Sample size determination

Ingeborg van der Tweel

dept. Biostatistics
Julius Centre for Health Sciences and Primary Care
University Medical Centre Utrecht
T: +31 88 755 9366
F: +31 88 756 8099
E: SecretariaatBiostatistiek@umcutrecht.nl
H: http://www.juliuscentrum.nl/julius/Services/Biostatistics/tabid/1005/Default.aspx

Intern report nr 4

October, 2006

CONTENTS

1 Introduction

Reflection upon the evaluation of sample size when planning a randomised clinical trial (RCT) is widely recognized as a matter of Good Statistical Practice. Less well known is that observational studies and pilot studies can also benefit from thinking about the required number of patients or animals in the design phase of the study. If the sample size is too small, important treatment differences can easily go undetected. This does not mean, however, that an investigator should enrol as many patients or animals as possible in the study. If the number of patients or animals exceeds the number required, the study will be unnecessarily expensive, prolonged and sometimes unethical. An investigator will have to strike a balance between enrolling sufficient patients or animals to detect relevant treatment effects, but not so many that important resources (patients or animals, time, money, etc.) are wasted.

The statistical concepts that play a role in sample size or power calculations are introduced in section 2. Sections 3, 4 and 5 discuss the determination of sample size for various outcome variables. First, sample size calculations are given for a qualitative, dichotomous outcome variable. Next, sample size calculations are given for a quantitative, continuous outcome variable. And third, sample size calculations are given for a survival outcome or time-to-event type variable. Section 6 addresses the ideas behind Cohen's effect size. Section 7 goes into sequential design and analysis as an efficient alternative for some studies. Section 8 offers some closing remarks and conclusions.

In the following sections we will use the term 'study' both for a RCT and for an observational or pilot study; the term 'patients' is used for reasons of convenience, but can be replaced by animals, plants, tissue samples, etc.

2 Statistical concepts

In the following sections we assume that there is one primary outcome variable in one or two samples. The case of more than two treatment groups will be discussed briefly in section 8.

Let the treatment difference or effect size to be estimated based on the results of the study be characterized by a parameter δ. The standard reasoning in calculating sample size proceeds as follows. Suppose some statistical test will be performed at the end of the study for the null hypothesis H_o: $\delta = 0$, i.e. the treatment difference or effect size equals zero. The quantity α, also called the type I error of a statistical test, is the probability of detecting a significant difference when in reality no difference exists, i.e. it represents the risk of a false positive (FP) result. Usually α is set at a value of 0.05, one-sided or two-sided. Next one has to choose a value δ_0, the smallest value of the treatment effect that should not go undetected. In other words, if δ_0 is the true treatment effect, there should be a high probability, say 1-ß, of rejecting the test of H_o: $\delta = 0$. The quantity δ_0 is often called the minimal (clinically) relevant treatment effect. The quantity 1-ß represents the degree of certainty with which the treatment effect δ_0 would be detected, and is called the *power* to detect a difference of magnitude δ_0. The quantity ß, commonly called the type II error, is the probability of not detecting a significant treatment effect when there is a true treatment effect of magnitude δ_0, i.e. ß is the risk of a false negative (FN) result. The foregoing is summarized in Table 1.

Table 1. Relation between the statistical test result and the (unknown) reality

		(unknown) reality	
		H_0 true	H_1 true
statistical test result	do not reject H_0	1-α	β (FN)
	reject H_0	α (FP)	1-β (power)

The relevant treatment effect is specified under the alternative hypothesis H_1: $\delta \geq \delta_0$. This is a one-sided alternative hypothesis, to be tested with a one-sided value of α. A two-sided alternative hypothesis would be formulated as H_1: $|\delta| \geq \delta_0$ and would be tested with a two-sided value of α.

When the outcome variable is continuous, its variability σ has to be specified, in addition to α, ß and δ_0.

Based on these specifications and the statistical test to be used, the minimal number of patients that is needed can be calculated. Note that there is a very close relation between sample size calculation in the planning phase of a study and the statistical test used in the analysis phase.

In the sample size formulae the dependence on α and β is expressed mathematically by quantities Z_α and $Z_ß$. For an arbitrary number γ between 0 and 1, $Z\gamma$ denotes the value such that a standard normal deviate has exactly the probability γ of exceeding that value. Table 2 gives the value of $Z\gamma$ for some frequently used values of γ.

Table 2. Values of $Z\gamma$ corresponding to specified values of γ

γ	$Z\gamma$
0.50	0
0.40	0.25
0.30	0.52
0.20	0.84
0.10	1.28
0.05	1.65
0.025	1.96
0.01	2.32
0.005	2.58

Unless otherwise stated, we assume that the statistical tests are two-sided. That means that for a statistical test with a two-sided α of 0.05, one should use $Z_{0.025} = 1.96$. For a one-sided test with $\alpha = 0.05$, $Z_{0.05} = 1.65$ is used. For a test with a power of 0.80, $Z_{0.20} = 0.84$ is used; for a power of 0.90, $Z_{0.10} = 1.28$ is used.

Figure 1 depicts the relation between the null hypothesis (H_0) and the alternative hypothesis (H_1), the type I error α and the power $1-\beta$. The sample size n depends on the variability σ in the outcome variable (when continuous), the effect size δ_0 to be detected, and the type I and type II errors. A larger variability requires a larger sample size to detect a given effect size. A large effect size requires a smaller sample size than a small effect size. When the effect size is fixed, a smaller value for α (i.e. a diminished probability of a false positive result) requires a larger sample size. The same can be said for the power: a large power (or a small probability of a false negative result) requires a larger sample size.

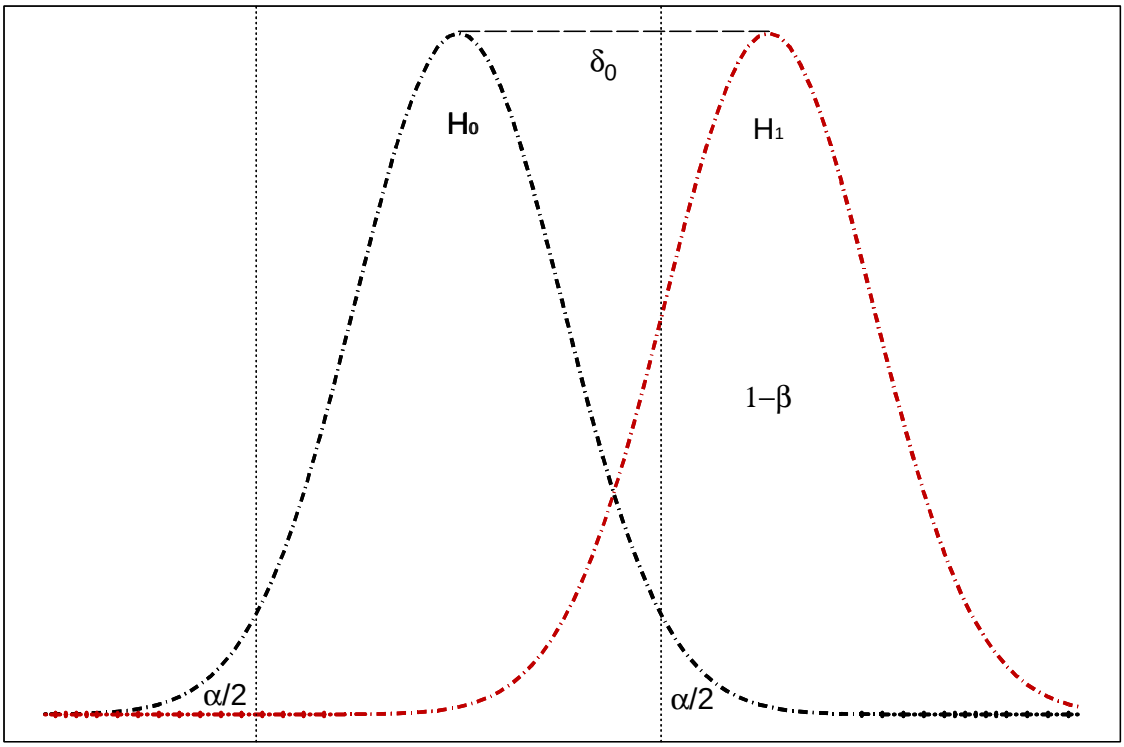

Figure 1. Relation between the null (H_0) and the alternative hypothesis (H_1), the two-sided type I error α and the power $1-\beta$

3 Dichotomous outcome

We consider the case of a dichotomous outcome variable, i.e. one that can be classified as ‘success’ or ‘failure’.

3.1 One sample

For one group with a dichotomous outcome variable, the sample size n can be estimated by

$$n \geq \frac{(Z_\alpha \sqrt{p_0(1-p_0)} + Z_\beta \sqrt{p_1(1-p_1)})^2}{(p_1 - p_0)^2}$$

with

p_0 = the ‘success’ rate assumed under the null hypothesis H_o;

p_1 = the ‘success’ rate assumed under the alternative hypothesis H_1.

(Note that one can also use ‘failure’ rate instead of ‘success’ rate.)

Example:
In a certain (high risk) population the incidence of thrombosis is assumed to be 30% (i.e. H_o: $p_0 = 0.3$). We want to sample a group large enough to detect an incidence of 40% (i.e. H_1: $p_1 = 0.4$), if that is the real, true incidence. We set the two-sided α at 0.05 and want a power of 0.90 to detect the difference in incidence. Then the sample size n should be at least 233 patients.

The appropriate statistical test in this situation is a binomial test for one proportion.

3.2 Two samples

When two independent groups are to be compared with respect to a dichotomous variable, the sample size per group can be estimated by

$$n \geq \frac{p_C(1-p_C)+p_E(1-p_E)}{\delta_0^2}(Z_\alpha + Z_\beta)^2$$

with

p_C = the 'success' rate in the control group;

p_E = the 'success' rate in the experimental group;

$\delta_0 = p_E - p_C$ = the relevant treatment effect to detect.

Example:
In a randomised, clinical trial the control treatment is assumed to have a 'success' rate of 0.2. The experimental treatment is expected to lead to a 'success' rate of at least 0.4. To detect this difference of 0.2 with a two-sided α of 0.05 and a power of 0.80, at least 79 patients per group are required.

The sample size calculation above can also be applied when the null hypothesis and the alternative hypothesis are expressed in terms of a relative risk (RR). If p_C was known or assumed to be known, then $p_E = RR \times p_C$.
Some general remarks should be made with regard to the formula above. First, for the sample size estimation in studies with two samples we assume equal group sizes, which is the statistically most efficient design. Second, the sample size is inversely proportional to δ_0^2, which means that, for fixed α and ß, halving the difference in success rates will require about a fourfold increase in the group size. Third, the factor p(1-p) is related to the variance of a dichotomous outcome variable. The variability is highest for p=0.5 and decreases if p tends to 0 or 1. Since p(1-p) is thus always smaller than or equal to 0.25, an upper bound for the sample size can be given by the simpler formula

$$n \geq \frac{(Z_\alpha + Z_\beta)^2}{2\delta_0^2}$$

This approximation can be very valuable in practice and is reasonably good when p does not differ greatly from 0.5 ($0.3 < p < 0.7$).

The appropriate statistical test in this situation is a chi-square test for the comparison of two proportions.

3.3 'Equivalence' in two samples

The sample size formulae given in section 3.2 cannot be applied when 'equivalence' is to be detected between two groups with respect to the dichotomous outcome variable. Real equivalence of two groups can never be concluded based on two samples, but one can specify the maximal difference δ_0 in the outcome variable for which the two treatments can be considered equivalent. Then the sample size per group can be estimated by

$$n \geq \frac{2p(1-p)\,(Z_\alpha + Z_\beta)^2}{\delta_0^2}$$

with p = the 'success' rate of both treatments ($p = p_E = p_C$).

Example:
The control treatment for hypertension is drug treatment. An investigator wants to compare this to an experimental intervention consisting of, amongst others, diet and relaxation therapy. She anticipates that both therapies will lead to control of blood pressure in about 80% of the patients. She regards the treatments as equivalent if the proportion of experimental patients with their hypertension under control is at most 10 % lower than in the drug treatment group. The required number of patients per group, with a two-sided α of 0.05 and a power 1-β of 0.80 is equal to 252.
If a difference of 10% is too large for 'equivalence' and 5% is considered more appropriate, the number of patients per group becomes at least 1005.
It is clear that 'true' equivalence requires a large number of subjects.

One should realize that in the equivalence situation described above, the roles of the null and the alternative hypothesis have, in fact, changed. Now the null hypothesis is formulated as "the two treatments differ by an amount of at least δ_0", while the alternative hypothesis becomes "the two treatments can be considered equivalent," or "the two

treatments differ by an amount of at most δ_0". Also the roles of α and β are reversed and they must be chosen with some thought. However, the impact of this on sample size may not be very large.

3.4 Paired samples

When two observations are made on the same individuals, these observations cannot be assumed to be independent of each other. The statistical test and the sample size estimation will have to take into account this dependency. For a dichotomous outcome variable with values 'positive' (+) and 'negative' (-), four combinations of observations are possible for each individual:

	observation nr 1	observation nr 2
combination 1	**+**	**+**
combination 2	**+**	**-**
combination 3	**-**	**+**
combination 4	**-**	**-**

Only the combinations 2 and 3 are informative, however, for a possible difference between the two observations. For a sample of individuals the results can be summarized as in Table 3.

Table 3. Combination of paired observations

		observation nr 2	
		+	**-**
observation nr 1	**+**		r
	-	s	

The expected difference can be expressed as a fraction of the total number of individuals with combination 2 or 3 or as an odds ratio (OR). The total number of informative pairs is equal to r + s = n. The OR is defined as r/s.

The necessary number of individuals with combination 2 or 3 can be estimated by

$$n \geq \frac{4(Z_\alpha + Z_\beta)^2}{\{\log(OR)\}^2}$$

with OR = the relevant odds ratio to detect.
Under the null hypothesis the OR is equal to 1.

Example:
We want to compare two tests A and B with respect to their dichotomous outcome 'positive' or 'negative' in a group of individuals.
To detect an OR equal to 2 with a two-sided α of 0.05 and a power of 0.80, at least 66 informative pairs of observations are necessary.

The expected difference between the two observations can also be expressed as a fraction r/n. Under the null hypothesis this fraction is assumed to equal 0.5. An OR of 2 corresponds to a fraction r/n of 0.667. For the example this means that test A is positive and test B is negative in two-thirds of the total number of informative individuals.

The appropriate statistical test in this situation is McNemar's test.

4 Continuous outcome

In this section we assume that the continuous outcome variable follows a normal distribution. A normal distribution is characterized by a mean μ and a standard deviation σ. The standard deviation σ is almost always unknown in practice. It must be estimated by a value s derived from one's previous research or from the literature.

4.1 One sample

The sample size necessary to detect a difference δ_0 in a normally distributed outcome variable can be estimated by

$$n \geq \frac{\sigma^2}{\delta_0^2}(Z_\alpha + Z_\beta)^2$$

with σ = the unknown (population) standard deviation which is estimated by a value s.

Example:
The dissolving time t of a drug in gastric juice is known to be normally distributed with a mean of 45 sec and a standard deviation s of 3 sec. Under an experimental condition, a difference in the average dissolving time of 2 sec is to be detected, if it is present. To detect this difference with a two-sided α of 0.05 and a power of 0.90, at least 24 observations are necessary.

The appropriate statistical test in this situation is a one-sample t-test.
It must be noted that the sample size estimation uses Z-values, derived from a normal distribution. But, because we substitute an estimated value s for the standard deviation σ, the statistical test is based on a t-distribution instead of the normal distribution. Thus, instead of the Z-values t-values should be used. These t-values depend on 'degrees of freedom'. Here the corresponding number of degrees of freedom is n-1. For large samples (n > 100) there is little difference between Z-values and t-values. For smaller samples the sample size n must be estimated iteratively using the appropriate t-values.

Example revisited:
The number of degrees of freedom (d.f.) for a sample size n of 24 is equal to 23. The t-value with 23 d.f. for a two-sided α of 0.05 is equal to 2.069; the t-value with 23 d.f. for a power of 0.90 is equal to 1.319. This leads, after a small number of iterations, to a re-estimate of n of at least 26 observations.

4.2 Two samples

When two samples are to be compared with respect to their means, we make two assumptions. First, we assume that the outcome variable follows a normal distribution. Second, we assume that the population standard deviations in the two groups are equal. Again the unknown (common) standard deviation σ is replaced by an estimate s. To detect a difference between the means of the two groups the sample size per group can be estimated by

$$n \geq \frac{2\sigma^2}{\delta_0^2}(Z_\alpha + Z_\beta)^2$$

with

σ = the unknown (population) standard deviation which is estimated by a value s;

$\delta_0 = \mu_E - \mu_C$ = the relevant difference in means to be detected;

μ_E = the population mean of the experimental treatment;

μ_C = the population mean of the control treatment.

Example:
In a biological experimental design the leaf production velocity (a measure for the regeneration of the tropical forest) in the dry season is compared to that in the wet season. In the dry season an average velocity of 0.67 is assumed; in the dry season an average velocity of 0.51 is assumed. The standard deviation is estimated from earlier research as 0.32. To detect this average difference of 0.16 with a two-sided α of 0.05 and a power of 0.80, at least 63 plants are needed in each season.

The statistical test appropriate in this situation is a two-sample t-test. Here, as in the one-sample design, instead of the Z-values t-values should be used. These t-values depend on

‘degrees of freedom’. With two groups of equal size the number of degrees of freedom is equal to 2(n-1). For large samples (total degrees of freedom larger than about 100) there is little difference between Z-values and t-values. For smaller samples the sample size n must be estimated iteratively using the appropriate t-values.

4.3 ‘Equivalence’ in two samples

To investigate ‘equivalence’ of two samples with respect to the mean of a continuous outcome variable the same formula and iterative process can be used as described in section 4.2. The same remarks can be made here as in the case with a dichotomous outcome (section 3.3).

4.4 Paired samples

When two observations are made on the same individuals, these observations cannot be assumed to be independent of each other; the statistical test and the sample size estimation will have to take into account this dependency. For a continuous outcome variable, the differences between the two observations per individual form the sample for the statistical analysis. Because of this we can use the sample size formula as given in 4.1 for one sample:

$$n \geq \frac{\sigma_d^2}{\delta_d^2}(Z_\alpha + Z_\beta)^2$$

with σ_d = the unknown (population) standard deviation of the differences which is estimated by a value s_d and δ_d = the mean difference to be detected.

Example:
In an animal experiment the ejection fraction of the pig heart is determined before and after an experimental condition. An increase in average ejection fraction from 0.20 before to 0.25 after the experiment is the relevant difference to be detected and the standard deviation of the differences is estimated to be 0.10. To detect this average difference of 0.05 with a two-sided α of 0.05 and a power of 0.80, at least 32 pigs are needed.

The appropriate statistical test is a t-test for paired samples. Again, as in the one- and two-sample design t-values should be used instead of Z-values leading again to an iterative process for determining the required sample size.

Note that s_d is not the same as s in the sample size formula for one sample. s_d is the standard deviation of the differences and depends on the correlation between the observation before and the observation after the intervention. When only the variance s^2 of the observations is known, the variance s_d^2 of the differences can be estimated using the formula $s_d^2 = 2s^2 (1-\rho)$, where ρ is the correlation between the observations. If $\rho = 0$ (no correlation), the observations before and after are in fact two independent samples, each with variance s^2. When $\rho = 1$ (perfect correlation), the differences have no variance. In practice, some value between 0 and 1 should be assumed, for example 0.2 when little correlation between the observations is expected, 0.8 when a large correlation is expected, or 0.5 if moderate correlation is expected.

If the correlation between the observations is large, the gain in efficiency from using a paired design and analysis is large compared to a design with two independent samples.

5 Survival outcome

5.1 Two samples

In this section the outcome variable is a time-to-event, for example when people are followed until death. In practice, not everyone can be observed until death; some people will die during the study, while others can be followed until the end of the study period. If these individuals are still alive at the end of the study, we call their observed follow-up times 'censored'. Thus the length of follow-up may vary between individuals and the outcome is dichotomous (event 'yes' or 'no'). To compare two groups of individuals with respect to their follow-up times on two different treatments (for example a control treatment C and an experimental treatment E), a log-rank test is used. The power of the log-rank test depends on the total number of events in the study, rather than on the number of individuals enrolled in the study. Therefore sample size and study duration are determined in two steps. First the number of events d is estimated as a function of the type I error α, the type II error β (or the power 1-β) and the treatment effect δ_0. The treatment effect δ_0 is characterized by the ratio of the hazard rates (i.e. the risks of the occurrence of the event in the two groups). It is assumed that this hazard ratio does not change with follow-up time.
The total number of events can be determined by

$$d = \left(\frac{1+\delta_0}{1-\delta_0}\right)^2 (Z_\alpha + Z_\beta)^2$$

where δ_0 is equal to the hazard ratio (HR)

$$\delta_0 = \frac{\log(p_E)}{\log(p_C)} = HR$$

with p_C = the estimated survival probability in the control group and p_E = the estimated survival probability in the experimental group after a certain follow-up time T.

Next the necessary total number of patients N can be determined based on the estimated number of events d and the estimated survival probabilities in the two treatment groups

$$N \geq \frac{2d}{2 - p_C - p_E}$$

Example

In a clinical trial two drugs are compared with respect to their one-year survival rates in cancer patients. The estimated survival probability in the group with the control treatment is 0.60. A difference in survival rate of 0.20 with the experimental group is the clinically relevant difference to be detected. The hazard ratio is then equal to log(0.8)/log(0.6) = 0.437. The total number of events d required to detect this treatment effect with a two-sided α of 0.05 and a power of 0.80 is equal to 52. The total number of patients N required to detect the number of events d and to be included in the trial is at least 174.

The above approach (described by Freedman in 1982) does not specifically take into account the consequences of loss-to-follow-up, non-compliance, etc. Particularly in long-term clinical trials, these complications may lead to considerable loss of power. Moreover, it is assumed that the relative risk or hazard ratio does not change over time, a rather restrictive assumption that is often not fulfilled in practice. More elaborate approaches can be found in Friedman et al (1998).

6 Cohen's effect size

The sample size calculations discussed in the previous sections depend on assumed values for e.g. the standard deviation of the outcome variable, the 'success' probability of the control treatment, etc. Especially in pilot studies this information is not always available from previous investigations or from literature. Despite this lack of information it still is a matter of Good Statistical Practice to make an educated guess of the sample size needed for one's study. For this purpose Cohen (1988,1992) introduced the concept of *'Effect Size'* (ES). Each statistical test has its own ES index. All these indexes are dimensionless and continuous, ranging upward from zero. For all tests, the null hypothesis is that ES=0. Cohen proposed for each ES index the operational definitions 'small', 'medium' and 'large', which are approximately consistent across the different ES indexes.
For the comparison of two groups, whether the outcome variable is continuous or dichotomous, a small ES is about 0.2, a medium ES corresponds to a value of about 0.5 and a large ES is about 0.8.
Note that in the following we revisit the examples just for illustrational purposes with the given or assumed values for the important parameters, such as the standard deviation or the response rate in the control group. In practice, an assumption only has to be made for the ES index, be it small, medium or large.

6.1 Two samples with a dichotomous outcome variable

For a difference between two proportions p_C and p_E the ES index is defined as $|\phi_E - \phi_C|$ with $\phi = 2 \arcsin \sqrt{p}$. In practice this ES index is approximately equal to $2|p_E - p_C|$.
The number of patients per group can be estimated using

$$n \geq \frac{2(Z_\alpha + Z_\beta)^2}{(ES)^2}$$

Example revisited:
In a randomised, clinical trial the control treatment leads to a 'success' rate of 0.2. The experimental treatment is expected to lead to a 'success' rate of at least 0.4. This

difference corresponds to an ES index of 0.442, which is about a medium effect size. To detect this ES with a two-sided α of 0.05 and a power of 0.80, *the necessary number of patients per group is 81.*

6.2 One sample with a dichotomous outcome variable

For this situation the same formula for the ES index can be applied as given in section 6.1 with p_E equal to p_1, the rate under the alternative hypothesis and p_c equal to p_0, the rate under the null hypothesis.
The sample size can be estimated using

$$n \geq \frac{(Z_\alpha + Z_\beta)^2}{(ES)^2}$$

Example revisited:
In a certain population the incidence of thrombosis is assumed to be 30% (i.e. H_o: p_0 = 0.3). We want to sample a group large enough to detect an incidence of 40% (i.e. H_1: p_1 = 0.4), if that is the real, true incidence. We set the two-sided α at 0.05 and want a power of 0.90 to detect the difference in incidence. The ES is equal to 0.21, which can be viewed as a small effect size. With a two-sided α of 0.05 and a power of 0.90, the sample size n necessary to detect this effect size should be at least 238 patients.

6.3 Two samples with a continuous outcome variable

For a difference δ_0 between two means the ES index is defined as δ_0 / σ with σ as the common standard deviation. The number of patients per group can again be estimated using

$$n \geq \frac{2(Z_\alpha + Z_\beta)^2}{(ES)^2}$$

Example revisited:
In a biological experimental design the leaf production velocity (a measure for the regeneration of the tropical forest) in the dry season is compared to that in the wet season. In the dry season an average velocity of 0.67 is assumed; in the dry season an average velocity of 0.51 is assumed. The standard deviation is estimated from earlier research as 0.32. This leads to an ES index of 0.16 / 0.32=0.5. To detect this medium scale ES with a two-sided α of 0.05 and a power of 0.80, at least 63 plants are needed in each season.

6.4 One sample with a continuous outcome or with paired observations

The ES index is defined in this case as δ_0 / σ for one sample and as δ_d / σ_d for paired observations. The sample size can again be estimated using

$$n \geq \frac{(Z_\alpha + Z_\beta)^2}{(ES)^2}$$

Example revisited:
The dissolving time t of a drug in gastric juice is known to be normally distributed with a mean of 45 sec and a standard deviation s of 3 sec. Under an experimental condition a difference in the average dissolving time of 2 sec is to be detected, if it is present.
The ES index is equal to 2 / 3 = 0.667. To detect this medium to large ES with a two-sided α of 0.05 and a power of 0.90, at least 24 observations are necessary.

Example revisited:
In an animal experiment the ejection fraction of the pig heart is determined before and after some experimental condition. An increase in average ejection fraction from 0.20 before to 0.25 after the experiment is relevant to detect and the standard deviation of the differences is estimated to be 0.10.
The ES index is equal to 0.05 / 0.10 = 0.5. To detect this medium ES with a two-sided α of 0.05 and a power of 0.80, at least 32 pigs are needed.

7 Sequential alternatives

Ludbrook (2003) notices: “Laboratory researchers often analyse their results before the final number of experimental units (humans, animals, tissues or cells) has been reached. If this is done in an uncontrolled fashion, the pejorative term ‘peeking’ has been applied. A statistical penalty must be exacted.” Although the methodology Ludbrook suggests to use in his paper is obsolete, his principal advice still stands up. The ‘penalty’ or adjustment that should be made is a correction of the overall type I error to a nominal value per interim analysis.

Interim analysis refers to the repeated analyses of data as they accumulate. Analysing the cumulative results of an experiment like a clinical trial or an observational study as they become available to the investigator is very natural and obvious.

Russell and Burch (1992) also mention that interim analysis or sequential analysis is a method of conducting experiments in stages and that this mode of analysis seems ready-made for batch testing of toxicity.

Although an adjustment must be made to the nominal type I error per interim analysis, on average sequential analysis requires fewer observations in order to come to a decision than does statistical analysis based on a predetermined fixed sample size.

This paper is not the place for an elaborate exposition of sequential analysis. For more on this topic the reader is referred to another technical report of the Centre for Biostatistics (Schipper and Van der Tweel), to Van der Tweel and Schipper (2002) (both in Dutch) or to, for example, Todd et al (2001).

To illustrate the efficiency gain that can be reached in theory the example in section 3.2 is revisited. An analysis with a predetermined fixed sample size would require two groups of 79 patients, that is at least 158 patients, to detect the relevant difference if it exists. For a sequential analysis the necessary number of subjects cannot be calculated beforehand. Only the average number that will be necessary to come to a decision can be estimated. For this example, 106 to 138 patients will be needed on average, an efficiency gain of 13 to 33 percent. However, sometimes more than 162 patients will be needed for enough evidence.

8 Miscellaneous remarks and conclusions

The first step in the process of determining the sample size of a study is usually the application of one of the sample size formulae of the previous sections. The resulting sample size provides a preliminary idea of the general order of magnitude that is needed. It is important to realize that sample size calculations will always be approximate. For example, the numbers resulting from the formulae are rather sensitive to misspecifications of the standard deviation of a continuous outcome variable or the incidence rate in the control group. In practice, often only poor estimates for these quantities are available.

The determination of sample size required in the comparison of more than two treatment groups is much more complicated than in the case of two groups. Greenland (1985) describes a general method that can be used if the outcome variable is dichotomous or polytomous, and the χ^2-test is used on multi-way contingency tables. Fleiss (1986) discusses the case that the outcome is continuous and one-way analysis of variance is used for testing the treatment effects.

If in the end the calculated sample size needed to detect relevant effect sizes is larger than is feasible, one should consider abandoning the study before it starts. This will avoid wasting money, time, efforts, etc. in a scientifically inadequate study.

It is a matter of Good Statistical Practice to estimate the minimal group size at the design phase of an experimental study. Quite rightly, institutional human and animal ethics committees insist on this.

Biostatistics can help with sample size estimation, but the investigator will have to provide necessary information. This information can be based on earlier studies in the same institute or laboratory or on published literature.

The important thing is to be realistic in one’s expectations. Expectation of large effect sizes may lead to feasibly small sample sizes, but a too small sample size is unethical.

Acknowledgement

For this report parts of a text written by prof. Theo Stijnen for a NIHES-course on sample size estimation in clinical trials were used.

I thank my colleague Rebecca K. Stellato for critically reading the manuscript and her comments on the statistical content and on the English language.

9 Literature

- Cohen J. Statistical power analyses for the behavioural sciences (2nd ed.) Erlbaum, Hillsdale, New Jersey (1988).

- Cohen J. *A power primer.* Psychological Bulletin, 1992, **112**: 155-159.

- Donner A. *Approaches to sample size estimation in the design of clinical trials- a review.* Statistics in Medicine, 1984, **9**: 199-214.

- Fleiss JL. The design and analysis of clinical experiments. J Wiley & Sons, New York 1986).

- Freedman LS. *Tables of the number of patients required in clinical trials using the logrank test.* Statistics in Medicine, 1982, **1**: 121-129.

- Friedman LM, CD Furberg, DL DeMets. Fundamentals of clinical trials (3d ed.). Springer-Verlag, New York (1998).

- Greenland S. Power, sample size and smallest detectable effect determination for multivariate studies. Statistics in Medicine, 1985, **4**: 117-127.

- Lachin JM. *Introduction to sample size determination and power analysis for clinical trials.* Contr Clin Trials, 1981, **2**: 93-113.

- Ludbrook J. Interim analyses of data as they accumulate in laboratory experimentation. BMC Medical research Methodology, 2003, **3**: 15.

- Pocock SJ. Clinical trials. A practical approach. J Wiley & Sons, Chichester (1983).

- Russell WLS and RL Burch. The principles of human experimental technique (special ed.). Methuen & Co Ltd, London (1992).

- Schipper M and I van der Tweel. Efficiënte analyse van accumulerende biologische en medische data. Intern rapport 2. Centrum voor Biostatistiek, Universiteit Utrecht.

- Todd S et al., *Interim analyses and sequential designs in phase III studies.* Br J Clin Pharmacol, 2001, **51**: 394-399.

- Van der Tweel I and M Schipper. Sequentiële analysen in klinisch en epidemiologisch onderzoek. NTvG, 2002, **146**: 2348-2352.