



This is a repository copy of *Tutorial in biostatistics - Sample sizes for clinical trials with Normal data.*

White Rose Research Online URL for this paper:  
<http://eprints.whiterose.ac.uk/145474/>

Version: Accepted Version

---

**Article:**

Julious, S.A. (2004) Tutorial in biostatistics - Sample sizes for clinical trials with Normal data. *Statistics in Medicine*, 23 (12). pp. 1921-1986. ISSN 0277-6715

<https://doi.org/10.1002/sim.1783>

---

This is the peer reviewed version of the following article: Julious, S.A. (2004) Tutorial in biostatistics - Sample sizes for clinical trials with Normal data. *Statistics in Medicine*, 23 (12). pp. 1921-1986, which has been published in final form at <https://doi.org/10.1002/sim.1783>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# **Tutorial in Biostatistics: Sample sizes for clinical trials with Normal data.**

Steven A. Julious

## **SUMMARY**

This article gives an overview of sample size calculations for parallel group and cross-over studies with Normal data. Sample size derivation is given for trials where the objective is to demonstrate: superiority, equivalence, non-inferiority, bio-equivalence and estimation to a given precision, for different Type I and Type II errors. It is demonstrated how the different trial objectives influence the null and alternative hypotheses of the trials and how these hypotheses influence the calculations. Sample size tables for the different types of trials and worked examples are given.

## **Key Words**

Sample size, Power, Type I error, Type II error, Superiority trials, Equivalence trials, Non-inferiority trials, Bio-equivalence, Precision, Baselines, Cross-over trials, Parallel group trials.

## CONTENTS

1. Introduction
  - 1.1. Estimation of the variance for calculations
2. Superiority trials
  - 2.1. Parallel group trials
    - 2.1.1. Worked Example
      - 2.1.1.1. Using the sample size tables
      - 2.1.1.2. Repeated using sample size software
    - 2.2. Cross-over trials
      - 2.2.1. Paired t-tests and period adjusted t-tests
      - 2.2.2. Sample size calculations
      - 2.2.3. Worked Example
        - 2.2.3.1. Using the sample size tables
        - 2.2.3.2. Repeated using sample size software
3. Equivalence trials
  - 3.1. General case
  - 3.2. Special case of no treatment difference
  - 3.3. Type I and setting the equivalence limit
    - 3.3.1. Choice of Type I error
    - 3.3.2. Choice of Equivalence Limit
  - 3.4. Parallel group trials
    - 3.4.1. General case
    - 3.4.2. Special case of no treatment difference
    - 3.4.3. Worked example
      - 3.4.3.1. Using the sample size tables

- 3.4.3.2. Repeated using sample size software
  - 3.5. Cross-over trials
    - 3.5.1. General case.
    - 3.5.2. Special case of no treatment difference.
    - 3.5.1. Worked example
      - 3.5.1.1. Using the sample size tables
      - 3.5.1.2. Repeated using sample size software
- 4. Non-inferiority trials
  - 4.1. Parallel group trials
    - 4.1.1. Worked example
      - 4.1.1.1. Using the sample size tables
      - 4.1.1.2. Repeated using sample size software
  - 4.2. Cross-over trials
    - 4.2.1. Worked example
      - 4.2.1.1. Using the sample size tables
      - 4.2.1.2. Repeated using sample size software
- 5. As good as or better trials
  - 5.1. A test of non-inferiority and a one sided test of superiority
  - 5.2. A test of non-inferiority and a two sided test of superiority
  - 5.3. Worked example and other considerations
- 6. Bio-equivalence trials
  - 6.1. Justification for log transformation
  - 6.2. Rational for using coefficients of variation
  - 6.3. Cross-over trials
    - 6.3.1. General case

- 6.3.2. Special case of ratio equalling unity
- 6.3.3. Replicate Designs
- 6.3.4. Use of quick formulae to estimate the sample size of a  
bioequivalence trial
- 6.3.5. Worked Example
  - 6.3.5.1. Using the sample size tables
  - 6.3.5.2. Repeated using sample size software
- 6.4. Parallel Group Studies
  - 6.4.1. General case
  - 6.4.2. Special case of ratio equalling unity
  - 6.4.3. Worked Example
    - 6.4.3.1. Using the sample size tables
    - 6.4.3.2. Repeated using sample size software
- 6.5. Individual and Population Bio-equivalence
- 7. Trials to a give precision
  - 7.1. Parallel group trials
    - 7.1.1. Worked Example
      - 7.1.1.1. Using the sample size tables
      - 7.1.1.2. Repeated using sample size software
  - 7.2. Cross-over trials
    - 7.2.1. Worked Example
      - 7.2.1.1. Using the sample size tables
      - 7.2.1.2. Repeated using sample size software
- 8. Design considerations
  - 8.1. Inclusion of Baseline or Covariates

8.2. Post dose measures summarised by summary statistics

8.3. Inclusion of Baseline or Covariates as well as Post Dose Measures

9. Summary

10. References

## **1. INTRODUCTION**

Since the first 'modern' randomised clinical trial was reported in 1948 [1], clinical trials have become a central component in the assessment of new therapies. The primary objective of any clinical trial is to obtain an unbiased and reliable assessment of a given regimen response independent of any known or unknown prognostic factors. First, by ensuring that the patients studied in the various regimen arms are objectively similar with reference to all predetermined relevant factors other than the regimens themselves. Second, by making sure that the assessment of the regimen response is independent of a given subject's regimen and finally through inclusion of an appropriate control to quantify a given regimen response [2].

Randomisation is important as it ensures that patients are objectively similar in the regimen groups being investigated for any demographic or prognostic factors that either known or unknown [3]. Randomisation achieves this by ensuring that each subject has a known chance of being given a given treatment in an allocation that can not be predicted [4].

Blinding is important as it removes any systematic bias there may be in treatment assessment and allocation during the conduct of the trial. It is important too once the trial has been completed during the cleaning and derivation of the data [5]. If there is any knowledge of treatment during the cleaning and querying of the data then this knowledge may affect how these data are consequently queried and cleaned [3].

The choice of an appropriate control is dependent on the objective of the trial being designed. For example a non-inferiority or equivalence trial will usually have a control which is active if the primary outcome is efficacy. The different types of trials will be described through this paper.

When planning a trial one essential step is the calculation of a sample size which will give the minimum sample size required to meet the given objectives of the study. Sample size issues are important for the planning of clinical trials. Studies that are either too small or too



large may be judged unethical [6]. For example, a study that is too large could have met the objectives of the trial before the actual study end had been reached, and so some patients may have unnecessarily entered the trial. A trial that is too small will have little chance of meeting the study objectives, and patients may be put through the potential trauma of a trial for no tangible benefit. The general approach to choosing sample size will be described in this article where a statistic can be assumed to take a Normal form and an estimate of the variance of that test statistic is available. The sections of the paper detail computation of sample sizes appropriate for:

1. Superiority trials.
2. Equivalence trials.
3. Non-inferiority trials.
4. As good as or better trials.
5. Bio-equivalence trials.
6. Trials to a given precision.

A distinction therefore is drawn to emphasise differences in trials designed to demonstrate 'superiority' and trials designed to demonstrate 'equivalence' or 'non-inferiority'. This is discussed with an emphasis on how differences in the null hypothesis can impact on calculations. The ICH guidelines E3 and E9 provide general guidance on selecting the sample size for a clinical trial [3, 7]. The ICH E9 guideline states that:

"The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed. This number is usually determined by the primary objective of the trial ....The method by which the sample size is calculated should be given in the protocol together with any quantities used in the calculations (such as variances, mean values, response rates, event rates, differences to be detected)."

This paper will go through the methods of sample calculation for studies with the six distinct objectives listed above. The paper will also, under the worked examples, give a brief description of how the calculations could be undertaken in the two packages PASS 2000 [8] and nQuery 4 [9]. Although PASS 2000 and nQuery 4 are the only packages described in detail this does not confer a recommendation as to their use by the author.

The paper is written on the premise that just two treatments are to be compared in the clinical trial and two study designs will be discussed: parallel group and cross-over designs.

With a parallel group design subjects are assigned at random to the two treatments to form two treatment groups which it is hoped are the same in all respects other than the treatment received.

With a cross-over trial all subjects receive both the treatments but it is the order that subjects receive the treatments which is randomised. The big assumption here is that prior to starting the second treatment all subjects return to baseline and that the order which subjects receive treatment does not affect their response to treatment. Cross-over trials can not be used therefore in degenerative conditions, where subjects get worse over time. Also, they are more sensitive to bias than parallel group designs [2].

Although this paper will concentrate on data that take a Normal form this does not limit its scope as trials where the primary endpoint is assumed to be Normal probably account for the majority of trials. Also, the discussion in each section on the null hypothesis for each trial and the sample size derivation is generalisable for other types of data. For superiority trials there is work for cross-over [10] and parallel group [11] trials where the data take other distributional forms as well as methodologies for parallel group non-inferiority [12] and equivalence trials [13] for binary data.

Conventions for multiple comparisons are not discussed in this paper, although the approaches for sample size calculation are applicable once appropriate adjustment has been

made. Koch and Ganksy give an overview of this topic [14] whilst the CPMP have issued guidelines [15].

Each section of the paper will walk through the derivation of the appropriate sample size formulae. Tables are given in each section which provide sample size estimates using these formulae and worked examples are described which use these tables. Also, within each section quick formulae are given which do not necessitate the use of tables for calculations.

### **1.1. Estimation of the variance for calculations**

Through out this paper one of the most important components in the sample size calculation is the variance estimate used. This variance estimate is usually estimated from retrospective data sometimes from a number of studies. To adjudicate on the relative quality of the variance one should consider the following aspects of the trial from which the variance is obtained

1. Design: is the study design ostensibly similar to the one you are designing? On the basic level is the data from a randomised controlled trial - observational or other data may greater variability. If you are undertaking a multi-centre trial is the variance estimated too from an similarly designed trial? Were the endpoints similar to those you plan to use – not just the actual endpoints but was the time relative to treatment of the outcome of interest similar to you own?
2. Population: is the study population similar to your own? The most obvious consideration is to ask is whether the demographics were the same but if the trial conducted was a multi centre one was it conducted in similar countries? Different countries may have different types of care (e.g. different concomitant medication) and so may have different trial populations. Was the same type of patient enrolled (the same mixed of mild, moderate and severe)? Was it conducted covering the same seasons (relevant for conditions such as asthma)?

3. Analysis: was the same statistical analysis undertaken? Not just the question of whether the same procedure was used for the analysis but were the same covariates fitted into the model? Was the same summary statistics used? Section 8 details how covariates and summary statistics impact on the variance.

The quality of the variance will obviously influence the strategy of an individual clinical trial - it has not been unknown to have next to no data on hand when designing a trial such that the range divided by four is taken as a variance estimate. Depending on the quality of the variance estimate (or even if one has a good variance estimate) it may be advisable to have some form of variance re-estimation during the trial. There is a developing literature on this topic although this paper will not go into any detail [16, 17, 18, 19, 20, 21, 22, 23, 24].

## 2. SUPERIORITY TRIALS

In a superiority trial the objective is to determine whether there is evidence of a statistical difference in the comparison of interest between the regimens with reference to the null hypothesis that the regimens are the same. The null ( $H_0$ ) and alternative ( $H_1$ ) hypotheses may take the form:

$H_0$ : The two treatments are not different with respect to the mean response ( $\mu_A = \mu_B$ ).

$H_1$ : The two treatments are different with respect to the mean response ( $\mu_A \neq \mu_B$ ).

In the definition of the null and alternative hypotheses  $\mu_A$  and  $\mu_B$  refer to the mean response on regimens A and B respectively. In testing the null hypothesis there are two errors one can make:

I. Rejecting  $H_0$  when it is actually true.

II. Not rejecting  $H_0$  when it is actually false.

These errors are usually referred to as Type I and Type II errors [25, 26, 27, 28, 29, 30]. The aim of the sample size calculation is to find the minimum sample size for a fixed probability of Type I error to achieve a value of the probability of a Type II error. The two errors are commonly referred to as the regulator's (Type I) and investigator's (Type II) risks and by convention are fixed at rates of 0.05 and 0.10 or 0.20 respectively. The Type I and Type II risks carry different weights as they reflect the impact of the errors. With a Type I error medical practice may switch to the investigative therapy with resultant costs whilst with a Type II error medical practice would remain unaltered.

In general, one usually thinks not in terms of the Type II error but in terms of the power of a trial (1-probability of a Type II error) which is the probability of rejecting the  $H_0$  when it is in fact false. Key trials should be designed to have adequate power for statistical assessment of the primary parameters with a Type I error of 5%. The power that should be considered as standard is 90% with the minimum to be considered for calculations being 80%. It should be

noted though that with 80% power one is doubling Type II error for only a 25% saving in sample size.

For a superiority trial there are two chances of rejecting the null hypothesis and thus making a Type I error. The null hypothesis can be rejected if  $\mu_A > \mu_B$  or if  $\mu_A < \mu_B$  by a statistically significant amount. As there are two chances of rejecting the null hypothesis the statistical test is referred to as a two tailed test with each tail allocated an equal amount of the Type I error (of 2.5%). The sum of these tails adds up to the overall Type I error rate of 5%. Thus, the null hypothesis can be rejected if the test of  $\mu_A > \mu_B$  is statistically significant at the 2.5% level or the test of  $\mu_A < \mu_B$  is statistically significant at the 2.5% level.

The purpose of the sample size calculation is hence to provide sufficient power to reject  $H_0$  when in fact some alternative hypothesis is true. One might therefore test that the two means are equal, against an alternative that they differ by an amount ' $d$ ' [31]. The amount  $d$  is chosen as a clinically important difference or effect size and is the main factor in determining a sample size. Reducing the effect size by half will quadruple the required sample size [32].

Formally the aim is to calculate a sample size suitable for making inferences about a certain function of given model parameter,  $\mu$ ,  $f(\mu)$  say. For data that take a Normal form  $f(\mu)$  will be  $\mu_A - \mu_B$  i.e. the difference in means of two populations A and B. Now let  $S$  be a sample estimate of  $f(\mu)$ . Thus  $S$  is defined as the difference in the sample means. As one is assuming that the data from the clinical trial are sampled from a Normal population, then, using standard notation,  $S \sim N(f(\mu), \text{Var}(S))$ , giving

$$\frac{S - f(\mu)}{\sqrt{\text{Var}(S)}} \sim N(0,1).$$

A basic equation can now be developed in general terms from which a sample size can be estimated. Let  $\alpha$  be the overall type I error level, with  $\alpha/2$  of this type I error equally

assigned to each tail of the two tailed test, and let  $Z_{1-\alpha/2}$  denote the  $(1 - \alpha / 2)$  100 percentage point of a standard Normal distribution.

Thus, an upper 2-tailed,  $\alpha$  -level critical region for a test of  $f(\mu) = 0$  is

$$|S| > Z_{1-\alpha/2} \sqrt{\text{Var}(S)}.$$

For this critical region one needs to test it against an alternative that  $f(\mu) = d$ , for some chosen  $d$  and specified power  $(1-\beta)$  [33]:

$$d - Z_{1-\beta} \sqrt{\text{Var}(S)} = Z_{1-\alpha/2} \sqrt{\text{Var}(S)}, \quad (2.1)$$

where  $\beta$  is the overall Type II error level and  $Z_{1-\beta}$  is the 100(1- $\beta$ )% point of the standard Normal distribution. Thus, in general terms for a 2-tailed,  $\alpha$  -level test one has:

$$\text{Var}(S) = \frac{d^2}{(Z_{1-\beta} + Z_{1-\alpha/2})^2} \quad (2.2)$$

where  $\text{Var}(S)$  will be unknown and depends on the sample size. Once  $\text{Var}(S)$  is written in terms of sample size, the above expressions can be solved to give the sample size.

In this section, and throughout the paper for parallel group trials, the assumption will be made that the variances in each group are equal i.e. that  $\sigma_A^2 = \sigma_B^2 = \sigma^2$ . This assumption is referred to as homoskedasticity. There are alternative derivations for the case of unequal variances [34, 35]. This paper will not go into further detail on this topic.

## 2.1 Parallel Group Trials

Suppose one wishes to design a two group study where the sample size in the second group,  $n_B$ , can be written as some multiple of the first,  $n_A$ , (say  $n_B = r n_A$ ). Then  $\text{Var}(S)$  can be written in terms of  $n_A$  and hence equation (2.2) can be solved for  $n_A$ . For example, for an  $r:1$   $\text{Var}(S)$  can be derived as:

$$\text{Var}(S) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} = \frac{r+1}{r} \cdot \frac{\sigma^2}{n_A} \quad (2.3)$$

Where  $\sigma^2$  is the population variance estimate. Substituting into equation (2.2) gives [35]:

$$n_A = \frac{(r+1)(Z_{1-\beta} + Z_{1-\alpha/2})^2 \sigma^2}{r d^2}, \quad (2.4)$$

where  $n_B = r n_A$ . Note:  $n = n_B + n_A$  is minimised when  $r = 1$ .

When the clinical trial has been conducted and the data has been collected and cleaned for analysis it is usually the case that for the analysis the population variance,  $\sigma^2$ , is considered unknown and a sample variance estimate,  $s^2$ , is used instead of  $\sigma^2$ . As a consequence of this a t-statistic as opposed to a Z-statistic is used for inference. This fact should be represented in the sample size calculation rewriting equation (2.4) so that t- as opposed to Z-values are used. Hence, the following equation should be used:

$$n_A \geq \frac{(r+1)(Z_{1-\beta} + t_{1-\alpha/2, n_A(r+1)-2})^2 \sigma^2}{r d^2} \quad (2.5)$$

where  $n_A$  is now defined as the least integer values that satisfies equation (2.5). As  $n_A$  appears on both the left and right side of the equation (2.5) it is best to re-write the equation in terms of power and then use an iterative procedure to solve for  $n_A$ :

$$1 - \beta = \Phi \left( \sqrt{\frac{r n_A d^2}{(r+1) \sigma^2}} - t_{1-\alpha/2, n_A(r+1)-2} \right) \quad (2.6)$$



where  $\Phi(\bullet)$  is defined as the cumulative density function of  $N(0,1)$ . Practically one could use equation (2.4) for the initial sample size calculation and then calculate the power for this sample size using equation (2.7), iterating the sample size up as necessary until the required power is reached. However, when a sample variance is being used in the analysis the power should be estimated from a cumulative t distribution as opposed to a cumulative Normal [35, 36, 37]. The reason for this is that by replacing  $\sigma^2$  with  $s^2$  equation 2.6 becomes:

$$1 - \beta = P\left(\sqrt{\frac{m_A d^2}{(r+1)s^2}} - t_{1-\alpha/2, n_A(r+1)-2}\right)$$

where  $P(\bullet)$  denotes a cumulative distribution defined below. This equation can in turn be rewritten as:

$$1 - \beta = P\left(\frac{\sqrt{m_A} d / \sqrt{(r+1)\sigma}}{\sqrt{s^2/\sigma^2}} - t_{1-\alpha/2, n_A(r+1)-2}\right)$$

by dividing top and bottom by  $\sigma^2$ . Thus, one has a Normal over a square root of a chi-squared which by definition is t-distribution. In fact as the power is estimated under the alternative hypothesis, and that under this hypothesis  $d \neq 0$ , Senn has shown specifically that instead of a t distribution the power should be estimated from a non-central t distribution with degrees of freedom  $n_A(r+1)-2$  and non-centrality parameter  $\sqrt{m_A/(r+1)\sigma^2}$  [35]. Thus, equation 2.6 should in fact be rewritten as:

$$1 - \beta = 1 - \text{Prob}\left(t_{1-\alpha/2, n_A(r+1)-2}, n_A(r+1) - 2, \sqrt{\frac{m_A d^2}{(r+1)\sigma^2}}\right) \quad (2.7)$$

where  $\text{Prob}(\bullet)$  is defined as the cumulative density function of a non central t distribution.

To further aid in these calculations a correction factor of  $Z_{1-\alpha/2}/4$  can be added to equation (2.4) to allow for the Normal approximation [38, 11]:

$$n_A = \frac{(r+1)(Z_{1-\beta} + Z_{1-\alpha/2})^2 \sigma^2}{r d^2} + \frac{Z_{1-\alpha/2}^2}{4}. \quad (2.8)$$

For quick calculations the following formula, to calculate a sample size, with 90% power and a two-sided 5% type I error rate, can be used:

$$n_A = \frac{10.5\sigma^2 (r+1)}{d^2 r}, \quad (2.9)$$

or for  $r=1$ :

$$n_A = \frac{21\sigma^2}{d^2}.$$

This,  $21/\delta^2$  ( $\delta = d/\sigma$ ) is a particularly useful result to remember for quick calculations. Equations (2.4) and (2.8) are close approximations to equation (2.7), giving estimates only one or two lower and thus provide quite good initial estimates. Table 2.1 gives sample sizes using equation 2.7 for various standardised differences ( $\delta = d/\sigma$ ).

### 2.1.1. Worked Example

#### 2.1.1.1. Using the sample size tables

An investigator wishes to design a hypertension trial with equal allocation between groups where the clinical effect of interest is a reduction in blood pressure, compared to control, of 8mmHg ( $d$ ). The expected standard deviation in the population in which the trial is to be undertaken is 40mmHg ( $\sigma$ ). Thus, the standardised difference equates to  $\delta = d/\sigma = 8/40 = 0.20$ . With the Type I and Type II errors fixed at 5% and 10% equation 2.8 gives a sample size of 526. Using this sample size to initiate iterations in Equation 2.7 one gets the following steps:

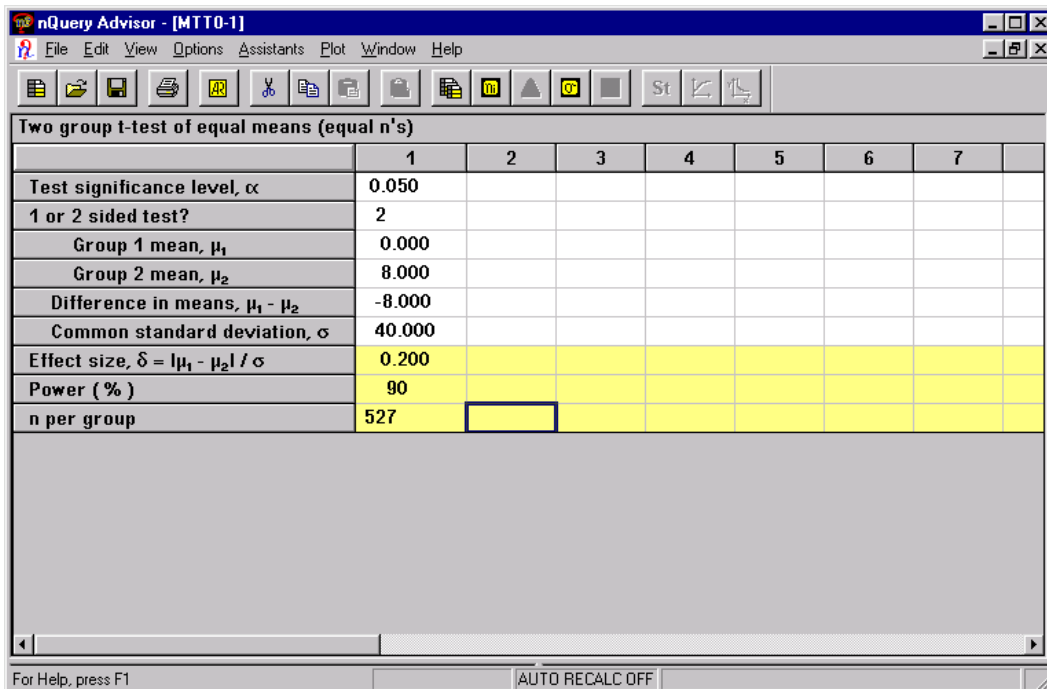
Iteration	n	Power
1	526	0.8993
2	527	0.9004

Thus, the sample size required is 527 subjects in each arm of the trial and a total sample size of 1054. Alternatively one could look up the standardised effect of 0.20 in table 2.1 which gives the same sample size.

If the trial was designed with an unequal allocation of 2:1 ( $r=2$ ) in favour of the control then one would required 395 subjects on the control arm and 790 in the investigation arm; a total sample size of 1185 patients.

### 2.1.1.2. Repeated using sample size software

To do the same calculations in nQuery one would need to click on File/New for Goal tick Means, Number of Groups tick Two and Analysis Method tick Test. Then select Two-sample t-test. There is an additional tick box depending whether wanted to have an equal or



unequal sample size. Above is the dialog box that subsequently comes from nQuery and the entries required to repeat the calculations given in Table 2.1. nQuery also returns a sample of 527 patients per group for an equal allocation ratio and 395 and 790 if the allocation ratio is 2:1

To repeat the calculations in PASS one needs to highlight Means and then t-test: 2 Groups. PASS gives a sample size of 526 one less than nQuery and Table 2.1 for equal allocation but gives the same sample sizes for an allocation ratio of 2:1. More details of the dialogue boxes of PASS will be given in the worked example of the next sub section on cross-over trials.

## 2.2 Cross-over trials

For the analysis of cross-over trial data this paper will concentrate on the case where an analysis of variance is the primary analysis (with a model with terms for subject, period and treatment). The additional assumption is that one is undertaking an AB/BA cross-over trial although the methodology described can be extended to a pair wise comparison in a multi-period cross-over trial (with appropriate adjustment to the degrees of freedom). With the analysis of variance approach the within subject residual errors are assumed to be sampled from a Normal distribution. This approach is equivalent to the period adjusted t-test which will be described on section 2.2.1 [35].

### 2.2.1. Paired t-tests and period adjusted t-tests

The difference between a period adjusted t-test and a standard paired t-test is that for a paired t-test one simply places the observed individual effects on the two treatments in two columns – ignoring any treatment ordering. For each subject a treatment difference is calculated and consequently a mean of these differences,  $\bar{d}$  (equivalent to a difference in the treatment means  $\mu_A - \mu_B$ ), and a n an estimate of the population standard deviation of the differences  $s_d$ . The test statistic is thus  $\bar{d}\sqrt{n}/s_d$ . This is compared to the t distribution on n-1 degrees of freedom. In comparison for a period adjusted t-test for each treatment sequence (AB or BA) a mean difference is calculated,  $\bar{d}_{AB}$  (equivalent to  $\mu_A - \mu_B$ ) and  $\bar{d}_{BA}$  (equivalent to  $\mu_B - \mu_A$ ).

Assuming that there is equal allocation to each sequence,  $n_{AB} = n_{BA} = n/2$ , and the within sequence variances,  $s_{d_{AB}}^2 = s_{d_{BA}}^2 = s_d^2$ , are the same then the mean difference of interest,

$(\bar{d}_{AB} - \bar{d}_{BA})/2$ , has the variance  $s_d^2(1/n_{AB} + 1/n_{BA})/4 = s_d^2/\sqrt{n}$ . Thus, the test statistic is

$$\frac{1/2(\bar{d}_{AB} - \bar{d}_{BA})}{s_d/\sqrt{n}}$$

which is compared to the t distribution on n-2 degrees of freedom.

If there is truly no period effect then,

$$\frac{1/2(\bar{d}_{AB} - \bar{d}_{BA})}{s_d/\sqrt{n}} \approx \frac{1/2((\mu_A - \mu_B) - (\mu_B - \mu_A))}{s_d/\sqrt{n}} \approx \frac{\bar{d}\sqrt{n}}{s_d}$$

and thus one would have an equivalent test to a paired t-test but with one less degree of freedom.

### 2.2.2. Sample Size Calculations

To estimate a sample size for a cross-over trial as well as quantifying the within subject estimate of the difference in treatment means that is of interest ( i.e. the effect size), one needs an estimate of the within- (intra-) subject standard deviation  $\sigma_w$ . The within-subject standard deviation is taken from the residual line of an ANOVA model and quantifies the expected variation among repeated measurements on the same individual [10].

Note that the within subject variability estimates from an ANOVA, model is related to the variability about the difference from a paired t-test through the following result  $\sigma_d^2 = 2\sigma_w^2$ .

With an estimate of both the within subject standard deviation and the effect size equation (2.2) can again be solved as per a parallel group study:

$$n = \frac{2(Z_{1-\beta} + Z_{1-\alpha/2})^2 \sigma_w^2}{d^2}, \quad (2.10)$$

where  $n$  here is the **total sample size**. Note that the allocation ratio has not been used as per equation (2.4) as in a cross-over trial the meaning of  $r$  here would be the allocation ratio to each treatment sequence AB and BA. The assumption here is that an equal number of subjects will be assigned to each sequence. For unknown variance one can rewrite equation (2.10) as:

$$n \geq \frac{2(Z_{1-\beta} + t_{1-\alpha/2, n-2})^2 \sigma_w^2}{d^2}, \quad (2.11)$$

where  $n$  now is the least integer value that satisfies equation (2.11). In turn equation (2.11) can be rewritten in terms of power to solve iteratively for  $n$ :

$$1 - \beta = \Phi\left(\sqrt{\frac{nd^2}{2\sigma_w^2}} - t_{1-\alpha/2, n-2}\right). \quad (2.12).$$

Similarly to parallel group trials, when the population variances is unknown, under  $H_1: d \neq 0$  the Type II error (and hence the power) should be calculated under the assumption of a non-central t distribution with degrees of freedom  $n-2$  and non-centrality parameter  $\sqrt{nd^2/2\sigma_w^2}$  [35]. Thus, equation 2.12 can be rewritten as:

$$1 - \beta = 1 - \text{Probt}\left(t_{1-\alpha/2, n-2}, n-2, \sqrt{\frac{nd^2}{2\sigma_w^2}}\right). \quad (2.13)$$

Again to solve for  $n$  in the same manner as for a parallel group study one can add a correction factor of  $Z_{1-\alpha/2}/2$  to equation (2.10) to allow for the Normal approximation, and use this for initial calculations in equation (2.13) [38]:

$$n = \frac{2(Z_{1-\beta} + Z_{1-\alpha/2})^2 \sigma_w^2}{d^2} + \frac{Z_{1-\alpha/2}}{2}. \quad (2.14)$$

For quick calculations one can adapt equation (2.10) for the calculation of sample sizes (estimated with 90% power and a two-sided 5% type I error rate):

$$n = \frac{21\sigma_w^2}{d^2}. \quad (2.15)$$

Equations (2.14) and (2.15) give slightly lower results than equation 2.13. Table 2.2 gives sample sizes using equation (2.13) for various standardised differences ( $\delta = d / \sigma$ ). The total sample for a cross-over trial are nearly equivalent to that for one arm of a parallel group study, for each standardised difference ( $\delta$ ). The slight differences are accounted for by the different degrees of freedom used in equations (2.7) and (2.13). Practically, though, they are the same.

It should be noted that the standardised differences in Tables 2.1 and 2.2 represent different quantities. The within- subject variance in a cross-over trial can be derived from  $\sigma_w^2 = \sigma^2(1 - \rho)$  - where  $\sigma^2$  is the population variance from a conventional parallel group design and  $\rho$  is the Pearson correlation coefficient estimated between two measures on the same subject. For a relatively modest correlation of 0.5, the within-subject variance would be half the population variance, and as a consequence for an equivalent mean difference the standardised difference would be 40% larger in a cross-over trial compared to a parallel group study. Parallel group and cross-over trials will only have an equivalent standardised difference for a zero correlation.

### **2.2.3. Worked Example**

#### **2.2.3.1. Using the sample size tables**

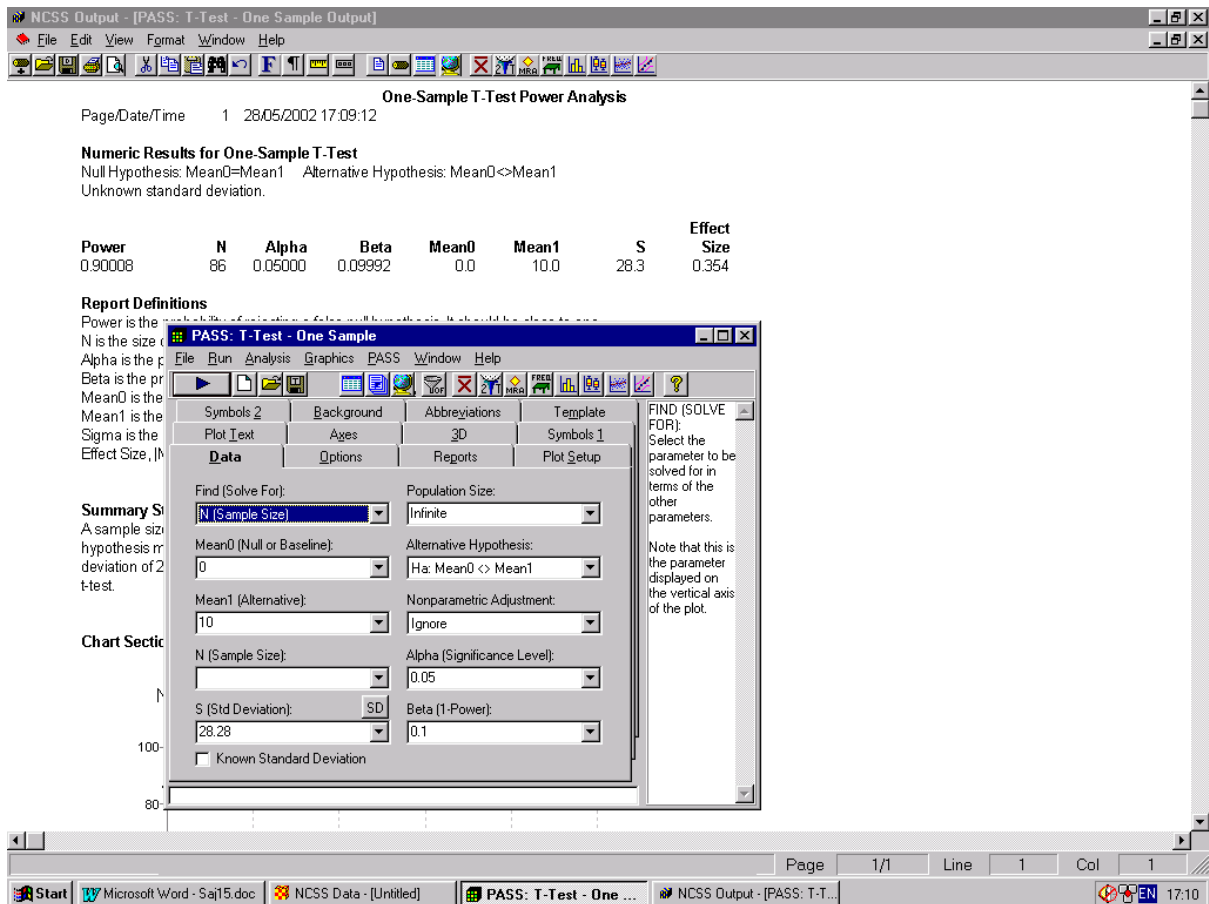
An investigator wishes to design a hypertension trial similar to that in Section 2.1.1. The clinical effect of interest is a reduction in blood pressure compared to control of 10mmHg (d). The expected within-subject standard deviation in the trial population the trial is expected to be half that of the between-subject standard deviation at 20mmHg ( $\sigma_w$ ). Thus, the standardised difference is  $\delta = d / \sigma_w = 10 / 20 = 0.50$ . For the Type I and Type II errors fixed at 5% and 10% respectively Table 2.3 gives a total sample size of 86.

### 2.2.3.2. Repeated using sample size software

For the sample size calculations in PASS and nQuery the assumption is that instead of doing an analysis of variance for the final analysis a paired t-test would be undertaken. As described in Section 2.2.1. for studies with paired data, one must specify the standard deviation of the difference of the outcome variable measured on the two treatments and the standard deviation of the difference can be calculated from the within subject standard deviation from the result  $\sigma_d = \sqrt{2}\sigma_w$ . Thus, for a paired t-test the standard deviation of the difference,  $\sigma_d$ , should be used instead of the within subject standard deviation and one should therefore replace  $2\sigma_w^2$  with  $\sigma_d^2$  in each of equations 2.10 to 2.15 and adjust the degrees of freedom to n-1 in equations 2.12 and 2.13.

To repeat the calculations in PASS one selects Means and then T-Test: 1 Group. The following dialogue box and consequent output then comes up:





The mean difference is still the same as in the worked example, 10, but the standard deviation for the calculations is now  $\sqrt{2} * 20 = 28.28$ . PASS gives the sample size as 86 as per table 2.2. To do the same calculations in nQuery one would need to click on File/New for Goal tick Means, Number of Groups tick One and Analysis Method tick Test. Then select Paired t-test. nQuery too returns a sample size of 86.

By looking at the two dialogue boxes for nQuery (given earlier for the parallel group case) and PASS one can see the two approaches to calculations in the two packages. nQuery works like a spread sheet where the inputs are entered into a column with the answer (i.e. the sample size) given at the bottom of the column. If one wishes to do several sample size calculations then one needs to fill in several columns. PASS works by entering the inputs into dialogue boxes with the answer going to separate window. One can enter several values into each

input, for example for Mean one can enter "5, 10, 15" or "5 to 15 by 5", and PASS will output the sample sizes for different values (or combination of values) in one Output window.

### 3. EQUIVALENCE TRIALS

In certain cases the objective of a clinical trial is not to demonstrate superiority but to demonstrate that two treatments have no clinically meaningful difference, i.e. that they are clinically equivalent. The null ( $H_1$ ) and alternative ( $H_0$ ) hypotheses for such equivalence trials take the form:

$H_0$ : The two treatment differences are different with respect to the mean response ( $\mu_A \neq \mu_B$ ).

$H_1$ : The two treatments are not different with respect to the mean response ( $\mu_A = \mu_B$ ).

Usually these hypotheses are written in terms of a clinical difference,  $d$ , and become:

$H_0: \mu_A - \mu_B \leq -d$  or  $\mu_A - \mu_B \geq +d$ .

$H_1: -d < \mu_A - \mu_B < +d$ .

These hypotheses are an example of an intersection-union test (IUT), in which the null hypothesis is expressed as a union and the alternative as an intersection. In order to conclude equivalence, one needs to reject each component of the null hypothesis.

Note that in an IUT, each component is tested at level  $\alpha$  giving a composite test which is also of level  $\alpha$  [39].

A common approach with equivalence trials to test each component of the null hypothesis with a  $t$  test - called the Two One-Sided Test (TOST) procedure. In practice, this is operationally the same as constructing a  $(1-2\alpha)100\%$  confidence interval for  $f(\mu)$  where equivalence is concluded provided that each end of the confidence interval falls completely within the interval  $(-d,+d)$  [40]. This is because the  $(1-2\alpha)100\%$  confidence interval is excluding two regions each of size  $\alpha$ , each of which must simultaneously preclude  $(-d, +d)$ . Hence, the overall significance level is  $\alpha$ .

Figure 1 highlights how equivalence can be demonstrated through confidence intervals and Figure 2 demonstrated how confidence intervals are used to test the different hypotheses in superiority and equivalence trials. The special case of bioequivalence is covered in Section 6. ICH E10 [41] goes into some detail in the description of equivalence trials, and the related non-inferiority trials (discussed in Section 4) whilst ICH E9 and E3 discuss the appropriate analysis of such trials [3, 7].

In this section the sample size formulae will initially be derived

- i) For the general case of inequality between treatments (i.e.  $f(\mu) = \Delta$ )
- ii) Adopting the same notation and assumptions as in Section 2
- iii) Under the assumption that the equivalence bounds  $-d$  and  $d$  are symmetric about zero

This section will then move on to the special case of no treatment difference replacing (i) with:

- i) For the special case of no mean difference (i.e.  $f(\mu) = 0$ ).

### 3.1. General case

As with Section 2:

$$\frac{S - f(\mu)}{\sqrt{\text{Var}(S)}} \sim N(0,1),$$

Hence, the  $(1 - 2\alpha)$  100% confidence limits for a non-zero mean difference would be:

$$S - \Delta \pm Z_{1-\alpha} \sqrt{\text{Var}(S)},$$

To declare equivalence the lower and upper confidence limit should be within  $\pm d$ :

$$S - \Delta - Z_{1-\alpha} \sqrt{\text{Var}(S)} > -d \text{ and } S - \Delta + Z_{1-\alpha} \sqrt{\text{Var}(S)} < d. \quad (3.1)$$

Thus, by extending the arguments for superiority trials, for the two one sided test procedure (TOST) with this critical region there are two opportunities against an alternative to have a Type II error for some chosen  $d$  and power  $(1-\beta)$

$$\Delta + d - Z_{1-\beta_1} \sqrt{\text{Var}(S)} = Z_{1-\alpha} \sqrt{\text{Var}(S)} \text{ and } \Delta - d - Z_{1-\beta_2} \sqrt{\text{Var}(S)} = Z_{1-\alpha} \sqrt{\text{Var}(S)}. \quad (3.2)$$

where  $\beta_1$  and  $\beta_2$  are the Type II errors associated with each one sided test from the TOST procedure and  $\beta = \beta_1 + \beta_2$ . Hence,

$$Z_{1-\beta_1} = \frac{-d - \Delta}{\sqrt{\text{Var}(S)}} - Z_{1-\alpha} \text{ and } Z_{1-\beta_2} = \frac{d - \Delta}{\sqrt{\text{Var}(S)}} - Z_{1-\alpha} \quad (3.3)$$

### 3.2. Special case of no treatment difference.

For the special case of no treatment difference  $\Delta = 0$  can be entered into (3.1). Thus, with the TOST procedure the Type II error for some chosen  $d$  and power  $(1-\beta)$  will come from

$$d - Z_{1-\beta} \sqrt{\text{Var}(S)} = Z_{1-\alpha} \sqrt{\text{Var}(S)} \text{ and } -d - Z_{1-\beta} \sqrt{\text{Var}(S)} = Z_{1-\alpha} \sqrt{\text{Var}(S)}.$$

Hence,

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

giving:

$$\text{Var}(S) = \frac{d^2}{(Z_{1-\alpha} + Z_{1-\beta/2})^2}. \quad (3.4)$$

### 3.3. Type I and setting the equivalence limit

#### 3.3.1. Choice of Type I error

Strictly speaking when undertaking two simultaneous one tailed tests setting  $\alpha=0.05$  would maintain an overall Type I error rate of 5%. However, the choice of the Type I error is a

controversial issue. The convention for equivalence trials is to set the Type I error rate at half of that which would be employed for a two sided test used in a superiority trial i.e.  $\alpha=0.025$ . That is, giving a Type I error rate of 2.5% [3]. However, setting the Type I error rate for equivalence trials at half that for superiority trials could be considered to be consistent. This is because although in a superiority trial one has a two sided 5% significance level in practice for most trials in effect what one has is a one sided investigation with a 2.5% level of significance. The reason for this is that one usually has an investigative therapy and a control therapy and it is only statistical superiority of the investigative therapy that is of interest.

Through the rest of the sections on equivalence and non-inferiority trials the assumption will be that  $\alpha=0.025$  and that 95% confidence intervals will be used in the final statistical analysis. This issue will be discussed again in the section on Bioequivalence.

### **3.3.2. Choice of Equivalence Limit**

The discussion on equivalence limits in this section can also be generalised to non-inferiority trials discussed in the preceding section. As with the choice of the Type I error the setting of the non-inferiority/equivalence limit is a controversial issue. The equivalence limit is defined as the "largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice" [42]. This difference also cannot be "greater than the smallest effect size that the active (control) drug would be reliably expected to have compared with placebo in the setting of the planned trial" [41].

However, beyond this there has not much formal guidance. Jones, Jarvis, Lewis et al [40] have recommended that the choice of limit be set at half the expected clinically meaningful difference between the active control and placebo. There are no hard regulatory guidance although the CPMP [43] in a concept paper state that for non-mortality studies it may be

acceptable to have an equivalence limit "of one half or one third of the established superiority of the comparator to placebo, especially if the new agent has safety or compliance advantages"

The definition of the acceptable level of equivalence or non-inferiority is made therefore with reference to some retrospective comparison to placebo [44, 45, 46]. In this context the definition of the non-inferiority and equivalence limits should address steps of the form [45, 46].

1. One must be confident that the active control would have been different from placebo had one been employed.
2. One should be able to determine that there is no clinically meaningful difference between investigative treatment and the control.
3. Through comparing the investigative treatment to control one should indirectly be able to determine that it is superior to placebo.

Steps 1. and 3. are important as there is a view that non-inferiority and equivalence trials reward "failed" studies i.e. if one conducted a poor trial where it would not have been possible to demonstrate the control to be superior to placebo then a poor investigative therapy may slip through comparison to this control. However, Julious and Zariffa [2] point out that this may not be the case as poor studies are poor for most objectives due to their higher statistical variability.

In summary therefore one can infer that the clinical difference used for the limits of equivalence and non-inferiority will be smaller than the difference used for placebo controlled superiority trials. There is no generic definition for its setting – its definition will need to be defined on a study by study basis with consultation with the appropriate agencies.

### 3.4. Parallel group trials

#### 3.4.1. General case.

For equivalence trials the sample size cannot be derived directly for the general case where the expected true mean difference is not fixed to be zero. This is because the total Type II error is the sum of the Type II errors associated with each one-tailed test.

As is the case with superiority trials  $\text{Var}(S)$  can be defined as :

$$\text{Var}(S) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B} = \frac{r+1}{r} \cdot \frac{\sigma^2}{n_A}. \quad (3.5)$$

From this (and the fact that  $\beta = \beta_1 + \beta_2$ ), equation (3.3) can be used to derive the power (and Type II error):

$$1 - \beta = \Phi\left(\sqrt{\frac{((\mu_A - \mu_B) - d)^2 n_A}{(r+1)\sigma^2}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{\frac{((\mu_A - \mu_B) + d)^2 n_A}{(r+1)\sigma^2}} - Z_{1-\alpha}\right) - 1. \quad (3.6)$$

To obtain the required sample size equation (3.6) until a sample size is reached which gives the required power (Type II error). For unknown variance equation (3.6) can be re-written as:

$$1 - \beta = \Phi\left(\sqrt{\frac{((\mu_A - \mu_B) - d)^2 n_A}{(r+1)\sigma^2}} - t_{1-\alpha, n_A(r+1)-2}\right) + \Phi\left(\sqrt{\frac{((\mu_A - \mu_B) + d)^2 n_A}{(r+1)\sigma^2}} - t_{1-\alpha, n_A(r+1)-2}\right) - 1. \quad (3.7)$$

As with superiority trials it is best to use a non-central t-distribution to calculate the Type II error and power. From a non-central t-distribution the power can be calculated using the following formula [37, 47, 48]

$$1 - \beta = \text{Probt}(-t_{1-\alpha, n_A(r+1)-2}, n_A(r+1) - 2, \tau_2) - \text{Probt}(t_{1-\alpha, n_A(r+1)-2}, n_A(r+1) - 2, \tau_1), \quad (3.8)$$

where  $\tau_1$  and  $\tau_2$  are non centrality parameters defined as:

$$\tau_1 = \frac{((\mu_A - \mu_B) + d)\sqrt{rn_A}}{\sqrt{(r+1)\sigma^2}} \quad \text{and} \quad \tau_2 = \frac{((\mu_A - \mu_B) - d)\sqrt{rn_A}}{\sqrt{(r+1)\sigma^2}}.$$



For quick calculations (and to provide an initial value for the sample size in the iterations), an estimate of the sample size can be obtained from the following equation

$$n_A = \frac{(r+1)\sigma^2(Z_{1-\beta} + Z_{1-\alpha})^2}{r((\mu_A - \mu_B) - d)^2}. \quad (3.9)$$

This provides reasonable approximations for the sample size when the mean difference is greater than zero ( $\mu_A - \mu_B > 0$ ), and approaches  $d$ . For very quick calculations (for 90% power and Type I error of 2.5%), the following formula can be used:

$$n_A = \frac{10.5\sigma^2(r+1)}{((\mu_A - \mu_B) - d)^2 r}, \quad (3.10)$$

or for  $r=1$ :

$$n_A = \frac{21\sigma^2}{((\mu_A - \mu_B) - d)^2}. \quad (3.11)$$

### 3.4.2. Special case of no treatment difference.

For the special case of no treatment difference ( $\mu_A - \mu_B = 0$ ), equation (3.5) can be substituted into equation (3.4) to obtain a direct estimate of the sample size

$$n_A = \frac{(r+1)\sigma^2(Z_{1-\beta/2} + Z_{1-\alpha})^2}{rd^2}. \quad (3.12)$$

For unknown variance equation (3.12) can be as

$$n_A \geq \frac{(r+1)\sigma^2 \left( Z_{1-\frac{\beta}{2}} + t_{1-\alpha, n_A(r+1)-2} \right)^2}{rd^2}, \quad (3.13)$$

Where  $n_A$  is the smallest integer value to satisfy equation (3.12). Equation 3.13 can in turn be rewritten to give power in terms of the sample size:

$$1 - \beta = 2\Phi \left( \sqrt{\frac{rd^2 n_A}{(r+1)\sigma^2}} - t_{1-\alpha, n_A(r+1)-2} \right) - 1. \quad (3.14)$$

Similarly to equation (3.8), under the assumption of a non-central t-distribution, the power can be derived from

$$1 - \beta = 2\text{Probt}\left(-t_{1-\alpha, n_A(r+1)-2}, n_A(r+1) - 2, \tau\right) - 1, \quad (3.15)$$

where  $\tau$  is defined as

$$\tau = \frac{-\sqrt{n_A}rd}{\sqrt{(r+1)\sigma^2}}.$$

For quick calculations (for 90% power and Type I error of 2.5%), the following formula can be used:

$$n_A = \frac{13\sigma^2(r+1)}{d^2r}, \quad (3.16)$$

or, for  $r=1$ ,

$$n_A = \frac{26\sigma^2}{d^2}. \quad (3.17)$$

It is worth noting here the difference between equations 3.16 and 3.17 and those given earlier equations 3.10 and 3.11. There is a difference in the coefficients (10.5 and 21 for equations 3.10 and 3.11 respectively and compared to 13 and 16 for equations 3.16 and 3.17) which is due to the non-symmetric allocation of the Type II error if the population mean is non zero.

Table 3.1 gives sample sizes using equation 3.8 for various standardised equivalence limits ( $\delta = d/\sigma$ ) and standardised mean differences, assuming equal allocation between groups.

### **3.4.3. Worked example**

#### **3.4.3.1. Using the sample size tables**

An investigator wishes to design a pain trial where the objective is to demonstrate equivalence between two treatments. The largest clinically acceptable effect for which equivalence can be declared is a change in visual analogue scale (VAS) assessed pain of 10mm (d). There is to be equal allocation between groups. The true mean difference between the treatments is thought to be zero and the expected standard deviation in the population in which the trial is to be undertaken is 50mm ( $\sigma$ ). Thus, the standardised equivalence limits are  $\pm \delta = \pm d / \sigma = \pm 10 / 50 = \pm 0.20$ . For the Type I and Type II errors fixed at 2.5% and 10% respectively Table 3.1 gives a sample size of 651 patients in each arm of the trial.

Suppose the true mean difference is thought to be 2mm. This equates to 20% of the standardised equivalence limits and would inflate the sample size to 827 patients in each arm of the trial.

#### **3.4.3.2. Repeated using sample size software**

To repeat the calculations in PASS one needs to select Means and then Equivalence-Means. The dialogue box below details the entries required to repeat both calculations in the worked example. One typographical issue to note is that PASS does not distinguish between bioequivalence and equivalence trials which as will be highlighted Section 6 are two difference concepts and so as a result PASS has as the heading in the output box "Bioequivalence Means Power Analysis"

NCSS Output - [PASS: Equivalence - Means Output]

Page/Date/Time 1 07/06/2002 16:41:55

**Bioequivalence Means Power Analysis**

**Numeric Results for Parallel-Group Design**

	Reference Mean	Treatment Mean	Lower Diff. Limit	Upper Diff. Limit	Standard Deviation	Sample Size 1 (N1)	Sample Size 2 (N2)	Alpha	Beta
Power 0.9005	0	0	-10	10	50	651	651	0.0250	0.0995
0.9003	0	2	-10	10	50	827	827	0.0250	0.0997

**Report Definitions**  
 Power is the probability of rejecting non-equivalence when they are equivalent.  
 The Reference Mean is the mean of the standard procedure.  
 The Treatment Mean is the mean of the treatment procedure.  
 The Upper Limit is the upper limit of the equivalence region.  
 The Lower Limit is the lower limit of the equivalence region.  
 The Standard Deviation is the standard deviation of the data.  
 N1 is the number of subjects in the first group.  
 N2 is the number of subjects in the second group.  
 Alpha is the probability of a Type I error.  
 Beta is the probability of a Type II error.

**Summary**  
 In an equivalence trial, the sample size is calculated for one-sided tests. The results in this report are for two one-sided tests (TOST).

**Chart Section**  
 N1  
 85  
 800

Microsoft Word - Saj15.doc NCSS Data - [Untitled] NCSS Output - [PASS: Equivalence - Means Output] PASS: Equivalence - Means

PASS gives a sample sizes respectively 651 and 827 respectively for the case of no treatment difference and a treatment difference of 2mm. The same as table 3.1.

To repeat the calculations in nQuery one would need to click on File/New, for Goal tick Means, Number of Groups tick Two and Analysis Method tick Equivalence. Then select Two one-sided tests (TOST) for two group or cross-over. nQuery too gives the same answers as table 3.1 for the two cases in the worked example.

### 3.5. Cross-over trials

The methodologies and assumptions for an equivalence trial with a cross-over design are the same as those for parallel group trials. This subsection will therefore only go briefly through the sample size calculations.

### 3.5.1. General case.

The power (and Type II error) can be estimated from

$$1 - \beta = \Phi \left( \sqrt{\frac{((\mu_A - \mu_B) - d)^2 n}{2\sigma_w^2}} - Z_{1-\alpha} \right) + \Phi \left( \sqrt{\frac{((\mu_A - \mu_B) + d)^2 n}{2\sigma_w^2}} - Z_{1-\alpha} \right) - 1. \quad (3.17)$$

For unknown variance equation (3.17) can thus be re-written as

$$1 - \beta = \Phi \left( \sqrt{\frac{((\mu_A - \mu_B) - d)^2 n}{2\sigma_w^2}} - t_{1-\alpha, n-2} \right) + \Phi \left( \sqrt{\frac{((\mu_A - \mu_B) + d)^2 n}{2\sigma_w^2}} - t_{1-\alpha, n-2} \right) - 1, \quad (3.18)$$

and under the assumption of a non-central t-distribution the power [Owen, Diletti et al, Chow et al) the power can be estimated from

$$1 - \beta = \text{Probt}(-t_{1-\alpha, n-2}, n-2, \tau_2) - \text{Probt}(t_{1-\alpha, n-2}, n-2, \tau_1), \quad (3.19)$$

where  $\tau_1$  and  $\tau_2$  are defined as

$$\tau_1 = \frac{((\mu_A - \mu_B) + d)\sqrt{n}}{\sqrt{2\sigma_w^2}} \quad \text{and} \quad \tau_2 = \frac{((\mu_A - \mu_B) - d)\sqrt{n}}{\sqrt{2\sigma_w^2}}.$$

For quick calculations one could use:

$$n = \frac{2\sigma_w^2 (Z_{1-\beta} + Z_{1-\alpha})^2}{((\mu_A - \mu_B) - d)^2}, \quad (3.20)$$

for sample size estimation and for very quick calculations (for 90% power and Type I error of 2.5%), one can use the following formula:

$$n = \frac{21\sigma_w^2}{((\mu_A - \mu_B) - d)^2}. \quad (3.21)$$

### 3.5.2. Special case of no treatment difference.

For the special case of no treatment difference ( $\mu_A - \mu_B = 0$ ), a direct estimate of the sample size can be estimated from

$$n = \frac{2\sigma_w^2(Z_{1-\beta/2} + Z_{1-\alpha})^2}{d^2}, \quad (3.22)$$

which, with unknown variance, can be re-written as

$$n = \frac{2\sigma_w^2\left(Z_{1-\frac{\beta}{2}} + t_{1-\alpha, n-2}\right)^2}{d^2}. \quad (3.23)$$

Equation 3.23 can in turn be re-written in terms of power for a given sample size

$$1 - \beta = 2\Phi\left(\sqrt{\frac{d^2 n}{2\sigma_w^2}} - t_{1-\alpha, n-2}\right) - 1, \quad (3.24)$$

which in turn (under the assumption of a non-central t-distribution), can also be rewritten as:

$$1 - \beta = 2\text{Probt}\left(-t_{1-\alpha, n-2}, n - 2, \tau\right) - 1, \quad (3.25)$$

where  $\tau$  is defined as

$$\tau = \frac{-\sqrt{nd}}{\sqrt{2\sigma_w^2}}, \text{,}$$

For quick calculations (for 90% power and Type I error of 2.5%), the following formula, can be used:

$$n = \frac{26\sigma_w^2}{d^2}. \quad (3.26)$$

The quick equations give reasonable estimates of the sample size, underestimating the sample size by just one or two subjects, and thus provides reasonable initial values for equations (3.19) and (3.25). Table 3.2 gives sample sizes using equation 3.19 for various standardised equivalence limits ( $\delta = d/\sigma$ ) and mean differences.

### 3.5.3 Worked example

#### 3.5.3.1. Using the sample size tables

An investigator wishes to design a pain trial similar to that in Section 3.4.3.1. Again the largest clinically acceptable effect for equivalence to be declared is a change in visual analogue scale (VAS) assessed pain of 10mm (d) and the true mean difference between the treatments is thought to be zero. The expected within-subject standard deviation in the trial population is 20mm ( $\sigma_w$ ). Thus, the standardised equivalence limits equate to  $\pm \delta = \pm d / \sigma = \pm 10 / 20 = \pm 0.50$ . For the Type I and Type II errors fixed at 2.5% and 10%, respectively, Table 3.2 gives a total sample size of 106 patients in the trial.

If the true mean difference is thought to be 2mm, equating to 20% of the standardised equivalence limits, the sample size would be inflated to a total of 135 patients in the trial.

#### 3.5.3.2. Repeated using sample size software

To do the same sample size calculations in nQuery one would need to click on File/New, for Goal tick Means, Number of Groups tick Two and Analysis Method tick Equivalence. Then select Two one-sided tests (TOST) for two group or cross-over. For equivalence trials nQuery does not use  $\sigma_d$ , as it does for superiority trials, or  $\sigma_w$ , as used in equation 3.9 but a new variance  $\sigma_w / \sqrt{2}$ , as described in the right hand dialogue box below under the heading "Suggestion". One rational for using this variance is that nQuery does not give the total sample size but the sample size per sequence (assuming one has two sequences AB and BA). By using  $\sigma_w / \sqrt{2}$  for the variance estimate (and by giving the sample size per sequence) it enables nQuery to use the same formula (equation 3.8), for sample size calculations for both cross-over and parallel group trials. This is because the degrees of freedom about the t

statistic in equation (3.8) will be correct for both cross-over and parallel group trials using the nomenclature of sample size per sequence.

The dialogue box below gives the entries to repeat the sample size calculations in nQuery

The screenshot shows the nQuery Advisor software window titled "nQuery Advisor - [MTE1-2]". The main window displays a table for "Two-group or crossover t-tests (TOST) of equivalence in means (equal n's)". The table has columns for 1, 2, 3, 4, and 5 groups. The rows include Test significance levels, Lower and Upper equivalence limits, Expected difference, Common standard deviation, Power, and n per group. The values for group 1 are 0.025, -10.000, 10.000, 0.000, 14.140, 90, and 53. The values for group 2 are 0.025, -10.000, 10.000, 2.000, 14.140, 90, and 68. A help panel on the right provides a definition of common standard deviation and a suggestion for crossover design, along with acceptable entries.

	1	2	3	4	5
<b>Test significance levels, <math>\alpha</math> (one-sided)</b>	0.025	0.025			
<b>Lower equivalence limit for <math>\Delta\beta - \Delta\alpha</math>,</b>	-10.000	-10.000			
<b>Upper equivalence limit for <math>\Delta\beta - \Delta\alpha</math>,</b>	10.000	10.000			
<b>Expected difference, <math>\Delta\beta - \Delta\alpha</math></b>	0.000	2.000			
<b>Common standard deviation, <math>\hat{\sigma}</math></b>	14.140	14.140			
<b>Power ( % )</b>	90	90			
<b>n per group</b>	53	68			

**Common standard deviation,  $\hat{\sigma}$**   
 The standard deviation is a measure of the variability between subjects within a group (it is the square root of the variance). Sigma,  $\hat{\sigma}$ , is assumed to be the same in each of the groups.

**Suggestion:**  
 When the study design is a crossover design, enter the standard deviation of differences divided by 2,  $\hat{\sigma}_D/2$ , or  $1/\sqrt{2}$  times the squareroot of the mean squared error from the crossover ANOVA,  $\hat{\sigma}_A/\sqrt{2}$ . Testing for equivalence of the ratio of means where data will be analyzed in the log scale, is the same as testing the difference in mean logs. When data are log-normally distributed, the variance in the log scale equals the log of  $(1+CV^2)$ . Select Estimate SD from the Assistants (or click on the button marked  $\hat{\sigma}$ ) and choose From Coefficient of Variation to compute an estimate of the standard deviation in the log scale from an estimate of the Coefficient of Variation. Or use table MTE2, Two-group or crossover t-tests (TOST) for ratio of means (using log scale).

**Acceptable entries:**  
 > 0.0

For the equivalent sample sizes to those given earlier for no mean difference and a mean difference of 2 nQuery gives a sample size per treatment sequence of 53 and 68 respectively or 106 and 136 in total. Taking account of rounding nQuery gives the same results for the total sample size as table 3.2. To do the same calculations in PASS one needs to select Means and then Equivalence-Means. the dialogue box is the same as that in Section 3.4.3.2. Now instead of "Parallel Group" though in the "Design Type" box one elects "Crossover"



It is worth noting that for the variance PASS uses  $\sigma_w^2$  (and not  $\sigma_w^2/2$  as with nQuery) but like nQuery it does give the sample size per sequence. PASS gives the same sample size per sequence as nQuery.

There is an issue with the approach of nQuery and PASS in calculating the sample size per sequence as this is assuming that one is investigating just 2 treatments in just 2 sequences (BA and AB). If one was simultaneously investigating 3 treatments say one may have 6 sequences. Another issue is that even if just two treatments are being investigated one may be applying a replicate design as described in Section 6.3.3 where again more than two sequences may be being used. It is more optimal therefore to calculate the total sample size and divide this by the number of sequences to get the sample size per sequence rather than vice versa.

#### 4. NON-INFERIORITY TRIALS

For certain trials the objective is not to demonstrate that two treatments are different or equivalent but rather to demonstrate that a given treatment is clinically not inferior compared to another. The null ( $H_0$ ) and alternative ( $H_1$ ) hypotheses for non-inferiority trials may take the form:

$H_0$ : A given treatment is inferior with respect to the mean response.

$H_1$ : A given treatment is non-inferior with respect to the mean response.

As with equivalence trials these hypotheses are written in terms of a clinical difference,  $d$ , which equates to the largest difference that is clinically acceptable [42]:

$H_0: \mu_A - \mu_B \leq -d$  .

$H_1: \mu_A - \mu_B > -d$  .

In the context of non-inferiority trials  $-d$  is known as the non-inferiority limit. Please see discussion in section 3.3.2 as to its definition. ICH E3 and E9 go into detail on the analysis of non-inferiority trials whilst ICH E10 discusses the definition of  $d$  [3, 7, 41].

In order to conclude non-inferiority, one needs to reject the null hypothesis. In terms of the equivalence hypotheses in Section 3 this is equivalent to testing just one of the two components of the TOST procedure. Thus, non-inferiority trials reduce to a simple one-sided hypothesis and test. In practice, this is operationally the same as constructing a  $(1-2\alpha)100\%$  confidence interval and concluding non-inferiority provided that the lower end of this confidence interval is greater than  $-d$ .

Usually non-inferiority trials (like equivalence trials) compare the investigative therapy to an active control. Statistically they could be considered a special case of equivalence trials. However, operationally non-inferiority trials are more often conducted since it is only the lower equivalence (now non-inferiority) limit that is usually of interest. For a non-inferiority trial a mean difference a long way from  $-d$ , in a positive sense, is not a negative outcome for

the study. Please see also the discussion on "as good as or better trials" for the context of non-inferiority studies with superiority studies.

Figure 3 highlights how non-inferiority can be demonstrated through a confidence interval and Figure 2 shows how confidence intervals are used to test the different hypotheses in superiority, equivalence and non-inferiority trials.

Adopting the same notation and assumptions as in Section 3 but with  $f(\mu) = -\Delta$  and the non-inferiority bound set at  $-d$ , the lower  $(1 - 2\alpha)$  100% confidence limit is

$$S - \Delta - Z_{1-\alpha} \sqrt{\text{Var } S}. \quad (4.1)$$

To declare non-inferiority the lower end of the confidence interval should lie above  $-d$ :

$$S - \Delta - Z_{1-\alpha} \sqrt{\text{Var}(S)} > -d. \quad (4.2)$$

For this critical region one therefore requires a  $(1 - \beta)$  100% chance that the lower limit lies above  $-d$  i.e.:

Hence:

$$Z_{1-\beta} = \frac{-d + \Delta}{\sqrt{\text{Var}(S)}} - Z_{1-\alpha}, \quad (4.3)$$

giving:

$$\text{Var}(S) = \frac{(d - \Delta)^2}{(Z_{1-\alpha} + Z_{1-\beta})^2}. \quad (4.4)$$

#### 4.1. Parallel group trials

As with superiority and equivalence trials  $\text{Var}(S)$  can be defined as

$$\text{Var}(S) = \frac{r+1}{r} \cdot \frac{\sigma^2}{n_A},$$

which can be substituted in to equation (4.4) (replacing  $\Delta$  with  $\mu_A - \mu_B$ ) giving a direct estimate of the sample size

$$n_A = \frac{(r+1)\sigma^2(Z_{1-\beta} + Z_{1-\alpha})^2}{r((\mu_A - \mu_B) - d)^2}. \quad (4.5)$$

Re-writing equation (4.5) to give power for a give sample size results in:

$$1 - \beta = \Phi \left( \sqrt{\frac{((\mu_A - \mu_B) - d)^2 r n_A}{(r+1)\sigma^2}} - Z_{1-\alpha} \right). \quad (4.6)$$

The equivalent formula to (4.6), in the case of unknown variance is

$$1 - \beta = \Phi \left( \sqrt{\frac{((\mu_A - \mu_B) - d)^2 r n_A}{(r+1)\sigma^2}} - t_{1-\alpha, n_A(r+1)-2} \right). \quad (4.7)$$

As with the sections on equivalence and superiority trials when the population variance is unknown it is best to calculate the power under the assumption of a non-central t-distribution [37]:

$$1 - \beta = 1 - \text{Prob}t(t_{1-\alpha, n_A(r+1)-2}, n_A(r+1) - 2, \tau), \quad (4.8)$$

where  $\tau$  is defined as

$$\tau = \left| \frac{((\mu_A - \mu_B) - d)\sqrt{r n_A}}{\sqrt{(r+1)\sigma^2}} \right|.$$

For quick calculations (for 90% power and Type I error of 2.5%), the following formula, similar to equation (2.9) can be used:

$$n_A = \frac{10.5\sigma^2(r+1)}{((\mu_A - \mu_B) - d)^2 r}. \quad (4.9)$$

In the case of  $r=1$  (4.9) resolves to:

$$n_A = \frac{21\sigma^2}{((\mu_A - \mu_B) - d)^2}. \quad (4.10)$$

Equations 4.9 and 4.10 are equivalent to equations 3.10 and 3.11 for the case of a non zero mean difference i.e. for  $\mu_A - \mu_B > 0$ . The quick equations give reasonable estimates of the

sample size, although with slight underestimation. Table 4.1 gives sample sizes using equation 4.8 for various standardised non-inferiority limits ( $\delta = d/\sigma$ ) and standardised mean differences assuming equal allocation between groups.

One important feature to highlight in tables 4.1 and 4.2 is the asymmetric effect on the sample size of different values for the true mean difference. In equivalence trials as there are two margins when one moves away from a zero mean difference - in any direction - the sample size is inflated. However, in non-inferiority trials the sample size is inflated only if the true mean difference moves towards the non-inferiority margin. If it is expected that the true mean difference is in favour of the comparator regimen (compared to control) then the sample size is significantly reduced.

The asymmetric effect of the mean difference on the sample size should be considered when designing non-inferiority trials as even only a small expected mean difference in favour of the comparator could have a marked effect on the sample size.

#### **4.1.1. Worked example**

##### **4.1.1.1. Using the sample size tables**

An investigator wishes to design an hypertension trial where the objective is to demonstrate that one treatment (an investigative therapy) is non-inferior to another (a standard therapy). As with the worked example in Section 3.2.3 the largest clinically acceptable effect to be able to declare non-inferiority is a change in blood pressure of 10mmHg ( $d$ ). The true mean difference between the treatments is thought to be zero with an expected standard deviation in the trial population of 40mmHg ( $\sigma$ ). There is to be equal allocation between groups. Thus, the standardised non-inferiority limits equate to  $-\delta = -d/\sigma = -10/40 = -0.25$ . For the Type I and Type II errors fixed at 2.5% and 10% respectively Table 4.1 gives a sample

size of 338 patients in each arm of the trial. The quick formula (equation 4.10) gives 336 patients in each arm.

Suppose, though, that one believes that the investigative therapy is a little superior to the standard such that the true mean difference is thought to be 2mmHg. This inflates the distance one expects the mean to be away from the non-inferiority margin by 20% and as a consequence reduces the sample size to required to 235 patients in each arm of the trial.

#### **4.1.1.2. Repeated using sample size software**

To do non-inferiority sample size calculations in nQuery one would need to click on File/New, for Goal tick Means, Number of Groups tick Two and Analysis Method tick Equivalence. Then click on Equivalence of Two Means.

Note that nQuery does not refer to these calculations as non inferiority but equivalence. However, it is clear from the instructions and the definition of the null hypothesis given in nQuery that the calculations are for a Non-inferiority trial (see the definition of the null hypothesis in the dialogue box below). The entries nQuery dialogue box are given below.

Two group t-test of equivalence in means [equal n's]

	1	2	3	4	5	6
Test significance level, $\alpha$	0.025	0.025				
Equivalence limit difference,	-10.000	-10.000				
Expected difference, $\hat{E}\hat{A}$	0.000	2.000				
$\hat{E}\hat{A} - \hat{E}\hat{A}$	-10.000	-12.000				
Common standard deviation, $\hat{\sigma}$	40.000	40.000				
Effect size, $\hat{E} =  \hat{E}\hat{A} - \hat{E}\hat{A}  / \hat{\sigma}$	0.250	0.300				
Power ( % )	90	90				
n per group	338	235				

**Equivalence limit,  $\hat{E}\hat{A}$**   
 The null hypothesis being tested is that the test and standard are not equivalent, that the test is "worse than" the standard by a specified amount or more, i.e. that  $\hat{D}\hat{S} - \hat{D}\hat{a}$  is equal to  $\hat{E}\hat{A}$  or is farther from zero in the same direction. If the null hypothesis is rejected it can be concluded that the test is "equivalent to" the standard or better.

**Acceptable entries:**  
 any value

For Help, press F1      -10.00000      AUTO RECALC OFF

For the same calculations in the worked example nQuery returns sample sizes that are the same as Table 4.1.

PASS can not do non-inferiority calculations. However, for the special case of no treatment difference one can get PASS to do the required calculations by going into the Equivalence dialogue box described in the worked examples in Section 3 and setting the Type II error to be twice what it should be - for example 0.20 for a non-inferiority Type II error of 0.10. With this trick equation (3.15) will give the same results as equation (4.8). With this trick PASS calculates the sample size to be 337.

Alternatively in PASS one could use the dialogue box for a superiority trial and a two group t-test described in Section 2. Now instead of setting the Type II error to twice what is required, one sets the Type I error to be twice what is required for the two sided alternative hypothesis

i.e. to 5% (note PASS also has the one sided alternative hypothesis too where one could use 2.5%). With this trick equation 2.7 would resemble equation 4.8. Now instead of entering non-inferiority limits one should enter distance of the true mean to the non-inferiority margin in the "Mean 2 (Mean of Group 2)" box. Thus, in the worked example for no mean difference enter 10 and for a mean difference of 2 enter 12. With the trick described PASS returns sample sizes of 337 and 234 respectively for mean differences of 0 and 2. Each sample size is 1 below the sample size given in Table 4.1 and by nQuery.

## 4.2. Cross-over trials

The equivalent sample size formula to equation (4.5) for cross-over trials is

$$n = \frac{2\sigma_w^2(Z_{1-\beta} + Z_{1-\alpha})^2}{((\mu_A - \mu_B) - d)^2}, \quad (4.11)$$

which when re-written in terms of power becomes

$$1 - \beta = \Phi\left(\sqrt{\frac{((\mu_A - \mu_B) - d)^2 n}{2\sigma_w^2}} - Z_{1-\alpha}\right). \quad (4.12)$$

The equivalent formula for unknown variance is

$$1 - \beta = \Phi\left(\sqrt{\frac{((\mu_A - \mu_B) - d)^2 n}{2\sigma_w^2}} - t_{1-\alpha, n-2}\right). \quad (4.13)$$

As with parallel group designs it preferable to calculate the power (and Type II error) under the assumption of a non-central t-distribution and thus equation 4.13 is rewritten as [37]:

$$1 - \beta = 1 - \text{Probt}(t_{1-\alpha, n-2}, n-2, \tau), \quad (4.14)$$

where  $\tau$  is defined as

$$\tau = \frac{|((\mu_A - \mu_B) - d)\sqrt{n}|}{\sqrt{2\sigma_w^2}}.$$



For quick calculations (for 90% power and Type I error of 2.5%), the following formula can be utilised:

$$n = \frac{21\sigma^2}{((\mu_A - \mu_B) - d)^2}. \quad (4.15)$$

As with parallel group estimation the quick equations give reasonable, although slightly underestimates of the sample size. Table 4.2 gives sample sizes using equation 4.14 for various standardised equivalence limits ( $\delta = d/\sigma$ ) and standardised mean differences assuming equal allocation between groups.

#### **4.2.1. Worked example**

##### **4.2.1.1. Using the sample size tables**

An investigator wishes to design a cross-over hypertension trial similar to that in Section 4.1.1 with the same clinically acceptable effect of non-inferiority of 10mmHg ( $d$ ). The true mean difference between the treatments is also thought to be zero and the expected within subject population standard deviation is 20mmHg ( $\sigma$ ). Thus, the standardised non-inferiority limits equate to  $-\delta = -d/\sigma = -10/20 = -0.50$ . For the Type I and Type II errors fixed at 2.5% and 10% respectively Table 4.2 gives a total sample size of 87 patients in the trial.

If the true mean difference is thought to be 2mmHg the sample size would be reduced to a total of 61 patients in the trial

##### **4.2.1.2. Repeated using sample size software**

To do non-inferiority sample size calculations in nQuery one would need to click on File/New, for Goal tick Means, Number of Groups tick One and Analysis Method tick

Equivalence. Then click on Paired t-test for Equivalence Means. The following dialogue box given below then comes up.

The screenshot shows the 'Paired t-test of equivalence of means' dialog box in nQuery Advisor. The table below represents the data entered in the dialog:

	1	2	3	4	5	6
Test significance level, $\alpha$ (one-sided)	0.025	0.025				
Equivalence limit difference, $\hat{E}\hat{A}$	-10.000	-10.000				
Expected difference, $\hat{E}\hat{A}$	0.000	2.000				
$\hat{E}\hat{A} - \hat{E}\hat{A}$	-10.000	-12.000				
Standard deviation of differences,	28.280	28.280				
Effect size, $\hat{E} =  \hat{E}\hat{A} - \hat{E}\hat{A}  / \hat{O}\hat{P}$	0.354	0.424				
Power ( % )	90	90				
n	86	61				

The right-hand panel of the dialog box contains the following text:

**Standard deviation of differences,  $\hat{O}\hat{P}$**   
 The standard deviation of differences is a measure of the variability of the difference between conditions within each subject.

**Suggestion:**  
 Enter a value observed in a pilot study or similar published studies.

**Acceptable entries:**  
 > 0.0

**Aid:**  
 Use the Assistants menu or the button marked [O] to estimate the standard deviation from a standard error, range, percentiles, coefficient of variation, upper confidence limit, or first and second condition standard deviations and correlation, or, use this menu to obtain the standard deviation for cluster sampling.

It is worth noting the changing of gears in nQuery here. For superiority trials the assumption was the final analysis would be with a paired t-test, thus, a standard deviation of the difference was used in the calculations and a total sample size given. For equivalence trials it was assumed that the final analysis would be an analysis of variance, thus, the within subject standard deviation (all be it divided by root two) was used in the calculations and a sample size per sequence given. Now for non-inferiority trials nQuery is back to the approach of superiority trials (despite the fact that non-inferiority is more akin to equivalence than superiority trials in concept) assuming that a paired t-test will be used in the final analysis. For the example give earlier for no mean difference (and a standard deviation of the difference

of 28.28) nQuery returns a total sample size of 86, one awry from table 4.2, but for a mean difference of 2 nQuery gives a total sample size of 61 which agrees with table 4.2.

As with parallel group trials given earlier PASS does not do non-inferiority sample size calculations for cross-over trials. However, using the trick of using the equivalence dialogue box described in the worked example in Section 4.1.1, for no treatment difference PASS gives 44 patients per sequence or 88 total.

Using the "One Sample T-Test" dialogue box for superiority trials (described in section 2.2.3) entering 12 and 10 in the "Mean 1 (Alternative)". For mean differences of 0 and 2 respectively PASS calculates the sample sizes to be 86 and 61.

## 5. "AS GOOD AS OR BETTER" TRIALS

For certain clinical trials the objective is to demonstrate either that a given treatment is clinically not inferior or that it is clinically superior when compared to the control i.e. that the treatment is "as good as or better" than the control. In non-inferiority trials two null and alternative hypotheses are investigated. First the non-inferiority null and alternative hypotheses:

$H_0$ : A given treatment is inferior with respect to the mean response.

$H_1$ : The given treatment is non-inferior with respect to the mean response.

If this null hypothesis is rejected then a second null hypothesis can be investigated:

$H_0$ : The two treatments have equal effect with respect to the mean response.

$H_1$ : The two treatments are different with respect to the mean response.

Practically these two null hypotheses are investigated through the construction of a 95% confidence interval to investigate where the lower (or upper as appropriate) bound lies. Figure 2 highlights how the two separate hypotheses for superiority and non-inferiority are investigated.

It should be noted that "As good as or better" trials are really a sub-category of either superiority or non-inferiority trials. However, in this article these trials are put into a separate section to highlight how as good as or better trials combine the null hypotheses of superiority and non-inferiority trials into one closed testing procedure whilst maintaining the overall Type I error.

To introduce the closed testing procedure this section will first describe the situation where a one-sided test of non-inferiority is followed by a one-sided test of superiority. The more general case where a one sided test of non-inferiority is followed by a two sided test of superiority is then described.

In describing as good as or better trials this paper draws heavily on the work of Morikawa and Yoshida [49]. The CPMP have recently issued a points to consider document [42].

### 5.1. A test of non-inferiority and a one sided test of superiority

The null ( $H_{1_0}$ ) and alternative ( $H_{1_1}$ ) hypotheses for a non-inferiority trial can be written as:

$$H_{1_0}: \mu_A - \mu_B \leq -d.$$

$$H_{1_1}: \mu_A - \mu_B > -d.$$

which alternatively can be written as:

$$H_{1_0}: \mu_A - \mu_B + d \leq 0.$$

$$H_{1_1}: \mu_A - \mu_B + d > 0.$$

Whilst the corresponding null ( $H_{2_0}$ ) and alternative ( $H_{2_1}$ ) hypotheses for a superiority trial can be written as:

$$H_{2_0}: \mu_A - \mu_B \leq 0.$$

$$H_{2_1}: \mu_A - \mu_B > 0.$$

What is clear from the definitions of these hypotheses is that if  $H_{2_0}$  is rejected at the  $\alpha$  level then  $H_{1_0}$  would also be rejected. Also, if  $H_{1_0}$  is not rejected at the  $\alpha$  level then  $H_{2_0}$  would also not be rejected. This is because  $\mu_A - \mu_B + d \geq \mu_A - \mu_B$ . Hence, both  $H_{1_0}$  and  $H_{2_0}$  are rejected if they are both statistically significant; neither  $H_{1_0}$  and  $H_{2_0}$  are rejected if  $H_{1_0}$  is not significant; and only  $H_{1_0}$  is rejected if only  $H_{1_0}$  is significant.

Based on these properties a closed test procedure can be applied to investigate both non-inferiority and superiority whilst maintaining the overall Type I error rate without  $\alpha$  adjustment. To do this the intersection hypothesis  $H_{2_0} \cap H_{1_0}$  is first investigated which, if rejected, is followed by a test of  $H_{1_0}$  and  $H_{2_0}$ . In this instance  $H_{2_0} \cap H_{1_0} = H_{1_0}$  and so both non-inferiority and superiority can be investigated through the following two steps [49].

1. First investigate the non-inferiority through the hypothesis  $H1_o$ . If  $H1_o$  is rejected then  $H2_o$  can be tested. If  $H1_o$  is not rejected then the investigative treatment is inferior to the control treatment.
2. If  $H2_o$  is then rejected in the next step one can conclude that the investigative treatment is superior to the control. Else if  $H2_o$  is not rejected then non-inferiority should be concluded.

## 5.2. A test of non-inferiority and a two sided test of superiority

The null ( $H3_o$ ) and alternative ( $H3_1$ ) hypotheses for a two sided test of superiority can be written as:

$$H3_o: \mu_A = \mu_B.$$

$$H3_1: \mu_A < \mu_B \text{ or } \mu_A > \mu_B.$$

These hypotheses are equivalent to two one-sided tests at the  $\alpha/2$  level of significance (summing to give an overall type I error of  $\alpha$ ) through the investigation of  $H2_o$  against the alternative of  $H2_1$  and the following null and alternative hypotheses:

$$H4_o: \mu_A \geq \mu_B.$$

$$H4_1: \mu_A < \mu_B.$$

In applying the closed test procedure in this instance it is apparent that the intersection hypothesis  $H1_o \cap H3_o$  is always rejected as it is empty and so both  $H1_o$  and  $H3_o$  can be tested. Due to there being no intersection the following steps can be applied steps [49]:

1. If the observed treatment difference is greater than zero and  $H3_o$  is rejected then  $H1_o$  is also rejected and one can conclude that the investigative treatment is statistically superior to control.

2. If the observed treatment difference is less than zero and  $H_{3_0}$  is rejected and  $H_{1_0}$  is not then the control is statistically superior to the investigative treatment. If  $H_{1_0}$  is also rejected then the investigative drug is worse than the control but is not inferior (practically though this may be difficult to claim).
3. If  $H_{3_0}$  is not rejected but  $H_{1_0}$  is, then the investigate drug is non-inferior compared to the control.
4. If neither  $H_{1_0}$  nor  $H_{3_0}$  are rejected then one must conclude that the investigative treatment is inferior to control.

Note that when investigating the  $H_{1_0}$  and  $H_{3_0}$  hypotheses,  $H_{3_0}$  will be tested at a two sided  $\alpha$  level of significance whilst  $H_{1_0}$  will be tested at a one sided  $\alpha/2$  level of significance. Thus, the overall level of significance is maintained at  $\alpha$ .

### **5.3. Worked example and other considerations**

To calculate the sample size required for an as good as or better trial one should apply the methodologies described in Sections 2 (Superiority) and 4 (Non-inferiority).

Supposed an investigator wished to design a parallel group trial to investigate a one sided test of non-inferiority and a two sided test of superiority trial. The trial will be designed about a standardised clinically meaningful difference (for superiority) and a standardised non inferiority margin of 0.25. The Type I error is fixed at 5% for the test of superiority and 2.5% for the test of non-inferiority, whilst the Type II error is fixed at 10%. From Table 2.1 for superiority one would require 338 patients in each arm. Whilst from table 4.1 for non-inferiority, assuming no treatment difference, again one would required 338 patients per arm. Note that here one is making the (probably unrealistic) assumption that the standardised non-inferiority limit and the standardised difference are the same.

On the face of it then one can switch between non-inferiority and superiority whilst maintaining the Type I error for no great cost in the sample size. However, if in the example above, there was anticipated to be a true mean difference between the two groups against the investigative therapy. If this mean difference equated to 20% of the standardised non-inferiority limit it would inflate the sample size, *mutatis mutandis*, to 527 patients per arm.

A more realistic scenario to the one described in the worked example is one where the non-inferiority margin is a fraction of the clinically meaningful difference (see Section 3.3.2). The sample size required to investigate non-inferiority would hence be a factor more than required to investigate superiority - the factor being the ratio of the clinically meaningful difference over the non-inferiority margin squared.

For as good as or better trials (given that one is also investigating superiority), it may be appropriate to power for non-inferiority (as this will usually be the large sample size estimate) but assuming a small difference between the two groups in favour of the investigative therapy (see section 4 and tables 4.1 and 4.2)

A further consideration in as good as better trials is the choice of data set to have as primary - the intent to treat (ITT) or per protocol (PP) data set.

The intent to treat population is the patient population evaluated on the basis of the treatment regimen patients were planned to receive as opposed to the actual treatment given. As a consequence "subjects allocated to a treatment group" are "followed up, assessed and analysed as members of that group irrespective of their compliance to the planned course of treatment" [3].

The per protocol patient population is the "subset of subjects who complied with the protocol sufficiently to ensure that these data would exhibit the effects of treatment according to the underlying scientific model" (ICH, E9).



For a superiority trial the primary data set would be that based on ITT data set; for a non-inferiority trial the primary data set would be both the PP and the ITT data set [42].

## 6. ASSESSMENT OF BIOEQUIVALENCE

Earlier in the paper calculations were given where the objective of the trial was to demonstrate that the two therapies are clinically equivalent. In an equivalence trial the comparators may be completely different, in terms of route of administration or even actual drug therapies, but the objective is to investigate whether they are clinically the same. However, in bioequivalence trials the comparators are ostensibly the same - the manufacturing site may have moved or a formulation altered slightly for marketing purposes. Bioequivalence studies are therefore conducted to demonstrate that two formulations of a drug have similar bioavailability i.e. does the same amount of drug get into the body for each formulation. The assumption in bioequivalence trials is that if the two formulations have equivalent bioavailability then one can infer that they have equivalent effect for both efficacy and safety. The pharmacokinetic bioavailability is therefore a surrogate for the clinical endpoints.

Equivalent bioavailability will be concluded if the drug concentration by time profiles for the test and reference formulations are super-imposable, see Figure 4 for an example. Through determining that the two profiles are super-imposable one can conclude that the two formulations are clinically the same.

In bioequivalence studies, therefore, one can determine *in vivo* whether the two formulations are bioequivalent by assessing the concentration time profiles for the test and reference formulations [50]. This is usually done by assessing if the rate and extent of absorption are the same, where the pharmacokinetic parameter AUC (area under the concentration curve) is used to assess the extent of absorption and the parameter C<sub>max</sub> (maximum concentration) is used to assess the rate of absorption. Figure 4 gives a pictorial representation of these parameters. If the two formulations are bioequivalent then they can be switched without reference to further clinical investigation and can be considered inter-changeable.

The null and alternative hypotheses are similar to those for equivalence studies:

H<sub>0</sub>: The test and reference formulations give different drug exposures ( $\mu_T \neq \mu_R$ ).

H<sub>1</sub>: The test and reference formulations give equivalent drug exposure ( $\mu_T = \mu_R$ ).

Similarly to other types of trials the objective of a bioequivalence study is to test the null hypothesis to see if the alternative is true. The 'standard' bioequivalence criteria is to demonstrate that average drug exposure on the test is within 20% of the reference on the log scale [51, 52, 53]. Thus, the null and alternative hypotheses can be rewritten as:

H<sub>0</sub>:  $\mu_T/\mu_R \leq 0.80$  or  $\mu_T/\mu_R \geq 1.25$ .

H<sub>1</sub>:  $0.80 < \mu_T/\mu_R < 1.25$ .

Two comparator formulations can thus be declared bioequivalent if it can be demonstrated that the mean ratio is wholly contained within 0.80 to 1.25. To test the null hypothesis two one-sided tests at the 5% level are constructed to determine whether  $\mu_T/\mu_R \leq 0.80$  or  $\mu_T/\mu_R \geq 1.25$ . If neither of these tests hold then the alternative hypothesis can be accepted of  $0.80 < \mu_T/\mu_R < 1.25$ . As one is performing two simultaneous tests on the null hypothesis, both of which must be rejected to accept the alternative hypothesis, the type I error is maintained at 5%. The convention is to represent the two one-sided tests as a 90% confidence interval around the mean ratio of  $\mu_T/\mu_R$  which neatly summarises the results of two one-tailed tests. Figure 5 highlights how average bioequivalence between two formulations can be demonstrated through 90% confidence intervals.

A test formulation of a drug can therefore said to be bioequivalent to its reference formulation if the 90% confidence interval for the ratio test:reference is wholly contained within the range 0.80 to 1.25, for both AUC and Cmax. As both AUC and Cmax must be equivalent to declare bioequivalence there is no need to allow for multiple comparisons.

For certain indications other parameters, such as C<sub>min</sub> (defined as the minimum concentration over a given period) or T<sub>mic</sub> (defined as time above a minimum inhibitory concentration over a given period), may also need to be simultaneously assessed.

Note that the criteria for acceptance of bioequivalence may vary depending on factors such as which regulatory authority's guidelines are being followed and the therapeutic window of the compound being formulated and so the 'standard' criteria may not always be appropriate.

The methodology described in this section can also be applied to other types of *in vivo* assessment such as the assessment of a food [54], drug interactions [55, 56] or special populations [57, 58]. The criteria for acceptance for other types of *in vivo* assessment may vary depending on either the guidelines [54] or an *a priori* clinical assessment [55, 56].

It may be worth noting the statistical difference between testing for equivalence and bioequivalence with reference to investigating the null hypothesis. In equivalence trials the convention is to undertake two one-sided tests at the 2.5% level which in turn are represented by a 95% confidence interval; in a bioequivalence trial two one-sided tests at the 5% level are undertaken, which are represented by a 90% confidence interval. Thus, in bioequivalence trials the overall type I error is maintained at 5% twice that of equivalence trials where the overall type I error is maintained at 2.5%.

As bioequivalence studies are usually designed as crossover studies. This section thus concentrates on this design first.

### **6.1. Justification for log transformation**

The concentration-time profile for a one compartment intravenous dose can be represented by the following equation:

$$c(t) = Ae^{(-\lambda t)} ,$$

where  $t$  is time,  $A$  is the concentration at  $t=0$  and  $\lambda$  is the elimination rate constant [59]. It is evident from this equation that a drug concentration in the body falls exponentially at a constant rate  $\lambda$ .

A test and reference formulation would be super-imposable when  $c_T(t) = c_R(t)$ . On the log scale this is equivalent to  $\log(A_T) - \lambda_T = \log(A_R) - \lambda_R$ , which for  $\lambda_T = \lambda_R$  (*a priori* one would expect this to be true) becomes  $\log(A_T) = \log(A_R)$ . Thus, on the log scale the difference between two curves can be summarised additively. Indeed it is upon this scale that such pharmacokinetic parameters as the rate constant,  $\lambda$ , and the pharmacokinetic half life,  $t_{1/2}$  ( $t_{1/2} = \log 2 / \lambda$ ), are derived [59]. This rationale also follows for the summary statistics used to measure exposure (AUC) and absorption (Cmax) as well as the variance estimates [59, 60]. From these arguments the standard assumption for pharmacokinetic data is that they follow a log Normal distribution with the default being to analyse  $\log_e(\text{AUC})$  and  $\log_e(\text{Cmax})$ . Any differences on the  $\log_e$  scale (test-reference) are then back-transformed to obtain a ratio on the original scale. It is this back transformed ratio and its corresponding 90% confidence interval that are used to assess bioequivalence.

## **6.2. Rational for using coefficients of variation**

As discussed in the previous sub-section all statistical inference for bioequivalence trials are undertaken on the log scale and back transformed to the original scale for interpretation. Thus, the within-subject estimate of variability on the log scale is used both for inference and sample size estimation. However, for the interpretation of the mean effect on the original scale it is optimal to have a measure of variability also on the original scale. A measure of variability that could be used is the Coefficient of Variability (CV) as this parameter is not scale dependent. Now, for log-Normally distributed data the following exact relationship

between the CV on the arithmetic scale and the standard deviation,  $\sigma$ , on the log scale holds [48, 59]:

$$CV = \sqrt{(e^{\sigma^2} - 1)} ,$$

For small estimates of  $\sigma^2$  [ $\sigma < 0.30$ ] the CV can be approximated by:

$$CV \approx \sigma .$$

Thus, both the measure of effect and its variability can both be interpreted on the original scale.

The derivation of this result is based on the following relationships for the log-Normal distribution [59]:

$$m = e^{\left(\frac{\mu + \sigma^2}{2}\right)} ,$$

$$s^2 = \left(e^{(2\mu + \sigma^2)}\right)(e^{\sigma^2} - 1) ,$$

where  $\mu$  and  $\sigma^2$  respectively relate to the mean and variance on the log-transformed scale and  $m$  and  $s$  the corresponding mean and variance on the non-transformed scales. Hence:

$$CV = \frac{s}{m} = \frac{\sqrt{\left(e^{(2\mu + \sigma^2)}\right)(e^{\sigma^2} - 1)}}{e^{\left(\frac{\mu + \sigma^2}{2}\right)}} = \sqrt{(e^{\sigma^2} - 1)} .$$

### 6.3 Cross-over trials

Calculations for two expected mean responses are described for the special case where the true mean is expected to be unity and the general case when the true mean ratio is not fixed to be unity ( $\mu_T/\mu_R = 1$ ). Although the calculations for the special case are more straight forward, it is recommended that even if one expects *a priori* the mean ratio to be unity, where practical, one should consider calculating sample sizes under the assumption of a small mean difference (of 5% say) as the power of a study is very sensitive to the assumption about the mean ratio (as a mean ratio of  $\mu_T/\mu_R \neq 1$  is closer to one of the boundaries, 0.80 to 1.25 say, and so it is more difficult to demonstrate bioequivalence).

#### 6.3.1 General case

The derivation of the sample size is similar to that for equivalence trials i.e. for the general case where the expected true mean ratio is not expected to be unity the sample size cannot be directly derived. Instead one has to iterate until a sample size is reached which gives the required Type II error (and power). Thus, to calculate the power for the bioequivalence acceptance limits of (0.80, 1.25), the following formula can be used:

$$1 - \beta = \Phi \left( \sqrt{\frac{(\log(\mu_T/\mu_R) - \log(1.25))^2 n}{2\sigma_w^2}} - Z_{1-\alpha} \right) + \Phi \left( \sqrt{\frac{(\log(\mu_T/\mu_R) - \log(0.80))^2 n}{2\sigma_w^2}} - Z_{1-\alpha} \right) - 1, \quad (6.1)$$

where  $\sigma_w$  is the within-subject variability on the log scale and n is the total sample size. For unknown variance equation (6.1) can be re-written as

$$1 - \beta = \Phi \left( \sqrt{\frac{(\log(\mu_T/\mu_R) - \log(1.25))^2 n}{2\sigma_w^2}} - t_{1-\alpha, n-2} \right) + \Phi \left( \sqrt{\frac{(\log(\mu_T/\mu_R) - \log(0.80))^2 n}{2\sigma_w^2}} - t_{1-\alpha, n-2} \right) - 1. \quad (6.2)$$

As with superiority, equivalence and non-inferiority trials, when the population variance is not being used it is best to calculate the power using a non-central t-distribution, as outlined by Owen [47, 48]:

$$1 - \beta = \text{Probt}(-t_{1-\alpha, n-2}, n-2, \tau_2) - \text{Probt}(t_{1-\alpha, n-2}, n-2, \tau_1), \quad (6.3)$$

where  $\tau_1$  and  $\tau_2$  are non centrality parameters:

$$\tau_1 = \frac{\sqrt{n}(\log(\mu_T/\mu_R) - \log(0.80))}{\sqrt{2\sigma_w^2}} \quad \text{and} \quad \tau_2 = \frac{\sqrt{n}(\log(\mu_T/\mu_R) - \log(1.25))}{\sqrt{2\sigma_w^2}}.$$

An estimate of the sample size for  $\mu_T/\mu_R$  greater than unity can be obtained from the following equation:

$$n = \frac{2\sigma_w^2(Z_{1-\beta} + Z_{1-\alpha})^2}{(\log(\mu_T/\mu_R) - \log(1.25))^2}, \quad (6.4)$$

which can be used to provide an initial value for the iterations. This equation provides reasonable approximations for  $\mu_T/\mu_R \neq 1$ , especially when the mean ratio becomes large relative to (0.80 to 1.25). This is because in such circumstances most of the Type II error comes from one of the two one sided tests. For quick calculations (for 90% power and a Type I error of 5%), the following formula can be used:

$$n = \frac{17\sigma_w^2}{(\log(\mu_T/\mu_R) - \log(1.25))^2}. \quad (6.5)$$

Obviously for true ratios less than unity  $\log(1.25)$  should be replaced by  $\log(0.80)$ .

### 6.3.2. Special case of the ratio equalling unity

For the special case where the expected true mean difference is expected to be unity the sample size can be directly derived from the following formula.

$$n = \frac{2\sigma_w^2 \left( Z_{1-\frac{\beta}{2}} + Z_{1-\alpha} \right)^2}{(\log(1.25))^2}. \quad (6.6)$$

For unknown variance equation 6.6 can be re-written to give the sample size:



$$n = \frac{2\sigma_w^2 \left( Z_{\frac{1-\beta}{2}} + t_{1-\alpha, n-2} \right)^2}{(\log(1.25))^2}. \quad (6.7)$$

In turn this can be re-written as:

$$1 - \beta = 2\Phi \left( \sqrt{\frac{(\log(1.25))^2 n}{2\sigma_w^2}} - t_{1-\alpha, n-2} \right) - 1.$$

Estimating the power from a non-central t-distribution, equation 6.3 can be re-written to

$$1 - \beta = 2\text{Probt}(-t_{1-\alpha, n-2}, n - 2, \tau) - 1, \quad (6.8)$$

where  $\tau$  is the non centrality parameters defined as

$$\tau = \frac{\sqrt{n}(\log(0.8))}{\sqrt{2\sigma_w^2}}.$$

Equation 6.6 can be used to obtain initial estimates of the sample size to use in Equation 6.8.

For quick calculations for 90% power, 5% Type I error rate and a 20% acceptance criteria on the log scale one could use:

$$n = 433\sigma_w^2. \quad (6.9)$$

This formula gives sample size estimated within 1 or 2 of equation 6.8. Table 6.1 gives sample size estimates using equation 6.3 for different CVs, mean ratios and acceptance criteria 10% (0.90 to 1.11), 15% (0.85 to 1.18), 20% (0.80 to 1.25) etc for a Type I error rate of 5% and 90% power.

### 6.3.3. Replicate Designs

For compounds with high variability the standard AB/BA can require relatively large sample sizes, especially if the mean ratio is not expected to be unity. Designs which can partially overcome this problem are replicate cross-over designs. Through adding an extra period arm to the study, such that the sequences are say ABB/BAA, the sample size is reduced by 25%

compared to a standard AB/BA design. An additional two periods and sequences say of ABBA/BAAB, can reduce the sample size by 50% [61]. The reasons for these reductions are due to the fact that the variances used in the contrast of the means in the final analysis are halved for a 4 period replicate design and reduced by 25% for a 3 period replicate design.

Note replicate designs may not be practical for certain compounds, for example those with a long half life, but it is a possible solution for compounds with high pharmacokinetic variability.

Another type of replicate design is a two period replicate design AA/AB/BA/BB. This design allows for an intra-subject estimate of variability for a given compound without increasing the number of periods beyond two. To consider the effect such a design has on the sample size one must consider the derivation of the total variance

$$\sigma^2 = \sigma_b^2 + \sigma_w^2,$$

where  $\sigma_w^2$  is the within-subject component of variation and  $\sigma_b^2$  is the between subject component of variation. Both these variance components can be estimated from previous cross-over trials with the test and reference compounds. See section 8 for a detailed description. Now suppose

$$\sigma_b^2 = k\sigma_w^2$$

it can be shown, assuming an equal allocation to each sequence, that the sample size required for a two period replicate design can be derived by multiplying the sample size for standard AB/BA design as follows:

$$n_{AA/AB/BA/BB} = \left( \frac{2k+1}{k+1} \right) * n_{AB/BA}.$$

Where  $n_{AB/BA}$  is the sample size derived from (6.3) The derivation number of this formula comes initially from imagining that the AB/BA and AA/BB sequences are from a cross-over trial and a parallel group trial respectively with n/4 subjects assigned to each sequence. For

each sequences the following total variance can thus be derived for the "parallel group" sequences:

$$\frac{4\sigma^2}{n} + \frac{4\sigma^2}{n}$$

If these sequences were from a parallel group study one would effectively take the average of the two sessions to compared A and B and so from equation 8.2.4 given later in the chapter:

$$\sigma^2 = \sigma_b^2 + \sigma_w^2/2$$

and with  $\sigma_b^2 = k\sigma_w^2$  this the variance becomes:

$$4(2k + 1)\sigma_w^2$$

which equals  $w_1$  say.

Now for the "cross-over" AB/BA sequences, the total variance can be derived as:

$$\frac{4\sigma_w^2}{n}$$

which equals  $w_2$  say. Now to combine the cross-over and parallel sequences into one overall variance one could use the following formula borrowed from meta-analysis methodology [62]:

$$\left( \sum_{i=1}^2 \frac{1}{w_i} \right)^{-1}$$

Thus, the overall variance is:

$$\left( \frac{n}{4\sigma_w^2} + \frac{n}{4\sigma_w^2(2k + 1)} \right)^{-1} = \frac{2\sigma_w^2(2k + 1)}{n(k + 1)}$$

From any of the sample size formulae given in this paper it is evident that one increases the sample size in direct proportion to any increase in the variance. If one is planning a simple AB/BA cross-over trial the overall variance would be  $2\sigma_w^2/n$ . Thus, the ratio of the variances is thus:

$$\frac{2\sigma_w^2(2k+1)}{n(k+1)} \cdot \frac{n}{2\sigma_w^2} = \frac{(2k+1)}{(k+1)}$$

and so the increase in sample sizes for doing a replicate cross-over is

$$n_{AA/AB/BA/BB} = \left(\frac{2k+1}{k+1}\right) * n_{AB/BA}$$

and the equation given earlier.

To verify this result 10,000 simulations for a fixed sample size of 48 and for various k were undertaken. Each simulation simulated AB/BA and AB/BA/AA/BB cross-over. The analysis for each simulation was done with all subjects entered into PROC MIXED with subject entered as random. The table below gives the results:

k	$\frac{2k+1}{k+1}$	Simulation
2	1.67	1.65
4	1.80	1.78
6	1.86	1.85
8	1.89	1.88
10	1.91	1.90

What is evident both from the table above and the equation is that a two period replicate design will always require more subjects than a standard AB/BA requiring the same sample size only for k=0. However, no matter how larger k becomes it will only require twice as many subjects at most. This is because as k becomes large virtually all the information, in the comparison of the mean ratio, comes from the AB/BA sequences and with twice as many subjects there will be as many people in these sequences as in a standard AB/BA design.

### **6.3.4. Use of quick formulae to estimate the sample size of a bioequivalence trial**

In this section on bioequivalence trials, as with the other sections, equations have been given which allow for quick calculation of the sample size. In absolute terms these formulae throughout the paper have been consistent – being at most just one or two off the sample sizes provided in the tables. For large trials such small differences may be considered minor – moving the sample size from 526 to 527 in the worked example in Section 2.1.1. However, for bioequivalence trials the absolute difference of one or two may equate to a large relative difference as the sample sizes are far smaller. For example for a CV of 15% (cross-over trial design, mean ratio assumed to be unity) table 6.1 gives the sample size as 12 whilst equation 6.3 returns a sample of 10. The sample size of 10 subjects equates, from equation 6.8, to a Type II error of 17%.

For bioequivalence trials it is therefore strongly recommended that the sample size tables only be used for final sample size estimation with the quick results only used for early ballpark calculations.

### **6.3.5. Worked Example**

#### **6.3.5.1. Using the sample size tables**

A bioequivalence trial to compare a test with reference formulation needs to be designed. The standard bioequivalence criteria (i.e. 0.80 to 1.25) will be used to demonstrate that the average drug exposure on the test is bioequivalent to the reference i.e. 0.80 to 1.25. The within-subject coefficient of variation is expected to be 25% (=CV) and the mean ratio is expected to be unity ( $\mu_T/\mu_R = 1$ ). The CV 25% equates to a within subject SD of 0.2462. The study design will be an AB/BA two period crossover. From Table 6.1 it can be seen that one would need at a minimum a total sample size of 28 subjects. Practically this would equate to at least 14 subjects on each sequence (AB and BA).

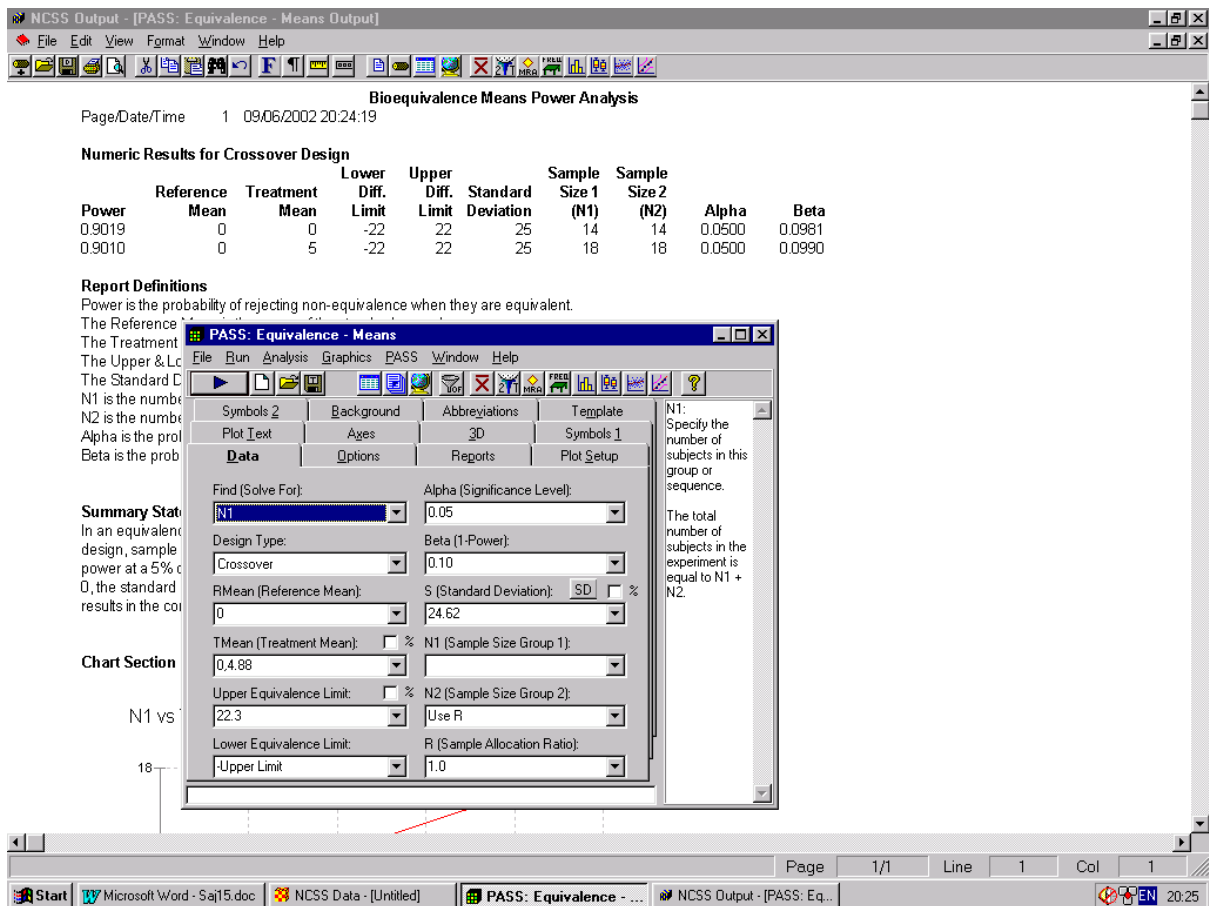
If the test formulation is expected, on average, to have exposures 5% greater than the reference ( $\mu_T/\mu_R = 1.05$ ) then the total sample size would increase to 36 subjects (or 18 per sequence).

Suppose though instead of an AB/BA design a replicate ABB/BAA or ABBA/BAAB designs was being considered for the case where exposures were expected to be 5% greater on test compared to reference. If one adopted a 4 period replicate design then one would multiply the total sample size calculated earlier by 0.50 to get  $36*0.5=18$  subjects in total required. If one adopted a 3 period replicate design then the total sample size calculated earlier should be multiplied by 0.75 to get  $36*0.75=27$  subjects in total required.

#### **6.3.5.2. Repeated using sample size software**

To repeat the calculations in PASS one needs to select Means and then Equivalence-Means. The dialogue box below details the entries required to repeat both calculations in the worked example.

Note that as discussed in section 3 on equivalence trials PASS does not calculate the total sample size but the sample size per sequence – assuming the design is an AB/BA cross-over trial



This dialogue box highlights one feature of PASS in that it only works on the arithmetic scale. Thus, the bioequivalence limit of 0.80 to 1.25 must be logged and a symmetric limit of -22.3 to +22.3 should be used in PASS calculations. Likewise instead of using a mean ratio of  $\mu_T/\mu_R = 1.05$  one should use  $\mu_T - \mu_R = 4.88$  instead.

Another feature to highlight from the output box is that, with the exception of the Type I error, Type II error and power, all the default outputs are reported with no decimal places. Thus, in the worked example 4.88 is reported as 5 and 22.3 as 22. This could cause problems if someone wished to replicate the results and only had the PASS dialogue box to work from. To amend the default one should click on "Reports" in the PASS output box and then amend the number of decimal places required as appropriate.

Repeating the calculations in PASS gives the same sample sizes as table 6.1 for mean ratios of 1.00 and 1.05.

To repeat the calculations in nQuery one would need to click on File/New, for Goal tick Means, Number of Groups tick Two and Analysis Method tick Equivalence. Then select TOST for ratio of means (log scale) for two group or cross-over. As may be evident from the dialogue box title selected nQuery allows one to do the calculations on the log scale (more detail as to what to enter in the dialogue box will be given in the worked example for parallel group trials in Section 6.4.3). As calculations are done the log scale nQuery also uses the Coefficient of Variation (CV) for the calculations, although as described for cross-over equivalence trials in Section 3.3.3 nQuery does not use  $\sigma_w$  to derive the CV for sample size calculations but  $\sigma_w/\sqrt{2}$ . nQuery also agrees with table 6.1 for the sample size calculations.

As noted in the worked example in section 3.3.3 both PASS and nQuery only give sample sizes per sequence assuming an AB/BA cross-over trial. It is in bioequivalence trials where this may be an issue as replicate designs with more than two sequences are not uncommon. Even for the worked example it is a little awkward. For a true ratio of 1.05 both PASS and nQuery returned a sample size of 18 per sequence. Supposing an ABB/BAA design is to be planned one would have to double 18 per sequence to get the total sample size and then multiply this by 0.75 to get the total sample size for this design. The sample size per sequence would then be 13.5 (14 rounded up).



## 6.4. Parallel Group Studies

Although cross-over trials are the 'norm' for the assessment of bioequivalence sometimes, particularly with very long half life compounds, these designs are not practical. This section briefly describes the methodology for sample size calculation for parallel group bioequivalence trials.

### 6.4.1 General case

The power for a bioequivalence trial with acceptance limits of (0.8, 1.25) is given by

$$1 - \beta = \Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R) - \log(1.25))^2 m_T}{(r+1)\sigma^2}} - Z_{1-\alpha}\right) + \Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R) - \log(0.80))^2 m_T}{(r+1)\sigma^2}} - Z_{1-\alpha}\right) - 1, \quad (6.10)$$

where  $\sigma$  is the between-subject variability on the log scale,  $r$  is the allocation ratio and  $n_T$  is the sample size in the test group. For unknown variance equation (6.10) can be re-written as

$$1 - \beta = \Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R) - \log(1.25))^2 m_T}{(r+1)\sigma^2}} - t_{1-\alpha, n_T(r+1)-2}\right) + \Phi\left(\sqrt{\frac{(\log(\mu_T/\mu_R) - \log(0.80))^2 m_T}{(r+1)\sigma^2}} - t_{1-\alpha, n_T(r+1)-2}\right) - 1. \quad (6.11)$$

and under the assumption of a non-central t-distribution the power is estimated from:

$$1 - \beta = \text{Probt}(-t_{1-\alpha, n_T(r+1)-2}, n_T(r+1) - 2, \tau_2) - \text{Probt}(t_{1-\alpha, n_T(r+1)-2}, n_T(r+1) - 2, \tau_1), \quad (6.12)$$

where  $\tau_1$  and  $\tau_2$  are non centrality parameters

$$\tau_1 = \frac{\sqrt{rn_T}(\log(\mu_T/\mu_R) - \log(0.80))}{\sqrt{(r+1)\sigma^2}} \quad \text{and} \quad \tau_2 = \frac{\sqrt{rn_T}(\log(\mu_T/\mu_R) - \log(1.25))}{\sqrt{(r+1)\sigma^2}}.$$

As with a cross-over trial a direct estimate of the sample size for a mean ratio greater than unity can be obtained from the following equation:

$$n_T = \frac{(r+1)\sigma^2(Z_{1-\beta} + Z_{1-\alpha})^2}{r(\log(\mu_T/\mu_R) - \log(1.25))^2}, \quad (6.13)$$

and for quick calculations (for 90% power and Type I error rate of 5%) :

$$n_T = \frac{17(r+1)\sigma^2}{r(\log(\mu_T/\mu_R) - \log(1.25))^2}. \quad (6.14)$$

If the mean ratio is expected to be less than unity then replace  $\log(1.25)$  with  $\log(0.80)$  in equations 6.13 and 6.14.

### 6.3.2. Special case of the ratio equalling unity

When the mean ratio is expected to be unity the sample size can be directly derived from:

$$n_T = \frac{(r+1)\sigma^2 \left( Z_{\frac{1-\beta}{2}} + Z_{1-\alpha} \right)^2}{r(\log(1.25))^2}. \quad (6.15)$$

For unknown variance equation 6.15 can be re-written as

$$n_T = \frac{(r+1)\sigma^2 \left( Z_{\frac{1-\beta}{2}} + t_{1-\alpha, n_T(r+1)-2} \right)^2}{r(\log(1.25))^2}. \quad (6.16)$$

Equation 6.16 can in turn can be re-written as

$$1 - \beta = 2\Phi \left( \sqrt{\frac{(\log(1.25))^2 n_T}{(r+1)\sigma^2}} - t_{1-\alpha, n_T(r+1)-2} \right) - 1,$$

and under the assumption of a non-central t-distribution the power can be derived from

$$1 - \beta = 2\text{Probt} \left( -t_{1-\alpha, n_T(r+1)-2}, n_T(r+1) - 2, \tau \right) - 1, \quad (6.17)$$

where  $\tau$  is the non centrality parameters defined as

$$\tau = \frac{\sqrt{n_T} r(\log(0.80))}{\sqrt{(r+1)\sigma^2}}.$$

Equation 6.15 can be used for initial estimates of the sample size to use in 6.17. For quick calculations of the sample size for 90% power, 5% Type I error rate and an acceptance criteria on the log scale of 20% one could use

$$10.75(k+1)\sigma^2/r. \quad (6.18)$$

Table 6.2 gives sample size estimates using equation 6.12 for different CVs, mean ratios and acceptance criteria 10% (0.90 to 1.11), 15% (0.85 to 1.18), 20% (0.80 to 1.25) etc for a Type I error rate of 5%, 90% power and an allocation ratio of one. As with cross-over trials the simpler equations provide good estimates for initial calculations.

### **6.4.3. Worked Example**

#### **6.4.3.1. Using the sample size tables**

A parallel group bioequivalence trial to compare a test with reference formulation needs to be designed using the standard bioequivalence criteria (i.e. 0.80 to 1.25). The between-subject standard deviation is expected to be 0.70 ( $=\sigma$ ) and the mean ratio is expected to be unity ( $\mu_T/\mu_R = 1$ ). This standard deviation of 0.70 equates to a between-subject CV of 80%. From Table 6.2 a minimum sample size of 216 subjects would be required in each arm. If the test formulation is expected, on average, to have exposures 5% higher than the reference ( $\mu_T/\mu_R = 1.05$ ) then the total sample size would be 282 subjects in each arm.

#### **6.4.3.2. Repeated using sample size software**

To repeat the calculations in nQuery click on File/New, for Goal tick Means, Number of Groups tick Two and Analysis Method tick Equivalence. Then select TOST for ratio of means (log scale) for two group or cross-over. The dialogue box below gives the entries required to repeat the calculations in nQuery. For the worked example given earlier nQuery concurs, giving sample sizes of 216 and 282 for mean ratios of 1.00 and 1.05 respectively.

The screenshot shows the nQuery Advisor interface. The main window displays a table titled "Two-group or crossover t-tests (TOST) for ratio of means [using log scale] [equal n's]". The table has columns for 1, 2, 3, 4, and 5 groups. The rows include Test significance levels, Lower and Upper equivalence limits, Expected ratio, Coefficient of variation, Power (%), and n per group. The Coefficient of variation is set to 0.800, and the power is 90%. The n per group is 216 for group 1 and 282 for group 2.

	1	2	3	4	5
<b>Test significance levels, <math>\alpha</math> (one-sided)</b>	0.050	0.050			
<b>Lower equivalence limit for <math>\Delta_B / \Delta_A</math>,</b>	0.800	0.800			
<b>Upper equivalence limit for <math>\Delta_B / \Delta_A</math>,</b>	1.250	1.250			
<b>Expected ratio, <math>\Delta_B / \Delta_A</math></b>	1.000	1.050			
<b>Coefficient of variation, <math>\hat{\sigma}_A / \hat{\sigma}_A</math></b>	0.800	0.800			
<b>Power ( % )</b>	90	90			
<b>n per group</b>	216	282			

The help panel on the right provides information about the Coefficient of variation,  $\hat{\sigma}_A / \hat{\sigma}_A$ . It states: "The coefficient of variation for the standard is its standard deviation divided by its mean. Sigma,  $\hat{\sigma}$ , is assumed to be the same in each of the groups." It also includes a suggestion: "When the standard deviation on the log scale is known use MTE1 Two-group or crossover t-tests (TOST) for equivalence in means. When it is not known we use the fact that when data are log-normally distributed, the variance in the log scale equals the log of (1+CV<sup>2</sup>). When the study design is a crossover design, compute CV using the standard deviation of differences divided by 2,  $\hat{\sigma}_D/2$ , or  $1/\sqrt{2}$  times the squareroot of the mean squared error from the crossover ANOVA,  $\hat{\sigma}_A/\sqrt{2}$ ." Acceptable entries are noted as  $> 0.0$ .

To repeat the calculations in PASS one needs to select Means and then Equivalence-Means.

For equivalent entries PASS concurs with nQuery and table 6.2.

## 6.5. Individual and Population Bio-equivalence

The assessment of bioequivalence as defined in this paper is based on average bioequivalence in which only the formulation means are required to be equivalent to declare bioequivalence. New paradigms for bioequivalence based on population and individual bioequivalence have also been proposed [63, 64] for which there are regulatory guidelines [52]. These alternative approaches also involve variabilities of the formulations as well as their means in the assessment bioequivalence. This paper will not go into any detail on these topics.

## 7. ESTIMATION TO A GIVEN PRECISION

In the previous sections of the paper calculations were described for the definitive investigation of specific objectives. However, there are cases when a preliminary, or pilot, investigation is conducted to estimate possible effects with view to doing a later definitive study [65, 66]. By definition such studies are held early in the drug development (or clinical investigation) paradigm. With estimation studies rather than formally testing null hypotheses it is more informative to give confidence intervals to estimate the unknown  $f(\mu)$ .

Recall that  $(1 - \alpha)$  100% confidence interval for  $f(\mu)$  has half-width

$$w = Z_{\alpha/2} \sqrt{\text{Var}(S)}. \quad (7.1)$$

Hence, if one is able to specify a requirement for  $w$  and write  $\text{Var}(S)$  in terms of 'n' then the above expression can be solved for  $n$  as before.

It should be noted though that if the sample size is based on precision calculations, then the protocol should clearly state this as the basis for the size of the study.

Precision calculations may also be undertaken when the sample size is determined primarily by practical considerations. In such cases one may quote the precision of the estimates obtained based on the half-width of the confidence interval, and provide this information in the discussion of the fixed sample size. Again it must be clearly stated in the protocol that the size of the study was determined based on practical, and not formal, considerations.

The estimation approach is also useful where one wishes to estimate possible effect across several doses. The overall context of such a study would be to assist in the selection of a dose to carry forward into a later study. CPMP [15] in their multiplicity guidelines says of such studies:

"Sometimes a study is not powered sufficiently for the aim to identify and recommend a single effective and safe dose (or dose range) but is successful only at demonstrating an overall positive correlation of the clinical effect with increasing dose. This is already a

valuable achievement.... Estimates and confidence intervals from pairwise comparisons of single doses are then used in an exploratory manner for planning of future studies".

In the context then of an overall clinical development (or investigation) an estimation study (or studies) could provide important cumulative evidence of the pharmacological benefit of a given drug asset. These studies can not prove a given affect but can valuably inform studies which can.

### 7.1 Parallel Group Trials

Defining  $\text{Var}(S)$  as per equation (2.3) one can solve equation (7.1) to give [36, 67]

$$n_A = \frac{(r+1)Z_{1-\alpha/2}^2 \sigma^2}{rW^2}, \quad (7.2)$$

where  $n_A = rn_B$  and  $n_A$ ,  $n_B$  and  $r$  are defined as per Section 2. For unknown variance equation (7.2) can be rewritten as

$$n_A \geq \frac{(r+1)t_{1-\alpha/2, n_A(r+1)-2}^2 \sigma^2}{rW^2}. \quad (7.3)$$

Equation 7.3 can be solved iteratively to find a value of  $n_A$  where the left hand side of the equation is greater than the right. An alternative equation to solve for  $n_A$  would be

$$\Phi\left(\sqrt{\frac{rn_A W^2}{(r+1)\sigma^2}} - t_{1-\alpha/2, n_A(r+1)-2}\right) \geq 0.5. \quad (7.4)$$

Equation 7.4 holds as:

$$\sqrt{\frac{rn_A W^2}{(r+1)\sigma^2}} - t_{1-\alpha/2, n_A(r+1)-2} \geq Z_{0.5} = 0,$$

and hence equation 7.3. Equation 7.4 is in fact the same as equation 2.6, given in the section in superiority trials, but with the Type II error set at 0.5 (although obviously as precision trials are not powered they can not have any Type II error). The practical application of this result is given later in the worked example using PASS to calculate sample sizes.

To allow for the Normal approximation equation 7.2 can have a correction factor added to assist in initial calculations [38]:

$$n_A = \frac{(r+1)\sigma^2 Z_{1-\frac{\alpha}{2}}^2}{rw^2} + \frac{Z_{1-\frac{\alpha}{2}}^2}{4}. \quad (7.5)$$

The following quick formula can be used (assuming one wishes to have a 95% confidence interval for the precision estimates):

$$n_A = \frac{4\sigma^2 (r+1)}{w^2 r}, \quad (7.6)$$

or for r=1:

$$n_A = \frac{8\sigma^2}{w^2}.$$

Table 7.1 gives sample sizes using equation 7.3 for various standardised widths ( $\delta = d / \sigma$ ).

The simpler equations slightly (by one or two) underestimate the sample size.

### 7.1.1 Worked Example

#### 7.1.1.1. Using the sample size tables

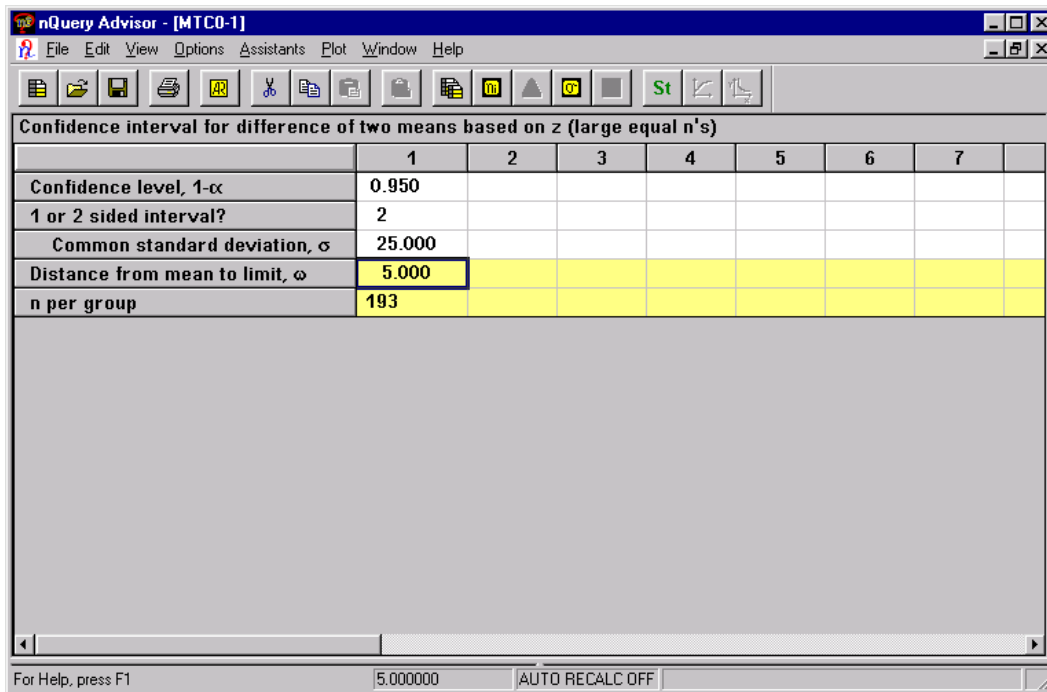
An investigator wishes to design a pilot safety cardiovascular trial with equal allocation between groups where the objective is to estimate any possible effect on QTc of new treatment compared to control with precision around the point estimate of  $\pm 5$  seconds ( $w$ ).

The expected standard deviation in the population in which the trial is to be undertaken is 25 seconds ( $\sigma$ ). Thus, the standardised width equates to  $\delta = d / \sigma = 5 / 25 = 0.20$ . Table 7.1 gives a sample size of 194 patients in each arm of the trial.

If an unequal allocation of 2:1 ( $r=2$ ) was necessary in favour of the investigative drug then one would required 145 subjects on the control arm and 290 in the investigation arm.

### 7.1.1.2. Repeated using sample size software

To do the same calculations in nQuery click on File/New for Goal tick Means, Number of Groups tick Two and Confidence Interval. Then select Confidence interval for difference of two means. The dialogue box below gives the entries required in nQuery.



nQuery gives a sample size of 193 which is one off table 7.1. This difference could be due to rounding error or, as it seems from the nQuery manual [9], nQuery uses equation (7.5) but without the correct factor from Guenther [38].

PASS does not do these calculations directly. A trick is to select Means and then T-Test: 2 Groups (see section 2 and Superiority Trials for a description of the dialogue boxes of this form). In the dialogue box enter 5 as the mean difference and 50% as type II error. These entries would get PASS to use a formula equivalent to equation 7.4 (see the discussion of equation 7.4 earlier in this section for the rational for this). Using this trick PASS returns a sample size of 192 per arm.



## 7.2 Cross-over group trials

Similarly to the parallel group case one can solve equation (7.1) to give

$$n = \frac{2Z_{1-\alpha/2}^2 \sigma_w^2}{w^2}, \quad (7.7)$$

where  $n$  is the total sample size. For unknown variance equation (7.7) can be rewritten as:

$$n \geq \frac{2t_{1-\alpha/2, n-2}^2 \sigma_w^2}{w^2}, \quad (7.8)$$

which can be solved iteratively. Alternatively as with parallel group trials the following formula could be used:

$$\Phi\left(\sqrt{\frac{nw^2}{2\sigma_w^2}} - t_{1-\alpha/2, n-2}\right) \geq 0.5. \quad (7.9)$$

To allow for the Normal approximation equation 7.7 can be amended to have a correction factor (Guenther):

$$n = \frac{2\sigma_w^2 Z_{1-\alpha/2}^2}{w^2} + \frac{Z_{1-\alpha/2}^2}{2}. \quad (7.10)$$

The following formula can be used for quick calculations (assuming one wishes to have a 95% confidence interval precision estimates):

$$n = \frac{8\sigma^2}{w^2}. \quad (7.11)$$

Table 7.2 gives sample sizes using equation 7.8 for various standardised widths ( $\delta = d / \sigma$ ).

As with parallel group trials the quick formula slightly under estimates the sample size.

### 7.2.1 Worked Example

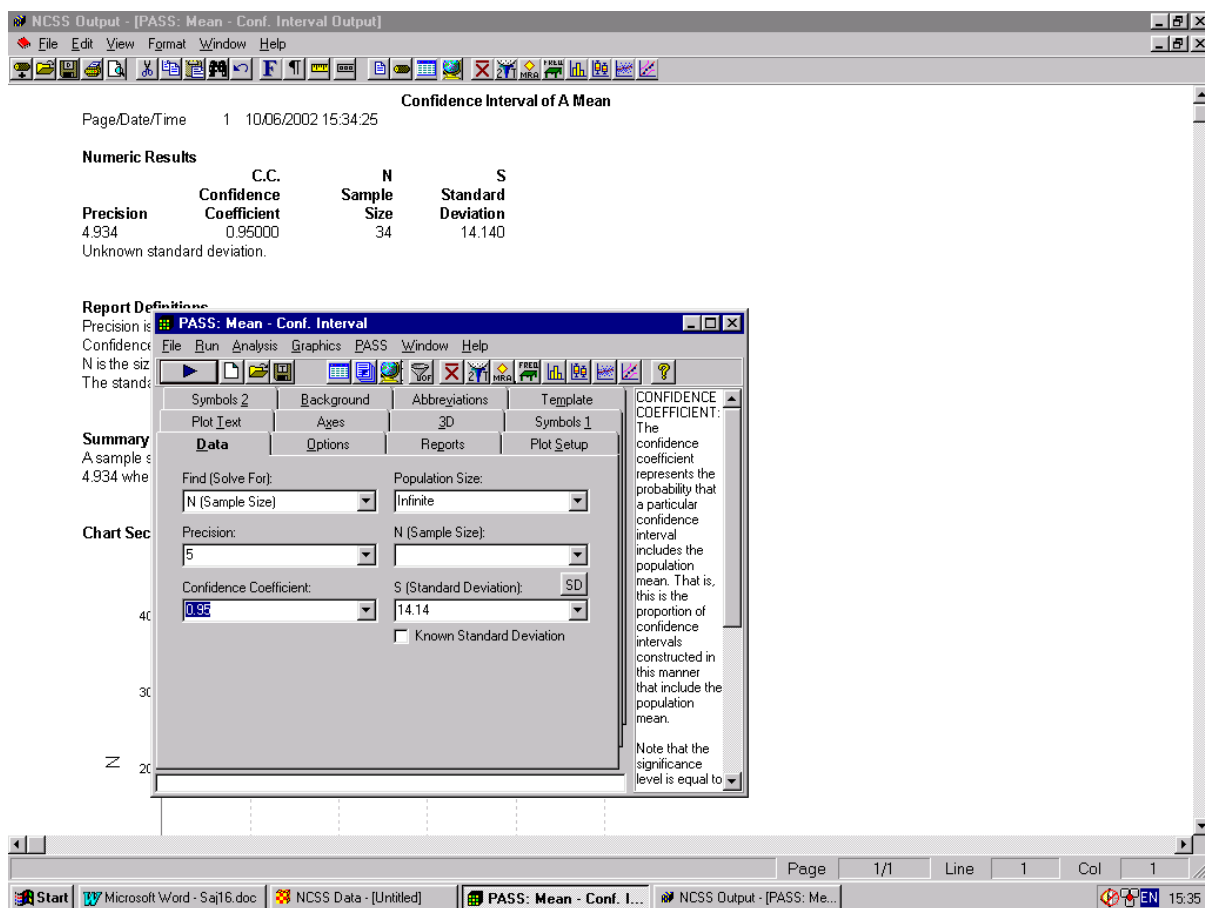
#### 7.2.1.1. Using the sample size tables

An investigator wishes to design a cross-over pilot cardiovascular safety study with equal allocation to estimate any possible effect on QTc with precision around the point estimate of

$\pm 5$  seconds ( $w$ ). The expected within-subject standard deviation in the population in which the trial is to be undertaken is 10 seconds ( $\sigma$ ). The standardised width equates to  $\delta = d/\sigma = 5/10 = 0.50$ . Table 7.2 gives a total sample size of 34 patients.

### 7.1.1.2. Repeated using sample size software

To do the same calculation in PASS one needs to select Means and then Confidence Interval – Mean. The following dialogue box gives the entries required.



As with superiority trials in Section 2 the standard deviation required for the calculations is the standard deviation of the difference. Thus, the standard deviation to use is  $14.14 = \sqrt{2} * 10$ . PASS returns a sample size of 34 which is the same as sample size obtained from Table 7.2. However, there is one anomaly in that the precision of 5 entered

into the dialogue box has metamorphosed into 4.934 in the output window. It is not clear why this is so.

To do the same calculations in nQuery click on File/New for Goal tick Means, Number of Groups tick One and then Confidence Interval. Then select Confidence interval for difference paired means. Like PASS nQuery uses the standard deviation of the difference in the calculations. nQuery returns a sample size of 31 subjects three short of table 7.2. This discrepancy could be due to rounding error, or, as with parallel group trials, it could be due to nQuery using equation (7.10) but without the correction factor.

## 8. DESIGN CONSIDERATIONS

### 8.1 Inclusion of Baseline or Covariates

In the final analysis of a clinical trial, the effect of treatment on the response of interest is often adjusted for predictive factors - such as demographic factors (like gender and age) or clinical covariates (such as baseline response) - by fitting them concurrently with treatment in the statistical model. This section will concentrate on the case where baseline is the predictive covariate of interest (although the results are generalisable to other factors), the design is a parallel group and an analysis of covariance (allowing for the baseline), is to be the final analysis. The CPMP have just issued draft notes for guidance on the design and analysis of studies with covariates [68].

Frison and Pocock [69] give a variance formula for various numbers of baseline measures:

$$\text{Variance} = \sigma^2 \left( 1 - \frac{p\rho^2}{1 + (p-1)\rho} \right). \quad 8.2.1$$

Here,  $\rho$  is the Pearson correlation coefficient between observations and  $p$  is the number of baseline measures taken per individual. From this equation a series of correction factors can be calculated [13] which give the variance reduction (and consequent sample size reduction) for different correlations and numbers of baselines.

From equation 8.2.1 it is clear that for fixed numbers of baseline measures the higher the correlation the greater the reduction in variance and consequent sample size. For example if three baseline measures were to be taken with the expected correlation between baseline and outcome of 0.5, the effect would be to reduce the variance to  $0.6250 * \sigma^2$ . However, for the same number of baseline measures if the expected

correlation between baseline and outcome was 0.7 then the effect would be to reduce the variance to  $0.3875 * \sigma^2$ .

Another result from equation 8.2.1 is that for a fixed correlation it seems that although there is incremental benefit with increasing numbers of baselines this incremental benefit approximately asymptotes at 3 baselines. The following table demonstrate this giving the correction factors for a fixed correlation between baseline and outcome of 0.50 and different numbers of baseline measures.

Number of baselines	Variance
1	0.7500
2	0.6667
3	0.6250
4	0.6000
5	0.5833
6	0.5714

The results in this sub-section demonstrate the importance, when estimating the sample size, in taking the variance estimate from the full model where all covariates are present. It also highlights how, if one ignores baseline and covariate information, one could potentially be overestimating the sample size. The variance allowing for covariates should therefore be used in the sample size equations given in previous sections.

## 8.2. Post Dose Measures Summarised by Summary Statistics

Often in parallel group trials, patients are followed up at multiple time points. Making use of all of the information obtained on a patient has the desirable property of increasing the precision for estimating the effects of treatment. Naturally as the precision is increased the

variability is decreased and one consequently needs to study fewer patients in order to achieve a given power.

Suppose one is interested in looking at the difference in the average of all of the post-dose measures:

$$H_0 : \bar{\mu}_A = \bar{\mu}_B \text{ versus } H_1 : \bar{\mu}_A \neq \bar{\mu}_B ,$$

where  $\bar{\mu}_A$  and  $\bar{\mu}_B$  represent the means of the average post-dose measures in the two treatment populations.

Frison and Pocock [69] explored several other summary measures for multiple post dose time points such as the rate of change of a particular endpoint across time. Diggle, Liang and Zeger [70] describe the hypothesis for such a trials. However, this paper will not get into detail on topic. The simplest approach of just taking the simple average of the post-dose assessments for each subject will be described.

Assuming one has  $r$  post-dose measures and that the correlation between these measures is  $\rho$  the variance can be calculated from

$$\text{Variance} = \frac{\sigma^2 \{1 + (r-1)\rho\}}{r}, \quad 8.2.2$$

where  $\sigma^2$  represents the variance obtained from a trial with a single post-dose measurement.

When looking at equation 8.2.2 it seems that as the correlation between post-dose measures increases the variance and consequent total sample size required increases. This is because, although it may seem counterintuitive, the advantage of taking additional measurements decreases as the correlation increases. This fact is due to how the total variance,  $\sigma^2$ , is constructed [71]:

$$\sigma^2 = \sigma_b^2 + \sigma_w^2, \quad 8.2.3$$

where  $\sigma_w^2$  is the within subject component of variation (as in cross-over trials) and  $\sigma_b^2$  is the between subject component of variation.

It is important here to distinguish between the within- (intra-) subject and the between-(inter-) subject components of variation. The within-subject component of variation quantifies the expected variation among repeated measurements on the same individual. It is a compound of true variation in the individual. Whilst the between-subject component of variation quantifies the expected variation of single measurements from different individuals. If only one measurement is made per individual it is impossible to estimate  $\sigma_w^2$  and  $\sigma_b^2$  and consequently only the total variation, given in equation 8.2.3, can be estimated

If one knows the between-subject variance and the correlation between measures the within-subject variance can be derived from:

$$\sigma_w^2 = \left( \frac{1-\rho}{\rho} \right) \sigma_b^2. \quad 8.2.4$$

Following on from this result the variance that takes account of the number of post dose measures can be defined as:

$$\text{Variance} = \sigma_b^2 + \frac{\sigma_w^2}{r} \quad 8.2.5$$

Thus, formula 8.2.2 is now actually quite intuitive. As for constant r the higher the correlation, from 8.2.4, the lower the within-subject variance and, from 8.2.5, the lower the total variance and consequent sample size. However, as  $\rho$  increases, and  $\sigma_w$  falls, the effect of taking repeated measures diminishes as  $\sigma_w$  already constitutes a small part of the overall variance.

Equation 8.2.2 also gives the incremental benefit of taking additional post dose measures. As with taking baselines it seems that although there is incremental benefit with increasing numbers of post dose measures this incremental benefit asymptotes at 4 post dose measures.

The following table demonstrates this result giving the correction factors for a fixed correlation between post dose measures of 0.50 and difference numbers of post dose measures measures.

Number of post dose measures	Variance
1	1.0000
2	0.7500
3	0.6667
4	0.6250
5	0.6000
6	0.5833

### 8.3 Inclusion of Baseline or Covariates as well as Post Dose Measures Summarised by Summary Statistics

As noted in the previous section further savings in sample size can be achieved by accounting for baseline as a covariate. Frison and Pocock [69] define an additional variance measure to account for the baseline (or multiple baselines) as a covariate and difference numbers of post dose measures. Assuming there are  $p$  baseline visits and  $r$  post dose visits the variance is defined as:

$$\text{Variance} = \sigma^2 \left[ \frac{1 + (r-1)\rho}{r} - \frac{p\rho^2}{1 + (p-1)\rho} \right]. \quad 8.2.6$$



## **9. SUMMARY**

This article walked through the calculations for cross-over and parallel group trials where the assumption is that the data are Normally distributed. The null and alternative hypotheses were described for the most common types of trial and it was highlighted how these hypotheses impact on sample size derivation.

For each type of trial, and trial design, a number of formulae were provided. From quick easy to use results to ones which required iteration to find a solution. It is recommended that the more complicated results be used generally and to assist in this recommendation sample size tables using these results have been provided.

Of the different types of clinical trial covered in this paper probably the ones that will become more prevalent in the future are non-inferiority or "as good as or better" trials. To facilitate their design greater work needs to be undertaken on defining non-inferiority margins. At present there is only general guidance.

An aspect of clinical trial design that is likely to grow is that of being adaptive in one's trials. An area only briefly mentioned in this paper. Through being adaptive one could modify one's sample size calculations during a trial's course and so optimise the trial's design. As such adaptive methodologies develop sample size calculations for Normal data, one will imagine, will be further adapted to accommodate them.

## 10. REFERENCES

1. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal* 1948; **2**: 769-82
2. Julious SA, Zariffa N. The ABC of pharmaceutical trial design: some basic principles. *The Journal of Pharmaceutical Statistics* 2002; **1**:45-53.
3. ICH E9. Statistical principals for clinical trials. September 1998. [http:// www.fda.gov/cder/guidance/ICH\\_E9-fnl.PDF](http://www.fda.gov/cder/guidance/ICH_E9-fnl.PDF) [last accessed 23 October 2003].
4. Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *British Medical Journal* 1999; **318**: 1209.
5. Day S. Blinding or masking. In: Armitage P and Colton T (Editors) *Encyclopedia of Biostatistics*, Chichester: John Wiley and Sons, 1998; 1: 410-417.
6. Altman DG. Statistics and ethics in medical research III - How large a sample? *British Medical Journal* 1980; **281**:1336-8.
7. ICH E3. Structure and content of clinical study reports. July 1996. [http:// www.fda.gov/cder/guidance/iche3.pdf](http://www.fda.gov/cder/guidance/iche3.pdf) [last accessed 23 October 2003].
8. Hintz JL. *PASS 2000 User's Guide*. Kaysville, 2000.
9. Elashoff JD. *nQuery advisor version 4 user's guide*. Los Angeles, 2000.
10. Julious SA, Campbell MJ, Altman DG.. Estimating sample sizes for continuous, binary and ordinal outcomes in paired comparisons: practical hints. *Journal of Biopharmaceutical Statistics* 1999; **9(2)**:241-51.
11. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *British Medical Journal* 1995; **311**:1145-8.
12. Donner A. Approaches to sample size estimation in the design of clinical trials – a review. *Statistics in Medicine* 1983; **3**:199-214.

13. Machin D, Campbell MJ, Fayers P, Pinol A. *Statistical tables for the design of clinical studies 2nd Ed.* Blackwell Scientific Publications: Oxford, 1997.
14. Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory trials. *Drug Information Journal* 1996; **30**:523-534.
15. CPMP. Points to consider on multiplicity issues in clinical trials. (CPMP/EWP/908/99) 19 September 2002 <http://www.emea.eu.int/pdfs/human/ewp/090899en.pdf> [last accessed 21 October 2003].
16. Gould AL. Interim analyses for monitoring clinical trials that do not materially effect the type I error rate. *Statistics in Medicine* 1992; **11**:55-66.
17. Gould AL. Planning and revising the sample size for a trial. *Statistics in Medicine* 1995; **14**:1039-51.
18. Zucker, DM and Denne, J (2002). Sample size re-determination for repeated measures studies. *Biometrics* **48(3)** 548-59.
19. Zucker, DM, Wittes, JT, Schabenberger, O and Brittain E (1999). Internal pilot studies II. Comparison of various procedures. *Statistics in Medicine* **18** 3493-3509.
20. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed data with unknown variance. *Communications in Statistics - Theory and Methods* 1992; **21**:2833-53.
21. Gould AL, Shih WJ. Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* 1998; **17**:89-100.
22. Birkett MA, Day SJ. Internal Pilot studies for estimating sample size. *Statistics in Medicine* 1994; **13**:2455-63.
23. Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* 2000; **19**:901-11.

24. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficacy of clinical trials. *Statistics in Medicine* 1990; **9**: 65-72.
25. Neyman J, Pearson ES. On the use and interpretation of test criteria. *Biometrika* 1928: **20(A)**;175-94.
26. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions Royal Society (London)* 1933; **23**:289-37.
27. Neyman J, Pearson ES. The testing of statistical hypotheses in relation to the probabilities a priori. *Proceeds of the Cambridge Philosophical Society* 1933: **29**;492-510.
28. Neyman J, Pearson ES. Contributions to the theory of testing hypotheses. *Journal Statistical Research Memoirs (University of London)* 1936; **1**:1-37.
29. Neyman J, Pearson ES. Sufficient statistics and uniformly most powerful test of statistical hypothesis. *Journal Statistical Research. Memoirs (University of London)* 1936; **1**:113-137.
30. Neyman J, Pearson ES. Contributions to the theory of testing statistical hypotheses. *Journal Statistical Research Memoirs (University of London)* 1938: **2**;25-57.
31. Fayers P, Machin D. Sample size: how many patients are necessary? *British Journal of Cancer*, 1995; **72**:1-9.
32. Armitage, P and Berry, G. *Statistical Methods in Medical Research*. Third Edition. Blackwell Science: Oxford.
33. Schouten HJA. Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine* 1999: **18**;87-91.
34. Singer, J. A simple procedure to compute sample size needed to compared two independent groups when the population variances are unknown. *Statistics in Medicine* 2001: **20**;1995.

35. Senn S. *Cross-over trials in clinical research*. John Wiley and Sons: Chichester, 1993.
36. Brush GG. How to choose the proper sample size. Volume 12, The ASQC Basic References in Quality Control, Statistical Techniques, American Society for Quality Control: Milwaukee, WI, 1988.
37. Chow SC, Shao J, Wang H. A note on sample size calculations for mean comparisons based on noncentral t-statistics. *Journal of Pharmaceutical Statistics* 2002; **12**:441-56.
38. Guenther WC. Sample size formulas for normal theory t tests. *The American Statistician* 1981; **35**:243-4.
39. Berger RL, Hsu, JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 1996; **11**:283-319.
40. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal* 1996; **313**:36-39
41. ICH E10 Choice of control group in clinical trials, 2000. May 2001 <http://www.fda.gov/cder/guidance/4155fnl.pdf> [last accessed 10 November 2003]
42. CPMP. Points to consider on switching between superiority and non-inferiority. (CPMP/EWP/482/99) 17 July 2000. <http://www.emea.eu.int/pdfs/human/ewp/048299en.pdf> [last accessed 21 October 2003].
43. CPMP. Concept paper on the development of a committee for proprietary medicinal products (CPMP) points to consider on biostatistical methodological issues arising from recent CPMP discussions on licensing applications: choice of delta. (CPMP/EWP/2158/99) 23 September 1999. <http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf> [last accessed 21 October 2003].
44. Hung HMJ, Wang SJ, Lawrence J, O'Neil RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine* 2003; **22**:213-25.

45. D'Agostino RB, Massaro J, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22**:169-86.
46. Wiens BL. Choosing an equivalence limit for noninferiority and or equivalence studies. *Controlled Clinical Trials* 2002; **23**: 2-14.
47. Owen DB. A special case of a bivariate non-central t-distribution. *Biometrika* 1965: **52**;437-446.
48. Diletti E Hauschke D, Steinijans VW. Sample size determination form bioequivalence assessment by means of confidence intervals. *International Journal of Clinical Pharmacology, Therapy and Toxicology* 1991; **29**: 1-8.
49. Morikawa T, Yoshida M. A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics* 1995: **5(3)**;297-306.
50. Senn S. In the blood: proposed new requirements for the registering of generic drugs. *The Lancet* 1998: **352**; 85-86.
51. FDA. Guidance for industry. Bioavailability and bioequivalence studies for orally administered drug products – general considerations. October 2000. [http:// www.fda.gov/cder/guidance/3615fnl.pdf](http://www.fda.gov/cder/guidance/3615fnl.pdf) [last accessed 23 October 2003].
52. FDA. Statistical approaches to establishing bioequivalence. January 2001. [http:// www.fda.gov/cder/guidance/3616fnl.pdf](http://www.fda.gov/cder/guidance/3616fnl.pdf) [last accessed 23 October 2003].
53. CPMP. Note for guidance on the investigation of bioavailability and bioequivalence. (CPMP/EWP/QWP1401/98) 26 July 2001. [http:// www.emea.eu.int/pdfs/human/ewp/140198en.pdf](http://www.emea.eu.int/pdfs/human/ewp/140198en.pdf) [last accessed 21 October 2003].

54. FDA. Guidance for industry. Food-effect bioavailability and fed bioequivalence studies. December 2000. <http://www.fda.gov/cder/guidance/5194fnl.pdf> [last accessed 23 October 03].
55. CPMP. Note for guidance on the investigation of drug interactions. (CPMP/EWP/560/95) 17 December 1997. <http://www.emea.eu.int/pdfs/human/ewp/056095en.pdf> [last accessed 21 October 2003].
56. FDA. Guidance for industry. In vivo drug metabolism/drug interaction studies - study design, data analysis, and recommendations for dosing and labelling. November 1999. <http://www.fda.gov/cder/guidance/2635fnl.pdf> [last accessed 23 October 2003].
57. FDA. Guidance for industry. Pharmacokinetics in patients with impaired renal function – study design, data analysis and impact on dosing and labelling. May 1998. <http://www.fda.gov/cder/guidance/1449fnl.pdf> pdf [last accessed 23 October 2003].
58. FDA. Guidance for industry. Pharmacokinetics in patients with impaired hepatic function: study design, data analysis and impact on dosing and labelling. May 2003. <http://www.fda.gov/cder/guidance/3625fnl.pdf> [last accessed 23 October 2003].
59. Julious SA, Debnarot CAM. Why are pharmacokinetic data summarised as arithmetic means. *Journal of Biopharmaceutical Statistics* 2000; **10(1)**:55-71.
60. Lacey JM, Keene ON, Pritchard JF, Bye A. Common non-compartmental pharmacokinetic variables: are they Normally or log-Normally distributed? *Journal of Biopharmaceutical Statistics* 1997; **7(1)**:171-8.
61. Liu JP. Use of the repeated cross-over designs in assessing bioequivalence. *Statistics in Medicine* 1995; **14**:1067-78.
62. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised trials. *Statistics in Medicine* 1991; **10**:1665-77.

63. Schall R, Williams RL. Towards a practical strategy for assessing individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* 1996; **24**:133-49.
64. Hauck WW, Anderson S. Types of bioequivalence and related statistical considerations, *International Journal of Clinical Pharmacology, Therapy and Toxicology* 1992; **30**:181-187.
65. Day S. Clinical trial numbers and confidence intervals of pre-specified size. *The Lancet* 1988; Dec 17:1427.
66. Wood J, Lambert M. Sample-size calculations for trials in health services research *Journal of Health Services and Research and Policy* 1999; **4**:226-9.
67. Desu MM, Raghavarao D. *Sample size methodology*. Academic Press: London, 1990.
68. CPMP. Points to consider on adjustment for baseline covariates. (CPMP/EWP/2863/99) 23 May 2003. [http:// www.emea.eu.int/ pdfs/ human/ ewp/ 286399en.pdf](http://www.emea.eu.int/pdfs/human/ewp/286399en.pdf) [last accessed 21 October 2003].
69. Frison LJ, Pocock SJ. Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and Its Implication for Design, *Statistics in Medicine* 1992; **11**:1685-1704.
70. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press, 1996.
71. Julious SA. Repeated measures in clinical trials: analysis using means summary statistics and its implications for design. *Statistics in Medicine* 2000; **19**:3133-35.



**Table 2.1. Sample sizes for one group,  $n_A$  ( $n_B=rn_A$ ) in a parallel group study for different standardised differences ( $\delta = d / \sigma$ ) and allocation ratios for 90% power and a two sided type I error rate of 5%.**

$\delta$	Allocation ratios			
	1	2	3	4
0.05	8407	6306	5605	5255
0.10	2103	1577	1402	1314
0.15	935	702	624	585
0.20	527	395	351	329
0.25	338	253	225	211
0.30	235	176	157	147
0.35	173	130	115	108
0.40	133	100	89	83
0.45	105	79	70	66
0.50	86	64	57	53
0.55	71	53	47	44
0.60	60	45	40	37
0.65	51	38	34	32
0.70	44	33	30	28
0.75	39	29	26	24
0.80	34	26	23	21
0.85	31	23	20	19
0.90	27	21	18	17
0.95	25	19	17	15
1.00	23	17	15	14
1.05	21	15	14	13
1.10	19	14	13	12
1.15	17	13	12	11
1.20	16	12	11	10
1.25	15	11	10	9
1.30	14	11	9	9
1.35	13	10	9	8
1.40	12	9	8	8
1.45	12	9	8	7
1.50	11	8	7	7

**Table 2.2. Total sample sizes (n) for a cross-over study for different standardised differences ( $\delta = d / \sigma$ ) and allocation ratios for 90% power and a two sided type I error rate of 5%.**

$\delta$	n
0.05	8408
0.10	2104
0.15	936
0.20	528
0.25	339
0.30	236
0.35	174
0.40	134
0.45	106
0.50	87
0.55	72
0.60	61
0.65	52
0.70	45
0.75	40
0.80	35
0.85	32
0.90	29
0.95	26
1.00	24
1.05	22
1.10	20
1.15	19
1.20	17
1.25	16
1.30	15
1.35	14
1.40	13
1.45	13
1.50	12

**Table 3.1. Sample sizes ( $n_1$ ) for one arm of a parallel group equivalence study with equal allocation ( $r=1$ ) for different standardised equivalence limits ( $\delta = d/\sigma$ ) and true mean differences (as a percentage of  $\delta$ ) for 90% power and a type I error rate of 2.5%.**

$\delta$	Percentage Mean Difference				
	0%	10%	15%	20%	25%
0.05	10397	11042	11915	13218	14960
0.10	2600	2762	2980	3306	3741
0.15	1157	1228	1325	1470	1664
0.20	651	691	746	827	936
0.25	417	443	478	530	600
0.30	290	308	332	369	417
0.35	214	227	245	271	307
0.40	164	174	188	208	235
0.45	130	138	149	165	186
0.50	105	112	121	134	151
0.55	87	93	100	111	125
0.60	74	78	84	93	105
0.65	63	67	72	80	90
0.70	55	58	62	69	78
0.75	48	51	54	60	68
0.80	42	45	48	53	60
0.85	37	40	43	47	53
0.90	34	36	38	42	48
0.95	30	32	34	38	43
1.00	27	29	31	35	39
1.05	25	27	29	31	35
1.10	23	24	26	29	32
1.15	21	22	24	26	30
1.20	20	21	22	24	27
1.25	18	19	21	23	25
1.30	17	18	19	21	24
1.35	16	17	18	20	22
1.40	15	16	17	18	21
1.45	14	15	16	17	19
1.50	13	14	15	16	18

**Table 3.2. Total sample sizes (n) for cross-over equivalence study for different standardised equivalence limits ( $\delta = d/\sigma$ ) and true mean differences (as a percentage of  $\delta$ ) for 90% power and a type I error rate of 2.5%.**

$\delta$	Percentage Mean Difference				
	0%	10%	15%	20%	25%
0.05	10398	11043	11916	13219	14961
0.10	2601	2763	2981	3307	3742
0.15	1158	1229	1326	1471	1665
0.20	652	692	747	828	937
0.25	418	444	479	531	601
0.30	291	309	333	370	418
0.35	215	228	246	272	308
0.40	165	175	189	209	236
0.45	131	139	150	166	187
0.50	106	113	122	135	152
0.55	88	94	101	112	126
0.60	75	79	85	94	106
0.65	64	68	73	81	91
0.70	56	59	63	70	79
0.75	49	52	55	61	69
0.80	43	46	49	54	61
0.85	39	41	44	48	54
0.90	35	37	39	43	49
0.95	31	33	36	39	44
1.00	29	30	32	36	40
1.05	26	28	30	33	36
1.10	24	25	27	30	33
1.15	22	23	25	28	31
1.20	21	22	23	26	29
1.25	19	20	22	24	27
1.30	18	19	20	22	25
1.35	17	18	19	21	23
1.40	16	17	18	20	22
1.45	15	16	17	18	20
1.50	14	15	16	17	19

**Table 4.1. Sample sizes ( $n_A$ ) for one arm of an parallel group non-inferiority study with equal allocation ( $r=1$ ) for different standardised equivalence limits ( $\delta = d/\sigma$ ) and true mean differences (as a percentage of  $\delta$ ) for 90% power and a type I error rate of 2.5%.**

$\delta$	Percentage Mean Difference										
	-25%	-20%	-15%	-10%	-5%	0%	5%	10%	15%	20%	25%
0.05	5381	5839	6358	6949	7626	8407	9316	10379	11636	13136	14945
0.10	1346	1461	1590	1738	1908	2103	2330	2596	2910	3285	3737
0.15	599	650	708	773	849	935	1036	1155	1294	1461	1662
0.20	338	366	399	436	478	527	584	650	729	822	935
0.25	217	235	256	279	306	338	374	417	467	527	599
0.30	151	164	178	194	213	235	260	290	325	366	417
0.35	111	121	131	143	157	173	192	213	239	270	306
0.40	86	93	101	110	121	133	147	164	183	207	235
0.45	68	74	80	87	96	105	116	130	145	164	186
0.50	55	60	65	71	78	86	95	105	118	133	151
0.55	46	50	54	59	64	71	78	87	98	110	125
0.60	39	42	46	50	54	60	66	74	82	93	105
0.65	33	36	39	43	47	51	57	63	70	79	90
0.70	29	31	34	37	40	44	49	54	61	68	78
0.75	25	27	30	32	35	39	43	48	53	60	68
0.80	23	24	26	29	31	34	38	42	47	53	60
0.85	20	22	23	26	28	31	34	37	42	47	53
0.90	18	20	21	23	25	27	30	34	37	42	48
0.95	16	18	19	21	23	25	27	30	34	38	43
1.00	15	16	17	19	21	23	25	27	31	34	39
1.05	14	15	16	17	19	21	23	25	28	31	35
1.10	13	14	15	16	17	19	21	23	26	29	32
1.15	12	13	14	15	16	17	19	21	23	26	30
1.20	11	12	13	14	15	16	18	20	22	24	27
1.25	10	11	12	13	14	15	16	18	20	23	25
1.30	10	10	11	12	13	14	15	17	19	21	24
1.35	9	10	10	11	12	13	14	16	17	20	22
1.40	8	9	10	10	11	12	13	15	16	18	21
1.45	8	9	9	10	11	12	13	14	15	17	19
1.50	8	8	9	9	10	11	12	13	14	16	18

**Table 4.2. Total sample sizes (n) for cross-over non-inferiority study for different standardised equivalence limits ( $\delta = d/\sigma$ ) and true mean differences (as a percentage of  $\delta$ ) for 90% power and a type I error rate of 2.5%.**

$\delta$	Percentage Mean Difference										
	-25%	-20%	-15%	-10%	-5%	0%	5%	10%	15%	20%	25%
0.05	5382	5840	6359	6949	7627	8408	9316	10380	11637	13137	14946
0.10	1347	1462	1591	1739	1909	2104	2331	2597	2911	3286	3738
0.15	600	651	709	774	850	936	1037	1156	1295	1462	1663
0.20	339	367	400	437	479	528	585	651	730	823	936
0.25	218	236	257	280	307	339	375	418	468	528	600
0.30	152	165	179	195	214	236	261	291	326	367	418
0.35	112	122	132	144	158	174	193	214	240	270	307
0.40	87	94	102	111	122	134	148	165	184	208	236
0.45	69	75	81	88	97	106	117	131	146	165	187
0.50	56	61	66	72	79	87	96	106	119	134	152
0.55	47	51	55	60	65	72	79	88	99	111	126
0.60	40	43	47	51	55	61	67	75	83	94	106
0.65	34	37	40	44	48	52	58	64	71	80	91
0.70	30	32	35	38	41	45	50	55	62	69	79
0.75	26	29	31	33	36	40	44	49	54	61	69
0.80	24	25	27	30	32	35	39	43	48	54	61
0.85	21	23	25	27	29	32	35	38	43	48	54
0.90	19	21	22	24	26	29	31	35	38	43	49
0.95	18	19	20	22	24	26	28	31	35	39	44
1.00	16	17	19	20	22	24	26	29	32	35	40
1.05	15	16	17	18	20	22	24	26	29	32	36
1.10	14	15	16	17	18	20	22	24	27	30	33
1.15	13	14	15	16	17	19	20	22	25	27	31
1.20	12	13	14	15	16	17	19	21	23	25	29
1.25	11	12	13	14	15	16	18	19	21	24	26
1.30	11	11	12	13	14	15	16	18	20	22	25
1.35	10	11	12	12	13	14	15	17	19	21	23
1.40	10	10	11	12	13	13	15	16	18	19	22
1.45	9	10	10	11	12	13	14	15	17	18	20
1.50	9	9	10	11	11	12	13	14	16	17	19

**Table 6.1. Total sample sizes (n) for a bio-equivalence cross-over study for different CVs, levels of bio-equivalence and true mean ratios for 90% power and a type I error rate of 5%.**

CV (%)	Ratio	Levels of Bio-equivalence				
		10%	15%	20%	25%	30%
10	0.80				43	12
	0.85			48	13	7
	0.90		54	14	8	5
	0.95	60	16	8	6	5
	1.00	21	10	7	5	5
	1.05	55	15	8	6	5
	1.10		40	13	7	5
	1.15			26	10	6
	1.20			104	17	8
15	0.80				93	23
	0.85			106	26	12
	0.90		119	29	14	8
	0.95	132	33	15	9	7
	1.00	45	20	12	8	6
	1.05	121	31	15	9	7
	1.10		86	25	12	8
	1.15			57	19	10
	1.20			231	36	15
20	0.80				163	40
	0.85			185	45	20
	0.90		207	50	22	13
	0.95	232	56	25	14	10
	1.00	78	34	19	12	9
	1.05	212	54	24	14	10
	1.10		151	43	20	12
	1.15			99	33	16
	1.20			405	62	24
25	0.80				251	60
	0.85			284	68	30
	0.90		320	77	33	18
	0.95	357	86	37	21	14
	1.00	120	52	28	18	12
	1.05	326	82	36	21	14
	1.10		232	65	30	17
	1.15			151	49	24
	1.20			625	95	36
30	0.80				356	85
	0.85			403	96	41
	0.90		454	108	46	25
	0.95	507	121	52	29	18
	1.00	170	73	39	25	17
	1.05	463	116	51	28	18
	1.10		329	92	42	24
	1.15			214	69	33
	1.20			888	135	50
35	0.80				477	113
	0.85			540	128	54
	0.90		608	145	61	33
	0.95	679	162	69	38	24
	1.00	227	97	52	32	22
	1.05	620	155	67	37	24
	1.10		440	123	55	31
	1.15			287	92	44
	1.20			1190	180	67

**Table 6.1 (Cont). Total sample sizes (n) for a bio-equivalence cross-over study for different CVs, levels of bio-equivalence and true mean ratios for 90% power and a type I error rate of 5%.**

CV (%)	Ratio	Levels of Bio-equivalence				
		10%	15%	20%	25%	30%
40	0.80				612	144
	0.85			694	164	69
	0.90		780	185	78	42
	0.95	871	207	88	48	30
	1.00	291	124	66	41	27
	1.05	796	198	86	47	30
	1.10		565	157	71	39
	1.15			367	118	56
	1.20			1527	231	86
45	0.80				760	179
	0.85			861	203	86
	0.90		969	230	97	52
	0.95	1082	257	109	60	37
	1.00	361	153	82	50	33
	1.05	989	246	106	59	37
	1.10		701	195	87	48
	1.15			456	146	69
	1.20			1897	286	106
50	0.80				919	216
	0.85			1042	246	103
	0.90		1172	277	117	62
	0.95	1309	311	132	72	45
	1.00	437	185	99	60	40
	1.05	1196	297	128	70	44
	1.10		848	236	105	58
	1.15			552	177	83
	1.20			2295	346	128
55	0.80				1089	256
	0.85			1233	291	122
	0.90		1387	328	138	74
	0.95	1550	368	156	84	52
	1.00	517	218	117	71	47
	1.05	1416	352	152	83	52
	1.10		1004	279	124	68
	1.15			653	209	98
	1.20			2718	410	151
60	0.80				1266	297
	0.85			1435	338	142
	0.90		1614	381	160	85
	0.95	1803	428	181	98	61
	1.00	601	254	136	82	54
	1.05	1648	409	176	96	60
	1.10		1168	324	144	79
	1.15			759	243	114
	1.20			3162	476	175
65	0.80				1451	340
	0.85			1644	387	162
	0.90		1849	437	183	97
	0.95	2067	490	207	112	69
	1.00	689	291	155	94	62
	1.05	1888	469	202	110	68
	1.10		1338	371	165	90
	1.15			870	278	130
	1.20			3624	546	200



**Table 6.2. Sample sizes ( $n_1$ ) for one arm of a bio-equivalence parallel group study for different CVs, levels of bio-equivalence and true mean ratios for 90% power and a type I error rate of 5%.**

CV (%)	Ratio	Levels of Bio-equivalence				
		10%	15%	20%	25%	30%
30	0.80				356	84
	0.85			403	95	40
	0.90		453	108	46	25
	0.95	506	121	51	28	18
	1.00	169	72	39	24	16
	1.05	462	115	50	28	17
	1.10		328	92	41	23
	1.15			213	69	33
	1.20			887	134	50
35	0.80				476	112
	0.85			540	128	54
	0.90		607	144	61	33
	0.95	678	161	69	37	23
	1.00	226	96	51	31	21
	1.05	620	154	67	37	23
	1.10		439	122	55	30
	1.15			286	92	43
	1.20			1189	179	66
40	0.80				611	144
	0.85			693	163	69
	0.90		779	184	78	41
	0.95	871	207	88	48	30
	1.00	291	123	66	40	26
	1.05	796	198	85	47	29
	1.10		564	157	70	38
	1.15			367	117	55
	1.20			1527	230	85
45	0.80				759	178
	0.85			861	203	85
	0.90		968	229	96	51
	0.95	1082	257	109	59	36
	1.00	361	152	81	49	33
	1.05	988	245	106	58	36
	1.10		700	194	87	47
	1.15			455	146	68
	1.20			1896	286	105
50	0.80				919	216
	0.85			1041	245	103
	0.90		1171	277	116	62
	0.95	1309	310	131	71	44
	1.00	436	184	98	60	39
	1.05	1195	297	128	70	43
	1.10		847	235	104	57
	1.15			551	176	82
	1.20			2295	345	127
55	0.80				1088	255
	0.85			1233	290	121
	0.90		1387	327	137	73
	0.95	1550	367	155	84	52
	1.00	516	218	116	70	46
	1.05	1416	351	151	82	51
	1.10		1003	278	124	68
	1.15			652	208	97
	1.20			2718	409	150

**Table 6.2 (Cont). Sample sizes ( $n_1$ ) for one arm of a bio-equivalence parallel group study for different CVs, levels of bio-equivalence and true mean ratios for 90% power and a type I error rate of 5%.**

CV (%)	Ratio	Levels of Bio-equivalence				
		10%	15%	20%	25%	30%
60	0.80				1266	297
	0.85			1434	337	141
	0.90		1613	381	160	85
	0.95	1803	427	180	97	60
	1.00	601	253	135	82	54
	1.05	1647	408	176	96	59
	1.10		1167	323	144	78
	1.15			759	242	113
	1.20			3162	476	174
65	0.80				1450	340
	0.85			1643	386	161
	0.90		1849	436	183	97
	0.95	2066	489	207	111	68
	1.00	688	290	154	93	61
	1.05	1887	468	201	109	68
	1.10		1337	371	164	90
	1.15			869	277	129
	1.20			3623	545	200
70	0.80				1641	384
	0.85			1860	437	182
	0.90		2092	494	207	109
	0.95	2338	553	234	126	77
	1.00	779	328	175	105	69
	1.05	2135	529	227	124	76
	1.10		1513	419	186	101
	1.15			984	313	146
	1.20			4100	616	226
75	0.80				1836	430
	0.85			2081	489	204
	0.90		2341	552	231	122
	0.95	2616	619	261	141	86
	1.00	871	367	195	118	77
	1.05	2390	592	254	138	85
	1.10		1693	469	208	113
	1.15			1101	351	164
	1.20			4588	690	253
80	0.80				2035	476
	0.85			2307	542	226
	0.90		2595	612	256	135
	0.95	2900	686	289	156	96
	1.00	966	407	216	131	85
	1.05	2649	656	282	153	95
	1.10		1877	520	230	125
	1.15			1220	388	181
	1.20			5086	764	280
85	0.80				2237	524
	0.85			2535	596	248
	0.90		2852	673	281	149
	0.95	3187	754	318	171	105
	1.00	1061	447	238	143	94
	1.05	2912	721	310	168	104
	1.10		2063	571	253	138
	1.15			1341	427	199
	1.20			5590	840	308

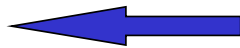
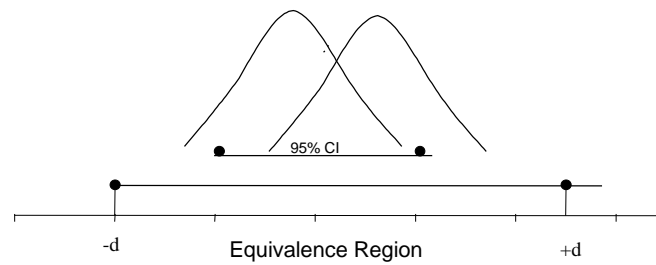
**Table 7.1. Sample sizes for one group,  $n_A$  ( $n_B=rn_A$ ) in a parallel group study for different standardised widths ( $\delta = w/\sigma$ ) and allocation ratios and 95% confidence intervals for the precision estimates**

$\delta$	Allocation ratios			
	1	2	3	4
0.05	3075	2306	2050	1922
0.10	770	578	513	481
0.15	343	257	229	214
0.20	194	145	129	121
0.25	125	94	83	78
0.30	87	65	58	54
0.35	64	48	43	40
0.40	50	37	33	31
0.45	40	30	26	25
0.50	32	24	22	20
0.55	27	20	18	17
0.60	23	17	15	14
0.65	20	15	13	12
0.70	17	13	12	11
0.75	15	12	10	10
0.80	14	10	9	9
0.85	12	9	8	8
0.90	11	8	7	7
0.95	10	8	7	6
1.00	9	7	6	6
1.05	9	7	6	5
1.10	8	6	5	5
1.15	8	6	5	5
1.20	7	5	5	4
1.25	7	5	4	4
1.30	6	5	4	4
1.35	6	5	4	4
1.40	6	4	4	3
1.45	6	4	4	3
1.50	5	4	3	3

**Table 7.2. Total sample sizes for a cross-over study for different standardised widths ( $\delta = w/\sigma$ ) and 95% confidence intervals for the precision estimates**

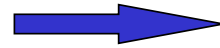
$\delta$	n
0.05	3076
0.10	771
0.15	344
0.20	195
0.25	126
0.30	88
0.35	66
0.40	51
0.45	41
0.50	34
0.55	28
0.60	24
0.65	21
0.70	19
0.75	17
0.80	15
0.85	14
0.90	13
0.95	12
1.00	11
1.05	10
1.10	10
1.15	9
1.20	9
1.25	8
1.30	8
1.35	8
1.40	7
1.45	7
1.50	7

**Figure 1: An illustration of average equivalence between two populations**



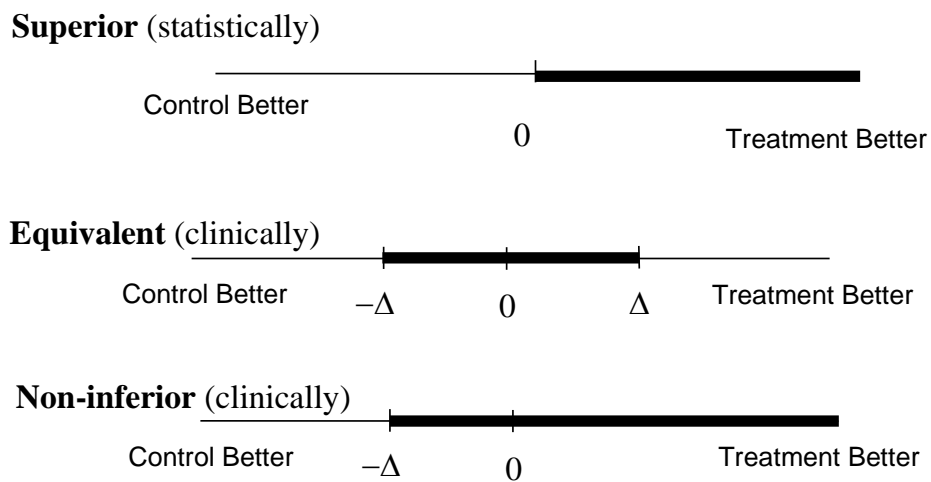
Reject that the difference is  $\leq -d$  with Type I error of 2.5%

and



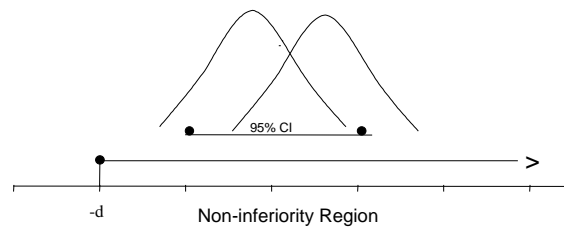
Reject that the difference is  $\geq +d$  with Type I error of 2.5%

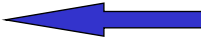
**Figure 2. An illustration of the difference between superiority, equivalence and non-inferiority.**



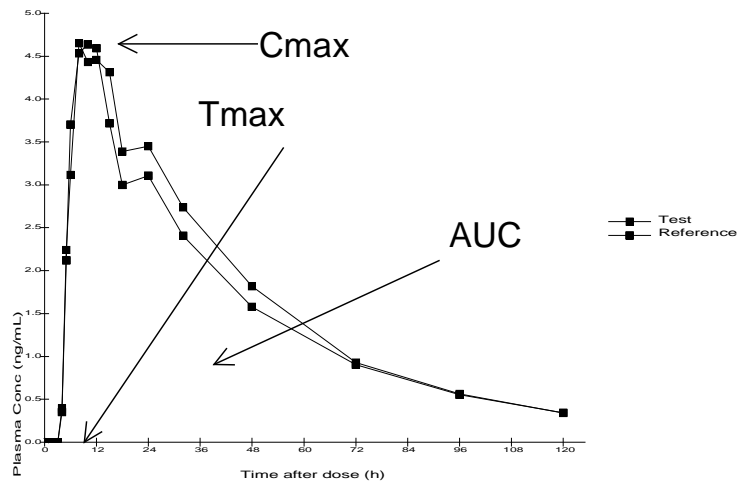
\*  $\Delta$  is variable

**Figure 3: An illustration of average non-inferiority between two populations**



  
Reject that the  
difference is  
 $\leq -d$  with Type I  
error of 2.5%

**Figure 4. An example of pharmacokinetic profiles for a test and reference formulation.**





**Figure 5: An illustration of average bioequivalence between two formulations**

