdot E^{-1}$, where H stands for the hypothesis matrix (which typically consists of t-1 orthogonal contrasts) and E for the error matrix (for more details see; Fleiss (1986), Hand and Taylor (1987), and Crowder and Hand (1990)).

The MANOVA model, without covariates, may be written as; $\mathbf{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_{ij}$, where $\mathbf{y}_{ij}$ is the observed t-dimensional vector of responses for subject j in treatment group i, $\boldsymbol{\mu}_i$ is the underlying mean vector for group i, and $\boldsymbol{\Sigma}_{ij}$ is the covariance matrix, assumed multivariate normal and identical between groups.

For this case, when no structure is imposed on $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$, the test for an overall treatment effect is identical to the corresponding test in a repeated measures ANOVA (and again identical to a summary statistic approach, e.g. ANCOVA when the pre-entry mean is included in the model). Further, when there are only two treatment groups, the MANOVA tests for treatment-by-time interaction, and for overall time effect, are identical to Hotelling T^2-tests (two-sample, and one-sample, versions, respectively).

In MANOVA a completely general structure is allowed for the covariance matrix (i.e. we do not have to bother about adjustments of degrees of freedom, as in the repeated measures ANOVA). The MANOVA is a very general approach which often is less powerful than its univariate counterpart, the reason being the losses in degrees of freedom caused by the need to estimate $(t+1)t/2$ parameters for the covariance matrix.

Using a standard statistical software package (like PROC GLM in SAS, SAS, 1992) one gets the output structured in much the same way as for repeated measures ANOVA. For the CPK-example the between-subjects part is again identical to what we get by using the summary statistics. The group-by-time interaction is also here non-significant, $F(7,144)=.97$, $p=.46$. For this example MANOVA obtains a more extreme F-statistic for the time main effect than repeated measures ANOVA, $F(7,144)=3.62$, $p=.001$. In summary, this approach, in its standard format, suffers from the same deficiencies as repeated measures ANOVA, though, many refinements are available.

Repeated measures ANOVA and MANOVA can differ markedly in the type of departures from the null hypothesis that they are able to detect. It has sometimes been recommended to perform both at the $\alpha/2$ level (Looney and Stanley, 1989).

6.1.2.5 Ante-dependence analysis

One further approach among the "unstructured" alternatives that will be mentioned is the one labelled ante-dependence analysis. This approach was developed by Kenward (1987) as an improvement on MANOVA. Realising that it is possible to decompose the MANOVA likelihood-ratio statistic (Wilks' lambda) into a product of independent univariate statistics, which are simple functions of analysis of covariance F-ratios. He developed the ante-dependence analysis as a way of saving degrees of freedom for a multivariate analysis. With MANOVA, all of the univariate statistics in the decomposition of the likelihood-ratio test are calculated having one of the individual time-points as dependent variable with all preceding time-points as covariates. This uses up a lot of degrees of freedom and causes, often, an unnecessary loss in power. Kenward suggested that usually it is only a few of the observations immediately preceding the dependent variable that make a real contribution as covariates, by omitting all the earlier ones many degrees of freedom can be saved.

Formally, we are saying that for some value g, $y_k | y_{k-1},\ldots,y_{k-g}$ is independent of $y_{k-g-1},\ldots,y_1$ (k>g). This is the definition of an ante-dependence covariance structure of order g (Gabriel, 1961). The basic principle of the ante-dependence approach to analysis is, first, to test for the order of ante-dependence. Than, significance tests similar in spirit to the likelihood-ratio test can be performed under the suggested degree of ante-dependence. For more details on this approach, see; Kenward (1987), and Crowder and Hand (1990).

6.1.2.6 General linear mixed models

The methods outlined so far have not made any assumptions at all, at the outset of the analysis, about the form or the shape of the profiles (mean treatment curves over time). We will now move on to classes of approaches, which, like the summary statistic approach, incorporates some feature(s) or structure of the profiles (like linear rate of change with time), formally or informally in the analysis.

More sophisticated methods have evolved from the simpler summary statistic approach. Most of these fall into the class of general linear mixed models, where "mixed" stands for a mixture of fixed and random effects in the linear model. General overviews covering this field may be found in Crowder and Hand (1990), Lindsey (1993), and Jones (1993).

Based on the work of Harville (1977), Laird and Ware (1982) proposed a very general linear mixed model for longitudinal data;

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\gamma_i + \varepsilon_i,$$

where $\mathbf{y}_i$ is an n_ix1 response vector for subject i, $\mathbf{X}_i$ is an $n_i$xb design matrix, β is a bx1 vector of regression coefficients assumed to be fixed, $\mathbf{Z}_i$ is an $n_i$xg design matrix for the random effects, γ_i, which are assumed to be independently distributed across subjects with distribution $\gamma_i \sim N(0,\sigma^2\mathbf{B})$, where $\mathbf{B}$ (for between-subjects) is an arbitrary covariance matrix.

The within-subject errors, ε_i, are assumed to be distributed as $\varepsilon_i \sim N(0,\sigma^2\mathbf{W}_i)$, where $\mathbf{W}_i$ (for within-subjects), is a covariance matrix which usually (because the random effects have removed many of the variance components) may be parameterized using a few parameters. Often it is assumed that $\mathbf{W}_i$ is equal to the identity matrix.

This model is very general since different subjects can have different numbers of observations, as well as different observation times. Most of the common statistical models for continuous measurements are special cases of this general linear mixed model.

A simple illustration of a linear mixed model will now be given. Assuming a repeated measures design with 2 treatment groups and 3 visits, the model (without covariates) for subject i in group A is;

$$\begin{bmatrix} y_{Ai1} \\ y_{Ai2} \\ y_{Ai3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \cdot \gamma_{Ai} + \begin{bmatrix} \varepsilon_{Ai1} \\ \varepsilon_{Ai2} \\ \varepsilon_{Ai3} \end{bmatrix},$$

and correspondingly for subject j in group B;

$$\begin{bmatrix} y_{Bj1} \\ y_{Bj2} \\ y_{Bj3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \cdot \gamma_{Bj} + \begin{bmatrix} \varepsilon_{Bj1} \\ \varepsilon_{Bj2} \\ \varepsilon_{Bj3} \end{bmatrix}.$$

With these models arbitrary mean response profiles are accommodated for the two groups, also a random subject effect (γ) is allowed for in the model. For group A the mean response curve over the three visits equals β_1, β_2 and β_3, respectively, while β_4 to β_6 represent the differences between the two mean curves over these three visits.

Attempting to heuristically explain what the Laird and Ware model is about, we might think of the profile from each subject as following some specified functional form, where the parameters of the function are allowed to vary among the subjects. Thus, the regression coefficient vectors, the γ_i, may be viewed as random drawings from some multivariate population.

In effect, the parameters of the function provide the summary statistics. The specification of the model for the profile, together with the error distribution and covariance structure of the repeated measurements, permits a complete formal analysis of the data for a wide variety of data structures. In particular, complex sets of data with irregular and unbalanced times of measurement can be accommodated. General linear mixed models also provide a convenient framework for prediction. For model-building, hypothesis testing, and estimation under this general frame-work, see the references given above.

Work in this area has been published under different headings, like; growth curves, random regression coefficient models (or two-stage models), and multi-level models.

The growth curve model with random effects was developed in a series of papers by C.R. Rao (1959,1965,1968,1987), see also Potthoff and Roy (1964), Grizzle and Allen (1969), and Lange and Laird (1989). Growth curve analysis emphasise the explanation of within-subject variability by the natural developmental or ageing process. In contrast, repeated measurements models typically assume that individual effects remain constant over the time period of interest (e.g. the "true underlying mean").

The multi-level model is an extension of the two-stage model described above. Goldstein (1987) and others have applied multi-level models to education data, where the random terms conveniently may be thought of as appearing at different levels, e.g. individual child, class, school, town, etc. Corresponding levels for a clinical trial might be; individual measurement, visit, period, subject, hospital, and country. For an extended account of these methods, see Goldstein (1987).

6.1.2.7 Other specific approaches

Another specific structured approach is the "latent class model" proposed by Skene and White (1992). They considered situations where several distinct modes of response within a group (e.g. "responders" and "non-responders") were anticipated. In such situations the effect of treatment can be characterized both by the shape of the fitted profiles and by estimating the proportion of cases who exhibit each particular response profile. They suggested how such experiments could be analysed through the introduction of a latent variable into the standard model. This approach has a certain appeal, however, categorizing all subjects as "responders" or "non-responders" introduces an often unrealistic dichotomy for an underlying continuous variable.

6.1.3 Modelling of within-subject dependencies

Repeated measurements on the same subject are inherently stochastically dependent. Ignoring this dependence when modelling the responses results in two problems: inefficient estimation of regression parameters, and, more important, inconsistent estimates of precision. Both need to be avoided.

There exists two useful ways of modelling the stochastic dependence; by random effects, and by variance components. These two strategies will be contrasted in this section.

Modelling with variance components is a direct way of accounting for the statistical dependence. This approach is based on the partition of variation among measurement into two basic types: between-subjects, and within-subjects. A certain structure is then specified for the within-subject covariance matrix. A simple example of such a structure is compound symmetry, which consists of two variance components, the variance, assumed homogeneous among time points, and a correlation, assumed identical between time points. In fact, compound symmetry is equal to a random intercept model, that is for a model with a random effect for the intercept, all correlations would be zero. This is a general relationship, any random effect model can be recast as a fixed-effect model with a complex response covariance pattern (Muller and LaVange, 1992).

To express this in another way, variance components can mop up unexplained variation, the covariance structure can be thought of as a surrogate for unmeasured (and perhaps unmeasurable) covariates (Louis, 1988).

Instead of directly modelling the stochastic dependence with variance components, one might assume that the subjects were chosen at random from some large population. Then, the parameters in the model which describe the differences among the subjects are taken to vary over this population, according to some distribution. This is known as a random effects model (Lindsey, 1993), and the correlations among responses for an individual are assumed to arise from natural heterogeneity in regression coefficients (for the effects included in the model) among people. Given knowledge of the true response pattern for an experimental unit, i.e. including all relevant covariates, the measurement process gives rise to independent errors. If there still exists autocorrelation among repeated measurements in a random effects model, this may be a consequence of under-specification of the mean structure for the experimental units of each treatment group (see; Skene and White, 1992, and Selwyn and DiFranco, 1993). The flexibility of random coefficients models may make them useful for certain data with peculiar covariance structures.

But due to the necessary arbitrariness in choosing which coefficients are random, they usually do not provide a satisfactory description of the underlying process generating the data. If some structure is known to exist in the covariance matrix, it is most often preferable to model it directly.

In summary, directly modelling the stochastic dependence structure, as in a variance components model, and assuming that some parameter distinguishing the units has a random distribution, as in the random effects model, can both result in essentially the same model. In fact, the random effects model is more limited, since for that model the covariances are restricted to be non-negative (Lindsey, 1993).

6.1.4 Relevance to practical research

The approaches categorized as "unstructured" have usually little to offer in relation to comparative clinical trials. When emphasis is on exploratory data analysis, and the generation of hypotheses, they are more relevant, but in most circumstances different types of graphical displays will serve these purposes better.

The overall test for a treatment-by-time interaction (e.g. F_I in section 6.1.2.3) has some appeal, but if different shapes of mean responses over time between groups is anticipated, this is more directly tested using a SLOPE or SLANC analysis (on transformed data if necessary). How one should find a scientifically meaningful interpretation of a significant difference in (for instance) a fifth degree polynomial contrast over time between groups goes beyond my imagination.

It is more meaningful to contrast the summary statistic approach with the more sophisticated general linear mixed models. The generalizability for this latter class of models is impressive. There may be an arbitrary mixture of fixed and random effects, the error structure may be taken to have any specified form, and if necessary non-linear models may be used (see, Berkey, 1982).

Obvious advantages are; allowance for missing values and irregular measurements, flexibility in modelling (interactions, time-dependent covariates, etc.), and the full range of statistical inference (goodness-of-fit, formal comparison of models, and so on). Among the disadvantages we have; greater reliance on asymptotic properties, difficulties in correct specification of the model, and complexity.

Maybe the complexity is the central issue, it is both a great strength (as noted above) and a definite weakness. Aspects contributing to making the complexity a negative characteristic include;

- How much of what these more sophisticated approaches do will it be possible to communicate to non-statisticians ?
- What is likely to be included in a clinical study report, accepted by authorities, or included in a medical journal ?
- What is really generalizable from a very complicated model ?
- In the next similar clinical trial will a completely different model be chosen when using the same model-selecting procedure ? i.e. how sensitive is the final model to relatively small changes in the observed data ?
- With such a wide range of possibilities in selecting a model and performing an analysis, is it not possible to squeeze the data until it confesses almost any desired p-value ?
- In effect, how many significance tests are really performed in the process of building and analysing a linear mixed model, what about type I and II errors ?
- Also, given that data in repeated measures designs often is very limited, model verification/discrimination is difficult. Probably these methods are best suited for large important data sets.

The trade-off between the simpler summary statistic approach and the more complex modelling approaches is a familiar one in statistics. However, in most standard applications the summary statistic approach is more than adequate, and should in general be the method of choice for repeated measures designs.

6.2 NEEDS FOR FURTHER METHODOLOGY

6.2.1 Extension of the summary statistic approach

A linear mixed model with one random effect (the intercept for each subject), for data which is complete and balanced, is equivalent to a summary statistic approach. When there are several random effects (e.g. intercept, slope and curvature) the linear mixed model (in the balanced case) might be termed a multivariate summary statistic approach. Thus, in many ways the natural extension of the summary statistic approach leads us to the class of general linear mixed models. However, in this subsection we will concentrate on discussions of possible useful extensions of the summary statistic approach, while trying to preserve simplicity.

6.2.1.1 Missing values

Missing values are common in any clinical trial, repeated measures designs are, by nature, especially susceptible to this problem. Missingness may occur for several reasons, like; occasionally missed visits, non-informative and informative drop-outs, treatment-related withdrawals, end-of-study censoring, observation missing due to technical failure, too sick to attend, etc. Whether the missingness is assumed to be random (ignorable) or not is an important distinction which will greatly affect our possibilities of making valid analyses.

Little and Rubin (1987) distinguish between observations that are "missing completely at random" (MCAR), where the probability of missing an observation is independent of both the observed responses and the missed responses, and "missing at random" (MAR), where the probability of missing an observation is independent of the missed observations.

Missing data is an extensive research topic on its own (see, e.g., Diggle and Kenward, 1994), and hence only some of the possible alternatives to the handling of missing values in the context of summary statistics for repeated measures will be mentioned. More explicitly, the following four alternative approaches will be given some consideration;

1) Include only patients with complete series of measurements.
2) Include only time-points where all (most) data are available.
3) Include all patients and time-points irrespective of missingness (through an all-encompassing model to estimate the summary statistics with missing data).
4) Use a stratification by missingness pattern.

These four approaches for the handling of missing data rely on slightly different assumptions for their validity. As a preliminary, we may note that it will usually still be possible to calculate the summary statistics in the presence of missingness. However, the assumption of identical distribution for the summary statistics in the population will not hold when they are derived using subjects with different numbers and choices of measurements. Departure from this assumption may often, but not always, be negligible. We also have to pay attention to the correctness of our underlying model. Often, missing data are more common towards the end of the study. Consider this to be the case, and suppose that we are using a mean summary statistic approach to the analysis. Then, if the treatment effect declines (or rises) towards the end of the study, and if the amount of missingness is unequal between groups, we will most likely end up with biased results.

Approach 1) is often used because of its simplicity. For some other approaches, like repeated measures ANOVA, it is used by necessity (unless we substitute some kind of interpolated or estimated values for the missing measurements). We do not have to worry about the identical distribution of the summary statistics, and calculations are simple, but for its strict validity we have to make the most restrictive missingness assumption, MCAR. MAR is not enough, since if missingness is allowed to depend on the observed values of a subject, and if missingness, as is often the case, is more common among low(poor)-scoring subjects, this will bias the results in favour of the less efficient treatment.

Approach 2) is similar in spirit to approach 1), but instead of excluding subjects with incomplete series of measurements, we exclude visits with missing observations. If there is a lot of end-of-study censoring, excluding one or a few measurements near the end is often a realistic approach. This may of course slightly alter the conclusions one will be able to draw from the study (if the effective study period decreases substantially). In principle, this approach only has to assume MAR (unless, perhaps, when the choice of exclusion is data dependent/provoked), since we are not making any conclusions relating to the time-points excluded. As a follow-up to excluding some time-point(s) it might still be necessary to use one of the other approaches on the remaining data, since there usually will be some missingness also for the earlier time-points.

One of the advantages with the summary statistic approach is that we do not have to throw out patients with a few missing measurements, we may still calculate our summary statistics (e.g. as long as there are any valid measurements post-randomisation for a subject we may calculate the post-treatment mean). Approach 3) is, thus, a possible alternative. Again, MAR is enough for validity, but as pointed out we have to pay attention to the correctness of our model. Also, with substantial missingness the assumption of identical distributions for the summary statistics will not apply. In these circumstances a weighted analysis might be appropriate (for instance, each subjects summary statistic could be given a weight according to its estimated precision).

Matthews (1993) considered this in relation to data on foetal distress during labour when analysing the rate of change (using SLOPE) over time for the subjects (babies). His results suggests that a weighted analysis can increase both efficiency and validity when the assumption of identically distributed summary statistics is violated. In particular, use of robust weights (trimmed linear estimators of the dispersions for the individual SLOPEs) seemed promising.

Approach 4), stratification by missingness pattern, has been suggested by Dawson and Lagakos (1991,1993). They suggested that one should divide the subjects into strata formed by the different missing data patterns, such that the summary statistics within each stratum all have the same distribution under the null hypothesis. For each strata, g, a standardized statistic

$$Z_g = \left(U_{gA\cdot} - \left\{ \frac{n_A}{n_A + n_B} \right\} U_{g\cdot\cdot} \right) / \hat{\sigma}_{U_{g\cdot\cdot}}$$

can be calculated. Here, $U_{gA\cdot}$ denotes the sum of the summary statistics for the subjects in groups A (arbitrary choice) for stratum g, and $U_{g\cdot\cdot}$ denotes the corresponding sum over both groups.

Further, $\hat{\sigma}_{U_{g\cdot\cdot}} = \frac{n_A \cdot n_B}{(n_A + n_B)(n_A + n_B - 1)} \sum_i \sum_j \left(S_{ij} - \frac{U_{g\cdot\cdot}}{n_A + n_B} \right)^2$,

where S_{ij} is the summary statistic for subject j in group i. These standardized statistics can then be combined into an overall statistic; $Z_{strat} = \sum_g w_g Z_g \Big/ \sqrt{\sum_g w_g^2}$, where g indexes the missingness strata, and the w_g's are the weights. The choice of the weights do not affect the validity of the test, but they will influence the power. Some choice reflecting the relative precision of estimates of treatment effects in each stratum is recommended. For this stratified approach to be valid we only have to assume that the summary statistics are equally distributed conditional on the missingness patterns.

Dawson and Lagakos (1991) gives two important cases when this applies; when the missingness in non-informative (i.e. MAR), and when the probability of missingness for a given level of the measurements is the same for the two groups. These validity criteria are less restrictive than what is required for an unstratified analysis. In conclusion, this stratified approach to analysis seems promising for hypothesis testing, however, it does not appear to be suited for estimation.

6.2.1.2 Graphical display of results from repeated measures trials

Graphical displays are a very effective way to visualize data and illustrate relationships. At the beginning of an analysis graphs may be used for: getting a feeling for the data, spotting potential outlying observations, suggesting relationships between variables, and for checking distributional assumptions. In this short account on graphical displays, emphasis will be on how to appropriately display the data and illustrate the conclusions at the reporting stage for a repeated measures trial.

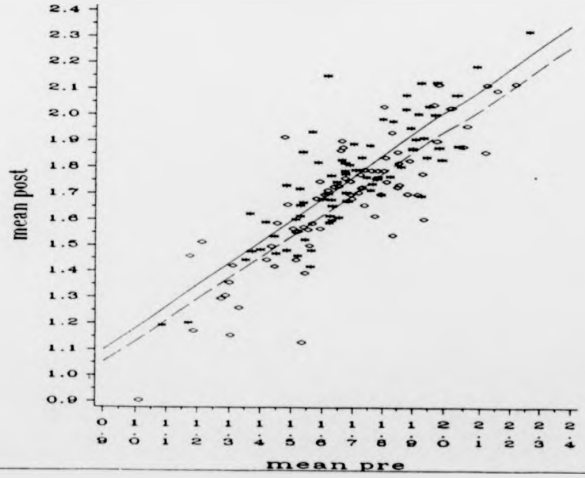
When dealing with typical multi-dimensional data, care is needed in reducing the dimensionality for illustrative purposes (as well as for the analysis), not to lose important aspects of the underlying relationships. When the sample sizes so allow (say, $n_A+n_B \leq 30$), individual graphs of response against time is a satisfactory option. Mostly this will not be feasible, certainly not for publication purposes. An alternative is to classify the curves into typical patterns, and to plot representative examples. However, care is needed to ensure that a biased selection has not taken place. This approach has a natural link to the latent class model referred to earlier in this chapter.

Most commonly mean curves for the treatment groups over time are given. This is often useful and enables large quantities of data to be plotted in a concise and meaningful way. A drawback is that correlations between time-points effectively are ignored. Also, caution is needed, since sometimes mean curves are not representative of any typical subjects.

Consider, for instance, peaked curves, where the peaks occurs at different time-points for different subjects. The mean of a sample of such curves is not likely to be meaningful, and may often be very misleading. This issue was considered by Matthews et al (1990). They suggested that a useful alternative in such peaked data is to plot the maximum for each subject against the time that the maximum occurred. This leads to the general recommendation that whatever summary statistics are felt appropriate for the analysis should also be illustrated graphically.

This implies that when a mean summary statistic approach seems appropriate, it should also be appropriate to display mean curves for treatment groups over time, and distributions of patient means by treatment groups. For instance, if ANCOVA is used, it is useful to plot post-treatment means versus pre-treatment means by groups with drawn in regression lines. This is illustrated, using the CPK-example, in figure 6.2.1.

Figure 6.2.1 :
CPK-example. post means versus pre means by treatment group. Separately fitted regression lines (i.e. not constrained to be parallel) are shown for group A (solid line, stars) and group B (dashed line, diamonds).



It is also valuable to illustrate the correlation structure and how the variability changes over time (see figures 2.5.2 and 1.5.2). One problem that may arise in relation to mean curves is when the number of subjects changes over time. Frequently subjects faring less favourably tend to withdraw at a higher rate. If this rate is non-negligible misleading conclusions (imposing a conservative bias towards the more effective treatment) are likely to be drawn from the figure.

One informative method, not least for skewed data, is to display boxplots over time. For each measurement occasion one boxplot is given for each treatment group, displaying, for instance; first, second and third quartile with a box, tenth and ninetieth percentile with the whiskers, and finally (when falling outside the whiskers) the five lowest and five highest measurements with stars. This way a feeling not only for how the centre of the distributions changes over time, but actually how the whole distributions changes over time, may be conveyed to the reader. An example of this type of graphical display is given in figure 6.2.2, utilizing data from the PD_{20}-example in table 1.5.1 (selecting the subjects with complete series of measurements up to the third visit post-randomisation). This graph is accompanied, for comparative purposes, by a "standard" mean curves graph. One drawback with the boxplots is that there are no links between occasions for individuals. A partial solution might be to substitute the stars, indicating the outliers, for subject numbers (or some other labels identifying individual subjects).

A final warning aimed at the plotting of arithmetically related variables will be given. Frequently figures displaying the changes (post-pre) versus the pre-entry measurements are given. The conclusion accompanying such a figure will almost without exception be that subjects with less favourable pre-entry levels experienced the largest improvements. This is potentially very misleading, the correlation between pre-entry values and changes will almost always have an (at least partly) arithmetic interpretation, with $-1/\sqrt{2}$ expected in the "null" case (a typical case of regression to the mean, see chapter 4 for more details).

Figure 6.2.2a : Boxplots for PD20, drug A (n=22), drug B (n=50)

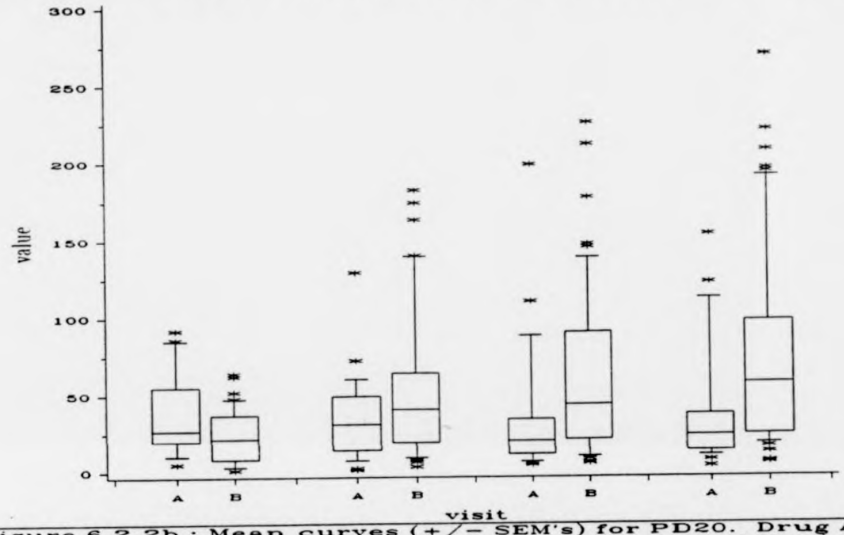
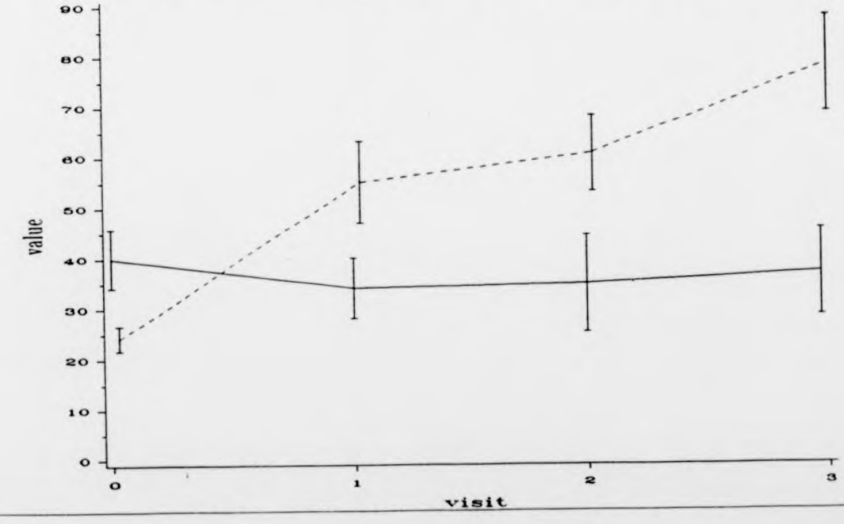


Figure 6.2.2b : Mean curves (+/- SEM's) for PD20. Drug A (n=22, solid line), drug B (n=50, dashed line).



A citation from Senn (1989) will finalize this warning, "The best advice is to avoid investigating relationships between consanguineous measures: in statistics as in biology, incest breeds freaks.".

6.2.1.3 Some further topics

Irregular measurements, i.e. when different subjects are evaluated at different points in time, poses problems similar to when there are missing values. We will (normally) still be able to calculate the summary statistics, but the validity of the analysis might be affected adversely. With very irregular patterns of measurement (unusual) the assumption of identically distributed summary statistics will be violated. Of possibly greater importance is the reliance we have to make on our model for the mean effects over time. Consider, for instance, a study where the linear rate of change over time is of interest, but where this rate maybe is not constant over time. If some subjects are measured primarily early on in the study, and others primarily during the later phases, then these subjects might certainly have different expected values for the summary statistics, irrespective of possible treatment effects. However, deliberately irregular measurements are unusual, but some variation in actual times around planned visits is common. It is important to consider when one may ignore these irregularities, and when one has to incorporate them into the analysis. How common irregular measurements are in practice, and how one should deal with them when they are of concern, is a subject for further research.

Cross-over experiments form an extensive topic in their own right. Even in their simplest version, the two-period cross-over design, they involve repeated measures on each subject. With more complex designs, this relationship become more obvious. Also, it is often the case that more than one measurement is taken on each subject in each period. There is much scope for work to be done on the use of summary statistics for this type of design. For an extensive overview on the topic of cross-over experiments, the book by Jones and Kenward (1989) is recommended reading.

Of the available armoury of summary statistics, this thesis has almost exclusively been aimed towards linear summary statistics. Further corresponding research is needed for the non-linear alternatives, like; maximum response, time to reach maximum, time above a certain threshold, etc. In particular, recommendations for the design of such studies, and ways to improve the efficiency of their analysis using non-linear summary statistics, is of great potential interest.

A topic that was touched upon in section 3.3 was the issue of non-equal covariance matrices and treatment-by-time interactions. Further research is needed in this area, both to investigate the sensitivity of common approaches to design and analysis on modest departures from these assumptions, and for finding remedies when covariance matrices obviously are not equal, like logarithmic transformations when effects are multiplicative.

Binary and categorical data appear frequently in repeated measures studies. Modelling treatment effects and within-subject dependencies are complicated by the fact that there is no "natural" equivalent to the multivariate normal distribution for these types of data. Most analyses are based, in one way or another, on the multinomial distribution, for which the same parameters occurs in both the first and second order moments of the distribution. This implies that no model can simultaneously achieve useful expressions for the joint, marginal, and conditional distributions, the interpretation of the model parameters will depend on this choice, and no approach will always be correct (Kenward and Jones, 1992). Relevant articles in this field have been written by; Agresti (1989), Ware, Lipsitz and Speizer (1988), and Kenward and Jones (1992). Considering the inherent problems in dealing with within-subject dependencies for repeated measures categorical data, there is much need for work on summary statistics in this area.

Finally, we have the issue of multiple repeated measures outcomes, e.g. when there are several response outcomes of interest being measured repeatedly over the course of the study. Examples of this are blood pressure studies with systolic BP and diastolic BP, and asthma studies with FEV_1 and PD_{20}.

The production of statistically valid and powerful analyses under these circumstances, that still are simple enough to be clinically meaningful, is a difficult balance and a true challenge.

6.3 CONCLUDING REMARKS

In most randomised clinical trials with a quantitative outcome measure subjects are assessed more than once to evaluate efficiency and safety aspects relating to the underlying medical question(s). One (or more) pre-randomisation visit(s) are usually performed, and they are accompanied by at least one (usually several) measurement(s) during treatment. Unfortunately, the analyses of such a repeated measures design in published reports of clinical trials are commonly rather inefficient, and sometimes even misleading. In particular a correct use of baseline measurements is often neglected. In some instances they are ignored, at other times simple post-pre changes are used, both these choices being clearly inferior to proper analysis of covariance adjustments.

When multiple post-treatment measurements are available, they are often analysed separately, thus, ignoring the repeated measures aspect of the design, and the within-subject dependencies. Apart from not properly addressing hypotheses relating to the time dimension, this approach also imposes a loss in power. Alternatively, repeated measures ANOVA is sometimes used, but this often fails to address the most relevant hypotheses in a sufficiently direct way.

The summary statistic approach has recently become increasingly popular, and I believe is the method of choice for most repeated measures designs. Among the attractive features, we find:

1) Validity; No assumptions are needed about the covariance structure among the repeated measurements for the validity of the analysis.

2) Sensitivity; By extracting information from all available repeated measurements into a summary statistic, and by reducing the random error by a covariance adjustment, powerful analysis are obtained.

3) Specificity; By an appropriate choice of summary statistic the primary objective of the trial may be addressed in a direct and meaningful way.

4) Simplicity; The results arrived at are readily interpretable and allow for an effective communication of the essential clinical trial findings.

This thesis has given specific advice on which summary statistic to choose under any given circumstances, to arrive at efficient and valid analyses. In particular, when a constant difference in treatment effects over time is anticipated, ANCOVA has been shown to be the method of choice. Similarly, under linear divergence between treatments over time, SLANC is the recommended approach to analysis. Further, explicit methods have been defined for comparing the relative efficiencies of different approaches to the analysis under any specified design considerations and anticipated alternative hypotheses. Also, recommendations have been made on how to design repeated measures trials with a view to maximizing statistical power and/or minimizing the required number of subjects, paying particular regard to the choice of the number of pre and post-treatment measurements.

In conclusion, I hope that the methods conveyed by this thesis will prove to be useful to many statisticians involved in the planning and analysis of repeated measures trials.

References

R.P. Abelson and J.W. Tukey. Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. Annals of Math Stat, 34:1347-1369, 1963.

A. Agresti. A survey of models for repeated ordered categorical response data. Stat in Med, 8:1209-1224, 1989.

C.S. Berkey. Comparison of two longitudinal growth models for pre-school children. Biometrics, 38:221-234, 1982.

N. Blomqvist. On the relation between change and initial value. JASA, 82:746-749, 1977.

G.E.P. Box. Some theorems on quadratic forms applied in the study of analysis of variance problems: I, Effects of inequality of variance in the one-way classification. Annals of Math Stat, 25:290-302, 1954.

T.E. Bradstreet. Using Orthogonal Polynomial Scores in Summarizing and Evaluating Longitudinal Data Collected in Phase I and II Clinical Pharmacology studies. Stat in Med, 12:633-643, 1993.

R.J. Caroll. Covariance analysis in generalized linear measurement error models (+ Discussion). Stat in Med, 8:1075-1093,1107-1108, 1989.

L.E. Chambless and J.R. Roeback. Methods for assessing difference between groups in change when initial measurement is subject to intra-individual variation. Stat in Med, 12:1213-1237, 1993.

C. Chatfield and A.J. Collins. Introduction to multivariate analysis. Chapman and Hall, London, 1980.

S. Chen and C. Cox. Use of baseline data for estimation of treatment effects in the presence of regression to the mean. Biometrics, 48:593-598, 1992.

E.M. Chi. Analysis of longitudinal data by models with random effects and AR(1) errors. Drug Inf J, 24:659-670, 1990.

W.G. Cochran. Analysis of covariance its nature and uses. Biometrics, 13:261-281, 1957.

D.R. Cox and D.V. Hinkley. Theoretical Statistics. Chapman and Hall, London, 1974.

D.R. Cox and P. McCullagh. Some aspects of analysis of covariance. Biometrics, 38:541-561, 1982.

M.J. Crowder and D.J. Hand. Analysis of repeated measures. Chapman and Hall, London, 1990.

G.R. Cutter. Some examples for teaching regression toward the mean from a sampling viewpoint. Am Stat, 30:194-197, 1976.

P. Das and P.G.H. Mulder. Regression to the mode. Stat Neerlandica, 37:15-20, 1983.

C.E. Davis. The effect of regression to the mean in epidemiologic and clinical studies. Am J of Epid, 104:493-498, 1976.

C.E. Davis. Regression to the mean. Enc of Stat Sc (Vol 7), eds. N.L. Johnson and S. Kotz, New York: John Wiley, pp. 706-608, 1986.

J.D. Dawson and S.W. Lagakos. Analyzing laboratory marker changes in AIDS clinical trials. J of AIDS, 4:667-676, 1991.

J.D. Dawson and S.W. Lagakos. Size and power of two-sample tests of repeated measures data. Biometrics, 49:1022-1032, 1993.

J.E. Diem and J.R. Liukkonen. A comparative study of three methods for analysing longitudinal pulmonary function data. Stat in Med, 7:19-28, 1988.

P.J. Diggle. An approach to the analysis of repeated measurements. Biometrics, 44:959-971, 1988.

P.J. Diggle and M.G. Kenward. Informative drop-out in longitudinal data analysis. JRSS-B, 43:49-93, 1994.

N.R. Draper and H. Smith. Applied Regression Analysis. Wiley, New York, 1981.

F. Ederer. Serum cholesterol changes: effects of diet and regression toward the mean. J Chron Dis, 25:277-289, 1972.

M.J. Egger, M.L. Coleman, J.R. Ward, J.C. Reading, and H.J. Williams. Uses and abuses of analysis of covariance in clinical trials. Contr Clin Tr, 6:12-24, 1985.

R.A. Fisher. The utilization of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179-188, 1936.

J.L. Fleiss. Design and analysis of clinical experiments. John Wiley, New York, 1986, Chapter 7.

B.W. Flury. Understanding partial statistics and redundancy of variables in regression and discriminant analysis. Am Stat, 43:27-31, 1989.

L. Frison and S.J. Pocock. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for designs. Stat in Med, 11:1685-1704, 1992.

K.R. Gabriel. The model of ante-dependence for data of biological growth. Bulletin Institut International Statistique (Paris), v39:253-264, 1961.

F. Galton. Typical laws of heredity. Nature, 15:512-514, 1877.

F. Galton. Regression towards mediocrity in hereditary stature. J of the Antr Inst, 15:246-263, 1885.

M.J. Gardner and J.A. Heady. Some effects of within-person variability in epidemiological studies. J Chron Dis, 26:781-795, 1973.

S. Geisser and S.W. Greenhouse. An extension of Box's results on the use of the F-distribution in multivariate analysis. Annals of Math Stat, 29:885-891, 1958.

H. Goldstein. Multilevel models. Griffin, Oxford, 1987.

S.W. Greenhouse and S. Geisser. On the methods in the analysis of profile data. Psychometrica, 24:95-112, 1959.

J.E. Grizzle and D.M. Allen. Analysis of growth and dose response curves. Biometrics, 25:357-381, 1969.

M. Gumpertz and S.G. Pantula. A simple approach to inference in random coefficient models. Am Stat, 43:203-210, 1989.

D.J. Hand and C.C. Taylor. Multivariate Analysis of Variance and Repeated Measures. Chapman and Hall, London, 1987.

Handbook of Mathematical Fuctions. Eds. M. Abramowitz and I.A. Stegun. Dover, New York, 1970.

D.A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. JASA, 72:320-340, 1977.

R.J. Hayes. Methods for assessing whether change depends on initial value. Stat in Med, 7:915-927, 1988.

B.E. Huitema. The Analysis of Covariance and Alternatives. John Wiley, New York, 1980.

H. Huynh and L.S. Feldt. Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. JASA, 65:1582-1589, 1970.

H. Huynh and L.S. Feldt. Estimation of the Box correction for degrees of freedom for sample data in randomised block and split-plot designs. J of Educational Statistics, 1:69-82, 1976.

K.E. James. Regression toward the mean in uncontrolled clinical studies. Biometrics, 29:121-130, 1973.

R.A. Johnson and D.W. Wichern. Applied Multivariate Statistical Analysis. Prentice-Hall, New Jersey, 1988.

W.D. Johnson and V.T. George. Effect of regression to the mean in the presence of within-subject variability. Stat in Med, 10:1295-1302, 1991.

B. Jones and M.G. Kenward. Design and Analysos of Cross-Over Trials. Chapman and Hall, London, 1989.

R.H. Jones. Longitudinal Data with Serial Correlation: A State-space Approach. Chapman and Hall, London, 1993.

M.G. Kenward. The use of fitted higher-order polynomial coefficients as covariates in the analysis of growth curves. Biometrics, 41:19-28, 1985.

M.G. Kenward. A method for comparing profiles of repeated measurements. Appl Stat, 36:296-308, 1987.

M.G. Kenward and B. Jones. Alternative approaches to the analysis of binary and categorical repeated measurements. J Biop Stat, 2:137-170, 1992.

P.A. Lachenbruch and M. Goldstein. Discriminant analysis. Biometrics, 35:69-85, 1979.

N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. Biometrics, 38:963-974, 1982.

N.M. Laird and F. Wang. Estimating rates of change in randomized clinical trials. Contr Clin Tr, 11:405-419, 1990.

N. Lange and N.M. Laird. The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. JASA, 84:241-247, 1989.

F.B. Leech and M.J.R. Healy. The analysis of experiments on growth rate. Biometrics, 15:98-106, 1959.

E.L. Lehmann. Theory of point estimation. John Wiley, New York, 1993.

J.K. Lindsey. Models for Repeated Measurements. Oxford Science Publications, Oxford, 1993.

R.J.A. Little and D.B. Rubin. Statistical Analysis with Missing Data. Wiley, New York, 1987.

S.W. Looney and W.B. Stanley. Exploratory repeated measures analysis for two or more groups. Review and update. Am Stat, 43:220-225, 1989.

T.A. Louis. General methods for analysing repeated measures. Stat in Med, 7:29-45, 1988.

G.A. MacGregor, G.A. Sagnella, and K.D. MacRae. Misleading paper about misleading statistics. Lancet, 1985:926-927, 1985.

J.N.S. Matthews, D.G. Altman, M.J. Campbell, and P. Royston. Analysis of serial measurements in medical research. Br Med J, 300:290-295, 1990.

J.N.S. Matthews. A refinement to the analysis of serial data using summary measures. Stat in Med, 12:27-37, 1993.

G.A. Milliken. Analysis of repeated measures designs. in D.A. Berry (ed.), Statistical methodology in the pharmaceutical sciences, Dekker, New York, 1990.

K.E. Muller, L.M. LaVange, S. Landesman-Ramey, and C.T. Ramey. Power calculations for general linear multivariate models including repeated measurements applications. JASA, 87:1209-1226, 1992.

A. Muñoz, V. Carey, J.P. Schouten, M. Segal and B. Rosner. A parametric family of correlation structures for the analysis of longitudinal data. Biometrics, 48:733-742, 1992.

P.C. O'Brien. Procedures for comparing samples with multiple endpoints. Biometrics, 40:1079-1087, 1984.

P.D. Oldham. A note on the analysis of repeated measurements of the same subjects. J Chron Dis, 15:969-977, 1962.

J.E. Overall and K.N. Magee. Directional baseline differences and type I error probabilities in randomized clinical trials. J Biop Stat, 2:189-203, 1992.

S.J. Pocock, N.L. Geller, and A.A. Tsiatis. The analysis of multiple endpoints in clinical trials. Biometrics, 43:487-498, 1987.

R.F. Potthoff and S.N. Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51:313-326, 1964.

C.R. Rao. Some problems involving linear hypotheses in multivariate analysis. Biometrika, 46:178-202, 1959.

C.R. Rao. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. Biometrika, 52:447-458, 1965.

C.R. Rao. A note on a previous lemma in the theory of least squares and some further results. Sankhya, Ser. A, 30:259-266, 1968.

C.R. Rao. Linear Statistical Inference and Its Applications. Wiley, New York, 1973.

C.R. Rao. Prediction of future observations in growth curve models. Statistical Science, 2:434-471, 1987.

A.C. Rencher. The contribution of individual variables to Hotelling's T^2, Wilks lambda, and R^2. Biometrics, 49:479-489, 1993.

D.J. Roe and E.L. Korn. Time-period effects in longitudinal studies measuring average rates of change. Stat in Med, 12:893-900, 1993.

J.R. Roeback, J.R. Cook, H.A. Guess and J.F. Heyse. Time-dependent variability in repeated measurements of cholesterol levels: clinical implications for risk misclassification and intervention monitoring. J Clin Epid, 1993 (in press).

H. Rouanet and D. Lépine. Comparison between treatments in a repeated measurement design: Anova and multivariate methods. Br J of Math and Stat Ps, 23:147-163, 1970.

J.G. Rowell and D.E. Walters. Analysing data with repeated observations on each experimental unit. J Agric Sc, 87:423-432, 1976.

P. Royston and S.G. Thompson. Comparing non-nested regression models. Biometrics, 1994 (in press).

SAS Institute Inc. SAS User's Guide: Statistics, Version 6 Edition. Cary, North Carolina, SAS Institute Inc., 1992.

SAS Institute Inc. SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07. Cary, North Carolina, SAS Institute Inc., 1992.

J.J. Schlesselman. Planning a longitudinal study: II. Frequency of measurement and study duration. J Chron Dis, 26:561-570, 1973.

M.R. Selwyn and D.M. Difranco. The application of large Gaussian mixed models to the analysis of 24 hour ambulatory blood pressure monitoring data in clinical trials. Stat in Med, 12:1665-1682, 1993.

S. Senn. Covariate imbalance and random allocation in clinical trials Stat in Med, 8:467-475, 1989.

S. Senn. Regression: A new mode for an old meaning?. Am Stat, 44:181-183, 1990.

S. Senn. Letters to the editor, Re : R.J. Caroll's paper (1989). Stat in Med, 9:583-586, 1990.

S. Senn. Using baselines in analysing clinical trials. Proceedings of the Fourteenth Meeting of the International Society for Clinical Biostatistics, Printed by Elitian Ltd, Mill Road, Cambridge, England, 1993.

A.M. Skene and S.A. White. A latent class model for repeated measurements experiments. Stat in Med, 11:2111-2122, 1992.

G.W. Snedecor and W.G. Cochran. Statistical methods. Iowa State Press, 1989, Chapter 14.

D-I Tang, N.L. Geller and S.J. Pocock. On the design and analysis of randomized clinical trials with multiple endpoints. Biometrics, 49:23-30, 1993.

S.G. Thompson and S.J. Pocock. The variability of serum cholesterol measurements: implications for screening and monitoring. J Clin Epid, 43:783-789, 1990.

J.W. Tukey. Tightening the Clinical Trial. Contr Clin Tr, 14:266-285, 1993.

E.E. Van-Essen-Zandvliet, M.D. Hughes, H.J. Waalkens, E.J. Duiverman, S.J. Pocock, and K.F. Keerebijn. Effects of 22 months of treatment with inhaled corticosteroids and/or beta-2-agonists on lung function, airway responsiveness, and symptoms in children with asthma. The Dutch Chronic Non-specific Lung Disease Study Group. American Review of Respiratory Disease, 146(3):547-554, 1992.

S. Wallenstein. Readers reaction: Regression models for repeated measurements. Biometrics, 38:849-850, 1982.

J.H. Ware, S. Lipsitz, and F.E. Speizer. Issues in the analysis of repeated categorical outcomes. Stat in Med, 7:95-107, 1988.

J. Wishart. Growth-rate determination in nutrition studies with the bacon pig, and their analysis. Biometrika, 30:16-28, 1938.

F. Yates. Readers reaction: Regression models for repeated measurements. Biometrics, 38:850-853, 1982.

ERRATA, to the thesis:

"Analysis of repeated measures in clinical trials using summary statistics", by Lars Frison 1994.

Page	Row	It says:	It should say:
3	-3	learning	teaching
19	2	ρ'	ρ^{t-1}
19	5	to	too
28	3	is the clinical objective	if the clinical objective is
35	7	k = -(p-1) ... 0	k = 1 ... r
35	7	l = 1 ... r	l = -(p-1) ... 0
38	-2	50 ≥ subjects	≥ 50 subjects
42	1	admittingly	admittedly
56	-7	to	too
73	3	covenient	convenient
74	4	than	then
77	-10	to	too
81	-4	figure 4.2.1b	figure 2.4.1b
93	-3	If we define	We now define
99	11	do	does
103	9	the variable	the covariate
106	8	are	our
118	-14	$\hat{p}$	p'
122	-1	wile	while
128	4	$E[\zeta_2]=\frac{2\cdot(n_A+n_B)}{n_A\cdot n_B\cdot(n_A+n_B-2)}\cdot\frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}\approx\frac{2}{n_A\cdot n_B}\cdot\frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}$	Should say: $E[\zeta_2]=\frac{2\cdot(n_A+n_B)}{n_A\cdot n_B\cdot(n_A+n_B-2)}\approx\frac{2}{n_A\cdot n_B}$
128	5	is at least twice	is expected to be twice
128	-3	16 subjects	11 subjects
128	-2	17 or more	12 or more
129	4	to 430	to 277
129	6	gives	given
130		Figure 3.6.2 should be replaced by the new figure 3.6.2 given below.	
131	4	γ_{31}	γ_3
133	10	give	given
138	-7	to	too

Page	Row	It says:	It should say:
139	-12	one's	ones
139	-13	one's	ones
147	-11	true	observed
156	-2	one's	ones
160	-7	Than	Then
215	-13	groups	group

Further, replace the passage of text starting on page 137 row -6 with; "To see this,..." and ending on row 7 on page 138 with "x1 .", with the following:

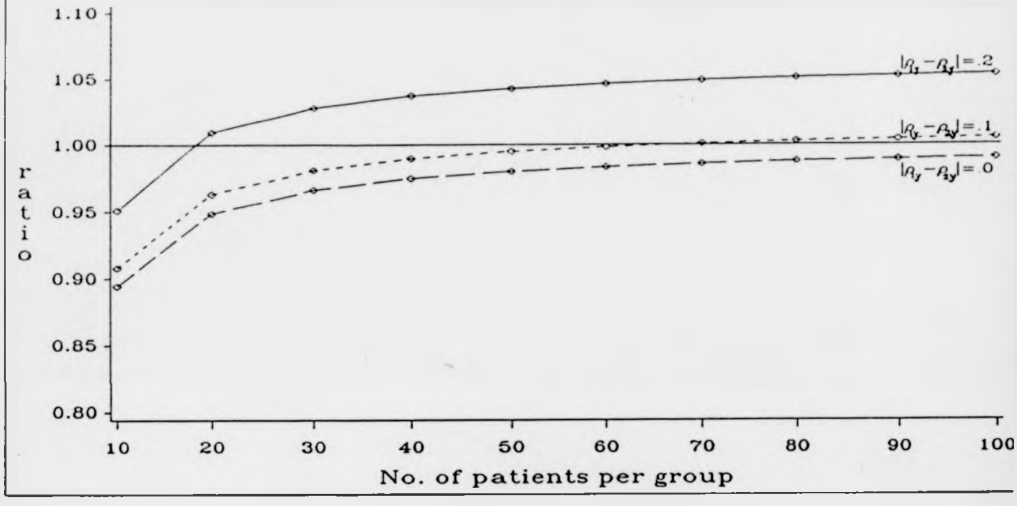
"The magnitude of attenuation that we expect, as a consequence of the use of a selection criterion under the current assumptions (e.g. when sampling from a bivariate normal distribution), is given by;

$$\rho_{(x_0,x_1|x_0\geq k)} = \rho_{(x_0,x_1)} \cdot \sqrt{\frac{1-\lambda}{1-\rho^2_{(x_0,x_1)} \cdot \lambda}} \quad .$$

Using this formula it may be seen to which extent the correlation between x_0 and x_1 is expected to be decreased due to regression to the mean, and hence how much less useful a covariate adjustment is likely to be. For instance, for the example above we would expect a correlation of 36/52=0.69 without the use of a selection criterion, and a correlation of 0.42 with the specific choice of k=95mmHg.

One way to improve the situation is to perform two pre-entry measurements, one for classification purposes (x_{01}), and an additional baseline not underlying the selection (x_{02}). Then, a more useful covariate adjustment may be based on the second unrestricted baseline."

Figure 3.6.2 : Variance ratio, Var(ANCOVA1) / Var(ANCOVA2), based on expected values for the correction factors. Depending on sample sizes and differences between "mixed" correlations



Mölndal, Sweden, 26 Aug 1994

Lars Frison