

# **Methods of Sample Size Calculation for Clinical Trials**

**Michael Tracy**

## **Abstract**

Sample size calculations should be an important part of the design of a trial, but are researchers choosing sensible trial sizes? This thesis looks at ways of determining appropriate sample sizes for Normal, binary and ordinal data.

The inadequacies of existing sample size and power calculation software and methods are considered, and new software is offered that will be of more use to researchers planning randomised clinical trials. The software includes the capability to assess the power and required sample size for incomplete block crossover trial designs for Normal data.

Following from on from these, the difference between calculated power for published trials and the actual results are investigated. As a result, the appropriateness of the standard equations to determine a sample size is questioned- in particular the effect of using a variance estimate based on a sample variance from a pilot study is considered.

Taking into account the distribution of this statistic, alternative approaches beyond power are considered that take into account the uncertainty in sample variance. Software is also presented that will allow these new types of sample size and Expected Power calculations to be carried out.

## **Acknowledgements**

I would very much like to thank Novartis for funding my tuition fees, and for providing a generous stipend. I would also like to thank Stephen Senn for all his support as my academic supervisor.

# Table of Contents

Chapter 1 .....	6
1.1 Introduction.....	6
1.2 Clinical trials and the importance of sample size.....	7
1.3 Power .....	8
1.4 The types of trial of interest .....	10
Superiority trials, Equivalence Trials, and Non-inferiority trials .....	10
Parallel and Cross-over .....	11
1.5 Sample size and power calculations for Normal data.....	13
1.6 Sample size and power calculations for binary data.....	18
1.7 Sample size and power calculations for ordinal data.....	22
Chapter 2 – SAS Programs for Calculating Sample Size.....	25
Computing approach to sample size calculation .....	25
Program 2.1. SAS program for Normal Data.....	27
Program 2.2: SAS Program for Normal Data 2.....	31
Program 2.3 SAS Program for Normal Data 3.....	33
Program 2.4 SAS Program for Normal Data 4.....	36
Program 2.5: SAS Program for binary Data .....	37
Program 2.6: SAS Program for ordinal data.....	39
Chapter 3 – R Programs for calculating sample size .....	40
Program 3.1: R panel program for Normal data .....	40
Program 3.2: R panel program for binary data .....	46
Program 3.3: R panel program for ordinal Data.....	49
Some Comparisons with other software and standard tables, with Discussion.....	53
1. Parallel Trial sample size, Normal Data.....	54
2. Crossover trial, Normal Data.....	56
3. Parallel Trial sample size, binary Data.....	56
4. Crossover Trial, Binary data .....	57
5. Parallel Trial, Ordinal data .....	57
6. Crossover Trial, Ordinal data .....	57
7. Incomplete Block Design, Normal data.....	58
8. Discussion of comparisons. ....	58
Chapter 4 .....	60
4.1 The use of sample size calculations.....	60
4.2 Alpha, beta and the treatment difference .....	60
4.3 Sample standard deviation as an estimator of population standard deviation.....	61
s given sigma.....	62
Sigma given s .....	66
4.4 Methods of incorporating uncertainty over variance of Normal data into sample size calculations.....	70
Expected Power compared to Power calculations using point estimates	81
4.5 Selecting pA .....	83

4.6 Methods of incorporating uncertainty over $p_A$ into sample size calculations .....	84
4.7 Simulation-based power estimation.....	85
Chapter 5 .....	88
Program 5.1: SAS Program for Normal data taking into account uncertainty in observed standard deviation.....	88
Program 5.2: SAS Program for Normal data with uncertainty 2 .....	89
Program 5.3: SAS Program for Normal data with uncertainty 3 .....	91
Program 5.4: SAS Program for Normal data with uncertainty 4 .....	92
Program 5.5: SAS program for binary data that takes into account uncertainty about true value of $p_A$ .....	93
Chapter 6 .....	95
Program 6.1: R program for Normal data taking into account uncertainty ..	95
Program 6.2: R panel Program for binary outcomes taking into account uncertainty in $p_A$ .....	107
Program 6.3.....	110
Some Comparisons with other software and standard tables, and Discussion .....	111
1. Parallel trial, Normal Data .....	112
2. Crossover trial, Normal Data.....	113
3. Parallel trial, Binary data .....	114
4. Parallel trial, Binary data .....	114
Chapter 7: Conclusion: Summary, and Discussion .....	115
7.1 Summary .....	115
Chapter 1 .....	115
Chapters 2 & 3.....	115
Chapter 4 .....	115
Chapter 5 & 6.....	116
7.2 Discussion, and Further Work .....	116
References .....	118
Appendix A .....	122

# Chapter 1

## *1.1 Introduction*

The purpose of this thesis is to look at the theories behind sample size calculations in a range of types of clinical trials, and to develop computer software that will be of practical use in dealing with some of the problems that a statistician may encounter. In particular, I intend to try and develop tools that will help calculate meaningful sample sizes and powers in situations with uncertain endpoint variances or unorthodox trial designs. In the first chapter I will give some background into the role of power and sample size calculation, and then show how these may be performed on a range of data types. In the next two chapters I intend to demonstrate that some new sample size calculation programs are needed, and that the resultant programs produce output consistent with currently used methods while being more user-friendly. Chapter 4 has a look at the assumption that the sample variance is a good estimator of the true variance for the purposes of sample size estimation, and when flaws are found then I try to describe some ways to deal with the situation. Similar uncertainty about  $p_A$  for binary data studies is dealt with, and again methods are suggested to cope. Finally, software that can implement the remedies of Chapter 4 shall be created, and described in Chapters 5 and 6.

This chapter will look at power and sample size, and some of the factors that they depend on. It will look at several different types of clinical trial where sample size calculations would be useful, and examine methods of determining power.

## ***1.2 Clinical trials and the importance of sample size***

Clinical trials are the formal research studies to evaluate new medical treatments. Before a possible new therapy is commercially available it usually must be shown to be acceptably safe, and the effectiveness of the therapy must be proven to the drugs company and regulatory bodies. The trials are vital to the process of bring through new drugs and finding new uses for existing drugs.

Clinical trials are a very expensive undertaking, consuming a great deal of time and resources. To compare the efficacy of different drugs, dosages, surgeries or combinations of these treatments can cost over \$500 million and take many years, so it is of great importance that the design of the clinical trial gives a good chance of successfully demonstrating a treatment effect. There are different ideas on how that chance should be calculated and interpreted, but in general the larger the number of participants in the trial the more chance there is of identifying a significantly different treatment effect. The more people tested, the more sure you can be that any observations of difference between therapies is due to a true underlying treatment effect and not just random fluctuations in the outcome variable.

However, there are factors that may lead us to limit the numbers on a trial. In the US alone there are over 40,000 clinical trials currently seeking participants and each of these may need up to thousands of subjects. With so many trials

seeking subjects, researchers are paying large bounties for potential recruits on top of what can be already expensive running costs. There is a financial concern to balance the desire to give a trial a high probability of identifying a treatment effect with the increasing cost of recruiting more test subjects. If a new treatment is for a condition which is already has a drug that improves the quality of life substantially for sufferers then it could be ethically unsound to place more patients on the new alternative than is necessary, as the trial participants may receive inferior treatment. The sample size of the trial must balance the clinical, financial and ethical needs of the sponsor, trial participants and potential future treatment recipients.

### **1.3 Power**

In statistics, the power of a test is the probability that it will reject a false null hypothesis. The power of a trial design or contrast between treatment effects in this thesis is the conditional probability of a resulting statistical analysis identifying a significant superiority of one treatment's effect on outcome over another's if a superiority of a stated magnitude truly existed.

To better understand the concept of power, consider the world as idealised in hypothesis testing.

A testable null hypothesis  $H_0$  and an alternative  $H_1$  are stated, they are logical opposites, one is completely true and the completely false. Data regarding



the hypotheses are statistically analysed. The null hypothesis is either accepted or rejected- rejection of  $H_0$  results in  $H_1$  being accepted.

There are four possible states of the world:

$H_0$  is actually true, and is correctly not rejected.

$H_0$  is actually true, and is wrongly rejected in favour of  $H_1$ , a Type I error.

$H_0$  is actually false, and is wrongly not rejected, a Type II error.

$H_0$  is actually false, and is correctly rejected in favour of  $H_1$ .

Table 1.1

	$H_0$ not rejected	$H_0$ rejected
$H_0$ is true	Correct to not reject Occurs with probability $1-\alpha$	Type I error Occurs with probability $\alpha$
$H_0$ is false	Type II error Occurs with probability $\beta$	Correct to reject Occurs with probability $1-\beta$

$\alpha$  is the probability of a type I error: The probability of saying there is a relationship or difference when in fact there is not. In other words, it is the probability of confirming our theory incorrectly.

$\beta$  is the probability of making a type II error: The probability of saying there is no relationship or difference when actually there is one. It is the probability of not confirming a theory when it's true.

1-  $\beta$  is power: The probability of saying that there is a relationship or difference when there is one. It is the probability of confirming our theory correctly, so a trial designer would generally want this to be as large as possible in order to be confident in detecting a hypothesised difference in treatment effects.

### ***1.4 The types of trial of interest***

#### **Superiority trials, Equivalence Trials, and Non-inferiority trials**

Superiority trials are trials that one treatment is better than another. Non-inferiority trials are intended to show that the effect of a new treatment is not worse than that of an active control by more than a specified margin [Snapinn, 2000]. Equivalence trials are attempts to establish if compared treatments differ by less a specified margin [Chi, 2002]. Non-inferiority trials do not truly attempt to show “non-inferiority”, because that is actually what superiority trials do. Instead, non-inferiority trials try to demonstrate a new treatment is at worst only inferior to a comparison by a (clinically) insignificant amount.

The hypotheses that the investigator would like to establish for each type of trial are

$H_1$ (superiority)	Effect A	< Effect B
$H_1$ (equivalence)	Effect A- $\delta$	< Effect B < Effect A+ $\delta$
$H_1$ (non-inferiority)	Effect A- $\delta$	< Effect B

where  $\delta$  is a clinically significant amount, and treatment B is a new treatment that is compared to existing treatment A.

In general, superiority trials are used when new advances in treatment therapy, or effect of active control is small, while non-inferiority trials are used when the new treatment has technical similarities with the existing, or if the active control has moderate to significant effect, or specific safety problems. It is always at least as difficult to show superiority rather than non-inferiority, so superiority trials will need more subjects.

Different regulatory authorities have different regulations as to when non-inferiority trials will be acceptable over superiority trials, and this has led to certain drugs being approved drugs in different territories [French JA 2004].

In this thesis the focus will be on superiority trials, and all formulae, trial designs, p values etc are for superiority.

### **Parallel and Cross-over**

The thesis is about ways of determining the correct sample size to achieve power in randomised controlled parallel and crossover superiority trials with one outcome variable.

In parallel group trials subjects are randomised to receive one of the treatments to be compared each. They stay on the one treatment they are allocated to.

Senn (2002) defines a crossover trial as “one in which subjects are given sequences of treatments with the object of studying differences between individual treatments (or sub-sequences of treatments)”.

The terms ‘crossover’, ‘cross-over’, ‘cross over’, ‘change-over’, ‘changeover’ and ‘change over’ are all terms for this type of trial design, and all have been used in recently published articles. I will stick with ‘crossover’ for the duration of this thesis, simply because it is the most popular.

The simplest crossover design is the AB/BA design. According to this design two treatments are applied to subjects over two periods. The trial participants should receive either treatment A or treatment B in the first period, and then be given the other treatment in the second period. Often in literature when the author uses the expression ‘crossover’ with no further elaboration they mean this AB/BA design. The hope in a crossover trial is that between-subject variation is cancelled out because each subject acts as their own control, so treatment effect should be a bigger part of any difference in observed results. This means fewer subjects might be needed than in a parallel trial to be sure if a difference in treatment effect exists.

The AB/BA design is not the only type of crossover. It is possible to design trials with more treatments or more periods. The most efficient from a statistical point of view is a trial where each subject receives all the treatments being compared, this would mean an equal number of treatments and

periods. An incomplete blocks crossover trial is a type of crossover trial in which each subject gets some but not all of the treatments. For an extreme example of this type see Vollmar et al (1997), wherein the process of deciding upon a 7 treatment, 5 period, 21 sequence design for comparing asthma remedies is described.

There are many possible reasons one might try an incomplete blocks design rather than a complete block. If the sponsor wants to compare several treatments with each other but ethical or financial concerns mean a long trial is not acceptable then an incomplete blocks design may be more attractive; incomplete blocks have fewer periods than a comparable complete blocks design, and should take less time to run. The fact that a subject does not receive all treatments means that only some of the uncertainty caused by between patient variation is removed, so it only has some of the benefits of a complete crossover design.

### ***1.5 Sample size and power calculations for Normal data***

The power of the trial is dependant on not only the trial design but also the intended method of statistical analysis. Crossover trials can be tested by OLS or by treating subject as a random effect, which can give differing results for incomplete block designs.

It is assumed that any observed difference in outcomes is a combination of treatment effect, 'period effect', pre-existing differences between subjects and random variation of the outcome variable.

Before the required sample size of a trial can be calculated the intended type of analysis must be decided, and values of  $\alpha$ ,  $\beta$  and  $\delta$  (size of relevant difference) chosen. Values for the variance of the outcome variable must be used, and if the nature of this variance can be split into variance due to a within-subject and between-subject type a better estimate may be made for crossover type designs. In parallel designs it is not relevant how this is split because there is no within-patient analysis, as each patient has only one result. For complete blocks designs only within-subject variation should be affecting the results, so identifying the proportion of variance that is of a 'within' nature will lower variance estimation used in calculation and give a higher power estimate. In incomplete blocks separating the sources of variance will allow to the statistician to try and recover between subject information and allow a more powerful random effects analysis [Cox DR and Reid N.2000; Senn 2002]. Note, in this thesis fixed effects analysis should be understood as analysis treating subject as a fixed effect, while random effect analysis means analysis where the subject term is treated as a random effect.

When S is the test statistic, then the relationship between  $\delta$  and S will be [Julious SA. 2005]

$$\text{Var}(S) = \frac{\delta^2}{(Z_{1-\beta} + Z_{1-\alpha/2})^2}$$

This is then basic equation from which all the sample size and power equations in this chapter are derived from, for binary and ordinal data as well as Normal data. Indeed, because of the centrality of this equation to sample size estimation Machin calls a slight rearrangement of this “the Fundamental Equation” [Machin 1997]. Because  $\text{Var}(S)$  for all these types of data can also be defined by a relationship between sample size and other parameters, an equation linking sample size to power can be derived dependent on the trial type. For normal data  $S$  is the difference between means. In the case of a parallel trial with two treatments, for example, variance of  $S$  can be defined as

$$\text{Var}(S) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}$$

where  $n_A$  is number of subjects randomised to treatment A, and  $\sigma$  is the population variance- it is usual to assume that the population variance is equal for each of the treatment groups when analysing the results [Julious SA. 2005]. With these results, a link between sample size and power can be established.

$$\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} = \frac{\delta^2}{(Z_{1-\beta} + Z_{1-\alpha/2})^2}$$

The formulae for power for crossover and parallel, complete and incomplete blocks, fixed effects analysis and random effects analysis can be unified in a generalised form of equation.

$$\left( \sqrt{\frac{\delta^2}{\text{Var}(S)}} - Z_{1-\alpha/2} \right)$$

The power of a trial to detect a treatment contrast can be calculated from a cumulative normal distribution in the form

$$1-\beta \approx \Phi(\Delta - t_{1-\alpha,df}) \quad (1.2)$$

where  $\Delta$  is the ratio of the treatment difference (usually  $\delta$ ) to the root of the covariance of the compared treatment effects, and  $df$  is the degrees of freedom used in analysis. The  $\Delta$  depends on the trial design, number of sequence repetitions, and the between (and within) subject variance.

But because any analysis would use the observed variance  $s^2$  not the true variance  $\sigma^2$ , the power is more accurately calculated from a cumulative non-central t distribution in the form

$$1-\beta = 1-pt(t_{1-\alpha,df}, df, \Delta) \quad (1.3)$$

where  $\Delta$  is used as the non-centrality parameter [Senn 2002, Julious 2005].

### Example 1.1

An investigator intends to run an AB/BA crossover trial to see if new drug B is superior to drug A. The clinical relevant difference is 1, the within subject and between subject standard deviations are both 1. They want to know the power of this trial to detect this difference if he gets 20 subjects enrolled, with



10 assigned to each sequence. What is the power when the one-sided alpha is 0.025?

The degrees of freedom used in the analysis of the power of a contrast from a crossover design is  $(R*S*P)-(R*S)-P-T+2$ , where S is the number of subjects needed to complete each block, R is the number of block repetitions, P is the number of periods and T is the number of treatments. In this example  $R=10$ ,  $S=2$ ,  $P=2$  and  $T=2$ , so  $df = 18$ .

$$\begin{aligned} \text{By (1.3)} \quad 1-\beta &= 1-\text{pt}(t_{1-\alpha,df},df, \Delta) \\ 1-\beta &= 1-\text{pt}(t_{0.975, 18}, 18, \Delta) \\ 1-\beta &= 1-\text{pt}(2.101,18, 3.162) \\ 1-\beta &= 1-(0.15156) \end{aligned}$$

The power by equation 1.3, our preferred method, is 0.84844. By formula 1.2 a slightly higher power, 0.85574, would have been calculated. Formula 1.3 gives the correct result, and that formula 1.2 gives results that are close.

So, formula 1.3 is the one preferred for use where possible in this thesis for power and sample size calculations for normally distributed data, but formula 1.2 has the advantage of being easier to manipulate- which becomes useful when we take into account uncertainties in standard deviation estimate in chapter 4 onwards.

## 1.6 Sample size and power calculations for binary data

In the analysis of binary type data the hypothesis  $H_0$  and the alternative  $H_1$  are not so obviously formulated for binary data as for Normal data. There are several possible ways of summarising the difference between treatments [Julious SA. 2005], but here we will be interested in just two: Odds Ratio and Absolute Risk Reduction. The incidence of a binary outcome of interest under the effects of treatment A and treatment B are labelled  $p_A$  and  $p_B$  respectively. They are probabilities, and have a value between 0 and 1 inclusive. With Odds Ratio (OR) the two treatment effects are described together in a ratio, defined as

$$OR = p_A(1-p_B) / p_B(1-p_A)$$

In Odds Ratio

the  $H_0$  would be  $\text{Log}(OR) = 0$

$H_1$  would be  $\text{Log}(OR) = d$

While in absolute risk

$H_0$ :  $\pi_A - \pi_B = 0$

$H_1$ :  $\pi_A - \pi_B = d$

The different hypothesis types have slightly different claims about the world and they would be statistically analysed differently. There are also arguments about how to analyse the data after deciding how to frame the hypothesis.

The power, no matter the hypothesis, will depend upon sample size, trial type,  $p_A$ , and the clinically relevant  $p_B$  or OR. For parallel trials there are two types of power calculation in the used later in the thesis, proportional difference (appropriate for testing an absolute risk type  $H_0$ ) and an Odds Ratio method. In general, the variance of the measure effect must satisfy

$$\text{Var}(S) = \frac{d^2}{(Z_{1-\beta} + Z_{1-\alpha/2})^2}$$

The variance of the log-odds ratio can be approximated as [Julious SA. 2005]

$$\text{Var}(S) = \frac{6}{n \left( 1 - \sum_{i=1}^2 \bar{p}_i^3 \right)}$$

From these the power for a proportional difference trial can be approximated as

$$1 - \beta = \Phi \left( \sqrt{\frac{n(p_A - p_B)^2}{(p_A(1 - p_A) + p_B(1 - p_B))}} - Z_{1-\alpha/2} \right)$$

and the odds ratio power approximated as

$$1 - \beta = \Phi \left( \sqrt{n(\log OR)^2 \left[ 1 - \sum_{i=1}^2 \bar{p}_i^3 \right] / 6} - Z_{1-\alpha/2} \right)$$

These two methods give similar results, the odds ratio method usually calculating the power a little higher than the alternative, and subsequently gives slightly lower sample size estimates.

For crossover trials there are more possible power calculation suggestions, including an approximate OR method. One may consider the expected results of a crossover trial with binary outcomes in terms of the proportions of the four possible combinations of outcomes for the two treatments.

Table 1.1: Expected proportions of combined treatment binary outcomes

		Response to treatment B	
		0	1
Response to treatment A	0	$\lambda_{00}$	$\lambda_{01}$
	1	$\lambda_{10}$	$\lambda_{11}$

The trialist will expect that a certain proportion of the subjects will either succeed on both treatments ( $\lambda_{11}$ ) or fail on both treatments ( $\lambda_{00}$ ), but it is in the proportion of subjects that have discordant outcomes ( $\lambda_{01}$  and  $\lambda_{10}$ ) that they shall find any evidence of a difference in treatment effect. The common way to analyse binary data is the McNemar test, for which the test statistic is only based upon the numbers of discordant pairs,  $n_{01}$  and  $n_{10}$ . It is possible to frame a sample size calculation in terms of the discordant sample size, that is the number of discordant results required. Approximating the conditional OR can allow a discordant sample size to be calculated

$$\frac{(Z_{1-\alpha/2}(\psi + 1) + 2Z_{1-\beta}\sqrt{\psi})^2}{(\psi - 1)^2}$$

One way of calculating the relationship between discordant ( $n_d$ ) and total crossover sample size ( $N_C$ ) is to divide the discordant sample size by the proportion expected to be discordant (using the notation of table 1.1)

$$N_c = \frac{n_d}{\lambda_{01} + \lambda_{10}}$$

to get an estimate of total sample size

$$\frac{(Z_{1-\alpha/2}(\lambda_{10} + \lambda_{01}) + 2Z_{1-\beta}\sqrt{\lambda_{10}\lambda_{01}})^2}{(\lambda_{10} + \lambda_{01})(\lambda_{10} - \lambda_{01})^2}$$

Conner [Conner, R.J. 1987] and Miettinen [Miettinen, OS 1968] both suggest similar ways to this approximate OR method to calculate sample size, but they allow fewer assumptions about the relation of the discordant to the total sample size, and consequently suggest higher sample sizes are required.

Juliuos [Julious SA. 2005] suggests another option, to base power and sample size calculations on the OR method for parallel trials by assuming that a crossover trial with n subjects will have about the same power as a parallel trial with 2n subjects. This highlights an issue about the relative merits of crossover trials between binary and continuous data- or at least about the analysis methods of binary crossover trials. With continuous data one would expect that a crossover trial would deliver the same power as parallel trial with somewhat less than half as many subjects, but Julious shows (backed up with mathematic derivation and also empirical evidence) that for the case of binary that approximately the same number of 'patient sessions' are required irrespective of the choice between parallel and crossover. There are still clear advantages to a crossover trial with binary outcomes, though. If a trialist intends to analyse the results of a binary crossover trial by the McNemar test, the efficiency gained over a parallel trial of equal power will be in and reducing patient numbers, but not patient sessions. Additionally the trialist can be assured that the treatment groups will not have differences in prognostic

factors, and can be more confident that any observed difference is due to treatment effect alone.

The four crossover methods give fairly similar results, but the Conner and Miettinen give slightly lower power estimates than the other two.

### ***1.7 Sample size and power calculations for ordinal data***

The situation with ordinal data is similar to that for binary data, and indeed binary could be thought of as a special case of ordinal. Like binary data, there are several ways to frame the hypotheses, but if we make a few reasonable assumptions about the treatment effect across levels and the analysis then an OR based method will be the easiest way to get power and sample sizes [Machin, 1997]. For ordinal, if the pA and pB are split into n levels, then we will assume proportional odds ratio, that is  $OR_1=OR_2=\dots=OR_{n-1}$ . This is the assumption that the Mann-Whitney U test for ordered categorical data with allowance for ties uses, and the Mann-Whitney U test is the usual type of analysis for this type of data [Conover WJ. 1980]. This would not be an appropriate assumption (and Mann-Whitney U would be an inappropriate test) if we believed that a new treatment had some non-constant effect compared to a comparator, for example pushing more people into extreme categories.

It is easier to state the hypothesis and to estimate the variance of the outcome statistic that all the sample size and power calculations depend upon for OR than for some absolute risk reduction based calculation.

For parallel trials a power based on an OR method can be used. Using the same arguments as for binary parallel trials the power can be calculated as

$$1 - \beta = \Phi \left( \sqrt{n \left[ 1 - \sum_{i=1}^k \bar{p}_i^2 \right]} (\log OR)^2 / 6 - Z_{1-\alpha/2} \right)$$

[Julious SA. 2005] When k=2 then this equation is identical to the Odds Ratio equation for binary trials. Binary trials are ordinal trials with two levels, and the binary OR equation is a specific case of the general ordinal formula.

For crossover trials there are two methods. Like for binary trials, one could assume crossover trial with n subjects will have about the same power as a parallel trial with 2n subjects. Alternatively a power can be estimated as

$$1 - \beta = \Phi \left( \sqrt{n (\log OR)^2 / \text{var}(\log(OR))} - Z_{1-\alpha/2} \right)$$

where

$$\text{var}[\log(OR)] = \frac{1}{n} \left( \frac{\sum_{i < j} (j-i)^2 p_{ij}}{\left[ \sum_{i < j} (i-j) p_{ij} \right]^2} + \frac{\sum_{i > j} (j-i)^2 p_{ij}}{\left[ \sum_{i > j} (j-i) p_{ij} \right]^2} \right)$$

The  $\text{var}(\log(\text{OR}))$  method gives lower power estimates than the parallel OR method usually, and gives higher sample size estimates.



## **Chapter 2 – SAS Programs for Calculating Sample Size**

### ***Computing approach to sample size calculation***

As the calculations for sample size and power can be quite complex, it would be useful to be able to use computer software to solve the equations. It was decided that SAS and R would be appropriate platforms for these types of program. It was also decided that the programs should be able to deal with incomplete block crossover trials, as there is no commonly used package that can deal with this type of design.

SAS, originally known as Statistical Analysis System, is a dedicated statistical software package that is commonly used in the pharmaceutical industry for dealing with clinical trial data. SAS claims to be used in 40,000 sites worldwide, including 96 of the top 100 companies on the FORTUNE Global 500. SAS is the standard software used in much of the pharmaceutical industry for the analysis of trial results, but often when sample size calculations are needed the statistician will put SAS aside and use a dedicated package for the job. It would be handier if there were programs for SAS that allowed the sample size calculations, so the user wouldn't have to switch between platforms.

SAS has a modular design, and SAS/IML is the SAS component that deals with data matrices. SAS/IML was used for the SAS programs because its matrix language has the capacity to multiply and invert matrices, which is necessary to complete some of the power and sample size calculations.

R also has a large user base, mainly in the academic world. This flexible software is improved constantly by users who write programs that add to the functionality. A particularly interesting add-on for R is the Rpanel package, which gives programmers the ability to create a graphical user interface for the activation of R functions. This will allow the creation of software that is easy for the user to dynamically change parameters of interest, making for a more enjoyable and useful experience. The R panel programs are designed to be more user-friendly to operate, and to have an accessible interface, and to provide more graphical output where appropriate.

The intention is to write a set of programs that will deal with many types of trial design for normal, binary and ordinal type data. The SAS programs are to be uncomplicated and quick to run, the R Panels are to be more accessible and take advantage of the superior graphical interface.

This chapter contains examples of SAS programs for calculating sample sizes and power for different types of superiority trial, and instructions for the programs' use.

The first four programs deal with Normally distributed outcome variables, and there are also programs to tackle problems with binary and ordinal outcomes.

### ***Program 2.1. SAS program for Normal Data***

First, a program to calculate power for a contrast between two treatments with Normal data for crossover or parallel trials. This program is based around formula 1.3.

The program needs the user to enter data representing a block of treatment sequences and enter a sample size and a required power size, and give information on parameters to use in calculation, as well as identifying which of the treatments to compare.

The program will calculate the covariance between the two treatment effects for both fixed effects and random effects, and from the inputted sigma and delta calculate the  $\Delta$ , initially for  $R=1$ . At this point the df is calculated by analysing the entered design to extract the S, P and T. All the variables to calculate  $1-pt(t_{1-\alpha,df},df, \Delta)$  are now in the memory, and it is possible to calculate the power for the sample size when  $R=1$ .

If the power calculated at  $R=1$  is not at least as big as the power required then the  $\Delta$  and df when  $R=2$  are computed and used in power calculations, and so on, until the minimum size  $R$  that gives an adequate power is found.

Figure 2.1.1: Output from Program 2.1

		The SAS System	13:45 Monday, July 9, 2007 119	
<b>A</b>	_____	Balanced Design		
<b>B</b>	_____	Treatment	F1 1 Versus	F2 2
<b>C</b>	_____	Number of Patients	2 (	1 repetitions)
	_____	one-sided alpha:	ALPHA 0.025	SIG 40
	_____	signa(within):	delta: 8	DELTA
<b>D</b>	_____	Power for fixed effects model	POWERFIX2 0	
	_____	Power for random effects model	POWERAN2 0 (when lambda is	LAM 0 )
<b>E</b>	_____	Fixed effects:	RREQ 264 reps (	520 subjects) needed to achieve power of 90 %
	_____	Random effects:	RREQRAN 264 reps (	520 subjects) needed to achieve power of 90 %

- A:** Displays whether design is balanced or not
- B:** The treatments to be compared. This is necessary because where there are more than two treatments the power of testing a specific contrast may, depending on design, be different from contrast to contrast.
- C:** The parameters that were used in the calculations
- D:** Power calculated for Fixed and Random Effects
- E:** The required sample size to achieve the requested power

### Example 2.1.1: AB/BA Crossover trial

An investigator intends to run a crossover trial to see if new drug B is superior to drug A. The clinical relevant difference is 1, the within subject and between subject standard deviations are both 1. He wants to know the power of this trial to detect this difference if he gets 20 subjects enrolled, with 10 assigned to each sequence. With a one-sided alpha set at 2.5%, what is the power?

Using program 2.1, he should set seq to {1 2, 2 1} for this design, and let delta, sig and lam all equal 1. R should be 10.

Running the program, he would find the design has a power of 0.84844 for both random effects and fixed effects.

### Example 2.1.2 Parallel trial

The investigator in Example 2.1.1 decides to change the trial to a parallel trial with equal numbers assigned to each treatment A and B. He wants to know how this would change the power if he kept the same sample size (20), and how large a sample size he would need to get at least as much power from this trial as from the crossover trial (0.84844)

Using Program 2.1 he should set seq to {1, 2} and beta to (1-0.84844). With this design the random effects power drops to 0.32175, and there is no fixed effects model possible. To get at least the same power as the previous trial the sample size must increase from 20 to 74 subjects, 37 per arm.

### Example 2.1.3

The investigator changes his mind, and would now like to run a trial with a more adventurous incomplete blocks design that will have 5 different treatment types (Table 2.1.2). What is the power to detect the clinically relevant difference between treatment A and E if sample size remains at 20, and how big should it be to achieve 90% power?

Table 2.1.2

<u>Period 1</u>	<u>Period 2</u>
Treat A	Treat E
Treat B	Treat A
Treat C	Treat B
Treat D	Treat C
Treat E	Treat D

Figure 2.1.2 shows how to code the sequence into Program 2.1. Beta should be set to 0.1.

Figure 2.1.2

```
seq={
1 5,
2 1,
3 2,
4 3,
5 4};
```

The power for the fixed effects is calculated as 0.316, while the power by random effects is greater at 0.384. 90 subjects would be required to have a power of 90% if analysis was by using fixed effects, whereas only 70 subjects are required if the analysis deals with between-subject variance by modelling subject as a random effect.

### ***Program 2.2: SAS Program for Normal Data 2***

The previous program works well, but the scope of trials it can deal with is limited. A program that could calculate the power of a contrast between a pair of treatments for crossover or parallel trials for normal data, but allowing the user to enter a more complex treatment sequence would be handy. This program would be useful for occasions where subjects are not randomised in equal numbers to treatment sequences either by accident or design.

The statistics used in this program are similar to the previous, power calculations again being from  $1 - pt(t_{1-\alpha, df}, df, \Delta)$ . However, the  $df$  and  $\Delta$  are calculated in different ways to cope with the different way the sequence is entered.

Figure 2.2.1: Output from Program 2.2

<b>A</b>	Unbalanced Design							
<b>B</b>	Treatment		1		Versus		2	
<b>C</b>	Sequence	Reps	Sequences	Period1	Period2	Period3	Period4	Period5
	1	Seq1		1	7	6	3	4
	1	Seq2		5	1	7	6	3
	7	Seq3		2	5	1	7	6
	1	Seq4		4	2	5	1	7
	10	Seq5		3	4	2	5	1
	1	Seq6		6	3	4	2	5
	1	Seq7		7	6	3	4	2
	1	Seq8		1	6	4	5	7
	1	Seq9		2	1	6	4	5
	1	Seq10		3	2	1	6	4
	1	Seq11		7	3	2	1	6
	1	Seq12		5	7	3	2	1
	1	Seq13		4	5	7	3	2
	1	Seq14		6	4	5	7	3
	1	Seq15		1	3	5	6	2
	1	Seq16		4	1	3	5	6
	1	Seq17		7	4	1	3	5
	1	Seq18		2	7	4	1	3
	1	Seq19		6	2	7	4	1
	1	Seq20		5	6	2	7	4
	1	Seq21		3	5	6	2	7
<b>D</b>	Number of Patients					36		
	One-sided alpha:	0.025	sigma(within):	1	delta:	1		
<b>E</b>	POWERFIX2							
	Power for fixed effects model 0.9550723							
	POWERRAN2							
	Power for random effects model 0.9611301 (when lambda is 0 )							

- A:** Displays whether design is balanced or not
- B:** The treatments to be compared
- C:** The sequences, and number of subjects assigned to each sequence
- D:** The parameters that were used in the calculations
- E:** Power calculated for Fixed and Random Effects



### Example 2.2.1

As in Example 2.1.1, an investigator intends to run an AB/BA crossover trial to see if new drug B is superior to drug A. The clinical relevant difference is 1, the within subject and between subject standard deviations are both 1. He wants to know the power of this trial to detect this difference if he gets 20 subjects enrolled, but due to clerical error 13 subjects were enrolled to one sequence and 7 to the other. What is the power?

Using program 2.2, he should set seq to {1 2, 2 1} for this design, and seqreps to {13,7}.

The power has dropped slightly from the optimal 0.848 to 0.814 for both random effects and fixed effects.

(NB: The output describes the design as 'balanced'. In these programs, 'balanced' should be understood as meaning 'the power of the contrasts is equal between all different treatments'. Any design with only two treatments will be balanced because there is only one contrast, 1 vs. 2)

### ***Program 2.3 SAS Program for Normal Data 3***

This program calculates power and required sample sizes for all possible contrasts of treatments used in a crossover or parallel trial, for Normally

distributed data. This program would be useful if the user wanted to see the power of all contrasts at once.

Figure 2.3.1: Output from Program 2.3

**A** ————— Balanced Design

(Fixed Effects) Power between pairs with 10 repetitions.  
**B** ————— Fixed Effects Power

	Treatment1	Treatment2	Treatment3	Treatment4	Treatment5	Treatment6	Treatment7
Treatment1	0	0.3864505	0.3864505	0.3864505	0.3864505	0.3864505	0.3864505
Treatment2	0.3864505	0	0.3864505	0.3864505	0.3864505	0.3864505	0.3864505
Treatment3	0.3864505	0.3864505	0	0.3864505	0.3864505	0.3864505	0.3864505
Treatment4	0.3864505	0.3864505	0.3864505	0	0.3864505	0.3864505	0.3864505
Treatment5	0.3864505	0.3864505	0.3864505	0.3864505	0	0.3864505	0.3864505
Treatment6	0.3864505	0.3864505	0.3864505	0.3864505	0.3864505	0	0.3864505
Treatment7	0.3864505	0.3864505	0.3864505	0.3864505	0.3864505	0.3864505	0

(Random Effects) Power between pairs with 10 repetitions.  
**C** ————— Random Effects Power

	Treatment1	Treatment2	Treatment3	Treatment4	Treatment5	Treatment6	Treatment7
Treatment1	0	0.4090771	0.4090771	0.4090771	0.4090771	0.4090771	0.4090771
Treatment2	0.4090771	0	0.4090771	0.4090771	0.4090771	0.4090771	0.4090771
Treatment3	0.4090771	0.4090771	0	0.4090771	0.4090771	0.4090771	0.4090771
Treatment4	0.4090771	0.4090771	0.4090771	0	0.4090771	0.4090771	0.4090771
Treatment5	0.4090771	0.4090771	0.4090771	0.4090771	0	0.4090771	0.4090771
Treatment6	0.4090771	0.4090771	0.4090771	0.4090771	0.4090771	0	0.4090771
Treatment7	0.4090771	0.4090771	0.4090771	0.4090771	0.4090771	0.4090771	0

(Fixed Effects) Required number of complete Replications to attain 0.9 power by pair  
**D** ————— Fixed Effects Repetitions Required

	Treatment1	Treatment2	Treatment3	Treatment4	Treatment5	Treatment6	Treatment7
Treatment1	0	38	38	38	38	38	38
Treatment2	38	0	38	38	38	38	38
Treatment3	38	38	0	38	38	38	38
Treatment4	38	38	38	0	38	38	38
Treatment5	38	38	38	38	0	38	38
Treatment6	38	38	38	38	38	0	38
Treatment7	38	38	38	38	38	38	0

(Random Effects) Required number of complete Replications to attain 0.9 power by pair  
**E** —————

- A:** Displays whether design is balanced or not
- B:** The power of contrasts between all pairs for Fixed Effects
- C:** The power of contrasts between all pairs for Random Effects
- D:** The number of repetitions required to achieve power between each of the pairs for Fixed Effects
- E:** The number of repetitions required to achieve power between each of the pairs for Random Effects

### Example 2.3.1

The investigator from Example 2.1.3 who was running a 2 period, 5 treatment trial (Table 2.1.2) would like to know the power of contrasts between all the treatments, and what sample size is required for each to reach 90%.

Figure 2.3.2: Output from Example 2.3.1

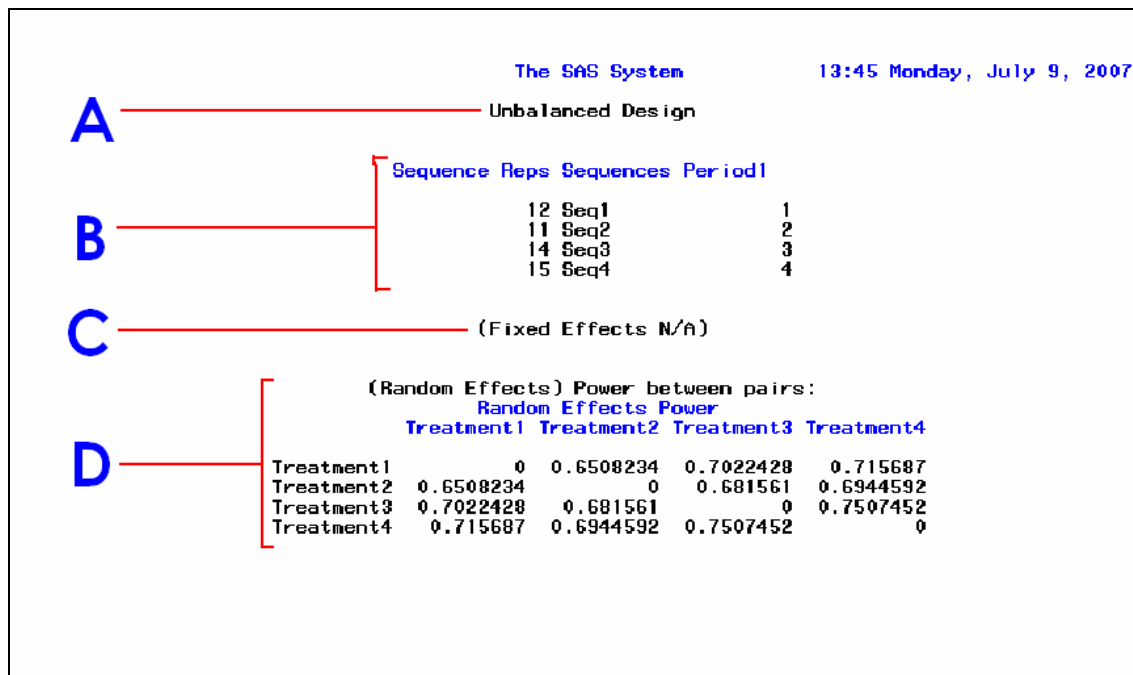
Unbalanced Design					
(Fixed Effects) Power between pairs with 2 repetitions.					
Fixed Effects Power					
	Treatment1	Treatment2	Treatment3	Treatment4	Treatment5
Treatment1	0	0.1384528	0.1066142	0.1066142	0.1384528
Treatment2	0.1384528	0	0.1384528	0.1066142	0.1066142
Treatment3	0.1066142	0.1384528	0	0.1384528	0.1066142
Treatment4	0.1066142	0.1066142	0.1384528	0	0.1384528
Treatment5	0.1384528	0.1066142	0.1066142	0.1384528	0
(Random Effects) Power between pairs with 2 repetitions.					
Random Effects Power					
	Treatment1	Treatment2	Treatment3	Treatment4	Treatment5
Treatment1	0	0.1634582	0.139711	0.139711	0.1634582
Treatment2	0.1634582	0	0.1634582	0.139711	0.139711
Treatment3	0.139711	0.1634582	0	0.1634582	0.139711
Treatment4	0.139711	0.139711	0.1634582	0	0.1634582
Treatment5	0.1634582	0.139711	0.139711	0.1634582	0
(Fixed Effects) Required number of complete Replications to attain 0.9 power by pair					
Fixed Effects Repetitions Required					
	Treatment1	Treatment2	Treatment3	Treatment4	Treatment5
Treatment1	0	18	26	26	18
Treatment2	18	0	18	26	26
Treatment3	26	18	0	18	26
Treatment4	26	26	18	0	18
Treatment5	18	26	26	18	0
(Random Effects) Required number of complete Replications to attain 0.9 power by pair					
Random Effects Repetitions Required					
	Treatment1	Treatment2	Treatment3	Treatment4	Treatment5
Treatment1	0	14	18	18	14
Treatment2	14	0	14	18	18
Treatment3	18	14	0	14	18
Treatment4	18	18	14	0	14
Treatment5	14	18	18	14	0

The output from program 2.3 shows the design is unbalanced. Power between each pair is either 0.107 or 0.138 for fixed effects and 0.140 or 0.163 for random effects. The random effects approach is more powerful, and leads to lower numbers required for 90% Power, between 14 and 18 reps (70 or 90 subjects). Fixed effects required sample size is 18 or 26 reps (90 or 130 subjects).

### **Program 2.4 SAS Program for Normal Data 4**

This program for crossover or parallel trials with normally distributed data gives power and required sample size for all pairs simultaneously, and allows the user to input an irregular treatment sequence.

Figure 2.4.1: Output from Program 2.4



- A:** Displays whether design is balanced or not
- B:** List of the sequences and the number of subjects assigned to each
- C:** The power of contrasts between all pairs for Fixed Effects. Here, fixed effects analysis is not possible, so a message explaining this is displayed instead
- D:** The power of contrasts between all pairs for Random Effects

### ***Program 2.5: SAS Program for binary Data***

This is a program to calculate powers and sample sizes for trials with binary outcomes, for AB/BA crossover designs or parallel trials with two treatments.

Like the Normal programs, it requires the parameters of the trial be entered and it will calculate power and sample size according to the four different binary methods mentioned in the previous chapter.

Figure 2.5.1: Output from Program 2.5

		The SAS System		13:45 Monday, July 9, 2007	
<b>A</b>	[	pA:	0.4	pB:	0.25
		OR:	2		
		alpha:	0.05	beta:	0.1
				<b>N</b>	
		CrossOver with	203 subjects:		
<b>B</b>	[	Power (OR Parallel Meth)	0.9032523 ,	202 subjects required for	90 % power
		Power (Approx OR Method)	0.9050409 ,	200 subjects required for	90 % power
		Power (Conner)	0.8957662 ,	206 subjects required for	90 % power
		Power (Miettinen)	0.9013489 ,	204 subjects required for	90 % power

**A:** The parameters used in the calculation

**B:** The power of the design, calculated by different methods, and the number of subjects required by each of the methods to achieve the requested power

### Example 2.5.1

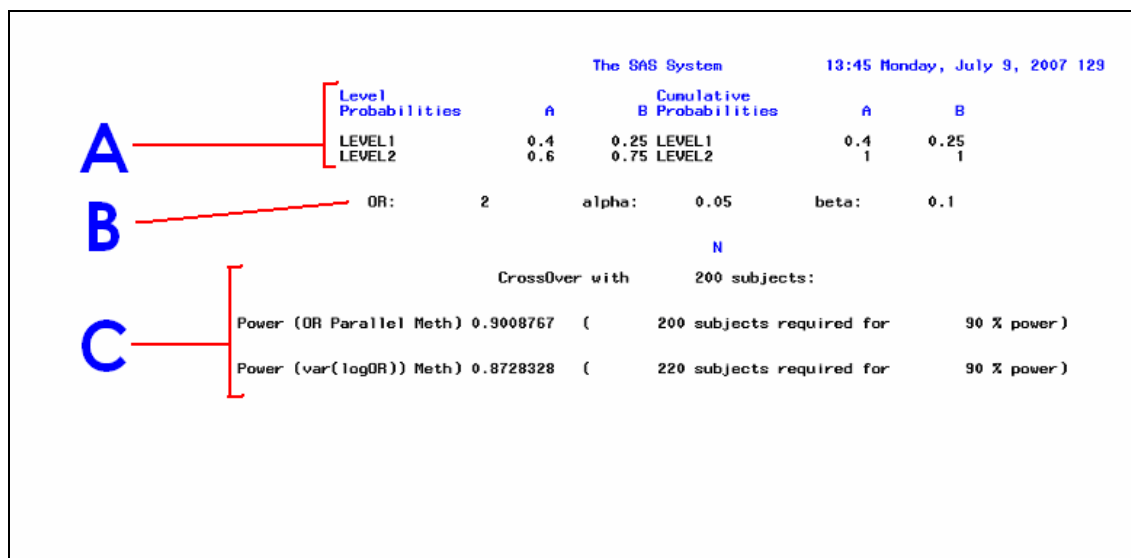
The incidence of a condition for subjects on treatment A is 0.4. A trial is set up to see if new drug B is would lower the incidence. With a two-sided alpha of 5%, what is the required sample size of an AB/BA crossover to achieve 90% power of detecting a clinically significant OR of 2?

Output from such a query is shown on fig 2.5.1. Between 200 and 206 subjects would be required, depending on method of calculation.

**Program 2.6: SAS Program for ordinal data**

This is a program to calculate powers and sample sizes for trials with ordinal outcomes, for AB/BA crossover designs or parallel trials with two treatments.

Figure 2.6.1: Output from Program 2.6



- A:** The level and cumulative probabilities each level assumed for both treatments
- B:** The OR, two-sided alpha and the beta used in calculations
- C:** The power of the design as calculated by the different methods and the minimum sample size required to achieve desired power

## **Chapter 3 – R Programs for calculating sample size**

This chapter is about R programs that could be used to calculate powers and sample sizes for a variety of trial designs with a variety of types of response variable. A justification for the use of R was given at the start of Chapter 2. These programs will make use of the Rpanels [Bowman, 2006] software add-on to R.

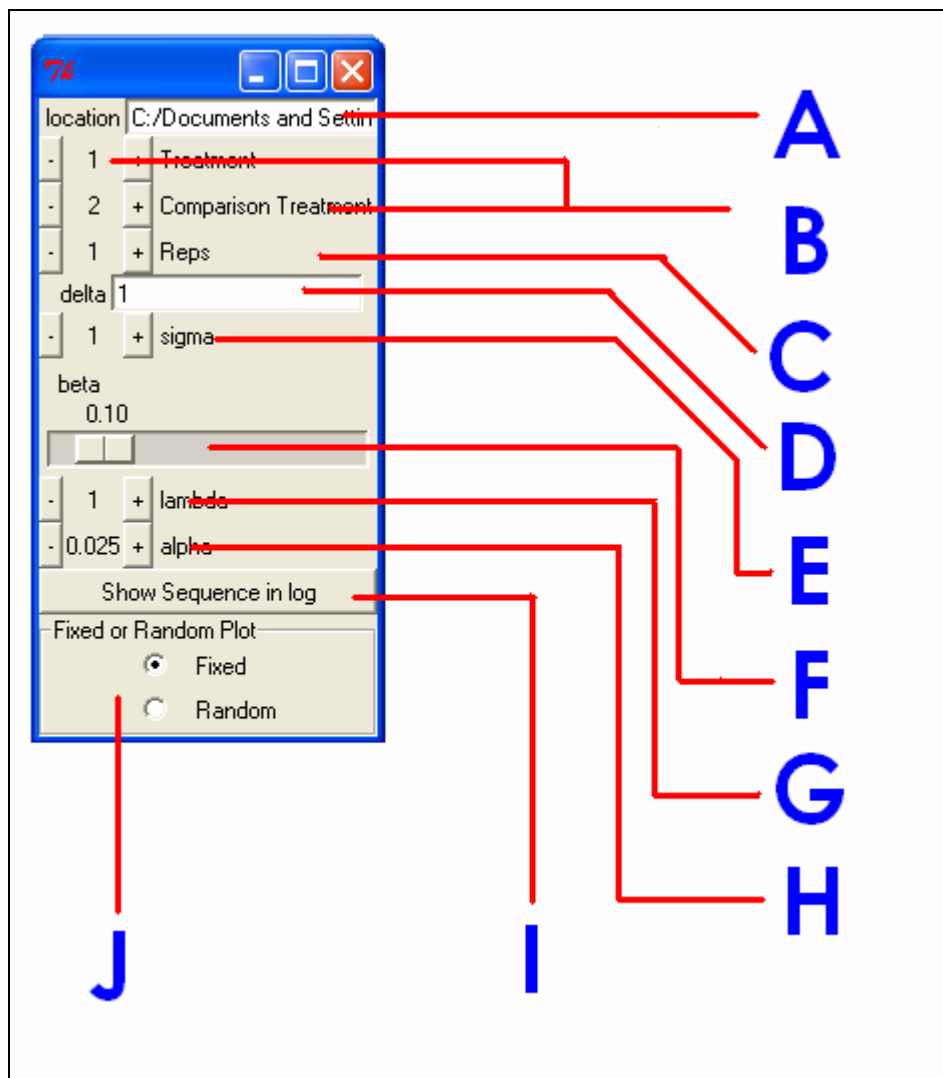
As for the SAS programs in the previous chapter, here there are programs to deal with Normal, binary and ordinal data.

### ***Program 3.1: R panel program for Normal data***

This program can be used to calculate required sample sizes and power of contrasts for crossover or parallel designs. The statistical steps taken are similar to those taken for program 2.1.



Figure 3.1.1: Interface for Program 3.1



**A:** Location of the txt file containing the information on the treatment sequences for the design. Press return after entering the location for the program to recognise the new instruction. Notice that for R the directory structure of a file location uses the “/” forward slash separator, not the “\” backslash.

- B:** The treatments, identified by number, whose contrasts are to be analysed. Press the “-” and “+” buttons to change the treatments compared.
- C:** Number of repetitions of the treatment sequences to be used in trial. Click on the “-” and “+” buttons to change the number of repetitions.
- D:** The  $\delta$  of the design, the size of a difference to be detected. Press return after entering the new delta for the program to recognise the new instruction.
- E:** The  $\sigma_{\text{within}}$ , the within-patient standard deviation. Click on the “-” and “+” buttons to decrease or increase the value.
- F:** The  $\beta$  of the trial, the size of the type II error. Click and drag the slider to change the  $\beta$ , and thus  $1 - \beta$ , the desired power of the trial.
- G:**  $\lambda$ , the ratio of  $\sigma^2_{\text{between}}$  to  $\sigma^2_{\text{within}}$ . Click on the “-” and “+” buttons to decrease or increase the value.
- H:** The one-sided  $\alpha$ , the size of the type I error. Click on the “-” and “+” buttons to decrease or increase the value.
- I:** Button to display information on the selected treatment sequence. Information is outputted to the R log.
- J:** Radio button to change between Fixed and Random effects plots and analysis.

Figure 3.1.2: Entering sequence information

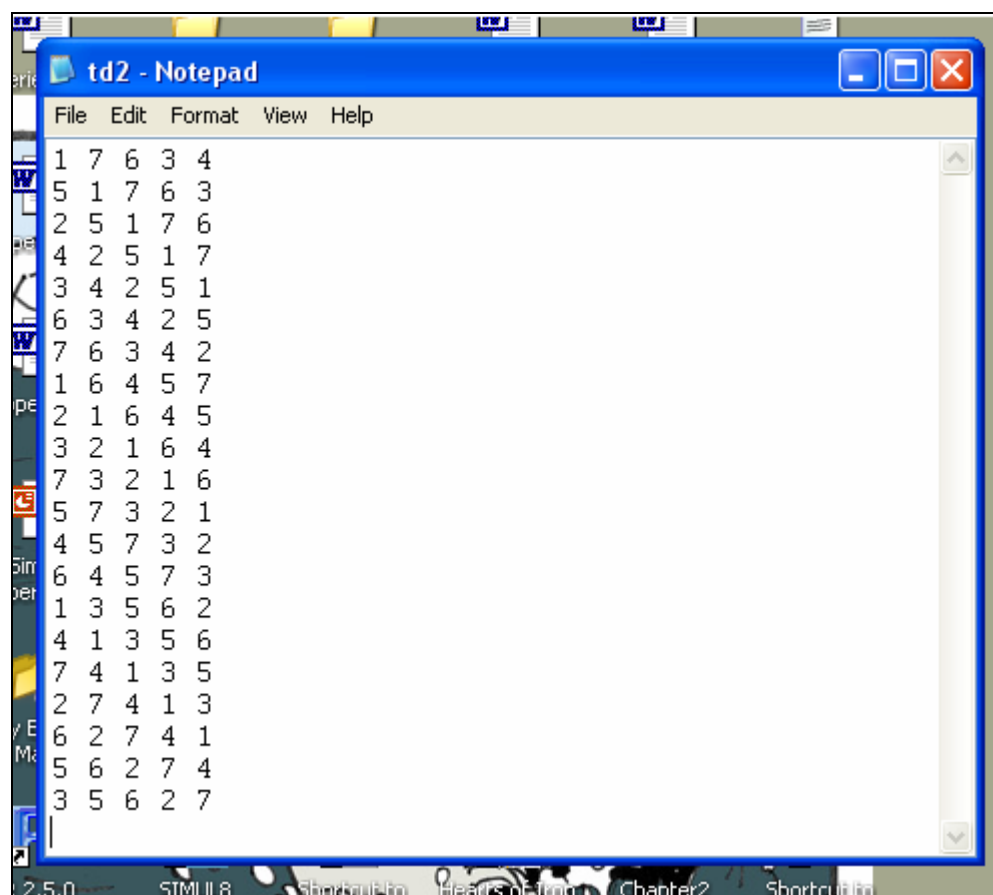


Fig 3.1.2 shows how the 5-period, 7-treatment, 21 patient incomplete block crossover design [Vollmar J, and Hothorn LA 1997] would be entered into a .txt file. Each row represents a different subject's regimen, and lines should be separated by a single carriage return (press of the return key). Each number represents a different treatment type, and one space should be left to separate the treatments assigned to each period.

Figure 3.1.3: Output from Program 3.1

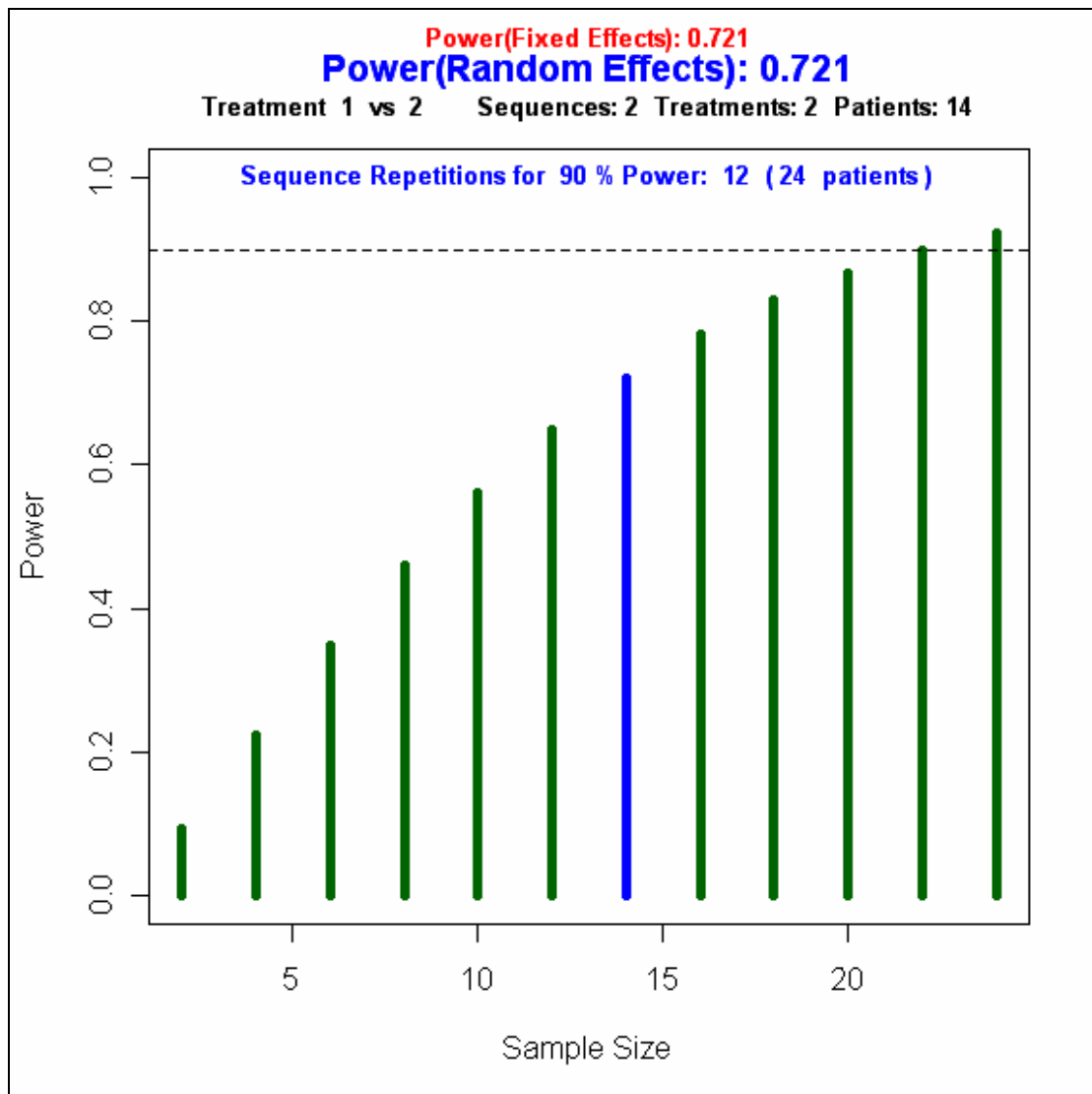
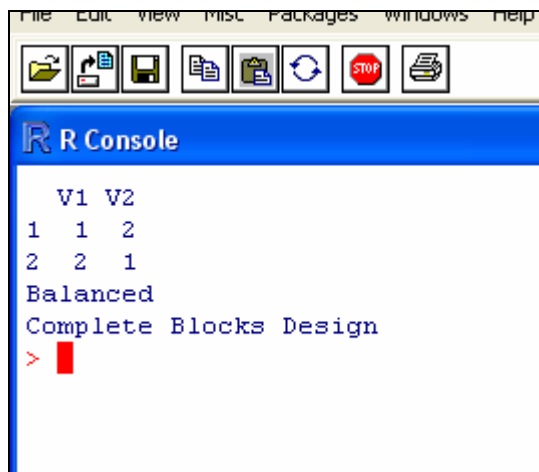


Figure 3.1.3 shows the graphical output from program 3.1 for an AB/BA crossover trial with 14 subjects analysed by random effects. A histogram shows the relation between sample size and power, for either fixed effects or random effects. The sample sizes are plotted from zero up to the number required to exceed the stated power, or up to size determined from 3.1.1C, whichever is greater. The bars are green except for the bar for the stated

reps. If Fixed Effects is selected in 3.1.1J, then the bar will be red. Random effects would be plotted in blue. The required power is shown by a dashed line. No matter which of Fixed or Random effects are to be analysed, the power of the sample size is displayed above the plot for both methods. Fixed effects is in red and Random is in blue, just like in the histogram proper. The selected model's results are shown in a larger font. In this example with a complete blocks crossover the power calculated is the same for each method, 72.1%. Also displayed are the treatments to compared, the sample size, the total number of treatments and the number of sequences. Below that the calculated sample size to achieve the power is displayed for the selected model using the same colour scheme to identify model type. Here 24 people are needed to achieve a power of 90%.

Figure 3.1.4: Output from Program 3.1



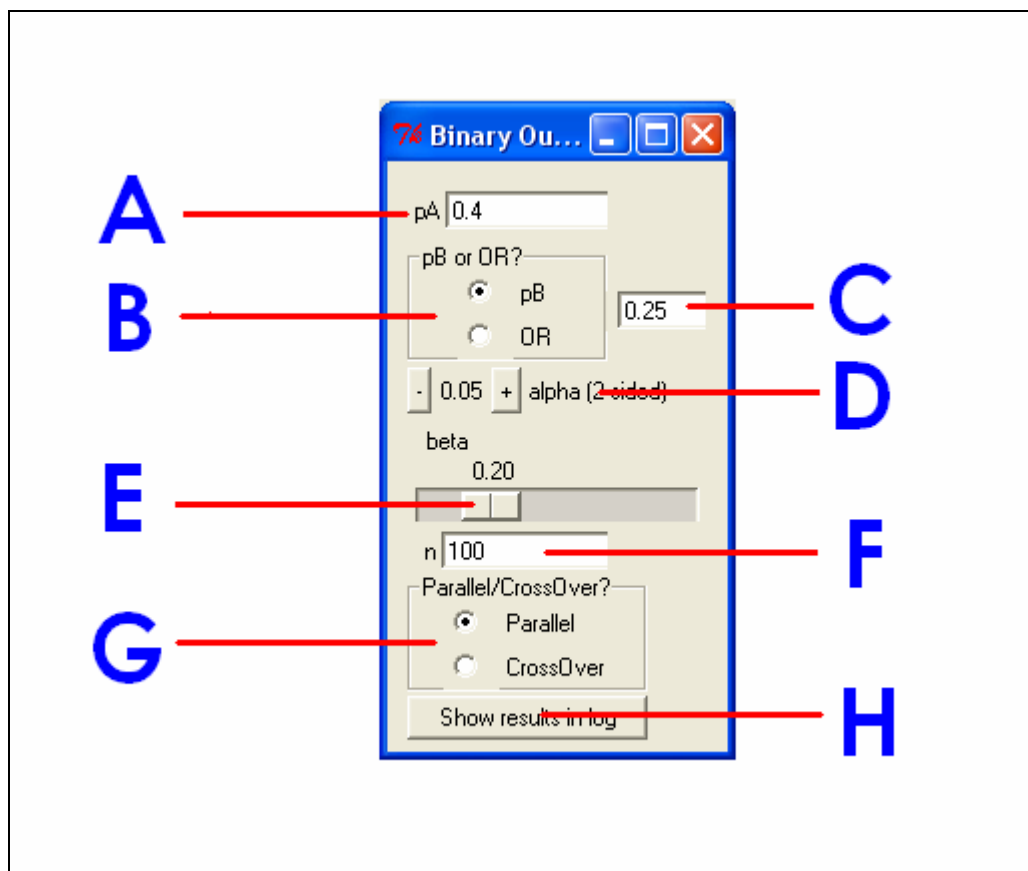
```
File Edit View Misc Packages Windows Help
[Icons]
R Console
  V1 V2
  1  1  2
  2  2  1
Balanced
Complete Blocks Design
> █
```

Figure 3.1.4 shows the output in the log for an AB/BA crossover trial, generated when button 3.1.11 is clicked. First, it displays the sequence that is being analysed, in the same format as it was entered (See the analysis for figure 3.1.2, above, for more details.). On the next line it is revealed if the design is balanced or not, and below that is a summary of the type of design. In this example the design is correctly assessed as a balanced complete blocks design.

### ***Program 3.2: R panel program for binary data***

This program allows the user to calculate sample sizes for trials with binary outcome response variables. It uses the sample size calculation methods mentioned in chapter 1.

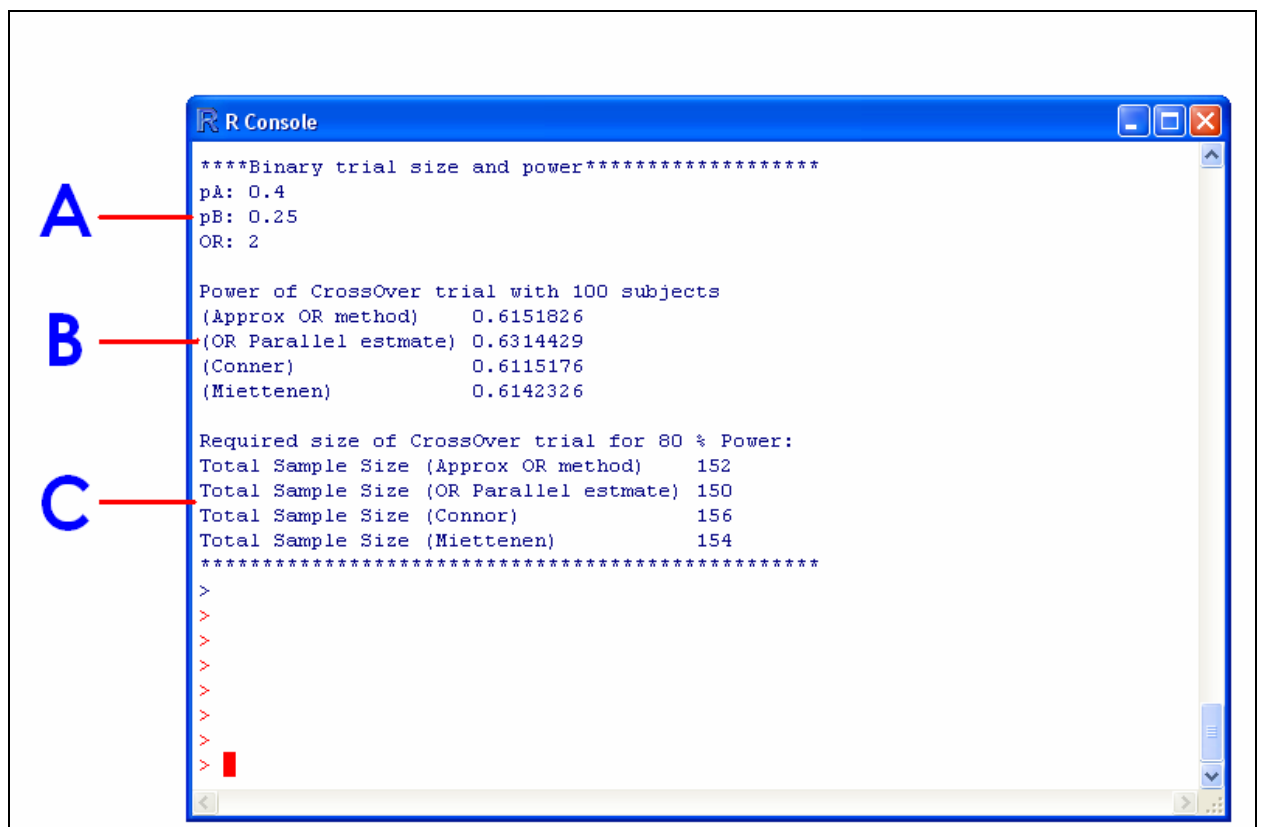
Figure 3.2.1: Interface for Program 3.2



- A:** pA, the response anticipated on treatment A.
- B:** Radio button to select between inputting Odds Ratio or pB, the response anticipated on treatment B.
- C:** Text entry box to enter either Odds Ratio or pB.
- D:** The two-sided  $\alpha$ , the size of the type I error. Click on the "-" and "+" buttons to decrease or increase the value.
- E:** The  $\beta$  of the trial, the size of the type II error. Click and drag the slider to change the  $\beta$ , and thus  $1 - \beta$ , the desired power of the trial.

- F:** Text entry box to enter n, the prospective size of the trial. For parallel trials n will be size per arm, and for crossover trials n will be the total size of the trial.
- G:** Radio button to choose between parallel trial and crossover trial.  
Parallel trials are trials with two different treatments or treatment levels with equal allocation to each arm. Crossover trials are AB/BA designs with equal allocations to each sequence.
- H:** Button to show results in log.

Figure 3.2.2: Output from Program 3.2





The output, as shown in figure 3.2.2, is to the log. First, it reminds the user of the variables used (**A**). The pA, pB, and OR are all shown- the pB, or the OR if not entered, is calculated.

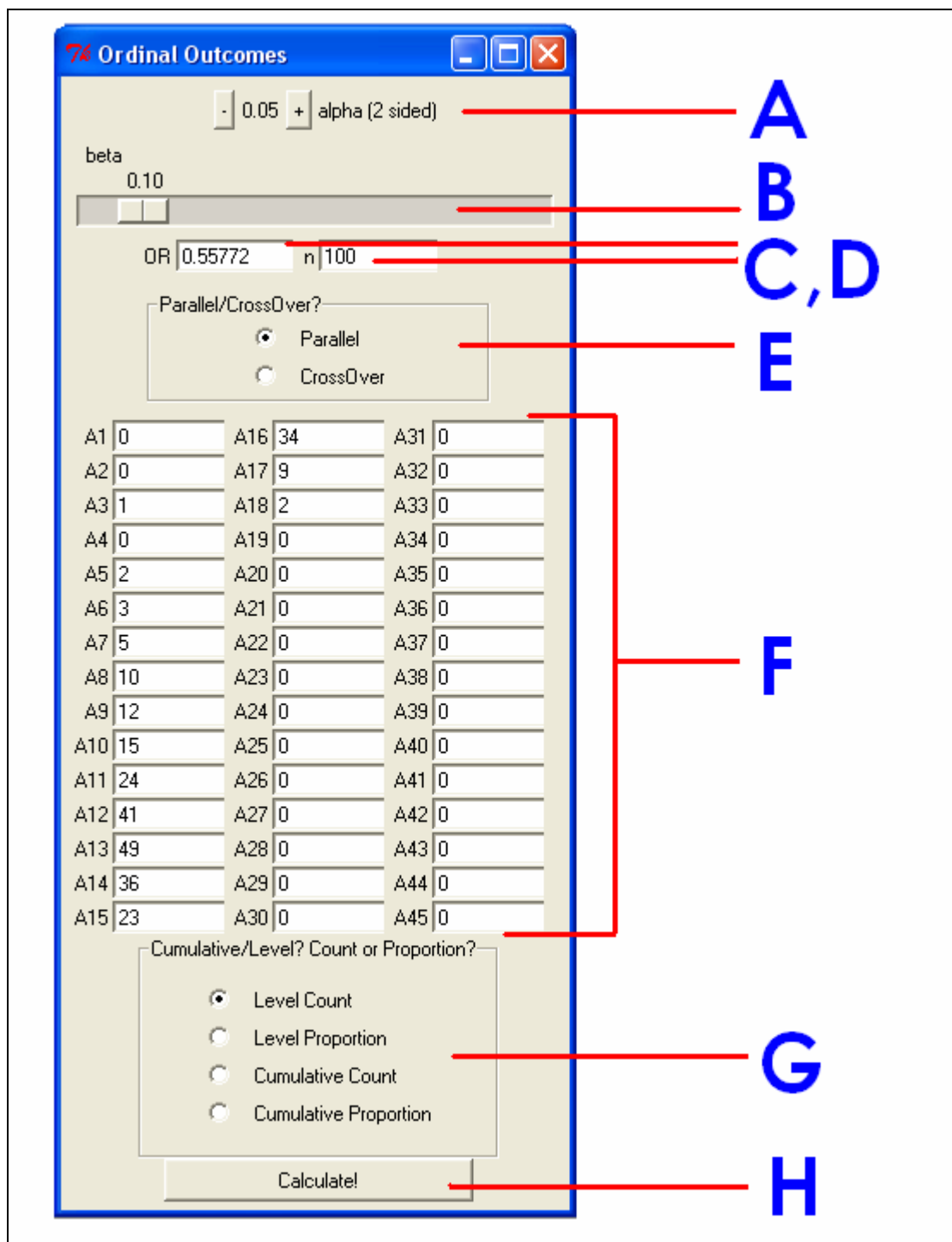
Next, (**B**), the power of the contrast between treatments A and B are displayed for the trial type and sample size. Here, there is a crossover trial with 100 subjects. For crossover type design there are four methods of power calculation used, and usually they give similar results; in this example they give powers of between 61.2% and 63.1%. Conner's method gives the lowest score, which is also typical. If a parallel trial type was selected there are two methods of power calculation used.

The required sample sizes to achieve the required power are shown for each of the power calculation methods. Like the power calculations, these are also typically similar for the different methods, but Conner is usually the highest. In this example the required sample size to get a power of at least 80% is estimated between 150 and 156.

### ***Program 3.3: R panel program for ordinal Data***

Program 3.3 can be used to calculate sample sizes and power of trial designs for treatments with an ordinal type response variable for up to 45 levels of response.

Figure 3.3.1: Interface of Program 3.3



- A:** The two-sided  $\alpha$ , the size of the type I error. Click on the “-” and “+” buttons to decrease or increase the value.
- B:** The  $\beta$  of the trial, the size of the type II error. Click and drag the slider to change the  $\beta$ , and thus  $1 - \beta$ , the desired power of the trial.
- C:** Text box to enter the Odds Ratio between the two treatments.
- D:** Text entry box to enter  $n$ , the prospective size of the trial. For parallel trials  $n$  will be size per arm, and for crossover trials  $n$  will be the total size of the trial.
- E:** Radio button to choose between parallel trial and crossover trial.
- Parallel trials for this program are trials with two different treatments or treatment levels with equal allocation to each arm. Crossover trials are AB/BA designs with equal allocations to each sequence.
- F:** A set of 45 text boxes, to enter the anticipated responses on treatment A at each level. The levels are ordered. There are a few different formats that the anticipated response can be entered, and the preferred format can be selected with dial G. For trials with fewer than 45 levels, simply set the probability of unused levels to zero.
- G:** Radio button to choose format of treatment effect in text boxes F. You can choose between Level Count, Level Probability, Cumulative Count and Cumulative Probability.
- H:** Button to activate the calculations. Information summarizing the calculations will be output to the R log.

The ability the program gives to plan a trial with over 40 levels of ordinal data shouldn't be taken as a recommendation that data with the number of levels be analysed as ordinal data. Some [Glass et al, 1972] recommend that ordinal data with as few as 7 levels be analysed as if it were continuous, others are less keen [Blair, 1981]. The number of levels available was deliberately chosen to be excessively high to allow any conceivable ordinal based trial to be planned for, while still maintaining a usable interface. Program 3.3's design philosophy was "Better to have options available you never need, than an option you need that is not available", and the user is still able to plan for far fewer levels. For those who would never need anywhere close to 45 levels in their planning there is a similar program (Program 3.3.15) with a more-reasonable 15 level maximum included in the appendix.

Figure 3.3.2: Output from Program 3.3

```

***Ordinal Outcome Sample Size and Power*****
alpha: 0.05
beta: 0.1
OR: 0.55772

Cumulative proportion A and B, Level proportion A and B
Level 3 A: 0.004 B: 0.007 A: 0.004 B: 0.007
Level 5 A: 0.011 B: 0.02 A: 0.008 B: 0.013
Level 6 A: 0.023 B: 0.04 A: 0.011 B: 0.02
Level 7 A: 0.041 B: 0.072 A: 0.019 B: 0.032
Level 8 A: 0.079 B: 0.133 A: 0.038 B: 0.061
Level 9 A: 0.124 B: 0.203 A: 0.045 B: 0.069
Level 10 A: 0.18 B: 0.283 A: 0.056 B: 0.081
Level 11 A: 0.271 B: 0.4 A: 0.09 B: 0.117
Level 12 A: 0.425 B: 0.57 A: 0.154 B: 0.17
Level 13 A: 0.609 B: 0.736 A: 0.184 B: 0.167
Level 14 A: 0.744 B: 0.839 A: 0.135 B: 0.103
Level 15 A: 0.831 B: 0.898 A: 0.086 B: 0.059
Level 16 A: 0.959 B: 0.977 A: 0.128 B: 0.078
Level 17 A: 0.992 B: 0.996 A: 0.034 B: 0.019
Level 18 A: 1 B: 1 A: 0.008 B: 0.004

Size of Parallel trial required for 90 % Power:
OR Method: 188 per arm ( 376 total)

Power of Parallel trial with 100 subjects per arm ( 200 total):
OR Method: 0.6577454

>

```

The output from this program will be similar to that shown in Figure 3.3.2. The alpha, beta and OR for the calculation is displayed first **(A)**.

From the pA level proportions or counts and the OR the cumulative and level proportions expected from treatments A and B are computed. They are displayed **(B)**, and they can be surveyed to check for errors in data entry. Attempts are made in computation to find logical inconsistencies, and warning messages will be displayed if problems are found. The output here can be useful but, unfortunately, untidy.

The sample size required to achieve the required power is displayed next **(C)**. In this example a parallel trial is calculated to need 376 subjects to get the desired 90% power. Crossover sample sizes can also be calculated, there are two methods used.

The power from a stated sample size is also shown. Here, **(D)**, a power of 65.8% results from a stated sample size of 200.

***Some Comparisons with other software and standard tables,  
with Discussion***

We shall now compare the results from the programs in the past two chapters with software and tables currently used for sample size by researchers today.

It is necessary to be sure of the results

nQuery Advisor, by Statistical Solutions Ltd., is the probably most popular dedicated sample size and power software used for trial design. User friendly and flexible, it can deal with a variety of trial designs and is used by trial designers around the world, but it cannot deal with the more complicated multi period or incomplete block designs discussed in this thesis. Machin (1997) gives a set of tables to help with trial size calculation, but again, they are not so useful for complex trials. We can only compare programs 2.1 & 3.1 with these sources, and only for parallel trials or AB/BA type crossover.

## **1. Parallel Trial sample size, Normal Data**

Here is an example from Sample Size Tables for Clinical Studies, 2<sup>nd</sup> Edition [Machin D et al, 1997]: “(Example 3.12) An investigator compares the change in blood pressure due to placebo with that due to a drug. If the investigator is looking for a difference between groups of 5 mmHg, then, with a between-subject SD as 10 mmHg, how many patients should he recruit? How is the calculation affected if the anticipated effect is 10 mmHg?”

The authors calculate the required sample size to detect 5 mmHg to be 172, and 44 for 10 mmHg.

The answers calculated by program 2.1 would also be 172 for 5 mmHg, but it calculates a larger sample of 46 should be used for 10 mmHg. Using nQuery also gets answers of 172 and 46.

Fig 3.4: Machin's sample size formula

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} + \frac{z_{1-\alpha/2}^2}{4}$$

The difference is due to Machin et al using a corrected-Normal approximation of the cumulative non-central t distribution to calculate sizes. Machin's formula is based on a version of the general equation 1.2 with a for the special case of a parallel trial with an equal number of subjects in each arm, but with a correction applied to better approximate equation 1.3. This correction means Machin's calculations are better than with raw 1.2 equations and still easy to calculate with pen and paper, but the slight underestimation of the required sample size for Normal data would be a fairly regular occurrence with this method.

Another example from Sample Size Tables for Clinical Studies, [Machin D et al, 1997], describing a trial carried out by Wollard and Couper (1981) for comparing moducren with propranolol as initial therapies in essential hypertension. "They proposed to compare the change in blood pressure due to the two drugs. Given that they can recruit only about 50 patients for each drug, and that they are looking for a 'medium' effect size of about  $\Delta = 0.5$  what is the power of the test, given a two-sided significance level of  $\alpha=0.05$ ?"

Using standard tables Machin et al get an answer of 'about 0.70'. By program 2.1, taking advantage of the fact that the effect of a two-sided alpha of 0.05 is virtually the same as a one-sided 0.025, we can calculate a power of 0.697, a result consistent with Machin's answer. This shows, for simple designs, that sets of sample size tables like Machin's are usually adequate. Researchers may also find it more convenient to flick through a tables book rather than boot up a computer and load a software package, and wonder if they have entered the correct data to get the desired result.

## **2. Crossover trial, Normal Data**

Julious (2005) has a list of sample sizes for different  $\Delta$  for 90% power for AB/BA crossover trials. These match results from Program 2.1 and 3.1. For example, for  $\Delta$  of 0.1 Julious gives a required sample size of 2104. Programs 2.1 and 3.1 also give that result. Again, a set of sample size tables is sufficient for this simple type of crossover trial, as long as the  $\Delta$  and type I and type II error rates are conventional.

## **3. Parallel Trial sample size, binary Data**

Again from Machin et al, in their example 3.1 they ask: With a two-sided alpha =0.10,  $\pi_1= 0.25$  and  $\pi_2=0.65$ , how many subjects are needed for a parallel trial with equal numbers in each arm to achieve a power of 0.9? Their answer is 25 in each arm, 50 in total.

From 2.5, by Proportional Difference method 23 each arm, totalling 46. By Odds Ratio, 24 are needed each arm, making 48 overall. These results are both fairly similar.



Machin goes on to ask how changing two-sided alpha to 0.05 effects sample size, before giving 62 as the answer. By 2.5 we get 56 by prop diff, and 58 by OR.

#### **4. Crossover Trial, Binary data**

From Julious (2005): “An investigator wishes to design a study where the marginal response anticipated on the control therapy is 40%. The effect of interest is 2.0 in favour of the control therapy and the investigator wishes to design the study with Type I and II errors fixed at 5% and 10% respectively.” Julious gets a result of “approximately 200”. With  $p_A = 0.4$ , Type I error = 5% and  $OR=2$ , both programs get results of between 200 and 206 for all the methods. This matches Julious.

#### **5. Parallel Trial, Ordinal data**

Machin gives an example (3.10) with control levels  $p_{A1} = 0.14$ ,  $p_{A2} = 0.24$ ,  $p_{A3} = 0.24$  and  $p_{A4} = 0.38$ . With an alpha of 5%, what sample size is required to have 80% power to detect an OR of 3? Machin’s answer is “approximately 90”. Using program 3.3, we would calculate a sample size of 92 subjects to achieve that power by the OR method. This is very close to Machin’s result, which is reassuring.

#### **6. Crossover Trial, Ordinal data**

Julious gives an example, with  $p_{A1} = 0.08$ ,  $p_{A2} = 0.191$ ,  $p_{A3} = 0.473$  and  $p_{A4} = 0.256$ . What sample size required, when alpha = 0.05, to detect an OR of 0.56 with 90% power? He gives answer of 213 or 229, depending on method.

By this thesis' programs almost identical results are achieved, 214 or 230 depending on preferred method.

## **7. Incomplete Block Design, Normal data.**

Because no other program or table could be found that can help calculate power for this type of design, another method was found to validate the programs.

With a sample size of 39, with 13 subjects on each possible sequence, what is the power of a 2-period, 3 treatment balanced incomplete blocks trial when  $\delta = 1$ ,  $\sigma = 1$ ,  $\lambda = 1$ , and a 1-sided alpha of 0.025, and subject is treated as a fixed effect?

This trial set-up was simulated 100,000 times, with the proportion of trials that detect superiority approximately equal to the power. On 85,923 occasions a significant difference was found, suggesting a power of around 85.9%. Using program 3.1 or 2.1, one would calculate a power of 86.0%, which is very close. We can now be confident that the program gives sensible results for incomplete block crossover trials.

## **8. Discussion of comparisons.**

The comparisons between the thesis's programs with nQuery software and the sample size tables show a very strong agreement. They may disagree slightly, but not so much as is likely to have a significant impact on trial

design. This shows the validity of the previous 2 chapter's programs. The inability of nQuery and Machin's tables to provide a comparison with more complicated trial design meant a simulation based comparison had to be used to check the validity of the sample size for an Incomplete Block design. That showed the limitations of those methods compared with this thesis's programs.

## **Chapter 4**

### ***4.1 The use of sample size calculations***

It has been shown that sample size calculations are of great use to the designer of a trial, potentially saving resources by making sure a non-excessive number of subjects are randomised, and stopping unrealistic trials from taking place. It has also been established, for simpler designs anyway, that tables and software to make these calculations possible are available. Despite this, sample size calculations are not universally utilized. In a recent survey of surgical trials [Maggard et al, 2003] only 38% of studies reported using any kind of sample size calculation before the trial. This leads to silly situations like many such trials running at one-tenth the size needed for a reasonable chance of detecting even a moderate treatment effect. Again, this is very poor from both a scientific and ethical viewpoint.

Perhaps almost as worrying, even when sample size calculations are known to have taken place they are often inaccurate [Freiman et al, 1978][Vickers, 2003]. This chapter will look at the conventions and problems with parameter selection and estimation for sample size calculations. It will consider the imperfect nature of estimates for variance, and make suggestions on how best to deal with this uncertainty with alternative sample size calculation methods.

### ***4.2 Alpha, beta and the treatment difference***

Acceptable levels of type I and type II error must be factored into sample size calculations. The convention is that it is much preferred that a type II error (failing to reject a false null hypothesis) is made than a type I error (rejecting a true null hypothesis), so  $\alpha$  will be smaller than  $\beta$ .

For sample size calculations for superiority trials a 1-sided  $\alpha$  of 0.025 is normally used. The  $\beta$  is more flexible, with values used in trials ranging between 0.4 and 0.1, meaning desired power of between 60% and 90%. There are regulatory and financial obligations to make sure clinical trials have a reasonable chance to achieve useful results, a trial with too low a power is judged unethical in some situations.

A clinical expert would determine how to quantify  $\delta$  or OR, the clinically significant difference. The smaller the magnitude of  $\delta$  or  $\log(\text{OR})$ , the larger the sample size required. The  $p_B$  can be calculated from  $p_A$  and OR if the absolute difference between  $p_A$  and  $p_B$  is of primary interest rather than the OR.

### ***4.3 Sample standard deviation as an estimator of population standard deviation***

Power and sample size calculations for Normally distributed data require that a value for population standard deviation be entered into an equation.

Equation 1.3 treats the true population standard deviation as being known, but this is not realistic. To know the true population variance for some endpoint

one would have to measure the attribute in every member of the population in question. For equations like 1.3 a point estimate must be made for sigma.

### **s given sigma**

Before judging how good an estimate of standard deviation is, consider the relationship between the true standard deviation ( $\sigma$ ) and the sample (s). The ratio of  $m*s^2/\sigma^2$  (where  $s^2$  is the sample variance and m is n-1, n being the number of subjects used in the variance calculation) follows a chi-squared distribution with m degrees of freedom.  $s^2/\sigma^2$  follows a chi-squared distribution divided by m,  $s/\sigma$  follows the square root of that distribution.

Figure 4.1

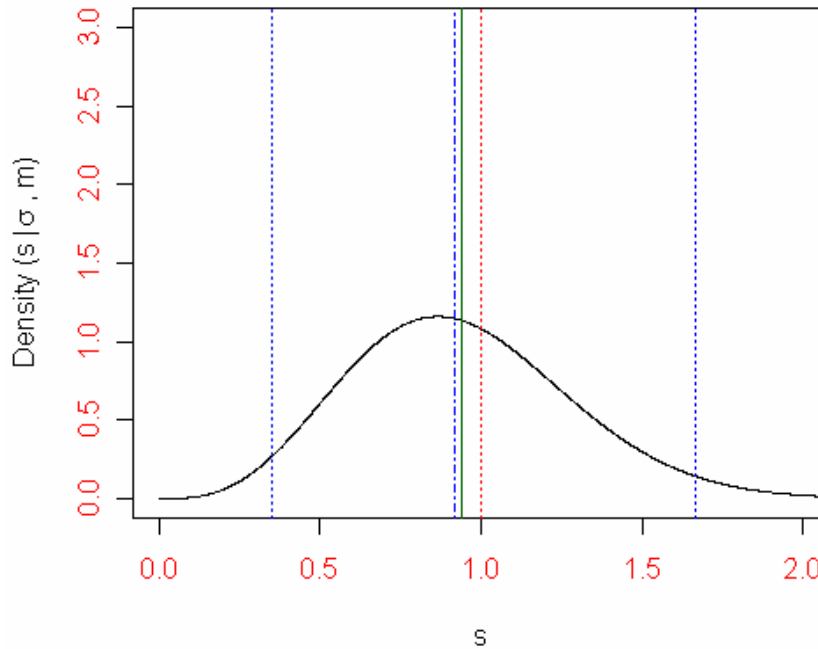


Figure 4.1 shows the pdf of distribution of  $s$ , sample standard deviation from a sample size 5. The red dotted line marks the true value of population standard deviation  $\sigma$ , the blue dotted lines show the 95% PI for  $s$ , the solid green line is the expected value of  $s$ , and the dot-dash blue line is the median  $s$ . The distribution looks almost normally distributed, but it is skewed slightly to the left with a long right tail. The 95% interval is wide but balanced around the true  $\sigma$ .

While  $s^2$  is an unbiased estimator for  $\sigma^2$ ,  $s$  is only an asymptotically unbiased estimator for  $\sigma$ . For any real sample it would be expected to underestimate  $\sigma$ , with a sample size of 5 the expected value of  $s$  is 94% of  $\sigma$ . At 4 degrees of freedom  $s$  will also be an underestimate approximately 59.4% of the time, so power calculations using  $s$  for  $\sigma$  will be overestimated 59.4% of the time. (The

sample  $s$  from a pilot study of size 5 will be less than that observed in a finite-sized trial a little less frequently, for example approximately 56.8% of the time for a trial of size 20).

Using the observed  $s$  from a sample size of 5 as an estimate of  $\sigma$  is undesirable, the prediction interval is very wide so there is no reassurance that  $s$  is close to  $\sigma$ .

Figure 4.2

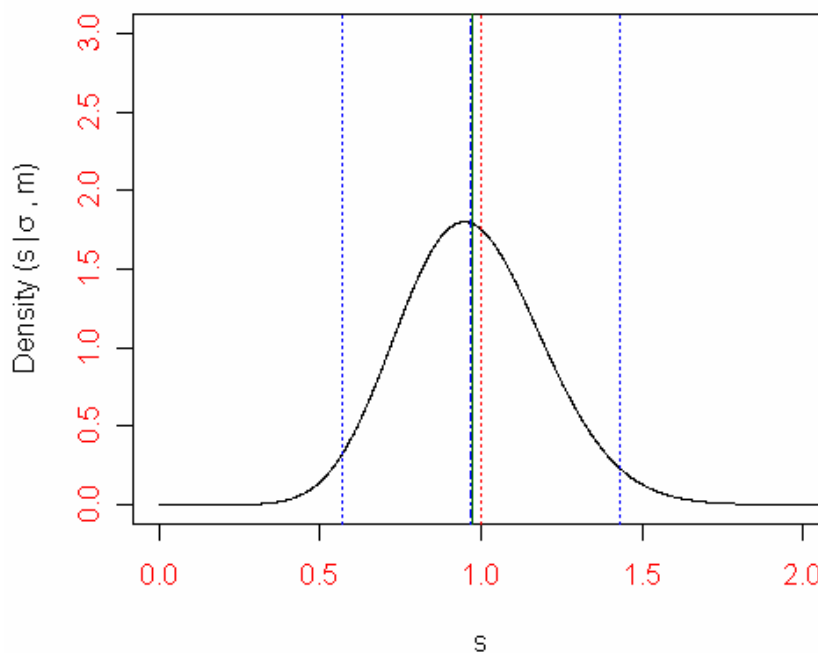
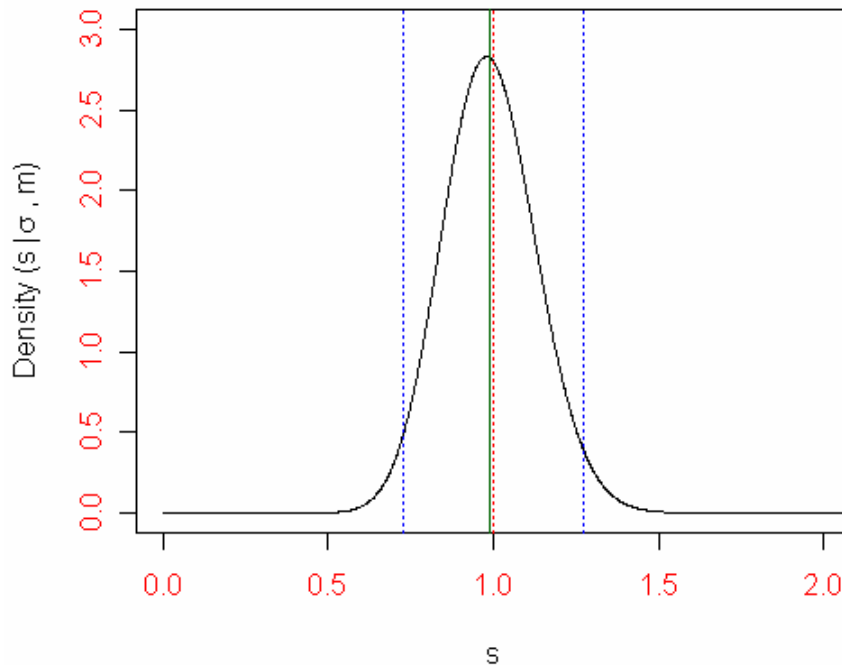


Figure 4.2 shows the distribution of  $s$  from a sample with 10 degrees of freedom. The expected value of  $s$  is 97.5% of  $\sigma$ , better than before, but the prediction interval is wide, a 95% chance  $s$  will be between 56.98% and 143.12% of  $\sigma$ . This large interval means  $s$  is still an unreliable estimator of  $\sigma$ .



Figure 4.3



When sample size increases to 26, the expected value of  $s$  is still under  $\sigma$  but within 1%. The 95% interval for  $s$  is (72.4%, 127.5%), still fairly wide. It is clear that as sample size increases, the more reliable and less biased an estimate  $s$  is of  $\sigma$ . The prediction intervals become very symmetric around the true value of  $\sigma$ . When there are 50 degrees of freedom the expected value of  $s$  is 99.5% of  $\sigma$ , and the prediction interval is about 19.5% either side of  $\sigma$ . By 100 degrees of freedom the prediction interval is about  $\pm 13.8\%$ . A sample size of about 2134 is needed to be 95% confident that observed  $s$  is within 3% of the true value. At this size  $s$  is hardly biased at all, its expected value is around  $0.9999\sigma$ .

The required sample size  $n$  to be  $(1-\beta)*100\%$  confident that  $s$  is within  $x\%$  of  $\sigma$  can be approximated

$$n = \frac{1}{2}(Z_{1-\beta})^2 / (x/100)^2$$

From this formula, one can see that to double accuracy one must quadruple sample size.

### Sigma given s

It is the probability characteristics of sigma given an observed s that are more important, as any estimate we make will be for sigma based on observed s.

Figure 4.4

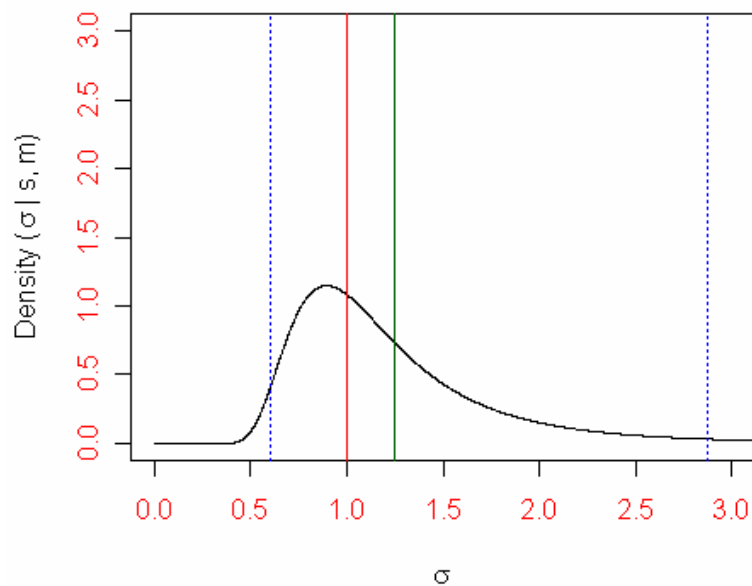


Figure 4.4 shows the pdf of sigma given observed s of 1 from a population of 5. Comparing with fig 4.1, which shows s given sigma, it can be seen that sigma distribution is far more skewed. The 95% confidence interval for  $(\sigma|s)$

for 4 degrees of freedom is much wider and asymmetrical than the 95% prediction interval for  $(s/\sigma|\sigma)$  but they are related- if the 95% PI for  $(s/\sigma|\sigma)$  is  $(a,b)$  then the 95% CI for  $(\sigma/s|s)$  is  $(1/b,1/a)$ .

$$\int_b^a \text{pdf}(s/\sigma|\sigma,m) = \int_{1/a}^{1/b} \text{pdf}(\sigma/s|s,m)$$

There is no such reciprocal relationship between  $E(s|\sigma)$  and  $E(\sigma|s)$ ; the expected value for sigma is 1.2518 times greater than s, while the expected value for s is much closer to  $\sigma$ . Overall, a sample of this size is not good enough to make accurate assumptions about  $\sigma$ . There is a 95% chance  $\sigma$  is between 60% and 287% of the value of s, a very large margin.

Figure 4.5

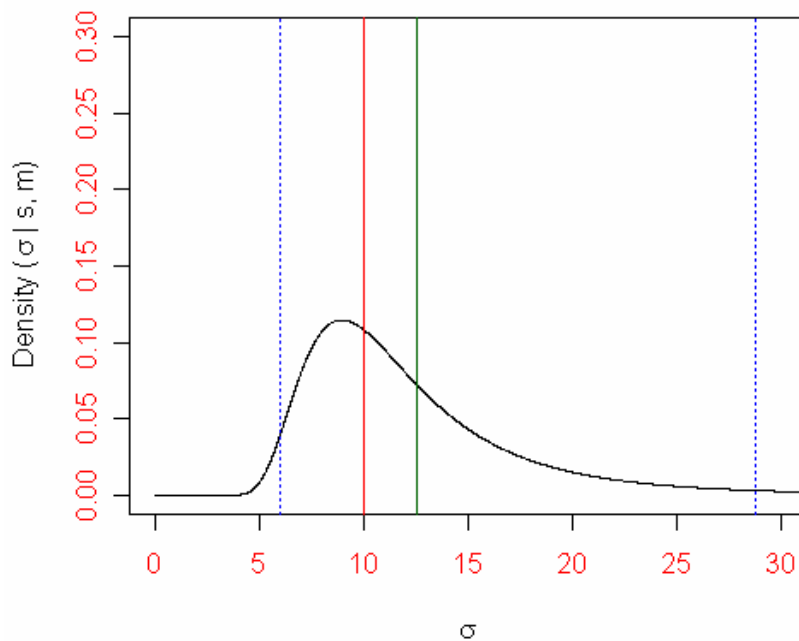
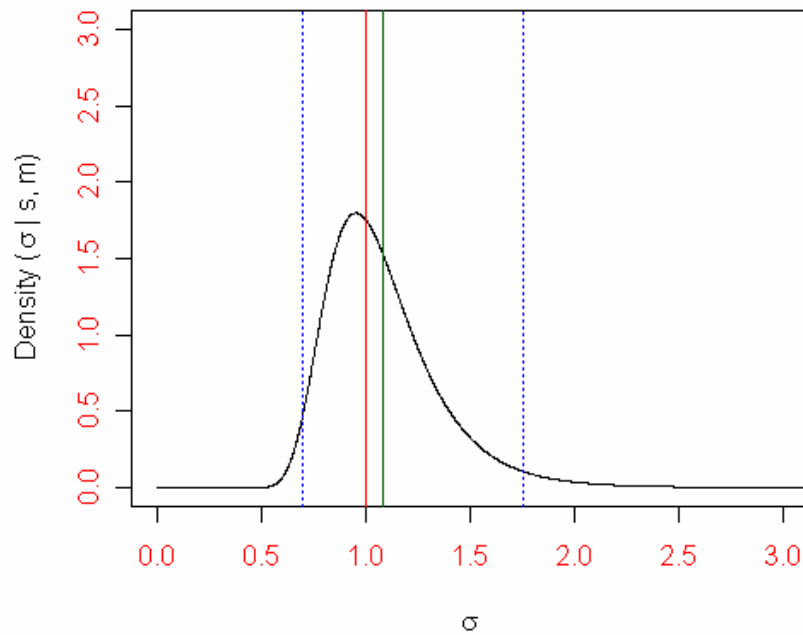


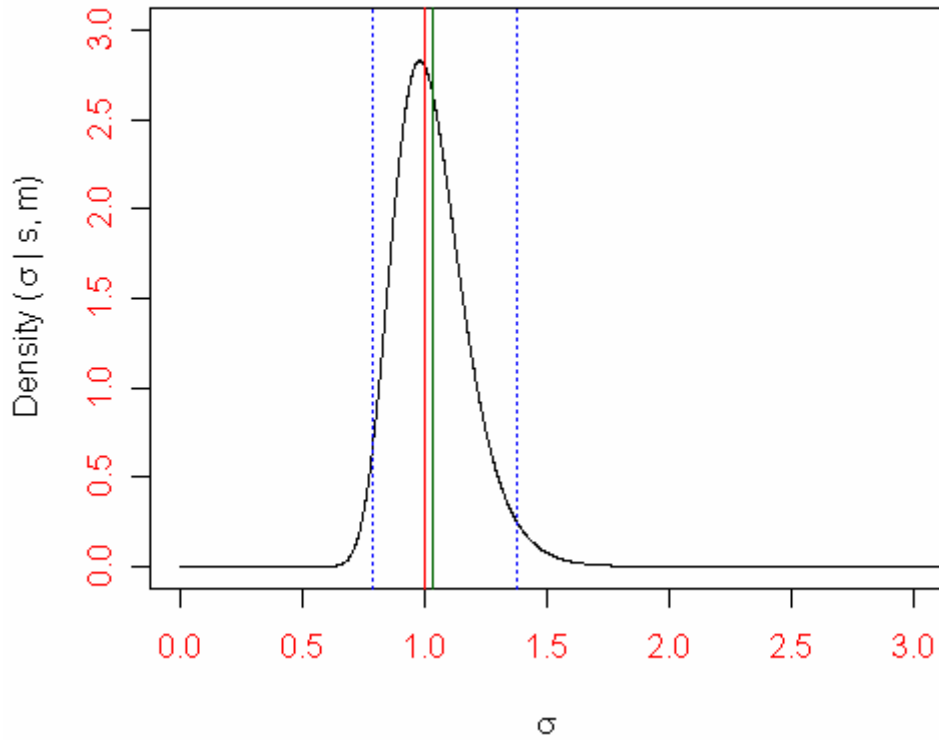
Figure 4.5 shows that the relationship between  $\sigma$  and  $s$  changes if  $s$  was 10 instead of 1 for the same sample size as before. The changes are all proportionate; CIs and Expected values just get multiplied by 10.

Figure 4.6



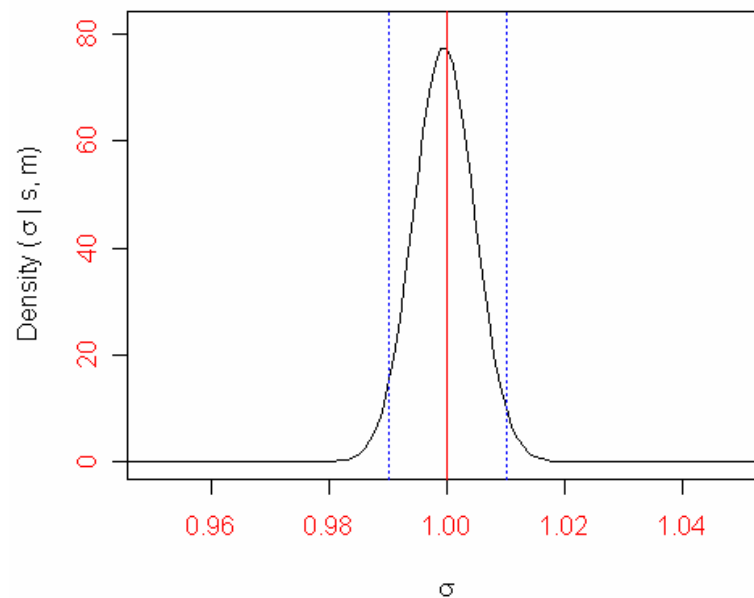
If the sample size were 11, the confidence intervals tighten and are a little more symmetric (0.698717 , 1.754934), and the expected value of sigma drops to 1.0837.

Figure 4.7



By 25 degree of freedom the distribution is looking less skewed and the CI is more symmetric. The expected value of sigma is only 3% greater than s. As the sample sizes increase the trends with skewness and confidence intervals continue. The distributions for  $(s|\sigma)$  and  $(\sigma|s)$  get very similar as m increases.

Figure 4.8



So, small sample sizes give unreliable estimates. It is possible to increase the accuracy of estimated  $\sigma$  by increasing the pilot study, but it takes a quadrupling of the sample size to double accuracy. A sample size of about 19,000 is needed to be confident of  $s$  being within 1% of  $\sigma$  (fig 4.8), so attempting to be this accurate with point estimates for  $\sigma$  will likely devour the resources of the trial designer.

#### ***4.4 Methods of incorporating uncertainty over variance of Normal data into sample size calculations***

It has been shown that  $s$  is not an ideal estimator for  $\sigma$ , because it tends to be an underestimate. An analysis [Vickers AJ 2003] of the power of trials

published in major journals found that sample size estimates turned out to be too low about 80% of the time, and placed some of the blame on this inadequacy of  $s$ .

The most obvious way to combat this is to multiply  $s$  by some amount based on how confident we are. If the size of the pilot study is known, we can calculate multiplication factors to be sure of a particular chance of choosing an estimate that is at least  $\sigma$ , or calculate  $E(\sigma/s|s,m)$  to make an unbiased estimate for  $\sigma$ . Table 4.1 shows the multiplicative factors for a range of sample sizes.

Table 4.1

Size of pilot sample	50 <sup>th</sup> percentile of $(\sigma/s s,m)$	95 <sup>th</sup> percentile of $(\sigma/s s,m)$	$E(\sigma/s s,m)$
3	1.20112	4.4154	1.6062
5	1.09163	2.3724	1.2518
7	1.05919	1.9154	1.1512
10	1.03864	1.6452	1.0942
15	1.02447	1.4597	1.0579
20	1.01790	1.3704	1.0418
25	1.01411	1.3165	1.0327
30	1.01165	1.2797	1.0268
50	1.00686	1.2017	1.0156
75	1.00453	1.1579	1.0103
100	1.00338	1.1336	1.0077

The choice of multiplicative factor should be made with the how much of risk of the trial being underpowered the sponsor is willing to take. The choice of being 95% confident of using at least  $\sigma$  will mean trial sizes that likely turn out to be much larger than was necessary, an expensive reassurance.

An alternative approach is to use calculation methods for sample size and power that calculate the expected power instead of using a point estimate that is not reliable. Taking into account the distribution of ( $\sigma$ |s) in calculations can produce estimates of the expected power.

Recall equations 1.2 and 1.3, the equations for when the population standard distribution is known.

$$1-\beta \approx \Phi(\tau - t_{1-\alpha,df}) \quad (1.2)$$

This is the normal approximation of the cumulative non-central t distribution correct form 1.3

$$1-\beta = 1-pt(t_{1-\alpha,df}, df, \tau) \quad (1.3)$$

Using the same terminology, it can be shown [Julious SA, 2005] that where the true variance is unknown but estimated from data with m degrees of freedom that

$$\text{Expected } (1-\beta) \approx pt(\tau, m, t_{1-\alpha,df}) \quad (4.1)$$

$pt(\tau, m, t_{1-\alpha,df})$  approaches  $\Phi(\tau - t_{1-\alpha,df})$  as m approaches infinity, so equation 4.1 can be considered the equivalent of equation 1.2 for unknown variance.



There is not an equally concise equation that gives the exact expected power, but arithmetic methods can be used to arrive at a more accurate estimate.

An arithmetic method could involve creating a pdf for the distribution of sigma given  $s$  and  $m$ , by determining the appropriate inverse square-root chi-squared distribution. A large number ( $Q$ ) of equally spaced (by probability) quantiles of this distribution are marked, and the values of sigma at each quantile are each used to make power calculations by equation 1.3.  $Q$  power estimates are made, and as  $Q$  approaches infinity the mean of the power estimates approaches the expected power. The arithmetic method used in the programs in later chapters is of this type, with a  $Q$  of 999. This  $Q$  is a compromise, as the higher the  $Q$  the more accurate the estimation but the longer computation must take. 999 is high enough for a good accuracy while also allowing for speedy calculation on a modern computer.

The method with a  $Q$  of 999 is referred to as Arithmetic Method 4.2 for the remainder of the thesis.

It would be useful to compare the results gained from equation 4.1 and arithmetic method 4.2 to see how they differ.

Table 4.2

<b>Reps</b>	<b>Equation 4.1 Df=10</b>	<b>Arithmetic method 4.2 Df=10</b>	<b>Equation 4.1 Df=100000</b>	<b>Arithmetic method 4.2 Df=100000</b>	<b>Equation 1.3</b>
1	0.00000	0.00000	0.00000	0.00000	0.00000
2	0.00269	0.13537	0.00194	0.13678	0.13678
3	0.15496	0.26351	0.14815	0.26658	0.26658
4	0.32418	0.38164	0.32747	0.39095	0.39095
5	0.45401	0.48336	0.47212	0.50245	0.50245
6	0.55485	0.56891	0.58759	0.59914	0.59914
7	0.63398	0.64003	0.67972	0.68093	0.68093
8	0.69723	0.69881	0.75289	0.74874	0.74874
9	0.74807	0.74728	0.81059	0.80402	0.80402
10	0.78923	0.78724	0.85573	0.84844	0.84845
11	0.82275	0.82024	0.89077	0.88370	0.88371
12	0.85019	0.84754	0.91776	0.91139	0.91139
13	0.87278	0.87020	0.93841	0.93292	0.93292
14	0.89138	0.88906	0.95411	0.94952	0.94952
15	0.90702	0.90483	0.96596	0.96222	0.96222

Table 4.2 shows different power estimations by the different power calculation methods for an AB/BA crossover trial with the setup from example 1.1, but with the variance estimate calculated with differing levels of certainty. The used observed variance from a pilot trial is in each case 1, and the power calculated in the cases where it is estimated from 10 and 100,000 degrees of

freedom for sample sizes from 2 to 30. The result by equation 1.3 is also shown.

The general trend is for 4.1 to give a somewhat lower estimate when 4.2 gives a power of about 40%, gives slightly lower but very close estimates where power is between 40% and 70%, and slightly higher when power is over 70%. The bigger disparity under 40% is a worry, but as it is unlikely a trial with such a low power would be designed it can be said that the two methods give very similar results for any realistic trial design. For a very high df it can be seen that 4.2 approaches the value by equation 1.3, but 4.1 approaches equation 1.2 and may exceed 1.3. It is not logical, from the distribution of sigma given s, for the expected power to be higher than the power, but the exceedance is very small.

Next, it is investigated if the difference in power estimation by the two methods actually affects calculated sample size for different effect sizes, defined as  $\Delta = \delta/\sigma$ . The required sample sizes calculated by the different equations for a range of  $\Delta$  for a selection of m are shown below.

Table 4.3: Sample Sizes calculated for 90% power with 2.5% one-sided alpha

<b>Beta=0.1</b>	<b>m</b>	<b><math>\Delta = 0.1</math></b>	<b><math>\Delta = 0.2</math></b>	<b><math>\Delta = 0.5</math></b>	<b><math>\Delta = 1</math></b>
Eq 4.1	10	1368	343	56	15
Am 4.2		1366	343	56	15
Eq 4.1	25	1167	293	48	13
Am 4.2		1166	293	48	13
Eq 4.1	100	1079	271	44	12
Am 4.2		1079	271	45	12
Eq 4.1	1,000,000	1052	264	43	12
Am 4.2		1052	264	44	12
Eq 1.3	infinite	1052	264	44	12

Table 4.4: Sample Sizes required for 80% power with 2.5% one-sided alpha

<b>Beta=0.2</b>	<b>m</b>	<b><math>\Delta = 0.1</math></b>	<b><math>\Delta = 0.2</math></b>	<b><math>\Delta = 0.5</math></b>	<b><math>\Delta = 1</math></b>
Eq 4.1	10	933	234	39	11
Am 4.2		933	234	39	11
Eq 4.1	25	841	211	35	10
Am 4.2		841	211	35	10
Eq 4.1	100	800	201	33	9
Am 4.2		800	201	33	10
Eq 4.1	1,000,000	786	198	33	9
Am 4.2		786	198	33	9
Eq 1.3	infinite	786	198	33	9

Table 4.5: Sample Sizes required for 70% power with 2.5% one-sided alpha

<b>Beta=0.3</b>	<b>m</b>	<b><math>\Delta = 0.1</math></b>	<b><math>\Delta = 0.2</math></b>	<b><math>\Delta = 0.5</math></b>	<b><math>\Delta = 1</math></b>
Eq 4.1	10	696	175	29	9
Am 4.2		696	175	29	9
Eq 4.1	25	648	163	27	8
Am 4.2		648	163	27	8
Eq 4.1	100	626	158	26	8
Am 4.2		626	158	26	8
Eq 4.1	1,000,000	619	156	26	8
Am 4.2		619	156	26	8
Eq 1.3	infinite	619	156	26	8

Table 4.6: Sample Sizes required for 60% power with 2.5% one-sided alpha

<b>Beta=0.4</b>	<b>m</b>	<b><math>\Delta = 0.1</math></b>	<b><math>\Delta = 0.2</math></b>	<b><math>\Delta = 0.5</math></b>	<b><math>\Delta = 1</math></b>
Eq 4.1	10	533	134	23	7
Am 4.2		533	134	23	7
Eq 4.1	25	507	128	22	7
Am 4.2		507	128	22	7
Eq 4.1	100	495	125	21	7
Am 4.2		495	125	21	7
Eq 4.1	1,000,000	491	124	21	7
Am 4.2		491	124	21	7
Eq 1.3	infinite	491	124	21	7

Table 4.7: Sample Sizes required for 50% power with 2.5% one-sided alpha

<b>Beta=0.5</b>	<b>m</b>	<b><math>\Delta = 0.1</math></b>	<b><math>\Delta = 0.2</math></b>	<b><math>\Delta = 0.5</math></b>	<b><math>\Delta = 1</math></b>
Eq 4.1	10	407	103	18	6
Am 4.2		406	103	18	6
Eq 4.1	25	394	100	17	6
Am 4.2		394	99	17	6
Eq 4.1	100	388	98	17	6
Am 4.2		388	98	17	5
Eq 4.1	1,000,000	386	98	17	6
Am 4.2		386	98	17	5
Eq 1.3	infinite	386	98	17	5

Table 4.8: Sample Sizes required for 40% power with 2.5% one-sided alpha

<b>Beta=0.6</b>	<b>m</b>	<b><math>\Delta = 0.1</math></b>	<b><math>\Delta = 0.2</math></b>	<b><math>\Delta = 0.5</math></b>	<b><math>\Delta = 1</math></b>
Eq 4.1	10	302	77	14	5
Am 4.2		302	77	14	5
Eq 4.1	25	297	76	14	5
Am 4.2		296	75	13	5
Eq 4.1	100	294	75	14	5
Am 4.2		294	74	13	5
Eq 4.1	1,000,000	293	75	14	5
Am 4.2		293	74	13	5
Eq 1.3	infinite	293	74	13	5

Table 4.9: Sample Sizes required for 30% power with 2.5% one-sided alpha

<b>Beta=0.7</b>	<b>m</b>	$\Delta = 0.1$	$\Delta = 0.2$	$\Delta = 0.5$	$\Delta = 1$
Eq 4.1	10	211	54	11	4
Am 4.2		211	54	10	4
Eq 4.1	25	209	54	10	4
Am 4.2		209	53	10	4
Eq 4.1	100	208	54	10	4
Am 4.2		208	53	10	4
Eq 4.1	1,000,000	208	54	10	4
Am 4.2		208	53	10	4
Eq 1.3	infinite	208	53	10	4

Table 4.10: Sample Sizes required for 20% power with 2.5% one-sided alpha

<b>Beta=0.8</b>	<b>m</b>	$\Delta = 0.1$	$\Delta = 0.2$	$\Delta = 0.5$	$\Delta = 1$
Eq 4.1	10	128	34	8	4
Am 4.2		127	33	7	3
Eq 4.1	25	128	34	8	4
Am 4.2		127	33	7	3
Eq 4.1	100	128	34	8	4
Am 4.2		127	33	7	3
Eq 4.1	1,000,000	128	34	8	4
Am 4.2		127	33	7	3
Eq 1.3	infinite	127	33	7	3

Table 4.11: Sample Sizes required for 10% power with 2.5% one-sided alpha

<b>Beta=0.9</b>	<b>m</b>	$\Delta = 0.1$	$\Delta = 0.2$	$\Delta = 0.5$	$\Delta = 1$
Eq 4.1	10	50	15	5	3
Am 4.2		48	13	4	2
Eq 4.1	25	50	15	5	3
Am 4.2		48	13	4	2
Eq 4.1	100	50	15	5	3
Am 4.2		48	13	3	2
Eq 4.1	1,000,000	50	15	5	3
Am 4.2		47	13	3	2
Eq 1.3	infinite	47	13	3	2

Table 4.12: Sample Sizes required for 2% power with 2.5% one-sided alpha

<b>Beta=0.98</b>	<b>m</b>	$\Delta = 0.1$	$\Delta = 0.2$	$\Delta = 0.5$	$\Delta = 1$
Eq 4.1	10	6	4	3	3
Am 4.2		2	2	2	2
Eq 4.1	25	6	4	3	3
Am 4.2		2	2	2	2
Eq 4.1	100	6	4	3	3
Am 4.2		2	2	2	2
Eq 4.1	1,000,000	6	4	3	3
Am 4.2		2	2	2	2
Eq 1.3	infinite	2	2	2	2



Tables 4.3 to 4.12 show the differing required sample sizes (in the form of reps, or pairs of subjects) calculated for an AB/BA crossover trial for a range of different beta values (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.98) and a range of different deltas (0.1, 0.2, 0.5 and 1). It is remarkable how similar the sample sizes calculated by the rival methods. It may have been guessed that the differences observed in the power in table 4.1 would translate into very different sample size calculations, but the two methods give almost identical results for all  $\beta$  under 0.9, and respectably close results at 0.9. The sample sizes only diverge significantly at a  $\beta$  of 0.98, where 4.1 gives much higher estimates. As it is unlikely that any trial will be designed with a desired power of 2% or under, we can be very confident that equation 4.1 will deliver sample size estimates that are practically the same as Arithmetic method 4.2 for any trial that would have a chance of taking place.

### **Expected Power compared to Power calculations using point estimates**

The expected power methods and the 's-adjustment' methods for sample size calculation do not presume to calculate the same thing, and the method used should depend on the objective of the power/sample size calculations. If it is more important to know how likely a trial is to be powerful enough then a point estimate based on the percentile of the distribution of  $(\sigma|s,m)$  should be used, and if it is important that the best possible estimate of power taking into account all the information available then the expected power should be used. The use of a point estimate of  $E(\sigma)$  falls in between the two objectives.

Using  $E(\sigma|s,m)$  as an estimate for  $\sigma$  does not result in power calculations of  $E(1-\beta)$ , in fact there is no multiplicative factor for  $s$  that will calculate expected power using the traditional methods of power calculation.

However, the multiplicative factor to  $s$  needed to achieve any particular expected power is only dependant on  $m$ . A statistician could make use of this to make sample size estimates for a particular expected power when using a calculator or computer package that don't use the cumulative non-central  $t$  distributions that equation 4.1 and Arithmetic method 4.2 require. If they wish to make a sample size calculation for an expected power of 80% using a traditional, Normal distribution based sample size equation like equation 1.2 as a rule-of-thumb they should adjust the observed  $s$  by multiplying it by about

$$m \sqrt{2.285 + 0.5/m}$$

For example, the required sample size for 80% expected power for AB/BA trial with  $s=1$ ,  $m=13$ ,  $\delta=0.2$ , by AM 4.2 is 450, and the sample size for 80% power by equation 1.3 when  $\sigma=(2.285+(0.5/13))^{**}(1/13)$  is also 450. Table 4.13 shows other rough adjustments that might be used for some other expected powers.

Table 4.13

90%	$m \sqrt{3.57 + 1.3/m}$
80%	$m \sqrt{2.285 + 0.5/m}$
70%	$m \sqrt{1.754 + 0.5/m}$

These are all just rule-of-thumb guidelines, and should only be used for rough estimates. These equations were derived using an iterative mathematical approach to approximate a solution in terms of  $m$  and power, rather than a precise calculation.

#### ***4.5 Selecting $p_A$***

Like the estimate for  $\sigma$  used for Normal data, the  $p_A$  is used in binary data type sample size calculations must be estimated from a pilot study or previous data. Unlike normal data, there are several competing ideas of how best to estimate or describe the distribution of the true  $p_A$  given an observed incidence, especially with small sample sizes.

Different ways to calculate confidence intervals include the Wilson Score [Wilson, E. B. 1927], Wald and the Exact method [Clopper, C. J., and Pearson, E. 1934]. For very large samples they give very similar results, but for smaller sample sizes the declared  $x\%$  confidence intervals are only nominal [Sauro J and Lewis JR 2005]. The Exact method is conservative, a confidence interval created has actually at least the nominal chance of containing  $p_A$ . The Normal-distribution based Wald method is often objectively less powerful than it claims to be, but is easy to calculate. An adjusted version of Wald method does give some of the best results for very small sample sizes [Agresti, A., and Coull, B 1998].

If a point estimate for pA is calculated from a sample, then a sample size of n when

$$n = \frac{1}{4}(Z_{1-\beta}^2 / (x/100)^2)$$

will mean there is at least a (1-β) chance that the observed pA will be within x percentage points of the true pA.

The larger the sample size the more confident we can be that the observed incidence is a good point estimate for pA. Cautious trial designers might use the upper bounds of a confidence interval to estimate pA if they are doubtful about the observed rate.

As for normally distributed data, it could be wiser to not place too much trust on a sample pA and instead make power and sample size calculations that take into account assumed conditional distributions of (pA|observed pA).

#### ***4.6 Methods of incorporating uncertainty over pA into sample size calculations***

There isn't an equivalent to equation 4.1 for binary type output, but the confidence interval style of describing uncertainty of pA is perfect for recovering quantiles of to plug into similar a similar arithmetic method to AM 4.2.

Confidence interval calculations with different levels of confidence can be carried out and the boundaries compared to find the quantiles of a pdf for the true  $p_A$  given observed  $p_A$ . Each quantile is then used in a power calculation to gain information on the distribution of power given the distribution of  $p_A$ . At this point the decision on whether the null hypothesis is a statement about  $\log(OR)$  or  $p_B$  has a large impact the calculated estimate of expected power. If  $OR$  is important, it should be held constant in power calculations, and if  $p_B$  is the important one it instead should be held constant. If  $OR$  is important then the quantiles of a pdf of  $p_B$  can be calculated from the pdf of  $p_A$  and the constant  $OR$ . Each quantile of  $p_A$  is compared with the equivalent quantile of  $p_B$ , and because the quantiles of  $p_A$  are equally spaced by probability then the mean of the power calculations is an approximation of the expected power. Results where  $OR$  are constant are more stable than where  $p_B$  is selected as constant, because in the former case some quantiles of  $p_A$  can have values very close or equal to  $p_B$ . This results in the pdf of power being extremely skewed.

Like AM 4.2, the method applied in my SAS and R programs uses 999 quantiles.

#### ***4.7 Simulation-based power estimation***

Sometimes with very unusual trial designs it can be difficult or impractical to frame the relationship between sample size and power. With modern computing power it can be simpler to run an extremely large number of

simulations of a proposed trial and look at the proportion that meet some desired criteria. To show how this simulation based approach can give good results re-look at Example 1.1. The question was: “An investigator intends to run an AB/BA crossover trial to see if new drug B is superior to drug A. The clinical relevant difference is 1, the within subject and between subject standard deviations are both 1. They want to know the power of this trial to detect this difference if he gets 20 subjects enrolled, with 10 assigned to each sequence. What is the power when the one-sided alpha is 0.025?”, and the answer given using equation 1.3 was 0.84844.

Using R, the simulated results of 450,000 clinical trials with the same variances and sizes as Example 1.1 with the true treatment effect of B 1 unit greater than A. Each of the simulated trials was analysed by ordinary least squares to see if a significant difference between treatment effects was detected (Table 4.14). The results also apply to the random effects model because the t-values and p-values of treatment effect calculated by each method are identical with a complete block crossover design.

Table 4.14

Treatment effect B > Treatment effect A	381,694	84.82%
No significant difference	68,306	15.18%
Treatment effect A > Treatment effect B	0	0.00%
Total	450,000	100.00%

If the true power of each individual trial was indeed 0.84844 it would be expected that approximately  $0.84844 * 450,000 (= 381,798)$  of the trials would show a significant difference in treatment effect. As can be seen in Table 4.14 above, a significant ( $p < 0.05$ ) difference was detected 381,694 times, very close. Assuming  $X \sim \text{Bi}(n, \theta)$  distribution, where  $X$  is the number of trials that have a significant difference in treatment effect and  $\theta$  is the true power, we can calculate the 95% [Clopper/Pearson] CI for  $\theta$  based on the observed results. The CI is (0.84724, 0.84917), which contains 0.84844.

The power of the trial has been estimated very precisely from the 'observed power' of this simulation, and more accurately than by the use of equation 1.2 (0.85574). The use of simulation to estimate power could be especially useful for non-Normally distributed data where formula-based approaches rely on some inherently inaccurate approximation of the variance of the test statistic. By brute force, you can get a better power estimate from simulation than from an imperfect formula. On the other hand, you may by chance get a very poor estimate with the simulation, the results will not be exactly repeatable, and regulatory authorities may frown upon the non-standardness of the approach. I would recommend that the trial designer should stick to regular calculation methods unless they have good reason to doubt the applicability of their sample size formula to the particular details of the trial. If the designer

## Chapter 5

This chapter deals with programs for SAS that can be used to calculate expected powers, and required sample sizes for expected power, for parallel and crossover trials taking into account uncertainty about the distribution of the outcome variable for Normally and binary data. It has been established that uncertainty about the variance or  $p_A$  means that the standard power-based equations are possibly not always appropriate, and these programs allow calculation about Expected power to be performed.

### ***Program 5.1: SAS Program for Normal data taking into account uncertainty in observed standard deviation***

Program 5.1 is used to calculate expected powers for a contrast between treatments, and required sample sizes for expected power, for parallel and crossover trials with a response that is known to be normally distributed but with uncertain variance. The program uses equation 4.2.

Figure 5.1.1: Output from program 5.1



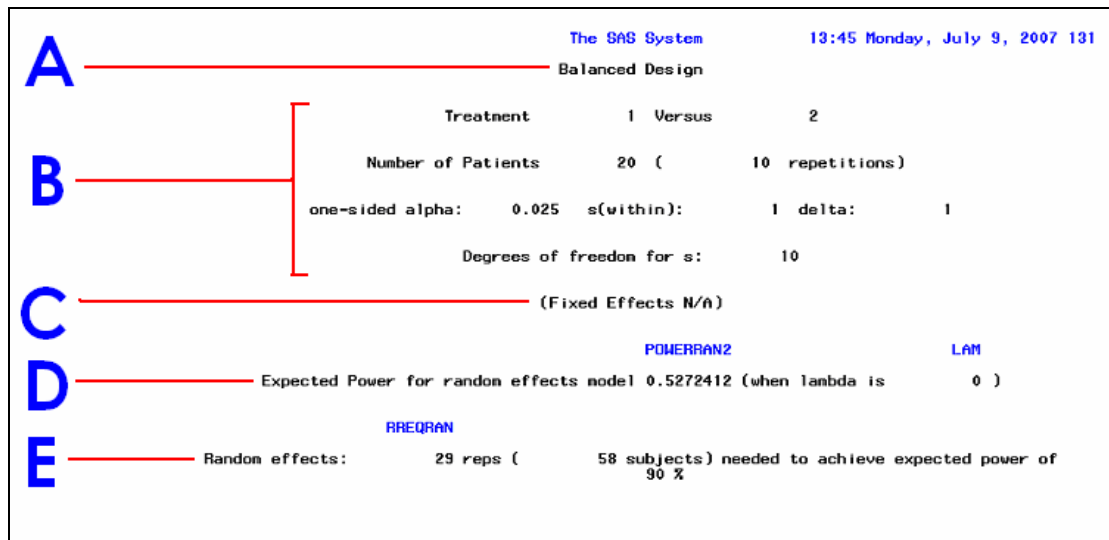
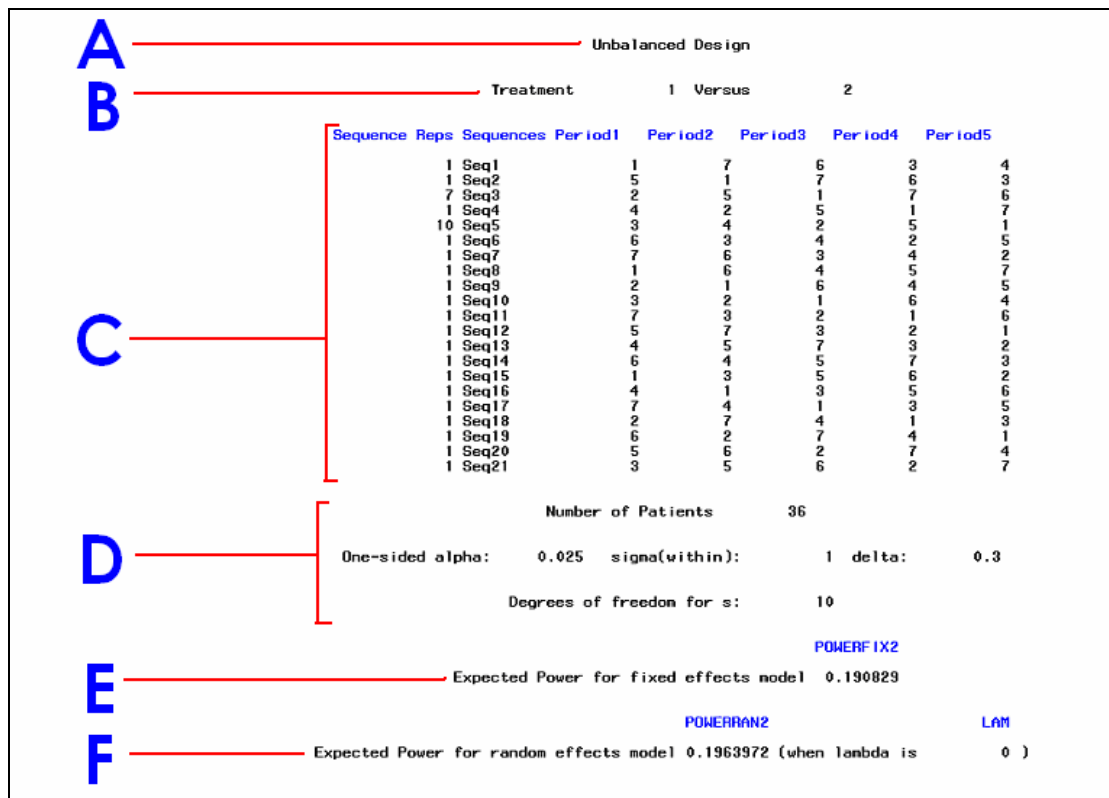


Figure 5.1.1 shows output from program 5.1. The output shows whether or not the design is balanced (**A**), and displays the values of the variables entered into the calculation (**B**). In this example a fixed effects analysis is not possible (it is a parallel design), so no calculations for fixed effects are shown. In their place (**C**) is a note that that analysis is not applicable. Random effects analysis is possible, and the expected power for the stated sample size (**D**) is next. The minimum required sample size to achieve the required expected power is also shown (**E**).

### ***Program 5.2: SAS Program for Normal data with uncertainty 2***

Program 5.2 allows the user to calculate the power of a contrast between a pair of treatments for crossover or parallel trials with an irregular sequence of treatments for normal distributed outcomes with uncertain variability. The program has a similar function to program 2.2 while allowing for uncertainty in estimate of variance. It uses equation 4.1, and can be found in Appendix A.

Figure 5.2.1: Output from Program 5.2



The design is checked to see if it is balanced or not. In the example shown in figure 5.2.1 the design is unbalanced (A). The treatments compared are at (B). The irregular sequence is displayed next (C).

The sample size is calculated from the sequence and displayed with some of the other important variables like alpha, delta, observed within sigma and the associated degrees of freedom. (D)

The expected power for fixed effects (if applicable) and random effects models are calculated. (E & F)

### Program 5.3: SAS Program for Normal data with uncertainty 3

This program calculates power and required sample sizes for all possible contrasts of treatments used in a crossover or parallel trial, for data that is known to be normally distributed with uncertain variance. This is like program 2.3 but taking into account uncertainty. It uses equation 4.1, and can be found in Appendix A.

Figure 5.3.1: Output from Program 5.3

The SAS System 13:45 Monday, July 9, 2007 135

Unbalanced Design

(Fixed Effects N/A)

R  
10 repetitions.

(Random Effects) Power between pairs with  
Random Effects Power

	Treatment1	Treatment2	Treatment3	Treatment4
Treatment1	0	0.5585659	0.5585659	0.6683054
Treatment2	0.5585659	0	0.5585659	0.6683054
Treatment3	0.5585659	0.5585659	0	0.6683054
Treatment4	0.6683054	0.6683054	0.6683054	0

(Random Effects) Required number of complete Replications to attain 0.9 power by pair  
Random Effects Repetitions Required

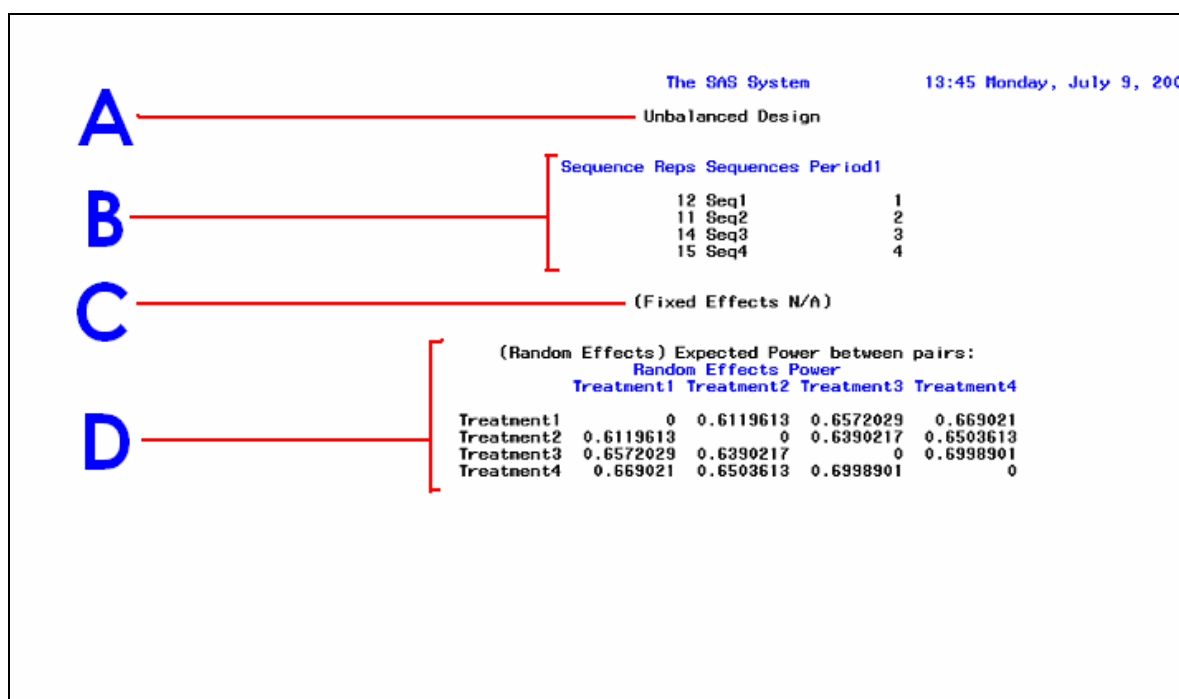
	TREATMENT1	TREATMENT2	TREATMENT3	TREATMENT4
TREATMENT1	0	28	28	21
TREATMENT2	28	0	28	21
TREATMENT3	28	28	0	21
TREATMENT4	21	21	21	0

Output from program 5.3 is shown in fig 5.3.1. The output shows whether or not the design is balanced (A), in this example a fixed effects analysis is not possible (it is a parallel design), so no calculations for fixed effects are shown. In their place (B) is a note that that analysis is not applicable. A matrix of powers between the pairs is shown next (C), and finally a matrix for minimum required sample size to achieve the required expected power for each contrast (D).

### Program 5.4: SAS Program for Normal data with uncertainty 4

This program allows the user to check the expected power of all contrasts for a design with unusual sequence repetitions. It is similar to program 2.4, but calculates expected power.

Figure 5.4.1: Output from program 5.4

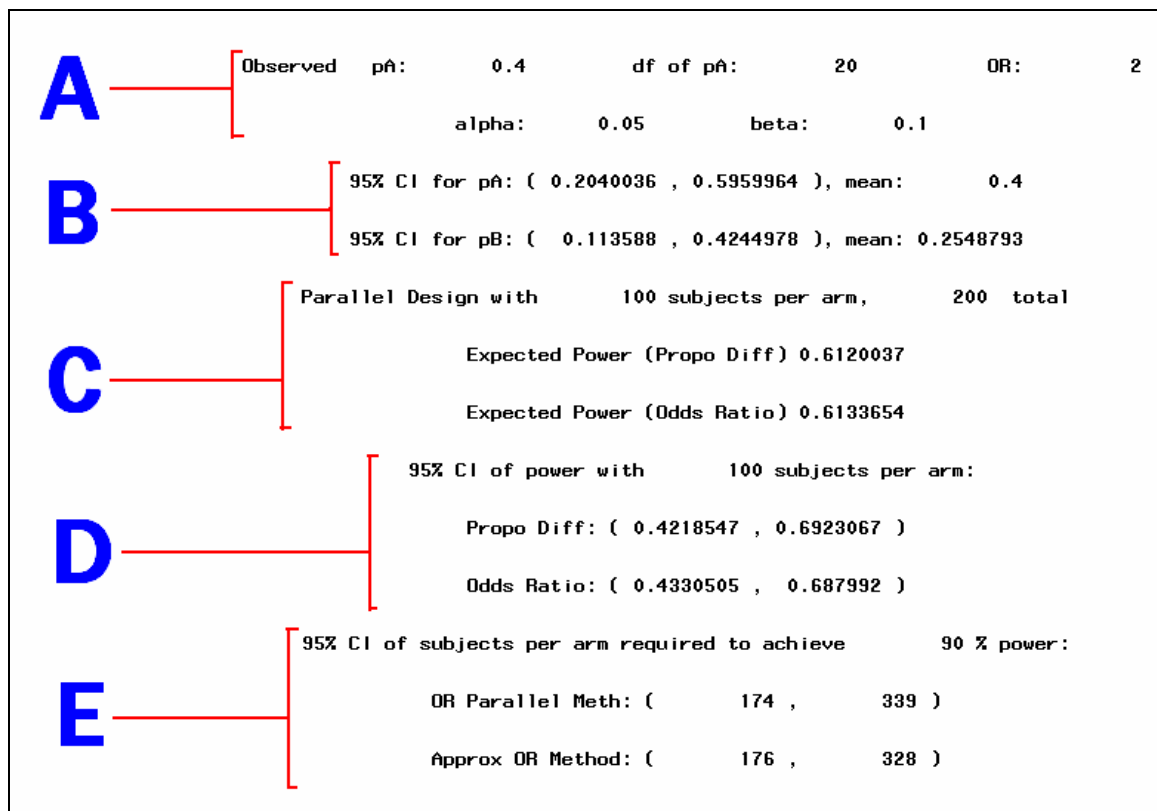


The trial design is analysed to see if it balanced, and the result displayed (A). The inputted sequences and numbers assigned to each is shown next (B), usually followed by a matrix of fixed effects expected powers (C). Finally, the matrix of random-effects expected powers between the pairs of treatments (D).

**Program 5.5: SAS program for binary data that takes into account uncertainty about true value of pA**

Using the method described in chapter 4, it is possible to make calculations for power that take into account uncertainty about pA for binary data. This program calculates expected power of parallel and AB/BA crossover trials for this type of data.

Figure 5.5.1: Output from program 5.5



The parameters entered into the program are displayed first (**A**), then confidence intervals for the uncertain variables are calculated (**B**). The expected power is shown for the sample size and trial type selected by the different methods (**C**), along with confidence intervals for the power of this

design (**D**). Finally (**E**), a 95% confidence interval is shown for required sample size.

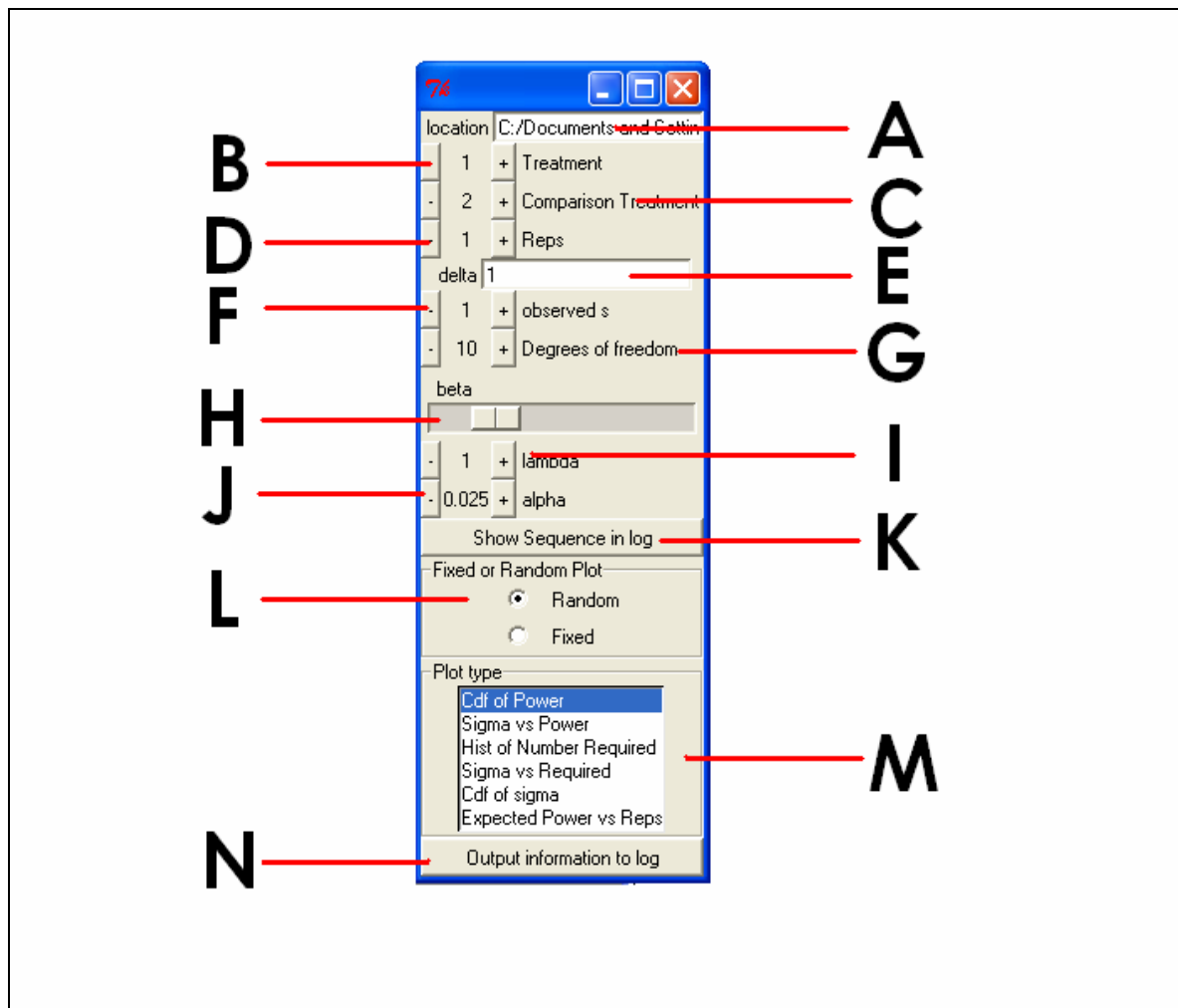
## Chapter 6

Like the previous chapter, this one deals with programs that can be used to calculate expected powers, and required sample sizes for expected power, for parallel and crossover trials taking into account uncertainty about the distribution of the outcome variable for Normally and binary data, but uses R panel.

### ***Program 6.1: R program for Normal data taking into account uncertainty***

This program is for calculating the expected power of contrasts where the true value of standard deviation is uncertain but is estimated from a pilot study. The program uses arithmetic method 4.1, and has several graphical options.

Figure 6.1.1: Interface of Program 6.1



- A:** Location of the txt file containing the information on the treatment sequences for the design. Press return after entering the location for the program to recognise the new instruction. Notice that for R the directory structure of a file location uses the “/” forward slash separator, not the “\” backslash.
- B, C:** The treatments, identified by number, whose contrasts are to be analysed. Press the “-” and “+” buttons to change the treatments compared.



- D:** The  $\delta$  of the design, the size of a difference to be detected. Press return after entering the new delta for the program to recognise the new instruction.
- E:** The  $\delta$  of the design, the size of a difference to be detected. Press return after entering the new delta for the program to recognise the new instruction.
- F:** The  $s_{\text{within}}$ , the sample within-patient standard deviation observed in a pilot study or previous data. Click on the “-” and “+” buttons to decrease or increase the value.
- G:** The degrees of freedom of used in the calculation for  $s_{\text{within}}$ .
- H:** The  $\beta$  of the trial, the size of the type II error. Click and drag the slider to change the  $\beta$ , and thus  $1 - \beta$ , the desired power (or Expected power) of the trial.
- I:**  $\lambda$ , the ratio of  $\sigma^2_{\text{between}}$  to  $\sigma^2_{\text{within}}$ . Click on the “-” and “+” buttons to decrease or increase the value.
- J:** The one-sided  $\alpha$ , the size of the type I error. Click on the “-” and “+” buttons to decrease or increase the value.
- K:** Button to display information on the selected treatment sequence. Information is outputted to the R log.
- L:** Radio button to change between Fixed and Random effects plots and analysis.
- M:** List Box to select the type of plot.
- N:** Button to output to log. Information summarizing the calculations will be output to the R log.

Figure 6.1.2: Output from Program 6.1

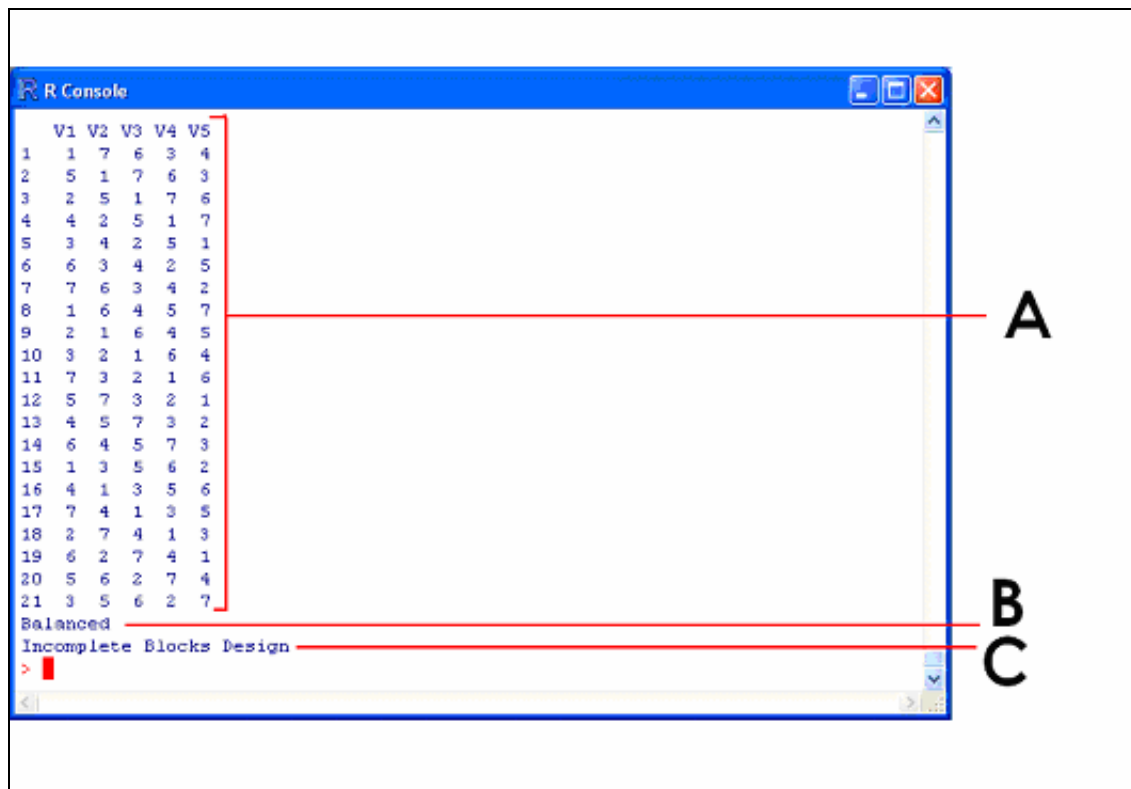


Figure 6.1.2 shows the summary or the design, outputted to log when button 6.1.1K is pressed. In this example the MTA/02 trial is analysed. The 21-sequence, 5 period design is displayed (**A**) according to the format first encountered in chapter 3. The design is calculated to be balanced, and this is displayed also (**B**). Finally, the design is correctly assessed to be an incomplete blocks crossover (**C**).

Figure 6.1.3: Output from Program 6.1

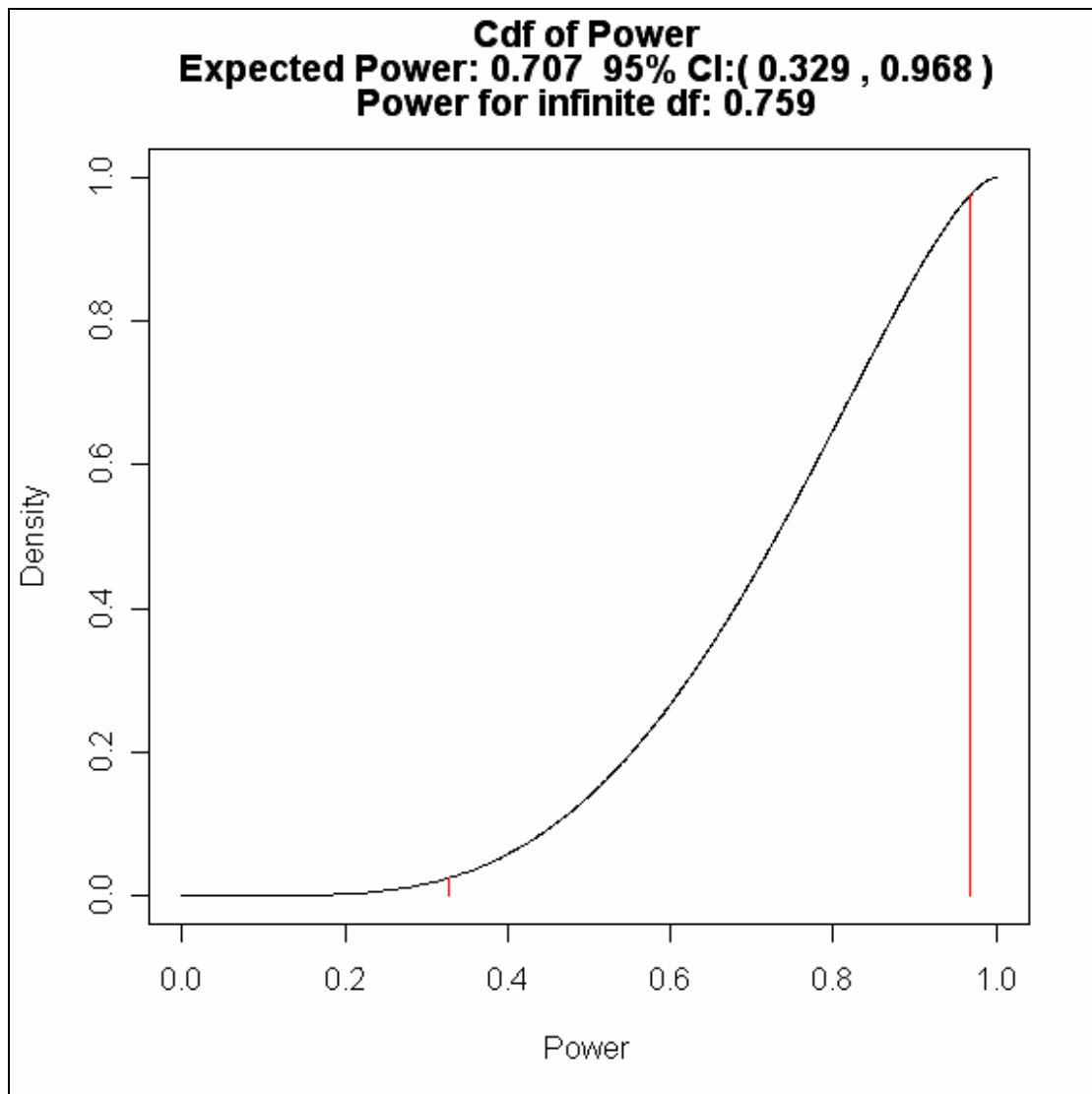


Figure 6.1.3 shows the output from program 6.1 when 'Cdf of Power' is selected in listbox 6.1.1M. The 95% CI for the true power for this example is 0.329 to 0.968. This interval is displayed in the title, and plotted with solid red lines on the graph. The expected power was calculated as 0.707.

Figure 6.1.4: Output from Program 6.1

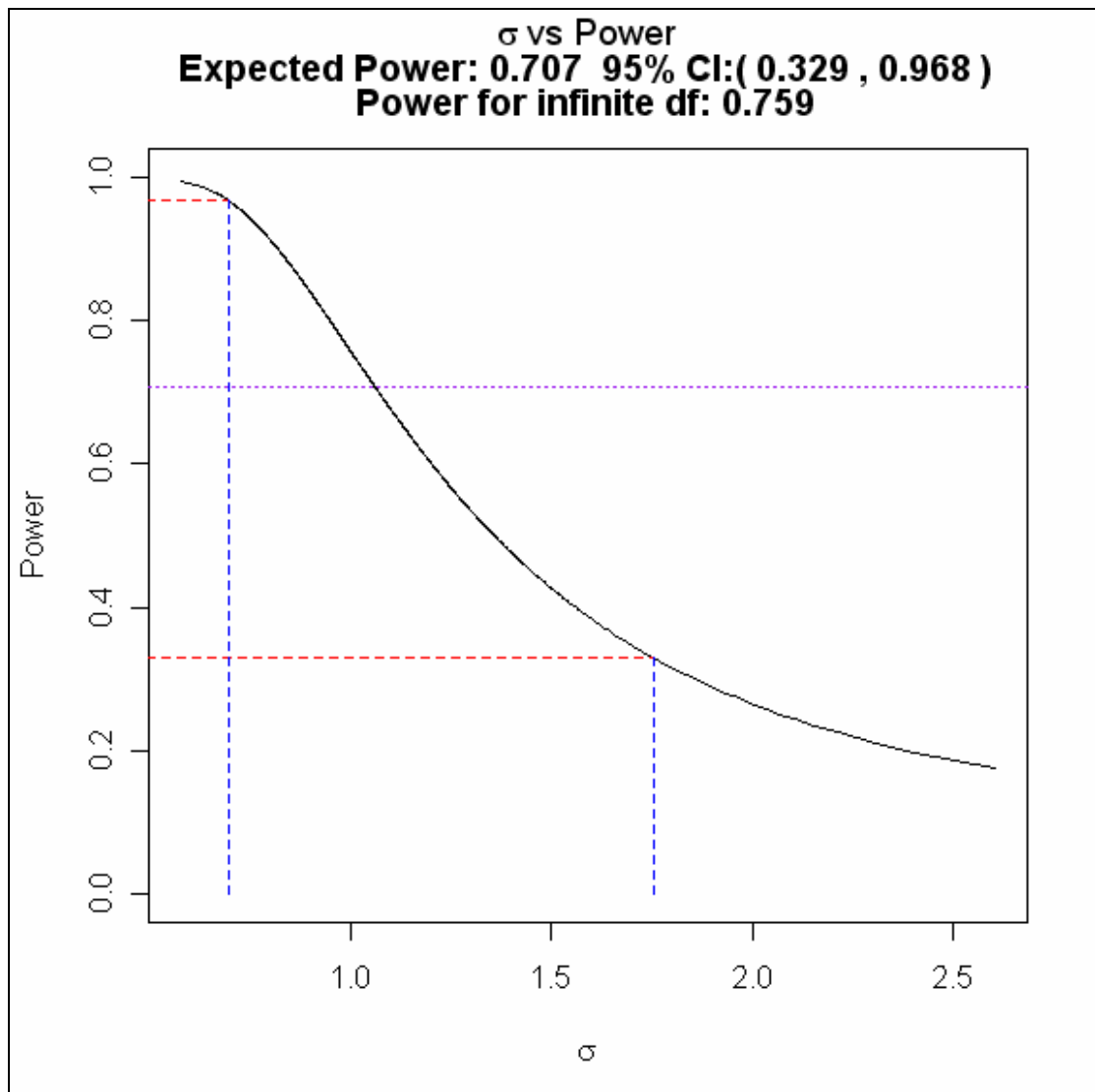
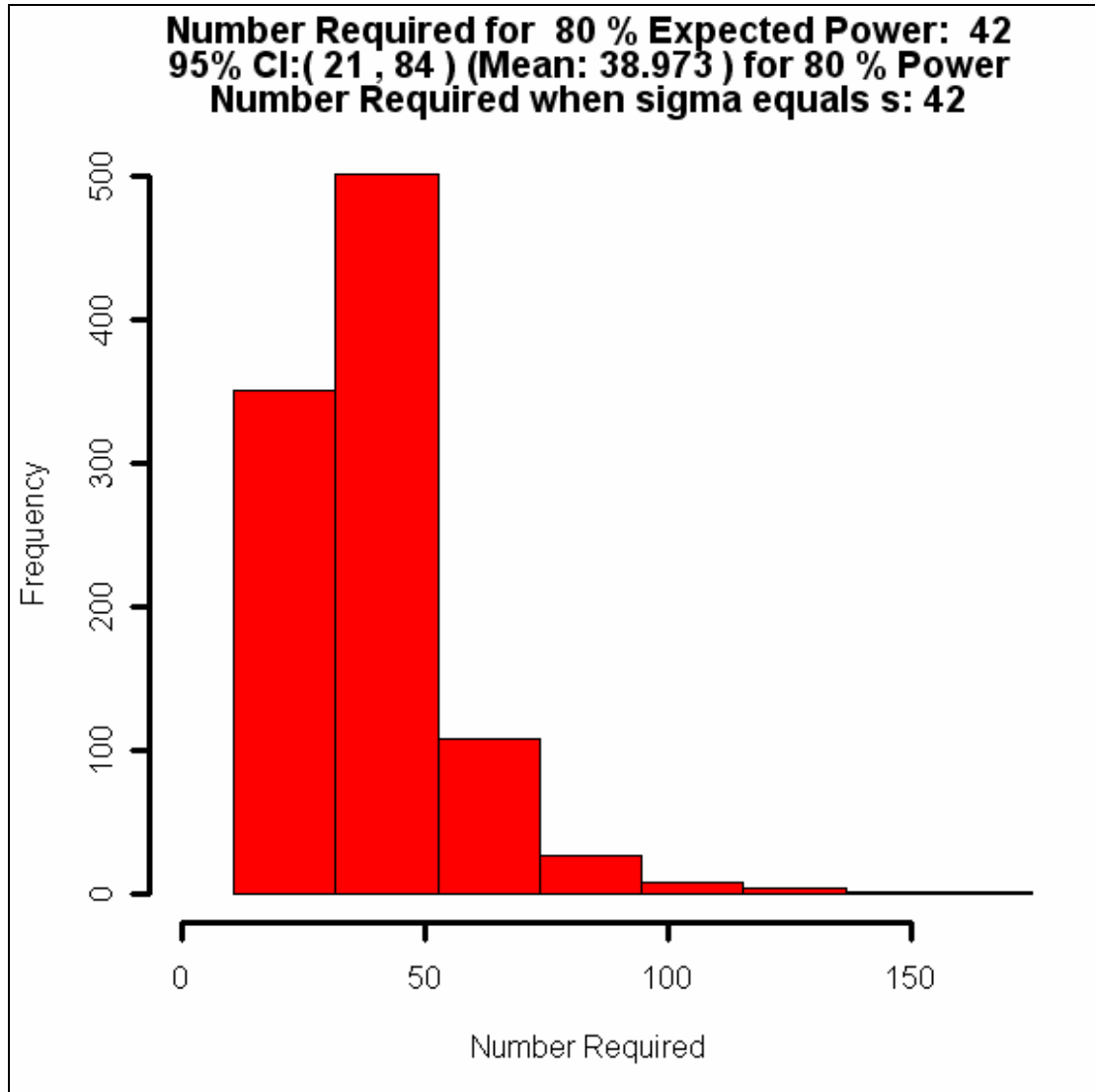


Figure 6.1.4 show an example output for when 'Sigma vs. Power' is selected in listbox 6.1.1M. While sigma could conceivably be any non-negative value, the values of sigma that constitute a 99.9% CI are plotted against the resultant power. The bounds of the 95% CI of sigma are drawn with blue dashed, the 95% confidence bound of power are red dashes. The expected power is shown by a purple dotted line.

Figure 6.1.5: Output from Program 6.1



Choosing 'Hist of Number Required' in listbox 6.1.1M will produce an output like figure 6.1.5. For each of the 999 values possible values of sigma a resulting minimum sample size is calculated to satisfy the requirement of power, in this case 80%. The histogram is of the 999 required sample sizes calculated. A 95% CI for sample size to achieve 80% power is displayed in the title, as well as the sample size in the case that sigma is known to be s.

Figure 6.1.6: Output from Program 6.1

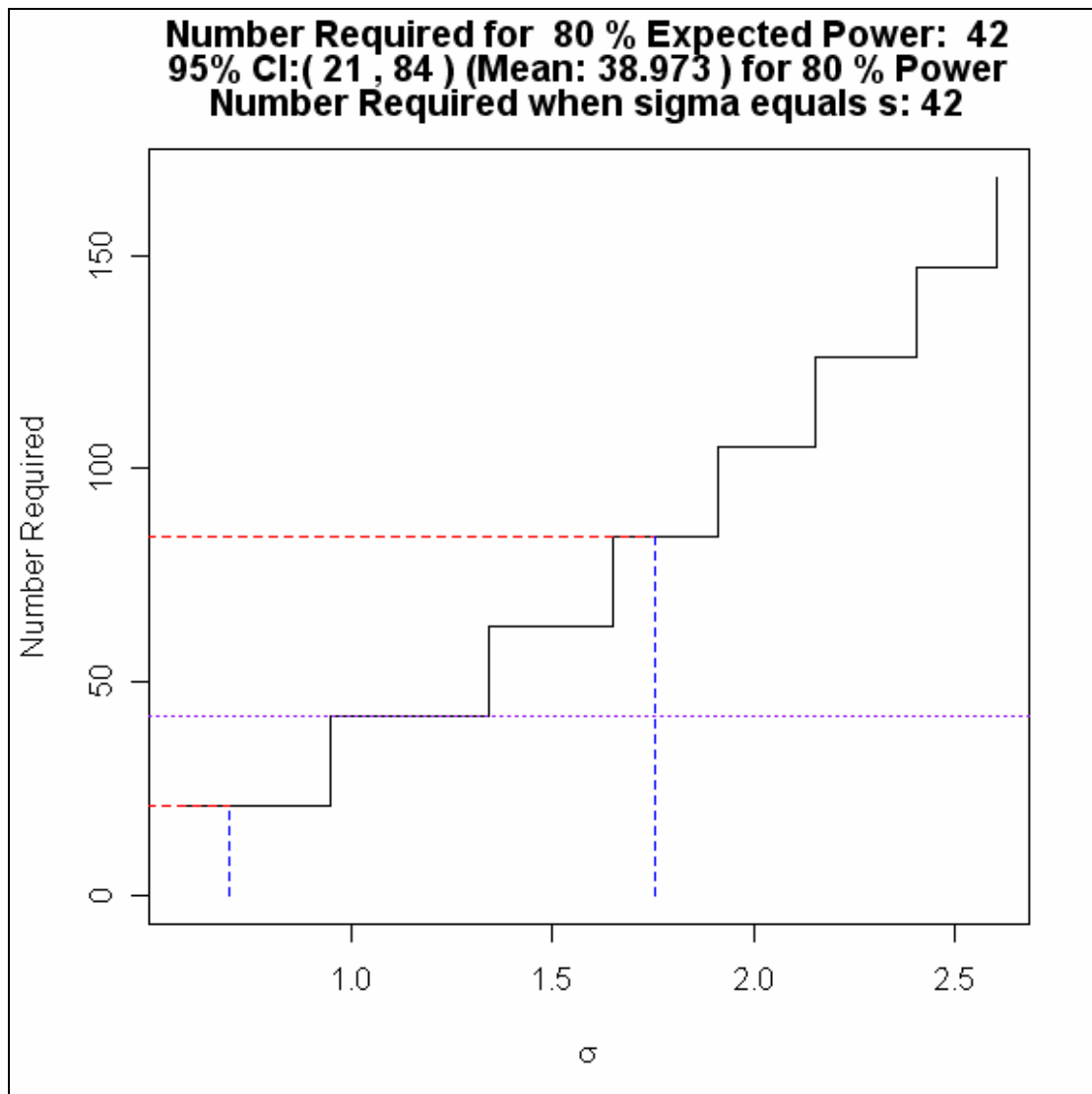
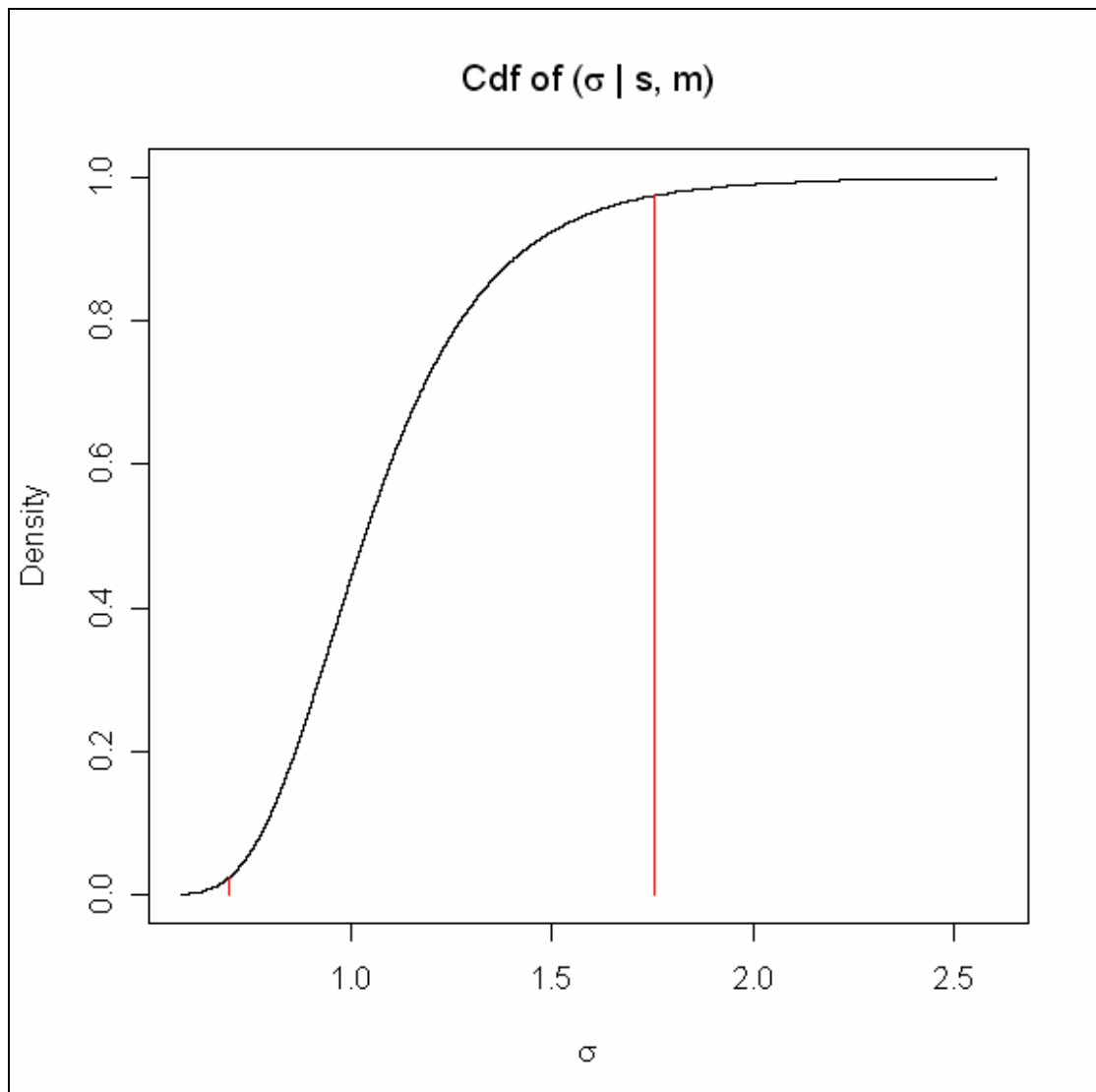


Figure 6.1.6 shows an output when 'Sigma vs Required' is chosen. This stepped graph shows the relationship between variance and required sample size. This graph can be used to get an idea of how sensitive the sample size is to variance estimates. A 95% CI for sigma is marked by vertical blue dashed lines, and a 95% CI for required sample size is demarcated by

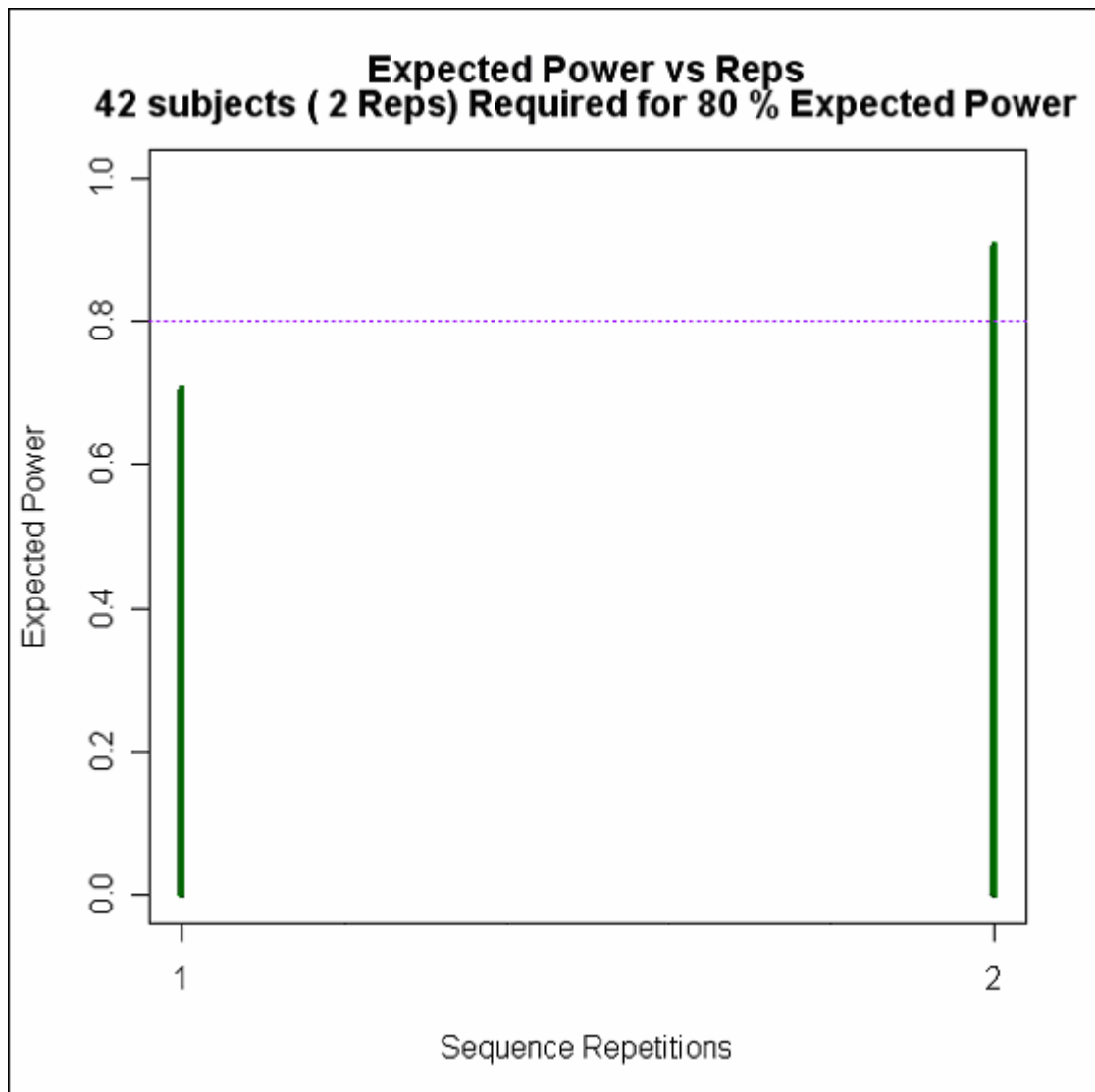
horizontal red dashes. A purple dotted horizontal line shows the expected required sample size.

Figure 6.1.7: Output from Program 6.1



Selecting 'CDF of sigma' will give the user a graph like the one shown in figure 6.1.7. It shows the cumulative probability density curve for the value of sigma, given s and m. The user can be 95% confident that the true value of sigma lies between the values marked by the solid red vertical lines.

Figure 6.1.8: Output from Program 6.1



This final graphical output (Figure 6.1.8) was produced by selecting 'Expected Power vs Reps', where "Reps" is an abbreviation of "Complete Sequence Repetitions". It is a histogram that shows the relationship between sample size (in the format of sequence repetitions) and expected power. The solid green vertical lines indicate the expected power at each increasing number of complete repetitions of the treatment sequences, up until a number that satisfies the minimum expected power required. The minimum expected



power required is marked by a dotted purple horizontal line.

Figure 6.1.9: Output from Program 6.1

```

*****
Treatment 1 vs Treatment 2
Random Effects Model
s          : 1
Degrees of Freedom: 10
lambda     : 1
alpha      : 0.025
beta       : 0.2
delta      : 1

95% CI for sigma: ( 0.698717 , 1.754934 )

For 21 subjects ( 1 Reps of sequence) Expected Power: 0.7071488
95% CI for power: ( 0.328735 , 0.9677778 )
(If sigma was known to be s: Power is 0.7585011 )

For 80 % Expected Power 42 subjects required ( 2 Reps)
(If sigma was known to be s: 42 subjects required ( 2 Reps) for 80 % Power)

95% CI for required size ( 21 , 84 ) for 80 % Power
mean required size: 38.97297
*****
>

```

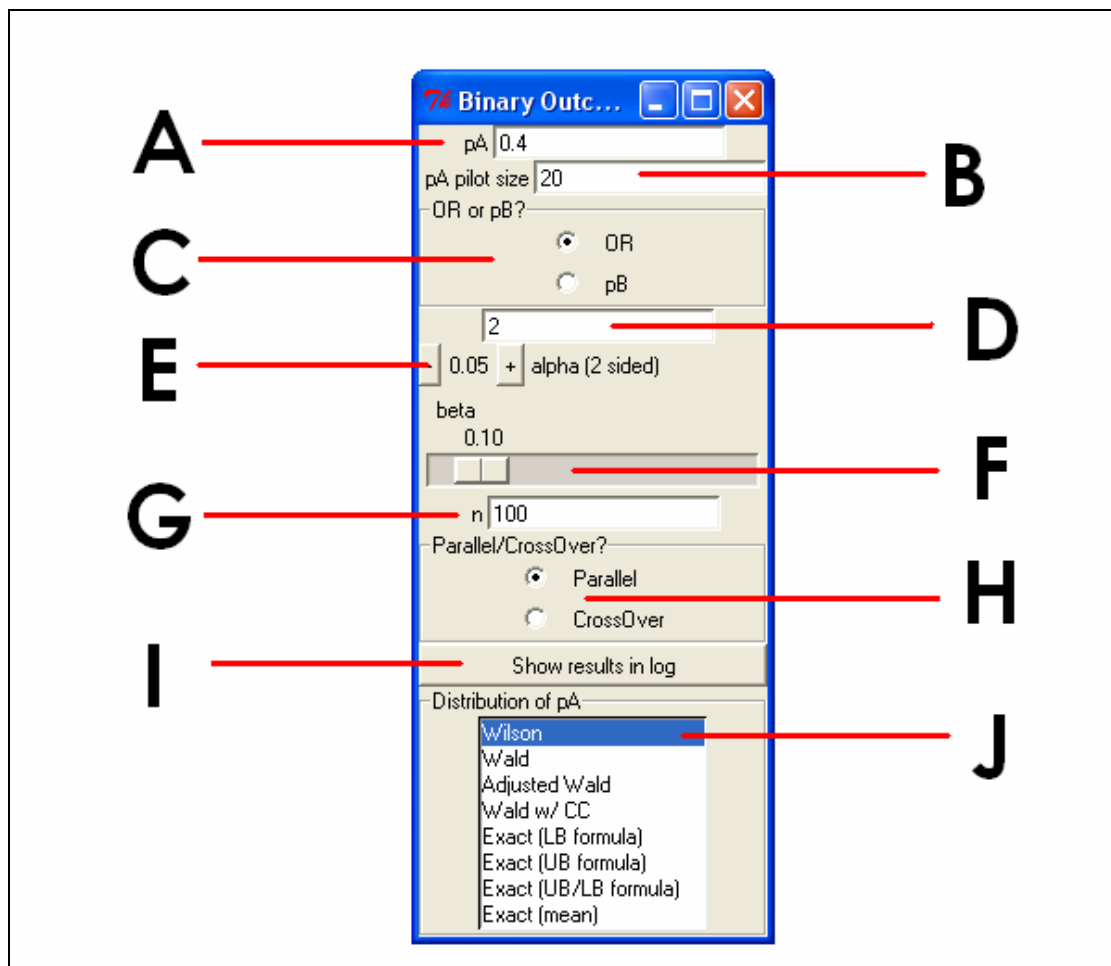
The screenshot shows the R Console output for Program 6.1. The output is annotated with letters A through F. A points to the treatment names, B points to the parameter values, C points to the 95% CI for sigma, D points to the expected power and its CI, E points to the required sample size and its CI, and F points to the 95% CI for the required size and the mean required size.

The final output type from program 6.1 is shown in fig 6.1.9. In the first part of this output (A) the treatments being compared are noted. Next (B), the other parameter values are listed. The 95% CI for the standard deviation is below that (C). Part (D) shows the expected power with this set-up, and gives a confidence interval for what the true power might be. It also gives the power if  $s$  was known to be  $\sigma$ , this is never less than the expected power. Part (E) gives the required sample size to achieve the desired expected power, and also the sample size required to get the power of a same value if  $s$  was known to be  $\sigma$ . Finally (F), a 95% CI for sample size required to get the wanted power. The mean required size is also shown, which gives a clue as to sensitivity of the required sample size.

**Program 6.2: R panel Program for binary outcomes taking into account uncertainty in pA.**

This program uses an arithmetic method to approximate the expected power for parallel and crossover trials while taking into account uncertainty about the true value of pA.

Figure 6.2.1: Interface of Program 6.2



- A:**  $p_A$ , the response anticipated on treatment A, as extrapolated from a pilot study.
- B:** Size of pilot study that decided  $p_A$ .
- C:** The choice of whether to input the OR or  $p_B$ . This is more important than with the previous program, because the value of the select variable will be held and the other will change with  $p_A$ .
- D:** Text box to enter desired OR or  $p_B$ .
- E:** The two-sided  $\alpha$ , the size of the type I error. Click on the “-” and “+” buttons to decrease or increase the value.
- F:** The  $\beta$  of the trial, the size of the type II error. Click and drag the slider to change the  $\beta$ , and thus  $1 - \beta$ , the desired power (or Expected power) of the trial.
- G:** Text entry box to enter  $n$ , the prospective size of the trial. For parallel trials  $n$  will be size per arm, and for crossover trials  $n$  will be the total size of the trial.
- H:** Radio button to choose between parallel trial and crossover trial.  
Parallel trials for this program are trials with two different treatments or treatment levels with equal allocation to each arm. Crossover trials are AB/BA designs with equal allocations to each sequence.
- I:** Button to show results in log.
- J:** List Box to select how the distribution of  $(p_A | \text{observed } p_A, df)$  is calculated. Possible selections include methods based on the Wilson score, the normal approximation, and Exact (Clopper-Pearson) confidence intervals.

Figure 6.2.2: Output from Program 6.2

```
R Console
****Binary trial size and power*****
Observed pA: 0.4   alpha: 0.05   beta: 0.1
Size pA pilot study: 20
Distribution of pA: Wilson
95% CI of pA: ( 0.2188065 , 0.6134185 )
95% CI of pB: ( 0.1228427 , 0.4423963 )
OR: 2

Expected Power of CrossOver trial with 100 subjects:
Approx OR method: 0.5980668
OR Parallel estimate: 0.6151126
Conner: 0.5951592
Miettinen: 0.597315

95% CIs for Power with 100 subjects:
Approx OR method: ( 0.4278517 , 0.6749417 )
OR Parallel estimate: ( 0.4540895 , 0.6868216 )
Conner: ( 0.4308322 , 0.6701747 )
Miettinen: ( 0.4286328 , 0.6737125 )

95% CI of Required size of CrossOver trial for 90 % Power:
Total Sample Size (Approx OR method) ( 178 , 314 )
Total Sample Size (OR Parallel estimate) ( 176 , 310 )
Total Sample Size (Conner) ( 182 , 326 )
Total Sample Size (Miettinen) ( 178 , 318 )
>
```

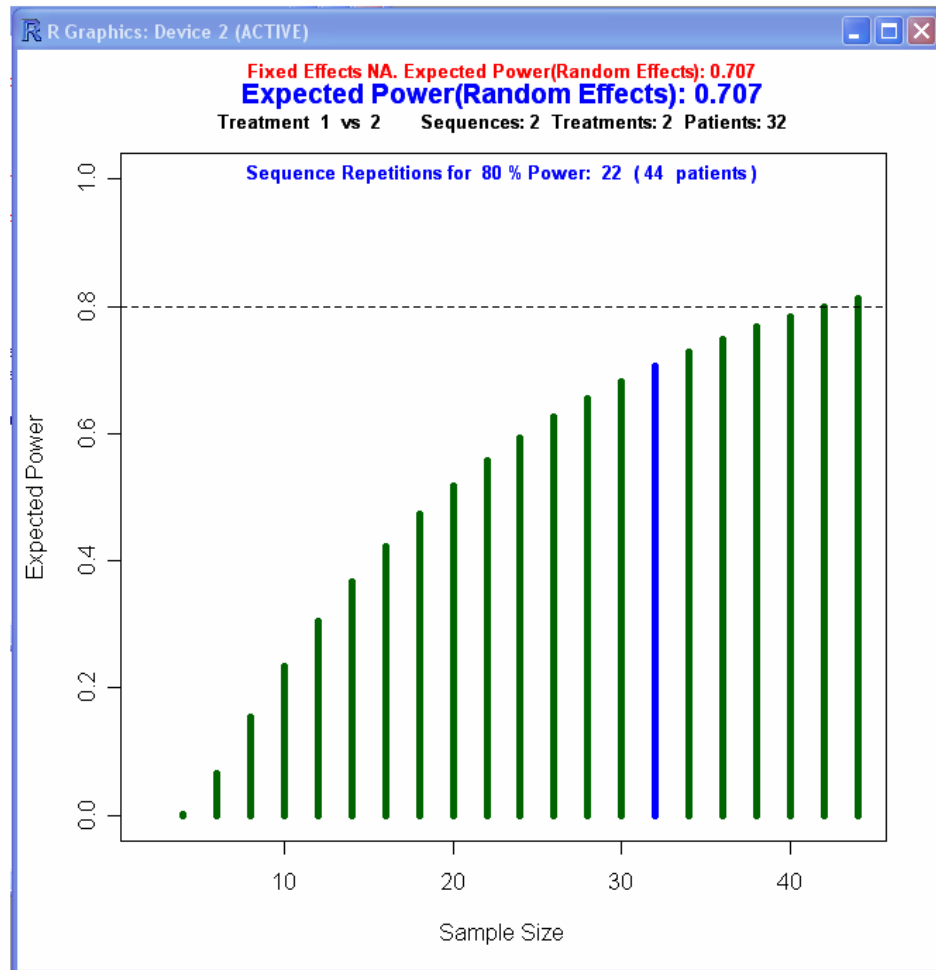
Figure 6.2.2 shows example output from program 6.2. In this example, the Wilson Score method has been chosen to model the estimation of the true value of  $p_A$  where the observed  $p_A$ , in a pilot with 20 subjects, was 0.4. Section (A) of the output shows this and other inputted variable values. 95% confidence intervals are given for  $p_A$ , and in this case  $p_B$  also. There is no confidence interval given for OR, showing that it has been held constant. If  $p_B$  had instead been held constant, then a 95% CI for OR would be displayed. The next part (B) shows the calculated expected power for a given sample size, in this case 100 subjects. A crossover design is being analysed, so the expected powers by the four different calculation methods are displayed. The results are quite close to each other, all within 2% points, with the Conner method giving the lowest power. Section (C) gives 95% confidence intervals

for power by the four different methods, the widths of each interval are similar in size. The final section (**D**) lists 95% confidence intervals of the required size of the trial type to achieve a desired power. The confidence intervals in this case are, again, all quite similar. To achieve 90% power it is likely between about 180 and 320 subjects would be required.

### ***Program 6.3***

Program 6.3 is also for calculating expected power for parallel and crossover trials with normally distributed endpoints where variance used in power calculations is unknown but estimated from a pilot study, but uses formula 4.1 instead of an arithmetic method. This means the calculations are quicker, but are less accurate. Program 6.2 can be a little slow, so this program can be less frustrating to use and gives close estimates to the true expected power and very similar sample size calculations. This program has a very similar interface to program 3.1. The program is contained in appendix A.

Figure 6.3.1



The output (as shown in figure 6.3.1) is of the same format as for program 3.1, but the expected power is calculated instead. The instructions for program 3.1 should be read to understand how to operate this program.

### ***Some Comparisons with other software and standard tables, and Discussion***

As previously discussed, there is no existing software that can make these types of calculation. The only source that has calculated sample sizes taking

into account the variance of a sample is Julious, 2005, and even then only for the simplest parallel and crossover trial design. We will compare examples.

### **1. Parallel trial, Normal Data**

Julious gives an example: "...[T]he clinical effect of interest is a reduction in blood pressure compared to control of 8mmHg (d) with an observed standard deviation from a pilot study 40mmHg (s) estimated with 10 degrees of freedom. Thus, the standardised difference equates to  $d = d / s = 10 / 40 = 0.20$ . For the Type I and Type II errors fixed at 5% and 10% respectively ... [the use of Julious's table of multiplication factors] ... gives a multiplication factor 1.301 for 10 degrees of freedom.

Previously the sample size, assuming the variance in the calculations to be a population variance, ... [was estimated as] ... 527 patients in each arm of the trial. To account for the imprecision in the sample variance therefore one needs to increase the sample size estimated earlier by 30% to 745 patients per arm. An inversion of this argument would be to say that by assuming that the standard deviation was a population estimate the sample size could be considered to be underestimated by 30%. This underestimation of the sample size would result in a reduction in the anticipated power by 6% to 84%."

*There is a slight error here,  $1.301 * 527 = 685.6$ . We will consider that to be the sample size Julious calculates.*

Using program 6.1 gives this output:



\*\*\*\*\*

95% CI for sigma:( 0.698717 , 1.754934 )

95% CI for power:( 0.03749447 , 0.06555857 )

For 90 % Expected Power 1368 subjects required ( 684 Reps)

(If sigma was known to be s: 1054 subjects required ( 527 Reps) for 90 %  
Power)

\*\*\*\*\*

So, we get very similar results. Julious gets a slightly higher result (c.686 per arm) than by 6.1 (684 per arm). There are two reasons for this. First, he uses a version of equation 4.1 to calculate the result which is slightly less accurate, and secondly, he uses a table entry to multiply which leads to rounding errors.

## **2. Crossover trial, Normal Data**

Julious has a table of sample sizes for a range of  $\Delta$ , m is 20. These results were manually compared with both 5.1 and 6.1. The results matched 5.1 100% of the time, and matched 6.1 almost all the time. When there was a disparity between Julious and 6.1, it was only by 1 subject. These results were as expected, as Julious uses a method based on the same assumptions behind program 5.1. For example, Julious gives as sample size of 30 being required when  $\Delta = 1$ ,  $m = 10$  to achieve 90% power in a crossover trial. 5.1 and 6.1 both give this same result.

### **3. Parallel trial, Binary data**

If the observed  $p_A$  was 0.4 in a pilot study of 50, with a two-sided alpha of 0.05, what is the expected power to detect an OR of 2 with 100 subjects on each arm?

Entering this into program 6.2, and selecting the Adjusted Wald method of distributing  $p_A$ , then the expected power will be almost 63% for both calculations.

### **4. Parallel trial, Binary data**

If the above trial was ran as an AB/BA crossover, keeping 200 subjects, what would the expected power be?

Again selecting the Adjusted Wald method of distributing  $p_A$ , then the expected power is around 89% for all four calculations types.

## **Chapter 7: Conclusion: Summary, and Discussion**

### ***7.1 Summary***

#### **Chapter 1**

In chapter 1 we established the importance of sample size and power calculations in clinical trials, and showed that there are moral, legal and financial reasons for an investigator to carry out these calculations. We looked at the basic equations behind the standard sample size and power calculations, and showed that they can be extended to the particular cases of Normal data, binary data, and ordinal data.

#### **Chapters 2 & 3**

In these chapters we showed desirability for new sample size and power software for SAS and R, and saw a necessity for incomplete block designs to be handled. New software was developed, and the results obtained from them compared with the results from established software and sample size tables currently used by experiment designers. We saw that the new software's results more-or-less matched existing methods and explained those slight differences. For incomplete-block crossover trials, with no previous method to compare with, we developed a simulation based method of validation that showed that our programs gave sensible results.

#### **Chapter 4**

We investigated one of the assumptions behind the standard power calculation, the idea that the sample standard deviation could be used as an estimator for the true standard deviation without problems. It was shown that

miscalculations result from that assumption, and an arithmetic method as well as an approximation based on a non-central t distribution were suggested as replacements. The same arguments were applied to the estimation of pA for binary data, and a solution offered for this case too. Sample size calculations based on Expected Power are seen as the solution.

## **Chapter 5 & 6**

After the revelation that the standard equations were inadequate, software is offered that allows the alternative calculations based on Expected Power. These are shown to match results with examples published elsewhere.

### ***7.2 Discussion, and Further Work***

Power calculations and sample size estimates are very important in the pharmaceutical industry, and computing methods can be used to make good estimates for a range of data types and trial designs. Uncertainty about important variables used in the calculations means that traditional sample size methods are unreliable, but this can be partly dealt with by either taking some conservative estimate for  $\sigma$ , or using all of the information available to calculate an Expected Power. The standard equations, the ones seen in the first three chapters, aren't wrong, exactly. The calculations that result from them are correct on their own terms, and the programs that result from them should help the trial planner, especially the facility to assess incomplete block crossover trials. So should the trial designer use the programs from chapters 5&6 to plan their trial? I would say they should. Ultimately, the statistician's role planning process should be to help the decision making process, and the

expected power based analysis will give a more appropriate result for decision making.

But we are now moving away from the traditional sample size calculations. At the start, in Chapter 1, we looked at what Machin called the Fundamental Equation of sample sizes, which depended on Z values, s and  $\delta$ . In chapter 1 we have shown that Z values should be replaced by quantiles from non-central t distributions, and in chapter 4 that s is not adequate without the addition of m to qualify it. The development of Assurance [O'Hagan, 2005] is a way around the ill-defined concept of clinically-relevant difference, by using a Bayesian approach to assessing the likelihood of outcomes of differing desirability. So  $\delta$  is being written out of the equation, too. Even the planned trial is being eroded, with a growing trend of clinical trials being the use of adaptive designs. A more flexible approach to trial design, using information as it becomes available to better direct resources, or to investigate endpoints that become interesting during the trial are of much interest to trial planners today. [Lehmacher and Wassmer 1999]. In general, it seems that the design and execution of trial is becoming less rigid, [Willan AR, Pinto EM, 2005] and an approach that integrates all information available decisions is preferred. If these developments grow in popularity then the traditional sample size calculations may soon be obsolete for cutting edge trials. This thesis has offered a way of dealing with at least some of the uncertainties, but new software that can deal with these less rigid designs and more nuanced end results will need to be developed to aid trial design in the future.

## References

- Agresti, A., and Coull, B. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, 52, 119-126.
- Bennett J.E, Powers J, de Pauw B, Dismukes W, Galgiani J, Glauser M, Herbrecht R, Kauffman C, Lee J, Pappas P, Rex J, Verweij P, Viscoli C, and Walsh T. (2003) Issues in the Design of Trials of Drugs for the Treatment of Invasive Aspergillosis *Clinical Infectious Diseases* 2003; 36(Suppl 3):S113–6
- Blair RC (1981) A Reaction to "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variance and Covariance"; *Review of Educational Research*, Vol 51, No. 4 pp 449-5
- Bowman A.W, Bowman R and Crawford E (2006) rpanel: Simple interactive controls for R functions using the tcltk package (<http://www.stats.gla.ac.uk/~adrian/rpanel/>)
- Campbell MJ, Machin D, Walters SJ. (2007) *Medical Statistics*. 4<sup>th</sup> ed. Wiley
- Campbell MJ, Julious SA, Altman DG. (1995) Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons *BMJ* 1995;311:1145-1148
- Chi GYH (2002) Active Control Non-Superiority Trial - What It Is About *Presentation at the 2002 ICOSA Applied Statistics Symposium*
- Clopper, C. J., and Pearson, E. (1934). The use of confidence intervals for fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

- Conover WJ. (1980) Practical nonparametric statistics. 2nd ed. New York:  
John Wiley
- Conner, R.J. (1987). Sample size for testing differences in proportions for the  
paired sample design. *Biometrics* 43:207-211.
- Cox DR and Reid N.(2000) The Theory of the Design of Experiments.  
*Chapman & Hall/CRC*
- Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR (1978). The importance of  
beta, the type II error and sample size in the design and interpretation  
of the randomized control trial. Survey of 71 "negative" trials. *N Engl J  
Med.* 1978 Sep 28;299(13):690-4.
- French JA (2004) Re: Docket ID # 2004-N-0181, *letter to FDA*
- Glass GV, Peckham PD, Sanders JR (1972) Consequences of Failure to  
Meet Assumptions Underlying the Fixed Effects Analysis of Variance  
and Covariance. *Review of Educational Research*, Vol. 42, No. 3, 237-  
288
- Julious SA. (2005) Designing Clinical Trials with Uncertain Estimates of  
*Variability. PhD thesis*
- Kim HS. (2004) TOPICS IN ORDINAL LOGISTIC REGRESSION AND ITS  
APPLICATIONS, PhD Thesis
- Lehmacher W and Wassmer G. (1999) Adaptive Sample Size Calculations in  
Group Sequential Trials *Biometrics*, Vol. 55, No. 4 (Dec., 1999), pp.  
1286-1290
- Machin D, Campbell MJ, Fayers P, Pinol A (1997). Sample size tables for  
clinical trials, second edition, Blackwell

- Maggard M.A, O'Connell J.B, Liu J.H, Etzioni D.A. and Ko C.Y. (2003) Sample size calculations in surgery: Are they done correctly? *Surgery Volume* 134, Issue 2, August 2003, Pages 275-279
- Miettinen, O.S. (1968). The matched pairs design in the case of all-or-none responses. *Biometrics* 24:339-353.
- Morgan CC, Stephen Coad D. (2007) A comparison of adaptive allocation rules for group-sequential binary response clinical trials. *Stat Med.* 2007 Apr 30;26(9):1937-54.
- O'Hagan A, Stevens JW, Campbell MJ. (2005) Assurance in clinical trial design *Pharmaceut. Statist.* 2005; 4: 187-201
- R Development Core Team (2007) R: A language and Environment for Statistical Computing, R Foundation for Statistical Computing {<http://www.R-project.org>}
- SAS Institute. (1999) SAS/IML user's guide: Version 8 SAS Publishing, SAS Institute
- Sauro J and Lewis JR. (2005) Estimating completion rates from small samples using binomial confidence Intervals: comparisons and recommendations *PROCEEDINGS of the HUMAN FACTORS AND ERGONOMICS SOCIETY 49th ANNUAL MEETING*, 2100-2104
- Senn, SJ. (2002) *Cross-over Trials in Clinical Research. 2<sup>nd</sup> Edition, Wiley*
- Senn, SJ. (2002) MathCAD 2001i Professional program to work out power of a cross-over design. *Unpublished MathCAD program*
- Snapinn SM (2000) Noninferiority trials *Curr Control Trials Cardiovasc Med.* 2000; 1(1): 19–21.



- Vickers AJ. (2003) Underpowering in randomized trials reporting a sample size calculation. *Journal of Clinical Epidemiology* 56 (2003) 717–720
- Vollmar J, and Hothorn LA(Ed). (1997) Cross-over clinical trials Biometrics in the Pharmaceutical Industry 7, Gustav Fischer
- Willan AR, Pinto EM.(2005) The value of information and optimal clinical trial design. *Stat Med.* 2005 Jun 30;24(12):1791-806.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209-212.

## Appendix A

Appendix A contains all the sample size calculation software referred to in chapters 2, 3, 5 and 6 of this thesis. This appendix can be found on an attached disc, or for more up-to-date versions email [michael@stats.gla.ac.uk](mailto:michael@stats.gla.ac.uk).

I intend to use and update these programs as part of my work, so improvements will be made to the presentation and functionality of the programs.

Prog\_2\_1.sas

Prog\_2\_2.sas

Prog\_2\_3.sas

Prog\_2\_4.sas

Prog\_2\_5.sas

Prog\_2\_6.sas

Prog\_3\_1.R

Prog\_3\_2.R

Prog\_3\_3.R

Prog\_3\_3\_15.R

Prog\_5\_1.sas

Prog\_5\_2.sas

Prog\_5\_3.sas

Prog\_5\_4.sas

Prog\_5\_5.sas

Prog\_6\_1.R

Prog\_6\_2.R

Prog\_6\_3.R