UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
Institute of Mathematics and Statistics
Mathematical Statistics Curriculum

Lisbeth Neevits

Sample Size Calculations in Clinical Trials

Bachelor's Thesis (9 ECTS)

Supervisors: Marju Valge, MSc
Pasi Antero Korhonen, PhD

Tartu 2016

Sample Size Calculations in Clinical Trials

The aim of this thesis is to give an overview of calculating sample size in clinical trials. First, a brief introduction to clinical trials and factors affecting sample size is given. This is followed by chapters on sample size calculations for three main trial types. Every design is then illustrated by a practical example and instructions for calculations in SAS and R software.
Keywords: clinical trials, R, sample size, SAS

P160 Statistics, operation research, programming, actuarial mathematics

Valimimahu arvutused kliinilistes uuringutes

Selle bakalaureusetöö eesmärgiks on anda ülevaade valimimahu arvutamisest kliinilistes uuringutes. Esmalt kirjeldatakse lühidalt kliinilisi uuringuid ning valimimahtu mõjutavaid tegureid. Seejärel käsitletakse valimimahu arvutusi kolme põhilise uuringutüübi korral. Igale disainile on lisatud praktiline näide ning SASi ja Ri koodid valimimahu arvutamiseks.
Võtmesõnad: kliinilised uuringud, R, valimimaht, SAS

P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Table of Contents

1. Introduction

The purpose of this Bachelor's thesis is to give an overview of calculating sample size in clinical trials. Clinical trials are studies carried out to test new experimental treatments before they can be made available for the market. These studies may have different purposes. The objectives of clinical studies may include one or more of the following four: (i) demonstrate/confirm efficacy, (ii) establish a safety profile, (iii) provide an adequate basis for assessing the benefit/risk relationship to support labelling, and (iv) establish the dose-response relationship. [1] Currently (according to data from 01.04.2016) there are 208 ongoing trials in Estonia. Most of them are conducted in the field of oncology, followed by neurological studies. [2]

The process of choosing the sample size is of great importance and it is equally inefficient to include either too few or too many subjects in a study. Clinical trials can be extremely expensive and subjects hard to find. Also, as there is no guarantee that the experimental treatment is better or safer than the already existing one, the subjects may be put at unnecessary risk. Due to these reasons smaller sample sizes are in favor. At the same time, it is unethical to include too few patients because that can lead to unreliable conclusions. Even worse is the increased probability of getting no conclusions at all when at the same time there are people who have been put at health risk.

There are three types of clinical trials that are clearly distinguishable: superiority, non-inferiority and equivalence. For each, a brief introduction is given which is followed by explanations on hypothesis testing and finding optimal sample size considering different study designs. Finally, to put theory into practice, different methods for statistical software – SAS and R – are presented throughout the thesis. In addition to providing commands for sample size calculations, simple explanatory cases are presented to illustrate the use of the programs.

The intention is to create a guidance for StatFinn Oy employees. Chapters are written in a way to make it possible quickly find the right formulas and/or syntax for software when the study type and design are determined. This is the main reason for having many repetitions in every paragraph. The aim is to be as clear as possible and not to overcomplicate the calculation progress for sample size.

2. Clinical Dictionary

A short dictionary of clinical terms used in this thesis is provided to make reading understandable. For Estonian translations English-Estonian medical dictionary by Birgit Parkson [3] was used.

active control *aktiivne võrdlusravi*
a marketed treatment used as a reference in a clinical trial

baseline *algandmed*
a data collected before subjects receive the first dose of treatment (pre-treatment)

bioavailability *biosaadavus*
the rate to which a treatment reaches the target organ or systemic circulation

bioequivalence *bioekvivalentsus*
an equivalent concentration of treatments in plasma and tissue when administered to the same patient

blinded study *pimemenetlusega uuring*
a strategy in clinical trials where one or more parties involved (e.g. participants, clinicians) do not know the treatment assigned for randomized groups

clinical trial *kliiniline uuring*
a research investigation where participants receive one or more treatments to answer questions about the safety and efficacy of these treatments

clinically meaningful difference *kliiniliselt oluline erinevus*
the smallest difference in treatment effects that is important for study conductors

control group *kontrollgrupp*
the subjects receiving control treatment (active control or placebo)

crossover trial *ristuvuuring*
a trial design where every subject serves as his/her own control

endpoint *tulemusnäitaja*
a variable that is of focus to evaluate safety and/or efficacy

equivalence trial *samaväärsusuuring*
a trial that is aiming to show that the results of experimental and control treatment differ by an amount that is clinically unimportant

experimental group *ekperimentaalgrupp*
the subjects receiving experimental treatment

experimental treatment *ekperimentaalravi*
a treatment of focus

intention-to-treat analysis *ravikavatsuse analüüs*
an analysis that is analyzing every randomized subject as assigned to their randomized group (including e.g. violators)

non-inferiority trial *mitte-halvemusuuring*
a trial that is aiming to show the results of experimental treatment are not much worse from control treatment

parallel trial *pralleelrühma uuring*
a trial design where every subject is randomized into an experimental group or a control group

per protocol analysis uuringuplaani analüüs
an analysis that is analyzing only those subjects that have followed the protocol

placebo *platseebo*
a treatment with no active ingredients used as a reference in a clinical trial

primary endpoint esmane tulemusnäitaja
the most important endpoint in the study

protocol *uuringuplaan*
a document describing the objectives, design, statistical considerations etc. of a certain clinical trial

protocol violator *uuringuplaani rikkuja*
a subject who has not been following the protocol

randomized controlled trial *randomiseeritud kontrollkatse*
an experiment where study participants are randomly allocated into different study groups

reference value *võrdlusnäitaja väärtus*
a baseline, pre-treatment value of a certain endpoint

sample size *valimimaht*
the number of subjects participating in the study

study treatment/drug *uuringuravi(m)*
a treatment/drug under investigation

subject *uuritav*
a person participating in the clinical trial

superiority trial *paremusuuring*
a trial that is aiming to show that the results of experimental treatment are better than the results of control treatment

treatment allocation *ravijaotus*
the desired proportion of subjects in each study group

treatment effect *raviefekt*
the true mean difference between a study drug and a control

3. Clinical Trials

3.1. Sample Size in Clinical Trials

In clinical trials the randomized controlled trial is a standard. It is a study design where subjects are randomized into a control or an experimental study drug group. The control group can be administered a standard treatment (active control) or placebo. Once the subjects have received the treatment, the aim of a randomized controlled trial is to measure and compare the outcomes of the study. This design reduces the effect of confounding factors that may lead to wrong interpretations considering the associations between study variables. [4]

Placebo control is usually used to prove the beneficial effect of a new treatment, but in many cases it is considered unethical. When beneficial standard treatment exists, it should always be used for comparison. The use of placebo can affect the sample size. For example, in non-inferiority trials the sample size needed for treatment difference orientated placebo-controlled studies is much smaller than for studies where active-control is needed. This is because the difference between no treatment and new treatment is much larger than the difference between new and already existing treatment.

Sample size should provide the right amount of subjects for different kinds of studies. In addition to considering the trial type (superiority, non-inferiority or equivalence) and study design (parallel or crossover), the sample size should be chosen so that the protocol violators, patient dropouts or subjects accidentally randomized into wrong groups would not have significant impact on the final result. The inclusion and exclusion criteria should be stated carefully and kept constant to avoid any misunderstandings and inaccurate conclusions. In long-term studies these criteria still may change because of findings during the study. [5]

There are two approaches in study group comparison that may affect the sample size chosen: Per Protocol analysis and Intention-To-Treat analysis. Patients who are not following the protocol are excluded from Per Protocol analysis. In Intention-To-Treat analysis the perfect scenario would be that none of the patients violate the protocol. Nevertheless, if such patients do exist they are still analyzed as if they followed the protocol perfectly. For superiority trials, Intention-To-Treat analysis is usually applied. This is mainly done to avoid overly positive conclusions caused by excluding all protocol violators and therefore reducing the probability of type I error. For non-inferiority and equivalence trials both Intention-To-Treat and Per Protocol analysis should be conducted to obtain proper conclusions. Both approaches should give similar results for these trials. When the results

differ a little, the least positive one is preferred. A good sample size should consider the analysis type used and try to minimize the effects it has on the results. [5] [6] [7]

Sample size calculations should be based on a single primary endpoint that is similarly evaluated for each study participant based on collected information on that subject. It should form the basis of the objectives of the trial and be of biological and/or clinical importance. For example, it could be death of the patient or occurrence of a symptom. A primary endpoint should be carefully considered. Despite the fact that it is not recommended, it may happen that there are several primary endpoints. In that case, all endpoints should be considered and sample size should be chosen so that every endpoint would have sufficient power in hypothesis testing. [1]

Planning sample size must be based on prior data. When parameters needed for the sample size calculation cannot be estimated, it is recommended to conduct a pilot study which is a preliminary study conducted to save resources in an inefficiently designed trial. [1] [8]

Therefore, sample size has to be determined by considering everything above-mentioned and balancing those factors.

3.1.1. Power and Hypothesis Testing

In the table below (Table 1) the concept of type I error and type II error can be seen. The null hypothesis (H_0) refers to a default position that there is no relationship or no difference among groups. There are four different possibilities: (i) H_0 is true and not rejected, (ii) H_0 is true but rejected (type I error), (iii) H_0 is false, but not rejected (type II error), and (iv) H_0 is false and rejected. Type I error is usually denoted by α and type II error by β. [1]

Table 1 The concept of type I error and type II error

	$\boldsymbol{H_0}$ **True**	$\boldsymbol{H_0}$ **False**
Reject $\boldsymbol{H_0}$	Type I Error (α)	Correct Rejection $(1-\beta)$
Fail to Reject $\boldsymbol{H_0}$	Correct Decision $(1-\alpha)$	Type II Error (β)

From the table it may be concluded that

$$\alpha = Pr\{type\ I\ error\} = Pr\{reject\ H_0\ when\ it\ is\ true\},$$

$$\beta = Pr\{type\ II\ error\} = Pr\{fail\ to\ reject\ H_0\ when\ it\ is\ false\}.$$

An upper bound for α is known to be the level of significance that is chosen for hypothesis testing. Most commonly the value 0.05 is chosen to show confidence concerning the parameter of interest. $1 - \alpha$ is called the level of confidence and it is the probability of not rejecting the null hypothesis (H_0) when it is true. In clinical trials, the aim is to decrease both type I error and type II error. With a fixed sample size, when α increases, β decreases and *vice versa*. The only way to decrease them both at the same time is to increase the sample size. [1]

The power of a test is the probability that the test will correctly reject the null hypothesis (H_0) when the alternative hypothesis (H_1) is true, i.e.,

$$Power = 1 - \beta = Pr\{reject\ H_0\ when\ it\ is\ false\}.$$

This probability should be as large as possible to be sufficiently comfortable of the likelihood of finding a statistically significant treatment effect when such exists. In practice, the desired power is commonly 80% or 90%. Lack of power can lead to various errors. For studies where null hypothesis (H_0) is not rejected, it is hard to distinguish between having no effect at all and failing to prove an adequately sized observed effect due to a too small sample size.

3.1.2. Objectives of a Typical Clinical Trial

In clinical trials, the general objective is proving superiority, non-inferiority or equivalence. These three study types are clearly distinguishable.

Superiority trials are aiming to prove that the experimental treatment is better than the active control or placebo. Nowadays, that is often hard to prove since there are so many effective drugs on the market. That is why the most common trial types are non-inferiority and equivalence trials. In those trials, the experimental treatment should have been proven superior to a placebo beforehand. [9]

Non-inferiority trials aim to prove that the new treatment is no less effective than the control. These studies may be conducted to show the increased safety of the experimental treatment instead of showing the superiority of treatment effect. [9]

Equivalence trials aim to prove that the experimental and control treatments are similar by showing that the difference in their treatment effects stays within an acceptable interval. Many of these trials are bioequivalence trials where it is desirable to show that the experimental treatment and control treatment have similar bioavailability. [9]

To prove superiority, non-inferiority or equivalence, the clinically meaningful difference is used. It is called a superiority margin, non-inferiority margin or equivalence margin (limit), respectively. The treatment effect, i.e. $\mu_E - \mu_C$ (μ_E and μ_C being the true mean responses of the experimental treatment and the control treatment, respectively), is compared with the margin in order to make conclusions. It is important to notice that the desired treatment effect may be negative or positive depending on the primary endpoint of the study. On one hand, if the value of the primary endpoint is expected to increase, e.g. haemoglobin concentration in patients with anaemia, the treatment effect is larger than zero. On the other hand, if the value of the primary endpoint is expected to decrease, e.g. the blood pressure in patients with hypertension, the treatment effect is less than zero. This concept can be seen in the figure below (Figure 1) where the visual summary of hypotheses for deciding whether the new treatment is superior, non-inferior or equivalent to control is presented. [1]

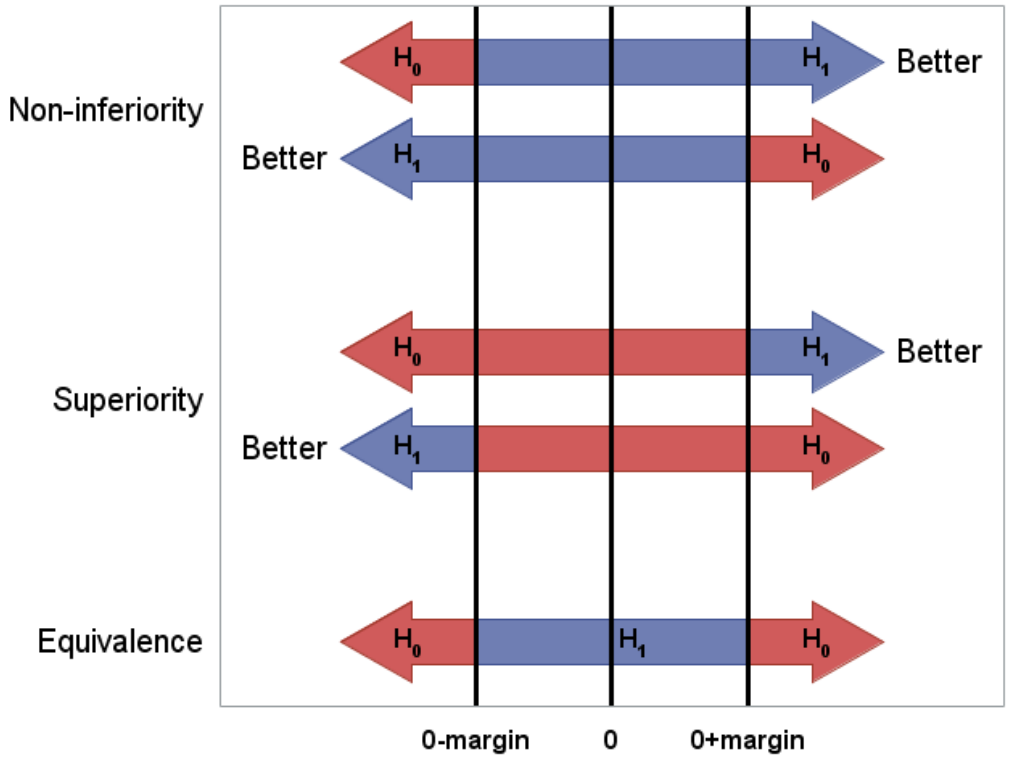

Figure 1 Hypotheses testing with different study types [10]

For non-inferiority and superiority, the upper hypotheses are used when the value of a primary endpoint is expected to increase and lower hypotheses are used when the value of a primary endpoint is expected to decrease. For example, when the larger value is better, the null hypothesis for superiority trial is $\mu_E - \mu_C \leq margin$, when smaller value is better, the null hypothesis is formulated as $\mu_E - \mu_C \geq -margin$.

From now on, for the convenience of notation and understanding and without loss of generality, it is assumed that the larger value is always better (the value of a primary endpoint is expected to increase).

3.1.3. Clinically Meaningful Difference

It is important to acknowledge that treatment effects may be statistically significant but are not always clinically important. For that reason, in randomized controlled trials, the general term clinically meaningful difference, is used. It is the smallest difference in treatment effects that is important for study conductors. The value of the clinically meaningful difference should never be zero as two treatments cannot be exactly equal. The term and notation differ for each trial type: for superiority trials it is the superiority margin (δ), in non-inferiority trials it is the non-inferiority margin (δ_{NI}) and in equivalence trials it is the equivalence limit (δ_E). Establishing δ is discussed in Chapter 5 and Chapter 6. Determining clinically meaningful difference is crucial prior to the study and it is rather difficult to find the best method for obtaining δ. Most often it is recommended to use data from previous placebo-controlled studies that were planned under similar conditions or use the information from a pilot study. For non-inferiority and equivalence studies, if the selected δ is too small, many effective drugs may be rejected. Contrary to that, when the δ is chosen to be too big, many inefficient drugs may be accepted. For superiority it is the opposite. [1] [4]

3.1.4. One-Sample Analysis

One-sample analysis is used to evaluate the effect within a particular study group. The hypotheses are defined to confirm whether there is a significant difference between pre- and post-treatment or mean change from baseline to endpoint. [1]

For sample size calculations, let x_j be the response of a treatment from the jth participant of the study group, $j = 1, 2, \dots, n$. It is assumed that x_j's are realizations of independent and identically distributed normal random variables X_j with mean μ and variance σ^2. The sample mean is then defined as

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j.$$

For one-sample analysis $\mu - \mu_0$ is being estimated where μ is the true mean response of an experimental treatment and μ_0 is the reference value. Reference value can be for example the pre-treatment value of an endpoint. [1]

In practice, one-sample analysis is not used in non-inferiority and equivalence trials and will therefore only be discussed for superiority trials in this thesis.

3.1.5. Two-sample Analysis

Two-sample analysis is used to compare efficacy or other factors of an experimental treatment to a control treatment. It can be also used for comparing different doses. There are two designs that are most often used in practice – parallel design and crossover design. In parallel study design, subjects are randomized into groups and get only one treatment (experimental or control) for the whole duration of the study. For crossover design the subjects receive several treatments over the course of the trial. [1]

Parallel design (Figure 2) is the most common one used in clinical trials and is quite easily conducted compared to other study designs. It is based on between-subject variability which means that differences are observed between different subjects. [1] [5]

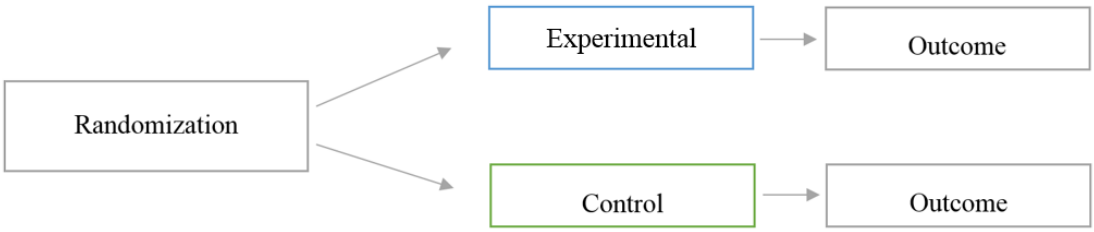

Figure 2 Parallel design

For two-sample parallel design let x_{jk}, $j = 1, 2, \dots, n_k$, $ki = 1, 2$, be the response that is observed from the jth participant in the kth treatment group. It is expected that x_{jk} are realizations of independent normal random variables with mean μ_k and variance σ_k^2. The sample mean for the kith treatment is defined as

$$\bar{x}_{k.} = \frac{1}{n_{ki}} \sum_{j=1}^{n_k} x_{jk} .$$

For two-sample analysis $\mu_E - \mu_C$ is being estimated, that is, the true mean difference between test drug and a control where μ_E is the true mean response of an experimental treatment and $\mu_C$is the true mean response of a control treatment. [1]

Crossover design is the most used one in equivalence trials. The simplest form of crossover trial is a standard 2x2 design (Figure 3) which means that there are two treatments and every study participant is getting them both in one of the two potential sequences ‘experimental first – then control’ or

'control first – then experimental'. It is based on within-subject variability which means that every participant is one's own control. [1] [5]

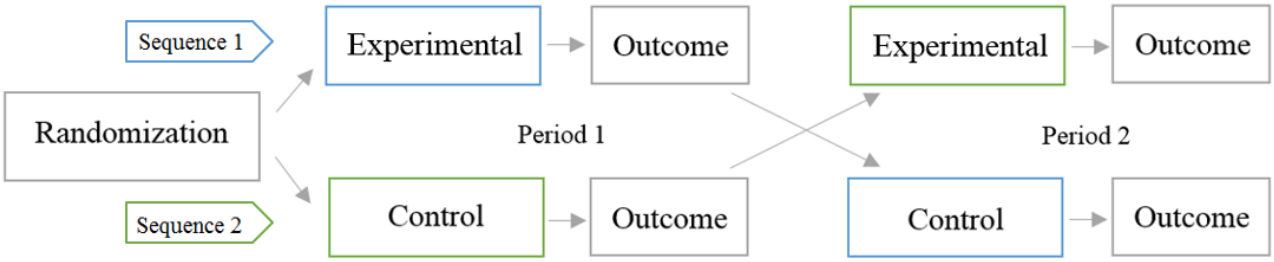

Figure 3 2x2 crossover design

Let y_{ijk} be the response from jth subject $(j = 1, \dots, n)$ in the ith sequence $(i = 1, 2)$ under the kth treatment $(k = 1, 2)$. A simplified 2x2 model for crossover trial is

$$y_{ijk} = \mu_k + \gamma_{ik} + s_{ij} + \varepsilon_{ijk},^{1}$$

where

- μ_k is the effect of the kth treatment;
- γ_{ik} is the fixed effect of the ith sequence under treatment k;
- s_{ij} is the between subject variability, $s_{ij} \sim N(0, \sigma_B^2)$, i.i.d.;
- ε_{ijk} is the within subject variability, $\varepsilon_{ijk} \sim N(0, \sigma_W^2)$, i.i.d. [1] [11]

3.1.6. Sample Size Calculations

In clinical trials, sample size calculations can be based on (i) precision analysis, (ii) power based analysis, (iii) probability assessment or (iv) some other statistical inferences. In practice, the most commonly used are the first two – precision analysis and power based analysis. This thesis discusses these analyses in separate paragraphs based on controlling type I error (or confidence level) and type II error (or power). [1]

Probability assessment relies on a probability statement and it is used when it is desirable to detect a small difference of rare events. This method helps to avoid the need for an extremely big sample size. Other methods for sample size calculations include reproducibility probability (an estimated power

[1] For detailed derivations of finding sample size for 2x2 crossover design, see B. Jones, M. G. Kenward, Design and Analysis of Cross-Over Trials, Chapman and Hall/CRC, 2014 (https://www.crcpress.com/Design-and-Analysis-of-Cross-Over-Trials-Third-Edition/Jones-Kenward/9781439861424).

approach for the second clinical trial[2]) and sample size re-estimation without unblinding (re-estimating sample size based on data collected up to a certain time point). These methods are out of the scope of this thesis and are not discussed any further. [1]

3.1.6.1. Power Based Sample Size Analysis

In hypothesis testing, type I error is considered to be the more serious error than type II error. Because of that an acceptable α is determined and the aim is to minimize β by choosing the right sample size. This kind of sample size determination where α and β are given, is referred to as power based analysis. It can be used to find the smallest possible sample size but still having a feasible probability of discovering an effect of given size. [1]

For power based analysis the following is needed:

- determination of an acceptable level of significance;
- selection of a desirable power;
- specification of a clinically meaningful difference;
- having the knowledge of the standard deviation of the primary endpoint. [1]

The formula used for sample size calculations when testing the following hypothesis

$$H_0: \mu_{\mathrm{E}} = \mu_{\mathrm{C}}$$

$$H_1: \ \mu_{\mathrm{E}} \neq \mu_{\mathrm{C}},$$

and assuming $n_1 = n_2 = n$, is

$$n = \frac{(\sigma_1^2 + \sigma_2^2)\left(z_{\alpha/2} + z_{\beta}\right)^2}{\delta^2},$$

where

- μ_{E} and μ_{C} are primary endpoint means for experimental and control study groups;
- n_1 and n_2 indicate the sample size for each study group;
- σ_k^2 is the standard deviation of the *k*th group of observations;

[2] The approval of an experimental treatment usually requires at least two clinical trials to be carried out.

- $z_{\alpha/2}$ is the upper $(\alpha/2)$th-quantile of the standard normal distribution, corresponding to two-tailed significance level;
- z_{β} is the upper βth-quantile of the standard normal distribution, corresponding to power;
- δ is the clinically meaningful difference. [1]

The formula above can also be solved for other parameters, e.g. power $(1-\beta)$.

3.1.6.2. Precision Analysis

Precision shows how consistent the measurements are when repeated. When population parameters are estimated it is important to do it with a certain level of precision. The aim is to find a sample size so that errors of estimation stay within certain limits. The precision of an interval depends on its width: the narrower the interval, the more precise the measurements, and the wider the interval, the more imprecise the measurements. The confidence interval approach is equivalent to the method of hypotheses testing and that allows us to apply it for sample size calculations. For these calculations, confidence interval $(1-\alpha)100\%$ is used. [1]

The precision analysis considers the maximum half width of the $(1-\alpha)100\%$ confidence interval of the unknown parameter that is considered sufficient. That half width of the confidence interval is also known as the maximum acceptable error of an estimate. [1]

When standard deviation σ is known, the confidence interval for the mean μ of the primary endpoint can be calculated as

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where

- $z_{\alpha/2}$ is the upper $(\alpha/2)$th quantile of the standard normal distribution;
- n is the sample size;
- $\bar{x}$ is the sample mean. [1]

When estimating μ, the maximum acceptable error (denoted by E_{max}) is defined as

$$E_{max} = |\bar{x} - \mu| = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ .[1]$$

Because of this, the sample size calculation can be carried out as

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E_{max}^2}.$$

3.2. Power and Sample Size Calculations with SAS

The POWER procedure in SAS can be used for finding the sample size for a given level of power and *vice versa*. The value of interest has to be denoted as a missing value (`.`). [12]

The basic syntax for the POWER procedure is

```
proc power;
<options>;
run;
```

For specifying one-sample and two-sample analysis, options `onesamplemeans` and `twosamplemeans` are used respectively. By default, both analyses use two-sided test, i.e. `sides=2`. For one-sided test, this command has to be carefully specified as shown in the table below (Table 2). [10] [12]

Table 2 The use of commands sides *and* nullmean *in SAS [10]*

Study type	**The direction of treatment change**	**`proc power;` `onesamplemeans`**
Superiority	Larger is better	`sides=U` `nullmean=`$\mu_0 + \delta$
	Smaller is better	`sides=L` `nullmean=`$\mu_0 - \delta$
Non-Inferiority	Larger is better	`sides=U` `nullmean=`$\mu_0 - \delta_{NI}$
	Smaller is better	`sides=L` `nullmean=`$\mu_0 + \delta_{NI}$

Both `onesamplemeans` and `twosamplemeans` perform sample size and power calculations using t-test by default. For `onesamplemeans` option, `mean` has to be given value μ and `nullmean` calculated as shown in the table above (Table 2).

For `twosamplemeans` option, it is necessary to choose `test=diff` for superiority/non-inferiority trials or `test=equiv_diff` for equivalence trials to calculate sample size for parallel and crossover designs. When `test=diff` is chosen, values `meandiff` and `nulldiff` must be defined as $\mu_E - \mu_C$ and δ, respectively. For `test=equiv_diff`, values `meandiff`, `upper`, and `lower` must be defined as $\mu_E - \mu_C$, $+\delta_E$, and $-\delta_E$, respectively. When sample size for parallel design is calculated, option `groupweights` is needed for defining treatment allocation $\kappa = n_1/n_2$. [12] Having unequal variances σ_1^2 and σ_2^2 in parallel design, `stddev` is given the value

$$\sigma_{pooled} = \sqrt{\frac{\frac{\sigma_1^2}{\kappa} + \sigma_2^2}{1 + \frac{1}{\kappa}}}.$$

In crossover design, sample size is calculated per sequence. Because of that, `npergroup` is used instead of `ntotal`. Also, for crossover design, standard deviation has to always be divided by two, i.e. $\sigma/2$, that is due to the crossover formula.

An example for sample size calculation with SAS:

One-sample analysis, where the true mean response $\mu = 2$ units, reference value $\mu_0 = 1.5$ units and clinically meaningful difference $\delta = 0.6$ units. The standard deviation $\sigma = 1$ units, alpha $\alpha = 0.05$ and power $1 - \beta = 0.8$.

```
proc power;
   onesamplemeans
   sides=2
   mean=2
   nullmean=2.1
   ntotal=.
   stddev=1
   alpha=0.05
   power=0.8;
run;
```

The procedure gives the following output (Figure 4) from where it can be seen that SAS performed two-sided t-test for mean where data is normally distributed. From the output it can be seen that the sample size for a given power is 787. The same syntax can be used for power calculations with a fixed sample size when replacing the value of power with a missing value (.).

The SAS System

One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Number of Sides	2
Null Mean	2.1
Alpha	0.05
Mean	2
Standard Deviation	1
Nominal Power	0.8

Computed N Total	
Actual Power	N Total
0.800	787

Figure 4 SAS output of sample size calculation

3.3. Power and Sample Size Calculations with R

For simple power calculations, R has a package called pwr [13]. Since population means are used in sample size calculations, the basic syntax to obtain power and sample size using t-test for one-sample and two-sample analysis where samples are of the same size, i.e. the number of subjects for experimental and control group is the same, is

pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL,
type = c("two.sample", "one.sample"),
alternative = c("two.sided", "less", "greater")).

When desired group sizes differ for two-sample analysis, i.e. an experimental or a control group is bigger than the other, another function should be used:

pwr.t2n.test(n1 = NULL, n2= NULL, d = NULL, sig.level = 0.05, power = NULL,
alternative = c("two.sided", "less","greater")).

The arguments in the pwr.test and pwr.t2n.test functions indicate the following:

- n (n1 and n2 for studies with unequal groups) stands for the number of observations in a group;
- d is the effect size, i.e. $d = \frac{|\mu_E - (\mu_C + \delta)|}{\sigma}$;
- sig.level is the significance level, i.e. α;
- power is the power of test, i.e. $1 - \beta$;
- type shows if it is one- or two-sample test;
- alternative specifies the alternative hypothesis, whether it is two-sided, greater or less. [13]

It can be noted that besides power, these functions can also be used to find sample size *n*, effect size *d* and the level of significance sig.level. The desired value (n, d, sig.level, or power) must be denoted by NULL. It has to be noticed that parameter sig.level is not NULL by default and must be specifically denoted so.

For sample size calculations in clinical trials, R has also a package TrialSize that is based on the book "Sample Size in Clinical Research" by S.C. Chow, J. Shao, H. Wang. It has its own function for every study type and design. OneSampleMean.NIS, TwoSampleMean.NIS, TwoSampleCrossOver.NIS for superiority and non-inferiority, and OneSampleMean.Equivalence, TwoSampleMean.Equivalence, TwoSampleCrossOver.Equivalence for equivalence trials. [14]

Arguments for these six functions are mostly the same:

(alpha, beta, sigma, k, delta, margin),

where

- alpha is the significance level, i.e. α;
- beta stands for type II error, i.e. β;
- sigma is the standard deviation, i.e. σ;
- k shows treatment allocation (needed for parallel design), i.e. $\kappa = n_1/n_2$;
- delta is the clinically meaningful difference, i.e. δ;
- margin is the true mean difference, i.e. $\mu_E - \mu_C$. [14]

Having unequal variances σ_1^2 and σ_2^2 in parallel design, sigma is given the value

$$\sigma_{pooled} = \sqrt{\frac{\frac{\sigma_1^2}{\kappa} + \sigma_2^2}{1 + \frac{1}{\kappa}}}.$$

An example of sample size calculations with R:

The same example is used as for sample size calculation with SAS (Chapter 3.2). One-sample analysis, where the true mean response $\mu = 2$ units, reference value $\mu_0 = 1.5$ units and clinically meaningful difference $\delta = 0.6$ units. The standard deviation $\sigma = 1$ units, alpha $\alpha = 0.05$ and power $1 - \beta = 0.8$. Here three variables (μ, μ_0 and δ) need to be combined to calculate $d = \frac{|\mu - (\mu_0 + \delta)|}{\sigma}$.

```
pwr.t.test(n=NULL, d=0.1 , sig.level = 0.05, power=0.8,
type = "one.sample",
alternative = "two.sided").
```

Result of the sample size calculation with R is given below (Figure 5).

```
One-sample t test power calculation

          n = 786.8089
          d = 0.1
  sig.level = 0.05
      power = 0.8
alternative = two.sided
```

Figure 5 R output of sample size calculation

It can be seen from the output that the sample size for a given power is $786.8089 \approx 787$. The same syntax can be used for power calculations with a fixed sample size when replacing the value of power with NULL.

4. Superiority Trials

In superiority trials, the purpose is to show that an experimental treatment is superior to another – active control and/or placebo. This is what the majority of randomized controlled trials aim for – the experimental treatment is hoped to appear superior to the control. With a statistically significant result, it can be concluded that the experimental treatment is more effective compared to the control treatment. When the result is not statistically significant, it cannot be claimed that the experimental is better than the control treatment. With a nonsignificant result it can also be wrongly concluded that two treatments are equal in effect. Two treatments cannot be identical and there is always some kind of slight difference in the results. Therefore, if that difference exists and it is in favour of the experimental treatment, it should always be possible to find the right sample size for a superiority trial to show that distinction. Often, when the null hypothesis is not rejected, it is concluded as an absence of evidence even though it cannot be proved that there is no difference in treatment effects. [9] [15] "Randomized controlled clinical trials that do not show a significant difference between the treatments being compared are often called "negative." This term wrongly implies that the study has shown that there is no difference, whereas usually all that has been shown is an absence of evidence of a difference. These are quite different statements [16]."

The hypotheses for superiority trials are:

$$H_0: \mu_E - \mu_C \leq \delta$$

$$H_1: \mu_E - \mu_C > \delta,$$

where

- μ_E is the mean of the primary endpoint for the experimental treatment;
- μ_C is the mean of the primary endpoint for the control treatment;
- δ is the clinically meaningful difference. [1]

When the null hypothesis is rejected, it indicates that there is a difference between experimental and the control treatment, i.e. the test drug is superior to standard therapy.

The above hypotheses are defined for one-sided test, in the case that the larger value of primary endpoint is better (healthier). In practice, two-sided test is often preferred for showing superiority. The hypotheses for that are:

$$H_0: |\mu_E - \mu_C| \leq 0$$

$$H_1: |\mu_E - \mu_C| > \delta,$$

where $\delta > 0$.

4.1. One-Sample Design

4.1.1. One-Sided Test

For one-sample design, let the hypothesis for the superiority trial be

$$H_0: \mu - \mu_0 \leq \delta$$

$$H_1: \mu - \mu_0 > \delta,$$

where

- μ is the true mean response of a test drug;
- μ_0 is a reference value, e.g. the pre-treatment value of an endpoint;
- δ is the clinically meaningful difference. [1]

The sample size is calculated as:

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{(\mu - \mu_0 - \delta)^2},$$

where z_β is the upper βth-quantile and z_α is the upper αth-quantile of the standard normal distribution. [1] For the derivation of this formula, see an example in Appendix 1.

Above formula assumes that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formula given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 s^2}{(\mu - \mu_0 - \delta)^2}.$$

4.1.2. Two-Sided Test

For one-sample design, let the hypothesis for the superiority trial be

$$H_0: |\mu - \mu_0| \leq 0$$

$$H_1: |\mu - \mu_0| > \delta,$$

where

- μ is the true mean response of a test drug;
- μ_0 is a reference value, e.g. the pre-treatment value of an endpoint;
- $\delta > 0$ is the clinically meaningful difference. [1]

The sample size corresponding to this test is

$$n = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \sigma^2}{(\mu - \mu_0 - \delta)^2},$$

where z_β is the upper βth-quantile and $z_{\alpha/2}$ is the upper $\alpha/2$th-quantile of the standard normal distribution. [1] For the derivation of this formula, see an example in Appendix 2.

Above formula assumes that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formula given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 s^2}{(\mu - \mu_0 - \delta)^2}.$$

4.1.3. SAS Calculations for One-Sample Design

Commands `sides` (only for one-sided test) and `nullmean` have to be determined following Table 2, other commands following Table 3. Option `onesamplemeans` is needed for one-sample analysis. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used.

One-sided:

```
proc power;
onesamplemeans
sides=
mean=
nullmean=
stddev=
alpha=
power=
ntotal=.;
run;
```

Two-sided:

```
proc power;
onesamplemeans
mean=
nullmean=
stddev=
alpha=
power=
ntotal=.;
run;
```

Table 3 SAS commands for one-sample design

Option	Value
mean	μ
nullmean	$\mu_0 + \delta$
stddev	σ or s
alpha	α
power	$1 - \beta$

4.1.4. R Calculations for One-Sample Design

For superiority one-sample design, R has a function OneSampleMean.NIS from package TrialSize. Having two-sided test instead of one-sided, alpha is divided by two, i.e. $\alpha/2$. When σ is unknown, it is replaced by s. Arguments have to be determined following Table 4.

OneSampleMean.NIS(alpha, beta, sigma, margin, delta)

Table 4 R arguments for one-sample design

Option	Value
alpha	α or $\alpha/2$
beta	β
sigma	σ or s
margin	$\mu - \mu_0$
delta	δ

4.1.5. An Example for One-Sample Superiority Trial

A pharmaceutical company is interested in having an 80% power for establishing superiority of their new treatment. The variance $\sigma^2 = 0.1$ units^2 and expected true mean difference, i.e. effect size $\mu - \mu_0$, is 0.3 units. Pre-treatment mean is 0.3 units and post-treatment mean is 0.6 units. The clinically meaningful difference for this study is 0.2 units.

The sample size is now found using three different methods (Table 5).

Table 5 Sample size calculations for one-sample superiority design

Method	Formula/syntax	Result
$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu - \mu_0 - \delta)^2}$	$\frac{(z_{0.05} + z_{0.2})^2 0{,}1}{(0.3 - 0.2)^2} = \frac{(1.6449 + 0.8416)^2 0.1}{(0.3 - 0.2)^2} = 61.82682$	$61.82682 \approx 62$

SAS	`proc power;`
	`onesamplemeans`
	`sides=U`
	`mean=0.6`
	`nullmean=0.5`
	`stddev=0.3162278`
	`alpha=0.05`
	`power=0.8`
	`ntotal=.;`
	`run;`
		64

R	OneSampleMean.NIS(alpha=0.05, beta=0.2, sigma=0.3162278, margin=0.3, delta=0.2)	$61.82559 \approx 62$

Hence for one-sided hypothesis testing with the type I error level set to 5%, a total of 64 patients would be required in order to detect a clinically meaningful difference of 0.2 units with 80% power when the expected effect size is 0.3 units and the variance $\sigma^2 = 0.1$ units^2 (based on calculations done with SAS software). The results differ slightly as a result of rounding etc.

4.2. Two-Sample Parallel Design

4.2.1. One-Sided Test

In two-sample parallel design subjects are randomized into two groups – experimental and control – and get the same treatment the whole time the trial is ongoing. Let the hypothesis for the superiority trial be

$$H_0: \mu_E - \mu_C \leq \delta$$

$$H_1: \mu_E - \mu_C > \delta,$$

where

- μ_E is the true mean response of the experimental treatment;
- μ_C is the true mean response of the control treatment;
- δ is the clinically meaningful difference. [1]

The formulas for obtaining the sample size for the experimental treatment group and the control group are then given by

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2 (1 + 1/\kappa)}{(\mu_E - \mu_C - \delta)^2},$$

where $\kappa = n_1/n_2$ shows treatment allocations, usually 1:1 or 2:1. z_β is the upper βth-quantile and z_α is the upper αth-quantile of the standard normal distribution. [1] For the derivation of these formulas, see an example in Appendix 3.

When the population variances σ_1^2 and σ_2^2 are not equal then the sample size is calculated as

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2 \left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)}{(\mu_E - \mu_C - \delta)^2},$$

where σ_1^2 is the variance of the experimental treatment group and σ_2^2 is the variance of the control treatment group.

Above formulas assume that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formulas given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2 s^2 (1 + 1/\kappa)}{(\mu_E - \mu_C - \delta)^2}.$$

4.2.2. Two-Sided Test

For two-sided test, let the hypothesis for parallel design be

$$H_0: |\mu_E - \mu_C| \leq 0$$

$$H_1: |\mu_E - \mu_C| > \delta,$$

where

- μ_E is the true mean response of the experimental treatment;
- μ_C is the true mean response of the control treatment;
- $\delta > 0$ is the clinically meaningful difference. [1]

Similarly to the one-sided test, the formulas of sample size for the experimental treatment group and the control are

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \sigma^2 (1 + 1/\kappa)}{(\mu_E - \mu_C - \delta)^2},$$

where $\kappa = n_1/n_2$ demonstrates treatment allocations, z_β is the upper βth-quantile and $z_{\alpha/2}$ is the upper $\alpha/2$th-quantile of the standard normal distribution. [1]

When the population variances σ_1^2 and σ_2^2 are not equal then the sample size is calculated by

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)}{(\mu_E - \mu_C - \delta)^2},$$

where σ_1^2 is the variance of the experimental treatment group and σ_2^2 is the variance of the control treatment group.

Above formulas assume that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formula given above can also be used when

σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 s^2 (1 + 1/\kappa)}{(\mu_E - \mu_C - \delta)^2}.$$

4.2.3. SAS Calculations for Two-Sample Parallel Design

Command `sides` (only for one-sided test) has to be determined following Table 2, other commands following Table 6. Option `twosamplemeans` is needed for two-sample analysis. With `test=diff` it is stated that there are two samples and means of those samples are compared. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used. For unequal variances σ is replaced by $\sigma_{pooled} = \sqrt{\left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)/\left(1 + \frac{1}{\kappa}\right)}$. Statement `groupweights` in needed for treatment allocation, for example 1:2 treatment allocation is denoted as `groupweights=1|2`. If it is desired to get sample size for the whole study, `ntotal` should be used, if getting sample size for each study group separately is of interest, then `npergroup` is used. It is important to notice that options `groupweights` and `npergroup` cannot be used together.

One-sided:

```
proc power;
   twosamplemeans
   test=diff
   sides=
   groupweights=
   meandiff=
   nulldiff=
   stddev=
   alpha=
   power=
   ntotal=.;
run;
```

Two-sided:

```
proc power;
   twosamplemeans
   test=diff
   groupweights=
   meandiff=
   nulldiff=
   stddev=
   alpha=
   power=
   ntotal=.;
run;
```

Table 6 SAS commands for two-sample parallel design

Option	Value
meandiff	$\mu_E - \mu_C$
nulldiff	δ
stddev	σ or s or σ_{pooled}
alpha	α
power	$1 - \beta$
groupweights	$n_1 \vert n_2$

4.2.4. R Calculations for Two-Sample Parallel Design

For superiority two-sample parallel design, R has the function TwoSampleMean.NIS from package TrialSize. This function calculates one-sided sample size by default. Having two-sided test instead of one-sided, alpha is divided by two, i.e. $\alpha/2$. It is important to notice that R gives sample size per group. Also, with two-sample design, treatment allocation $\kappa = n_1/n_2$ has to be considered. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used. For unequal variances σ is replaced by $\sigma_{pooled} = \sqrt{\left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)/\left(1 + \frac{1}{\kappa}\right)}$. With unequal study groups, e.g. $\kappa = n_1/n_2 = 1/2 = 0.5$, R calculates n_1 and the total sample size $n = n_1 + n_2 = n_1 + \frac{n_1}{\kappa}$. For the usual parallel study with equal study groups ($\kappa = n_1/n_2 = 1$), the sample size obtained with R simply has to be doubled to get the total sample size. Arguments have to be determined following Table 7.

TwoSampleMean.NIS(alpha, beta, sigma, k, delta, margin)

Table 7 R arguments for two-sample parallel design

Option	Value
alpha	α or $\alpha/2$
beta	β
sigma	σ or s or σ_{pooled}
k	n_1/n_2
delta	δ
margin	$\mu_E - \mu_C$

4.2.5. An Example for Two-Sample Parallel Superiority Trial

A pharmaceutical company is interested in having an 80% power for establishing superiority of their new headache treatment. It is expected that the new treatment improves headache symptoms by 60% while the same indicator for the control treatment is 30%. The clinically meaningful difference for this study is 20%, the variance $\sigma^2 = 0.1$. Let both treatment groups be equal. The sample size is now found using three different methods (Table 8).

Table 8 Sample size calculations for two-sample parallel superiority design

Method	Formula/syntax	Result
$n_1 = \kappa n_2,$ $n_2 = \frac{(z_\alpha + z_\beta)^2 \sigma^2 (1+1/\kappa)}{(\mu_E - \mu_C - \delta)^2} =$	$n_2 = \frac{(z_{0.05} + z_{0.2})^2 0.1(1+1/1)}{(0.6-0.3-0.2)^2} = \frac{(1.6449 + 0.8416)^2 0.1(1+1/1)}{(0.6-0.3-0.2)^2} = 123.6536$ $n_1 = 1 * 123.6536 = 123.6536$	$123.6536 + +123.6536 = = 247.3073 \approx \approx 248$
SAS	`proc power; twosamplemeans test=diff sides=U groupweights=1|1 meandiff=0.3 nulldiff=0.2 stddev=0.3162278 alpha=0.05 power=0.8 ntotal=.; run;`	250
R	TwoSampleMean.NIS(alpha=0.05, beta=0.2, sigma=0.3162278, k=1, delta=0.2, margin=0.3)	$123.6512 + +123.6512 = = 247.3023 \approx \approx 248$

Hence for one-sided hypothesis testing with the type I error level set to 5%, a total of 250 patients – 125 patients in each group – would be required in order to detect a clinically meaningful difference of 20% with 80% power and the variance $\sigma^2 = 0.1$ (based on calculations done with SAS software). The results differ slightly as a result of rounding etc.

4.3. Two-Sample Crossover Design

4.3.1. One-Sided Test

In two-sample 2x2 crossover design subjects are randomized into two groups – experimental and control. During the first period, one group gets experimental and the other one gets control treatment. During the second period, the subjects who got experimental treatment before now get control treatment and *vice versa*. That way each participant is one's own control. Let the hypothesis for the superiority trial be

$$H_0: \mu_E - \mu_C \leq \delta$$

$$H_1: \mu_E - \mu_C > \delta,$$

where

- μ_E is the true mean response of the experimental treatment;
- μ_C is the true mean response of the control treatment;
- δ is the clinically meaningful difference. [1]

The formula for calculating sample size in each sequence is

$$n_1 = n_2 = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{2(\mu_E - \mu_C - \delta)^2},$$

where z_β is the upper βth-quantile and z_α is the upper αth-quantile of the standard normal distribution. [1] [11]

Above formula assumes that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formula given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n_1 = n_2 = \frac{(z_\alpha + z_\beta)^2 s^2}{2(\mu_E - \mu_C - \delta)^2}.$$

4.3.2. Two-Sided Test

For two-sided test, let the hypothesis for parallel design be

$$H_0: |\mu_E - \mu_C| \leq 0$$

$$H_1: |\mu_E - \mu_C| > \delta,$$

where

- μ_E is the true mean response of the experimental treatment;
- μ_C is the true mean response of the control treatment;
- $\delta > 0$ is the clinically meaningful difference. [1]

Similarly to the one-sided test, the sample size in each sequence is obtained with two-sided test:

$$n_1 = n_2 = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{2(\mu_E - \mu_C - \delta)^2},$$

where z_β is the upper βth-quantile and $z_{\alpha/2}$ is the upper $\alpha/2$th-quantile of the standard normal distribution. [1] [11]

Above formula assumes that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formula given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n_1 = n_2 = \frac{(z_{\alpha/2} + z_\beta)^2 s^2}{2(\mu_E - \mu_C - \delta)^2}.$$

4.3.3. SAS Calculations for Two-Sample Crossover Design

Command `sides` (only for one-sided test) has to be determined following Table 2, other commands following Table 9. Option `twosamplemeans` is needed for two-sample analysis. The sample size needed is obtained by finding the sample size per sequence (`npergroup`), for total sample size that has to be doubled. For crossover design, standard deviation has to be divided by two, i.e. $\sigma/2$. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used.

One-sided:

```
proc power;
   twosamplemeans
   test=diff
   sides=
   meandiff=
   nulldiff=
   stddev=
   alpha=
   power=
   npergroup=.;
run;
```

Two-sided:

```
proc power;
   twosamplemeans
   test=diff
   meandiff=
   nulldiff=
   stddev=
   alpha=
   power=
   npergroup=.;
run;
```

Table 9 SAS commands for two-sample crossover design

Option	Value
meandiff	$\mu_E - \mu_C$
nulldiff	δ
stddev	$\sigma/2$ or $s/2$
alpha	α
power	$1-\beta$

4.3.4. R Calculations for Two-Sample Crossover Design

For superiority two-sample crossover design, R has the function TwoSampleCrossOver.NIS from package TrialSize. This function calculates one-sided sample size by default. For two-sided calculation, alpha needs to be divided by two, i.e. $\alpha/2$. It is important to notice that R gives sample size per sequence. Arguments have to be determined following Table 10.

TwoSampleCrossOver.NIS(alpha, beta, sigma, delta, margin)

Table 10 R arguments for two-sample crossover design

Option	Value
alpha	α or $\alpha/2$
beta	β
sigma	σ or s
delta	δ
margin	$\mu_E - \mu_C$

4.3.5. An Example for Two-Sample Crossover Superiority Trial

The same example is used as for the parallel design. A pharmaceutical company is interested in having an 80% power for establishing superiority of their new headache treatment. It is expected that the new treatment improves headache symptoms by 60% while the same indicator for the control treatment is 30%. The clinically meaningful difference for this study is 20%, the variance $\sigma^2 = 0.1$.

The sample size is now found using three different methods (Table 11).

Table 11 Sample size calculations for two-sample crossover superiority design

Method	Formula/syntax	Result
$n_1 = n_2 = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{2(\mu_E - \mu_C - \delta)^2}$	$\frac{(z_{0.05} + z_{0.2})^2 0.1}{2(0.6 - 0.3 - 0.2)^2} = \frac{(1.6449 + 0.8416)^2 0.1}{2(0.6 - 0.3 - 0.2)^2} = 30.91341 \approx 31$	$30.91341 \approx \approx 31$

SAS	`proc power;`
	`twosamplemeans`
	`test=diff`
	`sides=U`
	`meandiff=0.3`
	`nulldiff=0.2`
	`stddev=0.1581139`
	`alpha=0.05`
	`power=0.8`
	`npergroup=.;`
	`run;`
		32

R	TwoSampleCrossOver.NIS(alpha=0.05, beta=0.2, sigma=0.3162278, delta=0.2, margin=0.3)	$30.91279 \approx \approx 31$

A one-sided hypothesis testing with the type I error level set to 5%, a total of 64 patients, 32 in each sequence, would be required in order to detect a clinically meaningful difference of 20% with 80% power and the variance $\sigma^2 = 0.1$ (based on calculations done with SAS software). The results differ slightly as a result of rounding etc.

5. Non-Inferiority Trials

In the past, experimental treatment was always believed to be better than the control treatment. As the treatments have developed a lot, nowadays it is quite difficult to find a drug that is vastly better than the existing one. Instead of proving increased effect of a drug, the focus can be, for example, on finding a drug that has the same efficacy but is safer, cheaper or easier to administer. These trials are called non-inferiority trials, which are aimed to prove that the experimental treatment is not much worse than the control treatment, i.e. the treatment difference is not less than a non-inferiority margin $\delta_{NI} > 0$. When the lower limit of the confidence interval is above $-\delta_{NI}$, i.e. the confidence interval of the difference $\mu_E - \mu_C$ lies within the interval $(-\delta_{NI}; +\infty)$, the non-inferiority of the experimental treatment is proven. [6] [9]

The non-inferiority margin δ_{NI} is defined as the maximum extent of clinical non-inferiority. It is crucial to define it correctly. Otherwise, when the margin is chosen to be too large, it may happen that the test drug is allowed to be remarkably less effective than active control or even placebo. With a margin that is too small, it may be concluded that the test drug is inferior to the control. Whether the selection of either too small or too large, δ_{NI} could become a problem, depends on the disease for what the experimental treatment is tested for. The margin value should be chosen to be smaller than is the minimum difference between active control and placebo to ensure that the test drug has a clinically relevant effect greater than zero. It is recommended to base the defining process on prior placebo-controlled studies and the margin should be substantially smaller – approximately no more than one half – than the clinically meaningful difference used for superiority trials. [17] [7]

Non-inferiority trials are often used for ethical reasons. It is chosen instead of a clearly interpretable superiority trial in situations where it is unethical to conduct a placebo controlled trial or give a low dose of an active control. [15] Those kinds of trials include situations where active control could prevent some kind of serious harm, e.g. death of a patient. There are more reasons why non-inferiority trials are chosen to be the most appropriate. Where the experimental drug is not expected to be better on a primary endpoint, but it is safer or easier to produce or administrate compared to active control, then that trial type is preferred. For example, when there is a control treatment that has to be injected or the oral intake of a medication is very frequent, but the experimental treatment can be administered orally instead of injecting or the medication is taken only once a day. During the randomized controlled trial, it may seem that an experimental treatment has a higher efficacy because the patients are being controlled and motivated. Outside the trial, when the patients will most probably not follow the intake plan so carefully, the experimental treatment with an easier drug administration could have

a higher efficacy. The drug used as an active control should be proven beforehand to be superior in a placebo-controlled trial. [9] [18] [19]

When the 95% confidence interval excludes both – the non-inferiority margin and zero –, it is considered acceptable to prove superiority within the same trial. This does not apply the other way round. [18]

The hypotheses for non-inferiority trials are:

$$H_0: \mu_E - \mu_C \leq -\delta_{NI}$$

$$H_1: \mu_E - \mu_C > -\delta_{NI},$$

where

- μ_E is the mean of the primary endpoint for the experimental treatment;
- μ_C is the mean of the primary endpoint for the control treatment;
- $\delta_{NI} > 0$ is the clinically meaningful difference. [1]

When the null hypothesis is rejected, it indicates that the experimental drug is not much worse than the control treatment. It can be seen that the hypotheses of superiority trials and non-inferiority trials are almost the same, the only difference is a minus sign in front of non-inferiority margin.

5.1. Two-Sample Parallel Design

In two-sample parallel design study subjects are randomized into two groups – experimental and control – and get the same treatment the whole time the trial is ongoing. Let the hypothesis for the non-inferiority trial be

$$H_0: \mu_E - \mu_C \leq -\delta_{NI}$$

$$H_1: \mu_E - \mu_C > -\delta_{NI},$$

where

- μ_E is the true mean response of the experimental treatment;
- μ_C is the true mean response of the control treatment
- $\delta_{NI} > 0$ is the clinically meaningful difference. [1]

The formulas of sample size for the experimental treatment group and the control group are

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2 (1 + 1/\kappa)}{(\mu_E - \mu_C + \delta_{NI})^2},$$

where $\kappa = n_1/n_2$ shows treatment allocations, usually 1:1 or 2:1. z_β is the upper βth-quantile and z_α is the upper αth-quantile of the standard normal distribution. [1] For the derivation of this formula, see an example in Appendix 3.

When the population variances σ_1^2 and σ_2^2 are not equal, the sample size is given by

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2 \left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)}{(\mu_E - \mu_C + \delta_{NI})^2},$$

where σ_1^2 is the variance of the experimental treatment group and σ_2^2 is the variance of the control treatment group.

Above formulas assume that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formulas given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2 s^2 (1 + 1/\kappa)}{(\mu_E - \mu_C + \delta_{NI})^2}.$$

5.1.1. SAS Calculations for Two-Sample Parallel Design

Command `sides` has to be determined following Table 2, other commands following Table 12. Option `twosamplemeans` is needed for two-sample analysis. With `test=diff` it is stated that there are two samples and means of those samples are compared. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used. For unequal variances σ is replaced by

$\sigma_{pooled} = \sqrt{\left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)/\left(1 + \frac{1}{\kappa}\right)}$. Statement `groupweights` in needed for treatment allocation, for example 1:2 treatment allocation is denoted as `groupweights=1|2`. If it is desired to get sample size for the whole study, `ntotal` should be used, if getting sample size for each study group separately is of interest, then `npergroup` is used. It is important to notice that options `groupweights` and `npergroup` cannot be used together. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used.

```
proc power;
   twosamplemeans
   test=diff
   sides=
   groupweights=
   meandiff=
   nulldiff=
   stddev=
   alpha=
   power=
   ntotal=.;
run;
```

Table 12 SAS commands for two-sample parallel design

Option	Value
`meandiff`	$\mu_E - \mu_C$
`nulldiff`	$-\delta_{NI}$
`stddev`	σ or s or σ_{pooled}
`alpha`	α
`power`	$1 - \beta$
`groupweights`	$n_1 \vert n_2$

5.1.2. R Calculations for Two-Sample Parallel Design

For superiority two-sample parallel design, R has the function TwoSampleMean.NIS from package TrialSize. It is important to notice that R gives sample size per group. Also, with two-sample design, treatment allocation $\kappa = n_1/n_2$ has to be considered. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used. For unequal variances σ is replaced by $\sigma_{pooled} = \sqrt{\left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)/\left(1 + \frac{1}{\kappa}\right)}$. With unequal study groups, e.g. $\kappa = n_1/n_2 = 1/2 = 0.5$, R calculates n_1 and the total sample size $n = n_1 + n_2 = n_1 + \frac{n_1}{\kappa}$. For the usual parallel study with equal study groups ($\kappa = n_1/n_2 = 1$), the sample size obtained with R simply has to be doubled to get the total sample size. Arguments have to be determined following Table 13.

TwoSampleMean.NIS(alpha, beta, sigma, k, delta, margin)

Table 13 R arguments for two-sample parallel design

Option	Value
alpha	α
beta	β
sigma	σ or s or σ_{pooled}
k	n_1/n_2
delta	$-\delta_{NI}$
margin	$\mu_E - \mu_C$

5.1.3. An Example for Two-Sample Parallel Non-Inferiority Trial

A pharmaceutical company is interested in having an 80% power for establishing non-inferiority of their experimental treatment. It is expected that the mean value for the experimental treatment is 0.3 units while the same indicator for the control treatment is 0.2 units. This means that the experimental treatment is supposed to increase the value of the primary endpoint by 0.1 units, i.e. effect size $\mu_E - \mu_C = 0.1$. The clinically meaningful difference for this study is 0.2 units, the variance $\sigma^2 = 0.2$ units^2. Let both treatment groups be equal.

The sample size is now found using three different methods (Table 14).

Table 14 Sample size calculations for two-sample parallel non-inferiority design

Method	Formula/syntax	Result
$n_1 = \kappa n_2,$ $n_2 = \frac{(z_\alpha + z_\beta)^2 \sigma^2 (1+1/\kappa)}{(\mu_E - \mu_C + \delta_{NI})^2}$	$n_2 = \frac{(z_{0.05} + z_{0.2})^2 0.2(1+1/1)}{(0.3 - 0.2 + 0.2)^2} =$ $= \frac{(1.6449 + 0.8416)^2 0.2(1+1/1)}{(0.1+0.2)^2} = 27.47859$ $n_1 = 1 * 248.3228 = 248.3228$	$27.47859 +$ $+27.47859 =$ $= 54.95718 \approx$ ≈ 55

SAS	`proc power;`
	`twosamplemeans`
	`test=diff`
	`sides=U`
	`groupweights=1|1`
	`meandiff=0.1`
	`nulldiff=-0.2`
	`stddev=0.4472136`
	`alpha=0.05`
	`power=0.8`
	`ntotal=.;`
	`run;`
		58

R	TwoSampleMean.NIS(alpha=0.05, beta=0.2, sigma=0.4472136, k=1, delta=-0.2, margin=0.1)	$27.47803 +$ $+27.47803 =$ $= 54.95607 \approx$ ≈ 55

With 80% power and a one-sided type I error level set to 5% a total of 58 patients, 29 in each group, would be required in order to show that the experimental treatment is at most 0.2 units inferior to the

control treatment when the expected effect size is 0.1 units and the variance $\sigma^2 = 0.2$ units^2 (based on calculations done with SAS software). The results differ slightly as a result of rounding etc

5.2. Two-Sample Crossover Design

In two-sample 2x2 crossover design subjects are randomized into two treatment sequences. During the first period, one group gets the experimental and the other one gets the control treatment. During the second period, the subjects who got experimental treatment before, now get control treatment and *vice versa*. That way each participant is one's own control. Let the hypothesis for the non-inferiority trial be

$$H_0: \mu_E - \mu_C \leq -\delta_{NI}$$

$$H_1: \mu_E - \mu_C > -\delta_{NI},$$

where

- μ_E is the true mean response of the experimental treatment;
- μ_C is the true mean response of the control treatment.
- $\delta_{NI} > 0$ is the clinically meaningful difference. [1]

The formula for calculating sample size for each treatment sequence is

$$n_1 = n_2 = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{2(\mu_E - \mu_C + \delta_{NI})^2},$$

where z_β is the upper βth-quantile and z_α is the upper αth-quantile of the standard normal distribution. [1] [11]

Above formula assumes that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formula given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n_1 = n_2 = \frac{(z_\alpha + z_\beta)^2 s^2}{2(\mu_E - \mu_C + \delta_{NI})^2}.$$

5.2.1. SAS Calculations for Two-Sample Crossover Design

Command `sides` has to be determined following Table 2, other commands following Table 15. Option `twosamplemeans` is needed for two-sample analysis. The sample size needed is obtained by finding the sample size per group (`npergroup`), for total sample size that has to be doubled. For crossover design, standard deviation has to be divided by two, i.e. $\sigma/2$. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used.

```
proc power;
   twosamplemeans
   test=diff
   sides=
   meandiff=
   nulldiff=
   stddev=
   alpha=
   power=
   npergroup=.;
run;
```

Table 15 SAS commands for two-sample crossover design

Option	Value
meandiff	$\mu_E - \mu_C$
nulldiff	$-\delta_{NI}$
stddev	$\sigma/2$ or $s/2$
alpha	α
power	$1-\beta$

5.2.2. R Calculations for Two-Sample Crossover Design

For non-inferiority two-sample crossover design, R has the function TwoSampleCrossOver.NIS from package TrialSize. It is important to notice that R gives sample size per sequence. Arguments have to be determined following Table 16.

TwoSampleCrossOver.NIS(alpha, beta, sigma, delta, margin)

Table 16 R arguments for two-sample crossover design

Option	Value
alpha	α
beta	β
sigma	σ
delta	$-\delta_{NI}$
margin	$\mu_E - \mu_C$

5.2.3. An Example for Two-Sample Crossover Non-Inferiority Trial

The same example is used as for the parallel design. A pharmaceutical company is interested in having an 80% power for establishing non-inferiority of their new treatment. It is expected that the mean value for the experimental treatment is 0.3 units while the same indicator for the control treatment is 0.2 units, the variance $\sigma^2 = 0.2$ units^2. That means that new treatment is supposed to increase value by 0.1 units, i.e. effect size $\mu_E - \mu_C$. The clinically meaningful difference for this study is 0.2 units.

The sample size is now found using three different methods (Table 17).

Table 17 Sample size calculations for two-sample crossover non-inferiority design

Method	Formula/syntax	Result
$n_1 = n_2 = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{2(\mu_E - \mu_C + \delta_{NI})^2}$	$\frac{(z_{0.05} + z_{0.2})^2 0.2}{2(0.3 - 0.2 + 0.2)^2} = \frac{(1.6449 + 0.8416)^2 0.2}{2(0.1 + 0.2)^2} = 6.869647$	$6.869647 \approx \approx 7$

SAS	`proc power;`
	`twosamplemeans`
	`test=diff`
	`sides=U`
	`meandiff=0.1`
	`nulldiff=-0.2`
	`stddev=0.2236068`
	`alpha=0.05`
	`power=0.8`
	`npergroup=.;`
	`run;`
		8

R	TwoSampleCrossOver.NIS(alpha=0.05, beta=0.2, sigma=0.4472136, delta=-0.2, margin=0.1)	$6.869508 \approx \approx 7$

With 80% power and a one-sided type I error level set to 5% a total of 16 patients, 8 in each sequence, would be required in order to show that the experimental treatment is at most 0.2 units inferior to the control treatment when the expected effect size is 0.1 units and the variance $\sigma^2 = 0.2$ units^2 (as determined by SAS software). The results differ slightly as a result of rounding etc.

6. Equivalence Trials

It has been mentioned before that it is practically impossible for two treatments to be exactly equal. In clinical trials, equivalence means that the effects of two treatments may not differ more than tolerable, i.e. the effect difference of treatments stays within a small determined interval. Therefore, the aim of equivalence trials is to show that the differences are not substantial in either direction. Conducting clinical trial as an equivalence trial is reasoned when the experimental drug is believed to be safer or cheaper to produce. This still applies when the therapeutic effect of the test drug is not assumed to be as large as for the active control. [6]

Defining an appropriate interval for equivalence can be difficult and controversial. When the difference in population means $\mu_E - \mu_C$ stays within the interval of equivalence limit, i.e. $\pm\delta_E$, the equivalence is shown. It is desirable to limit the approval of a test drug that is inferior to a standard drug as much as possible. Therefore, the defined interval should be rather narrow. It is recommended to define δ_E so that it is no more than one-half of the value that would be used in a superiority trial. [6] [19]

With equivalence trial there is no internal control for validity because equivalence does not indicate that one of the treatments – experimental or control – is superior to another. The drug used as an active control should be proven beforehand to be superior in a placebo-controlled trial. [19]

Nowadays, many of the equivalence trials are bioequivalence trials. Bioequivalence describes the relationship between two products when they are pharmaceutically equivalent and in the same dosage have the similar bioavailability – in both rate and extent of which is demonstrated with peak concentration and area under the time-concentration curved being equivalent. The aim of bioequivalence trials is to compare a generic drug to an already existing commercial drug that is going off-patent. After the approval of a generic drug, it can be used as a substitute to a commercial drug. These trials are most commonly carried out using crossover design. [1]

In equivalence trials the hypotheses are defined a little differently. It is standard that when the null hypothesis is not rejected there is no difference between comparable means. For equivalence trials, null hypothesis states that there is at least a difference of δ_E while alternative hypothesis states that the difference is smaller.

$$H_0: |\mu_E - \mu_C| \geq \delta_E$$

$$H_1: |\mu_E - \mu_C| < \delta_E,$$

where

- μ_E is the mean of the primary endpoint for the experimental treatment;
- μ_C is the mean of the primary endpoint for the control treatment;
- δ_E is the equivalence limit. [1]

6.1. Two-Sample Parallel Design

For two-sample parallel design, let the hypothesis for the equivalence trial be

$$H_0: |\mu_\mathrm{E} - \mu_\mathrm{C}| \geq \delta_E$$

$$H_1: |\mu_\mathrm{E} - \mu_\mathrm{C}| < \delta_E,$$

where

- $\mu_\mathrm{C}$is the true mean response of a control treatment;
- $\mu_\mathrm{E}$is the true mean response of an experimental treatment;
- δ_E is the equivalence limit. [1]

The sample size formulas for the treatment group and the control group are

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + \mathrm{z}_{\beta/2}\right)^2 \sigma^2 (1 + 1/\kappa)}{(\delta_E - |\mu_\mathrm{E} - \mu_\mathrm{C}|)^2},$$

where $\kappa = n_1/n_2$ shows treatment allocations, usually 1:1 or 2:1. $z_{\beta/2}$ is the upper $\beta/2$th-quantile and z_α is the upper αth-quantile of the standard normal distribution. [1] For the derivation of this formula, see an example in Appendix 3.

When the population variances σ_1^2 and σ_2^2 are not equal then the sample size is given by

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + \mathrm{z}_{\beta/2}\right)^2 \left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)}{(\delta_E - |\mu_\mathrm{E} - \mu_\mathrm{C}|)^2},$$

where σ_1^2 is the variance of the experimental treatment group and σ_2^2 is the variance of the control treatment group.

Above formulas assume that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formulas given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + \mathrm{z}_{\beta/2}\right)^2 s^2 (1 + 1/\kappa)}{(\delta_E - |\mu_\mathrm{E} - \mu_\mathrm{C}|)^2}.$$

6.1.1. SAS Calculations for Two-Sample Parallel Design

Command have to be determined following Table 18. Option `twosamplemeans` is needed for two-sample analysis. With `test=equiv_diff` it is stated that means of two samples are compared. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used. For unequal variances σ is replaced by $\sigma_{pooled} = \sqrt{\left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)/\left(1 + \frac{1}{\kappa}\right)}$. The equivalence limit is specified by `lower` and `upper`. Statement `groupweights` is needed for treatment allocation, for example 1:2 treatment allocation is denoted as `groupweights=1|2`. If it is desired to get sample size for the whole study, `ntotal` should be used, for getting sample size for each study group separately, `npergroup` is used. It is important to notice that options `groupweights` and `npergroup` cannot be used together. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used.

```
proc power;
   twosamplemeans
   test=equiv_diff
   groupweights=
   meandiff=
   upper=
   lower=
   stddev=
   alpha=
   power=
   ntotal=.;
run;
```

Table 18 SAS commands for two-sample parallel design

Option	Value
meandiff	$\mu_E - \mu_C$
stddev	σ or s or σ_{pooled}
alpha	α
power	$1 - \beta/2$
lower	$-\delta_E$
upper	$+\delta_E$
groupweights	$n_1 \vert n_2$

6.1.2. R Calculations for Two-Sample Parallel Design

For equivalence two-sample parallel design, R has the function TwoSampleMean.Equivalence from package TrialSize. It is important to notice that R gives sample size per group. Also, with two-sample design, treatment allocation $\kappa = n_1/n_2$ has to be considered. It is assumed that σ is known. If not, σ is replaced by s and the same syntax can still be used. For unequal variances σ is replaced by $\sigma_{pooled} = \sqrt{\left(\frac{\sigma_1^2}{\kappa} + \sigma_2^2\right)/\left(1 + \frac{1}{\kappa}\right)}$. With unequal study groups, e.g. $\kappa = n_1/n_2 = 1/2 = 0.5$, R calculates n_1 and the total sample size $n = n_1 + n_2 = n_1 + \frac{n_1}{\kappa}$. For the usual parallel study with equal study groups ($\kappa = n_1/n_2 = 1$), the sample size obtained with R simply has to be doubled to get the total sample size. Arguments have to be determined following Table 19.

TwoSampleMean.Equivalence(alpha, beta, sigma, k, delta, margin)

Table 19 R arguments for two-sample parallel design

Option	Value
alpha	α
beta	β
sigma	σ or s or σ_{pooled}
k	n_1/n_2
delta	δ_E
margin	$\mu_E - \mu_C$

6.1.3. An Example for Two-Sample Parallel Equivalence Trial

A pharmaceutical company is interested in having an 80% power for proving the equivalence of an experimental and a control treatment. The true mean difference is thought to be 0.01 units and the equivalence limit is considered to be 0.05 units, the variance $\sigma^2 = 0.01$ units^2. Let both treatment groups be equal.

The sample size is now found using three different methods (Table 20).

Table 20 Sample size calculations for two-sample parallel equivalence design

Method	Formula/syntax	Result
$n_1 = \kappa n_2,$ $n_2 = \frac{(z_\alpha + z_{\beta/2})^2 \sigma^2 (1 + 1/\kappa)}{(\delta_E - \lvert\mu_E - \mu_C\rvert)^2}$	$n_2 = \frac{(z_{0.05} + z_{0.1})^2 0.01(1 + 1/1)}{(0.05 - \lvert 0.01\rvert)^2} =$ $= \frac{(1.6449 + 1.2816)^2 0.01(1 + 1/1)}{(0.05 - \lvert 0.01\rvert)^2}$ $= 107.0516$ $n_1 = 1 * 107.0516 = 107.0516$	$107.0516 +$ $+107.0516 =$ $= 214.1033 \approx$ ≈ 216

SAS	`proc power;`
	`twosamplemeans`
	`test=equiv_diff`
	`groupweights=1|1`
	`meandiff=0.01`
	`upper=0.05`
	`lower=-0.05`
	`stddev=0.1`
	`alpha=0.05`
	`power=0.9`
	`ntotal=.;`
	`run;`
		218

R	TwoSampleMean.Equivalence(alpha=0.05, beta=0.2, sigma=0.1, k=1, delta=0.05, margin=0.01)	$107.0481 +$ $+107.0481 =$ $= 214.0962 \approx$ ≈ 216

With 80% power and a type I error level set to 5% a total of 218 patients, 109 in each group, would be required in order to detect the clinically meaningful difference of 0.05 units when the expected effect size is 0.01 units and the variance $\sigma^2 = 0.01$ units^2 (as obtained with SAS software). The results differ slightly as a result of rounding etc.

6.2. Two-Sample Crossover Design

In two-sample 2x2 crossover design subjects are randomized into two groups – experimental and control. During the first period, one group gets experimental and the other one gets control treatment. During the second period, the subjects who got experimental treatment before now get control treatment and *vice versa*. That way each participant is one's own control. Let the hypothesis for the equivalence trial be

$$H_0: |\mu_E - \mu_C| \geq \delta_E$$

$$H_1: |\mu_E - \mu_C| < \delta_E,$$

where

- μ_C is the true mean response of a control treatment;
- μ_E is the true mean response of an experimental treatment;
- δ_E is the equivalence limit. [1]

The formula for calculating sample size for each sequence is

$$n_1 = n_2 = \frac{\left(z_\alpha + z_{\beta/2}\right)^2 \sigma^2}{2(\delta_E - |\mu_E - \mu_C|)^2},$$

where $z_{\beta/2}$ is the upper $\beta/2$th-quantile and z_α is the upper αth-quantile of the standard normal distribution. [1] [11]

Above formula assumes that the data is normally distributed and σ^2 is known. In practice, σ^2 is often unknown and hence the test statistic which is used to test hypotheses and derive sample size, has t-distribution. However, having n sufficiently large, the formulas given above can also be used when σ^2 is unknown. The required sample size is then calculated by replacing σ^2 by the sample variance s^2:

$$n_1 = n_2 = \frac{\left(z_\alpha + z_{\beta/2}\right)^2 s^2}{2(\delta_E - |\mu_E - \mu_C|)^2}.$$

6.2.1. SAS Calculations for Two-Sample Crossover Design

Commands have to be determined following Table 21. Option `twosamplemeans` is needed for two-sample analysis. The sample size needed is obtained by finding the sample size per group (`npergroup`), for total sample size that has to be doubled. The equivalence limit is specified by `lower` and `upper`. For crossover design, standard deviation has to be divided by two, i.e. $\sigma/2$.

```
proc power;
   twosamplemeans
   test=equiv_diff
   meandiff=
   upper=
   lower=
   stddev=
   alpha=
   power=
   npergroup=.;
run;
```

Table 21 SAS commands for two-sample crossover design

Option	Value
meandiff	$\mu_E - \mu_C$
stddev	$\sigma/2$
alpha	α
power	$1 - \beta/2$
lower	$-\delta_E$
upper	δ_E

6.2.2. R Calculations for Two-Sample Crossover Design

For equivalence two-sample crossover design, R has a function TwoSampleCrossOver.Equivalence from package TrialSize. It is important to notice that R gives sample size per sequence. Right now, the value of β has to be divided by two[3], i.e. $\beta/2$. Arguments have to be determined following Table 22.

TwoSampleCrossOver.Equivalence(alpha, beta, sigma, delta, margin)

Table 22 R arguments for two-sided crossover design

Option	Value
alpha	α
beta	$\beta/2$
sigma	σ
delta	δ_E
margin	$\mu_E - \mu_C$

6.2.3. An Example for Two-Sample Crossover Equivalence Trial

The same example is used as for the parallel design. A pharmaceutical company is interested in having an 80% power for proving the equivalence of a new experimental treatment. The expected true mean difference between experimental and control treatment is 0.01 units and the equivalence limit is considered to be 0.05 units, the variance $\sigma^2 = 0.01$ units^2.

The sample size is now found using three different methods (Table 23).

[3] The author of this thesis discovered a contradiction in the package TrialSize related to equivalence crossover design and contacted the R team who confirmed the issue. The change will be made with the next package update (planned in May 2016). Due to that, the syntax given may change.

Table 23 Sample size calculations for two-sample crossover equivalence design

Method	Formula/syntax	Result
$n_1 = n_2 = \\ = \frac{(z_\alpha + z_{\beta/2})^2 \sigma^2}{2(\delta_E - \lvert\mu_E - \mu_C\rvert)^2}$	$\frac{(z_{0.05} + z_{0.1})^2 0.01}{2(0.05 - 0.01)^2} = \\ = \frac{(1.6449 + 1.2816)^2 0.01}{2(0.05 - 0.01)^2} = 26.76376$	$26.76376 \approx \\ \approx 27$

SAS	`proc power;`
	`twosamplemeans`
	`test=equiv_diff`
	`meandiff=0.01`
	`upper=0.05`
	`lower=-0.05`
	`stddev=0.05`
	`alpha=0.05`
	`power=0.9`
	`npergroup=.;`
	`run;`
		28

R	TwoSampleCrossOver.Equivalence(alpha=0.05, beta=0.1, sigma=0.1, delta=0.05, margin=0.01)	$26.76202 \approx \\ \approx 27$

With 80% power and a type I error level set to 5% a total of 56 patients, 28 in each sequence, would be required in order to detect the clinically meaningful difference of 0.05 units when the expected effect size is 0.01 units and the variance $\sigma^2 = 0.01$ units^2 (as obtained with SAS software). The results differ slightly as a result of rounding etc.

Conclusion

The aim of this thesis was to give guidelines for determining the sample size in clinical trials. First three chapters gave a short overview of clinical trials and what has to be considered when finding sample size. The three following chapters discussed the main trial types and distinct methods for calculations.

The progress of obtaining sample size is of great importance in pharmaceutical industry and the determination progress should be a collaboration of clinicians and statisticians. Every trial type requires its own approach and points to consider. Superiority trials are the easiest to conduct and to interpret. For those trials, both one-sample and two sample (parallel and crossover) designs were discussed. For more complicated trials – non-inferiority and equivalence –, only two-sided designs were explained as the one-sample design is not used in practice for those two. Also, the progress of choosing non-inferiority margin and equivalence limit were more thoroughly discussed.

Working with SAS or R, two most important aspects have to be considered: the choice of proper parameters and the correct interpretation as these may differ depending on the program used. For every method, a short example was included for the best understanding of using the directions provided.

References

[1] S.-C. Chow, J. Sao and H. Wang, Sample Size Calculations in Clinical Research, Chapman and Hall/CRC, 2008.

[2] "Estonian Agency of Medicines," [Online]. Available: www.ravimiamet.ee. [Accessed 18 April 2016].

[3] B. Parkson, "Inglise-eesti seletav kliiniliste uuringute sõnastik," 2012.

[4] B. Zhong, "How to Calculate Sample Size in Randomized Controlled Trial?," *Journal of Thoracic Disease,* 2009.

[5] ICH, "Guidance for Industry: E9 Statistical Principles for Clinical Trials," 1998.

[6] E. Christensen, "Methodology of Superiority vs. Equivalence Trials and Non-Inferiority Trials," *Journal of Hepatology,* 2007.

[7] A. Dasgupa, K. A. Lawson and J. P. Wilson, "Evaluating Equivalence and Noninferiority Trials," *American Journal of Health-System Pharmacy*, 2010.

[8] B. Röhrig, J.-B. d. Prel, D. Wachtlin, R. Kwiecien and M. Blettner, "Sample Size Calculations in Clinical Trials," *Deutsches Ärzteblatt International*, 2010.

[9] E. Lesaffre, "Superiority, Equivalence, and Non-Inferiority Trials," *Bulletin of the Hospital for Joint Disease,* 2008.

[10] "SAS® Customer Support," [Online]. Available: http://support.sas.com/kb/48/616.html. [Accessed 20 April 2016].

[11] B. Jones and M. G. Kenward, Design and Analysis of Cross-Over Trials, Chapman and Hall/CRC, 2014.

[12] SAS Institute Inc., "The POWER Procedure," in *SAS/STAT® 9.2 User's Guide*, 2008.

[13] S. Champely, C. Ekstrom, P. Dalgaard, J. Gill, J. Wunder and H. D. Rosario, "Package 'pwr'," 2015.

[14] E. Zhang, V. Q. Wu, S.-C. Chow and H. G. Zhang, "Package 'TrialSize'," 2015.

[15] European Medicines Agency, "Points to Consider on Switching Between Superiority and Non-Inferiority," 2000.

[16] D. G. Altman and J. M. Bland, "Absence of Evidence is not Evidence of Absence," *The BMJ,* 1995.

[17] European Medicines Agency, "Guidline on the Choice of the Non-Inferiority Margin," 2005.

[18] S. Hahn, "Understanding Noninferiority Trials," *Korean Journal of Pediatrics,* 2012.

[19] "Lesson 6: Sample Size and Power - Part B.3 Equivalence Trials," [Online]. Available: https://onlinecourses.science.psu.edu/stat509/node/52. [Accessed 18 April 2016].

Appendices

Appendix 1

For one-sample superiority design, the one-sided hypotheses are

$$H_0: \mu - \mu_0 \leq \delta$$

$$H_1: \mu - \mu_0 > \delta,$$

where

- μ is the true mean response of a test drug;
- μ_0 is a reference value.

For sample size calculations, z test is used. When σ is known, the distribution of test statistic given null hypothesis is

$$Z \sim N(0,1).$$

The null hypothesis is rejected when the calculated test statistic z

$$z = \frac{\bar{x} - \mu_0 - \delta}{\sigma_n} > z_\alpha,$$

where

- z_α is the upper αth quantile of the standard normal distribution;
- $\bar{x}$ is sample mean;
- σ_n is the standard error of mean.

The power formula for given hypotheses is

$$Power = 1 - \beta = Pr\left(\frac{\bar{X} - \mu_0 - \delta}{\sigma_n} > z_{1-\alpha} \middle| H_1\right).$$

Definition of standard normal distribution function being used

$$1 - \beta = \Phi\left(\frac{\mu - \mu_0 - \delta}{\sigma_n} - z_{1-\alpha}\right).$$

Definition of standard normal quantiles being used

$$z_{1-\beta} = \frac{(\mu - \mu_0 - \delta)}{\sigma_n} - z_{1-\alpha}.$$

A little algebra and that $z_\alpha = -z_{1-\alpha}$

$$-z_\beta = \frac{\mu - \mu_0 - \delta}{\sigma_n} + z_\alpha$$

$$\frac{1}{\sigma_n} = \frac{-z_\beta - z_\alpha}{\mu - \mu_0 - \delta}$$

Having $\sigma_n = {}^{\sigma}/_{\sqrt{n}}$, the sample size equals

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{(\mu - \mu_0 - \delta)^2}$$

Appendix 2

For one-sample superiority design, the two-sided hypotheses are

$$H_0|: \mu - \mu_0| \leq \delta$$

$$H_1: |\mu - \mu_0| > \delta,$$

where $\delta > 0$ and

- μ is the true mean response of a test drug;
- μ_0 is a reference value.

For sample size calculations, z test is used. When σ is known, the distribution of test statistic given null hypothesis is

$$Z \sim N(0,1).$$

The null hypothesis is rejected when calculated test statistic z

$$z = \left|\frac{\bar{x} - \mu_0 - \delta}{\sigma_n}\right| > z_{\alpha/2},$$

where $z_\alpha$is the upper αth quantile of the standard normal distribution and $\bar{x}$ is sample mean.

The power formula for given hypotheses is

$$Power = 1 - \beta = Pr\left(\left|\frac{\bar{X} - \mu_0 - \delta}{\sigma_n}\right| > z_{1-\alpha/2}\middle| H_1\right).$$

Definition of standard normal distribution function being used

$$1 - \beta = \Phi\left(\frac{\mu - \mu_0 - \delta}{\sigma_n} - z_{1-\alpha/2}\right) + \Phi\left(-\frac{\mu - \mu_0 - \delta}{\sigma_n} - z_{1-\alpha/2}\right).$$

Because of using two-sided test

$$1 - \beta = \Phi\left(\frac{|\mu - \mu_0 - \delta|}{\sigma_n} - z_{1-\alpha/2}\right).$$

Definition of standard normal quantiles being used and $z_\alpha = -z_{1-\alpha}$.

$$-z_\beta = \frac{|\mu - \mu_0 - \delta|}{\sigma_n} + z_{\alpha/2}$$

$$\frac{1}{\sigma_n} = \frac{-z_\beta - z_{\alpha/2}}{|\mu - \mu_0 - \delta|}$$

Having $\sigma_n = {\sigma}/{\sqrt{n}}$, the sample size equals

$$n = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \sigma^2}{(\mu - \mu_0 - \delta)^2}.$$

Appendix 3

For two-sample parallel superiority design, the one-sided hypotheses are

$$H_0: \mu_E - \mu_C \leq \delta$$

$$H_1: \mu_E - \mu_C > \delta,$$

where

- μ_E is the mean of the primary endpoint for the experimental treatment;
- μ_C is the mean of the primary endpoint for the control treatment.

For sample size calculations, z test is used. When σ is known, the distribution of test statistic given null hypothesis is

$$Z \sim N(0,1).$$

In parallel design, two sample means – $\bar{x}_1$ and $\bar{x}_2$ (estimates of population means μ_E and μ_C, respectively) – are being observed. To reject the null hypothesis using z test, the following has to be true:

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_\alpha,$$

where n_1 and n_2 are the sample sizes for each group.

The power formula for given hypotheses is

$$Power = 1 - \beta = Pr\left(\frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{1-\alpha} \middle| H_1\right).$$

Definition of standard normal distribution function being used

$$1 - \beta = \Phi\left(\frac{\mu_E - \mu_C - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - z_{1-\alpha}\right).$$

Definition of standard normal quantiles being used

$$z_{1-\beta} = \frac{\mu_E - \mu_C - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - z_{1-\alpha}.$$

A little algebra and that $z_\alpha = -z_{1-\alpha}$.

$$-z_\beta = \frac{\mu_E - \mu_C - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} + z_\alpha$$

$$\frac{1}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{-z_\beta - z_\alpha}{\mu_E - \mu_C - \delta},$$

which leads to

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2 (1 + 1/\kappa)}{(\mu_E - \mu_C - \delta)^2},$$

where $\kappa = n_1/n_2$ demonstrates treatment allocations.

Appendix 4

For two-sample parallel superiority design, the one-sided hypotheses are

$$H_0|: \mu_E - \mu_C| \leq \delta$$

$$H_1: |\mu_E - \mu_C| > \delta,$$

where $\delta > 0$ and

- μ_E is the mean of the primary endpoint for the experimental treatment;
- μ_C is the mean of the primary endpoint for the control treatment.

For sample size calculations, z test is used. When σ is known, the distribution of test statistic given null hypothesis is

$$Z \sim N(0,1).$$

To reject the null hypothesis using z test, the following has to apply:

$$z = \left| \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > z_{\alpha/2},$$

where n_1 and n_2 are the sample sizes for each group.

The power formula for given hypotheses is

$$Power = 1 - \beta = Pr\left(\frac{\bar{X}_1 - \bar{X}_2 - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{\alpha/2} \middle| H_1 \right).$$

Definition of standard normal distribution function being used

$$1 - \beta = \Phi\left(\frac{\mu_E - \mu_C - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - z_{1-\alpha/2} \right) + \Phi\left(-\frac{\mu_E - \mu_C - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - z_{1-\alpha/2} \right).$$

Because the one-sided test is used

$$1 - \beta = \Phi\left(\frac{|\mu_E - \mu_C - \delta|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} - z_{1-\alpha/2} \right).$$

A little algebra and that $z_\alpha = -z_{1-\alpha}$.

$$-z_\beta = \frac{|\mu - \mu_0 - \delta|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} + z_{\alpha/2}$$

that leads to

$$n_1 = \kappa n_2,$$

$$n_2 = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \sigma^2 (1 + 1/\kappa)}{(\mu_E - \mu_C - \delta)^2},$$

where $\kappa = n_1/n_2$ demonstrates treatment allocations.

Non-exclusive licence to reproduce thesis and make thesis public

I, Lisbeth Neevits,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

"Sample Size Calculations in Clinical Trials",

supervised by Marju Valge and Pasi Antero Korhonen

2. I am aware of the fact that the author retains these rights

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **29.04.2016**