Paper AS02

Sample size: a couple more hints to handle it right using SAS and R

Andrii Artemchuk, Experis Clinical Solutions, Kharkiv, Ukraine

ABSTRACT

One of the key points in clinical trials is sample analysis. During the development of the protocol, the team must determine the target sample size for the trial and take multiple possible scenarios into consideration. What to do in the case of rare diseases where very few patients pass the inclusion criteria for the study? How many of these patients should be enrolled to make so the results achieve statistical significance? This paper is going to answer these questions.

Two main topics are covered by the paper: the algorithms that calculate optimal sample size and their statistical background, and a check of the hypothesis statement that a sample analysis will yield a statistically significant result. The calculations are performed using (but not limited to) PROC POWER, PROC GLMPOWER in SAS and functions from the PWR package in R; comparison of the results is performed.

INTRODUCTION

Sample size determination is a very important step in planning a clinical trial. If the sample size is too small, it is likely that the effect of the treatment (if it does exist) will not be detected. It is equally important that the trial not be "too big", because the excess of the sample size uses more financial, human and time resources than needed to answer a medically important question, formulated it the trial.

Therefore, it is necessary to optimize the sample size, making a balance between factors that affect it. This paper will consider the most popular method of estimating the sample size - power analysis - which includes the analysis of tested hypotheses, formulated in the clinical trial. Power analysis makes it possible to control the statistical significance of the results obtained.

Formulas and recommendations given in the paper will help to calculate the optimal sample size.

WHAT AFFECTS THE SAMPLE SIZE

In this section the factors that affect the sample size will be listed. Some of them are specified manually before the study begins, and others need to have an idea what the expected results of the clinical trial are.

LEVEL OF SIGNIFICANCE

Sample size grows as the level of significance decreases, since reducing it decreases the probability of making type I error, i.e. probability of rejecting the tested hypothesis when it is correct.

POWER OF THE CRITERION

The higher power of the criterion is, the more likely it is to detect differences between the compared groups, so the greater statistical significance of the test criterion is. Roughly speaking, high power of the criterion requires large sample size. Ideally, it would be desirable that the power of the criterion be 100%, but this is impossible, since in reality there is always a chance, albeit minor, to make type II error.

DATA VARIABILITY

The sample size necessary to achieve the desired statistical significance is directly proportional to the variability of the outcome measure in the population under study. It is often necessary to know the spread of the data, which is necessary to estimate the variance or the standard deviation. The bigger they are the larger sample size may be required.

THE LEAST EXPECTED TREATMENT EFFECT

Treatment effect is the magnitude of the clinical benefit of the treatment, which is clinically important and can't be ignored. The effect is often expressed through the difference in statistics, observed in the treatment groups – for example through the difference between means, standard deviations or proportions.

From one side it is relatively easy to choose power (usually the acceptable minimum level is assumed to be 80%) and significance level of the criterion (usually 0.05 or 0.01) that meet the specific requirements of the clinical trial. From the other side it's difficult to define data variability and the expected treatment effect. The exact value of the observed effect may be unknown in advance. Similarly, data variability is also unknown, because the trial data have not been analyzed yet.

Some of these factors will be considered further in the paper in more detailed way.

LEVEL OF SIGNIFICANCE ALPHA

In practice, alpha is assumed to be 0.05 or 0.01. The smaller alpha is the larger sample size will be required to achieve the given power. Also, the number of hypotheses tested can affect the trial size. Assume the conclusion about the statistical population is made on the basis of several hypotheses. In this case the probability that at least one of m hypotheses is false will be equal to $1-(1-\alpha)^m$. If only 3 hypotheses are tested and $\alpha = 0.05$ the probability will be 14.3%, and will grow with the increase of m.

The Bonferroni correction can be applied to decrease this probability. For each hypothesis tested, when the total number of hypotheses is m, $\frac{\alpha}{m}$ should be used in formulas instead of α. This method has a drawback: the loss of power with the increase of m. To keep the power at a desired level the Holm–Bonferroni method should be used (not covered in this paper).

TREATMENT EFFECT

One of the steps in the sample size determination is eliciting the treatment effect in advance. In general, the smaller difference between the groups to be detected, the larger sample size is required. There are two options possible, depending on the direction of the treatment effect.

- One-tailed test: when the direction of the treatment effect is known - for example, the study of a drug that reduces blood pressure compared to placebo. The tested hypothesis asserts that the mean pressure values are equal for both the treatment and placebo, and the alternative – that they are different. A desired treatment effect is a decrease of blood pressure by 15 mm hg in the treatment group, so the direction of effect is known.
- Two-tailed test: this is a more general approach, because it takes into account every opportunity of the direction of the effect. It is preferable when there is no certainty in the direction of difference between treatment groups, if one exists.

In the formulas it is expressed as follows: for a two-tailed test in the standard normal distribution it is necessary to use $Z_{1-\frac{\alpha}{2}}$ - the value of the inverse distribution function for the level of significance $1-\frac{\alpha}{2}$. If the direction of treatment effect is known, then values Z_{α} or $Z_{1-\alpha}$ should be taken depending on the direction of the effect. Usually the treatment effect value enters the formulas in the square, so small effect changes lead to significant changes in the sample size. Basically, the smaller the magnitude of the effect should be detected, the larger the sample size will be required. There are several approaches to assessing the treatment effect value. In some clinical trials (for example, in oncology trials) it makes sense for ethical reasons to repulse from some minimal value that needs to be detected. For example, a new chemotherapy regimen will be considered effective only if it increases the likelihood of remission by more than 20% compared to the standard regimen. Another approach is to determine the best effect that a drug can produce – suggestion of that kind can be based on hypothetical assumptions, analysis of literature or the past clinical trials.

There are some ways how to act in real life when there's no exact information on the desired treatment effect that needs to be discovered:

- try to determine direction of the effect and define its minimum and maximum values;
- consider several options for the possible effect, estimate the sample size for each of them and choose the most optimal one;
- find out if it's possible to allocate a budget for the inclusion of several more patients, in order to detect the effect, that can't be seen at a current stage of a clinical trial;
- assess whether the cost of including new patients is worthwhile to detect a difference between groups if at the current stage of a trial either the effect of the treatment is already present (positive or negative), or there is no effect at all;
- do the opposite, and find out what effect will be unreasonable for detection.

Based on this, it can be concluded that it is necessary to balance the sample size, the desired treatment effect, ethical and economic factors.

VARIANCE, STANDARD DEVIATION

The main problem is that it is impossible to predict in advance what the exact variance will be before a specific sample analysis is done. Instead it is possible to evaluate it by using the data from similar published clinical trials, from literature, or from a pilot study. If we are talking about the observed characteristics, like SBP and DBP, growth or weight of patients, then it is possible to estimate the minimum and maximum values of these characteristics that were observed in the clinical experience. In the clinical trials, where time to event is investigated, or parameter of interest has a binary form, the problem of estimating the average event rates arises instead of estimating the variance.

If there are no exact methods or it is completely impossible to collect data from previous clinical trials, standard deviation can be estimated as the difference between the maximum and minimum known values of the parameter, divided by 4. This method is applicable, since all sample values are within 2 standard deviations from the standard one.

WITHDRAWALS, MISSING DATA AND LOSSES TO FOLLOW-UP, NON-COMPLIANCE

In practice, patients often happen to be withdrawn from the clinical trial, they may forget to take the drug, or take the wrong dose, can be lost to follow-up, as well as important clinical data can be filled in with errors, that may lead to the impossibility of including this patient in the final analysis.

In order to provide for this, it is possible to make a correction of the sample size: $N' = \frac{N}{1-q}$, where q is the estimated proportion of patients who will be lost to follow-up, or to be more general – proportion of patients, who for some reason will not be included in the final analysis. This proportion will not be known before a clinical trial starts, and a rough estimate can be obtained by analyzing previous trials, or using the information from the pilot study.

Another problem may occur due to the noncompliance. Suppose that during a clinical trial a patient from the placebo group begins to take some concomitant medication (because of poor health, adverse event etc.). This medication might result in effect similar to the effect of the study drug. On the other hand, a patient from the treatment group may stop taking the study treatment (again, due to adverse events). In this case, for this patient, the expected effect of the drug will be close to the expected effect of placebo. Both cases negatively affect the power of the observed criterion, since they reduce the difference between the studied groups.

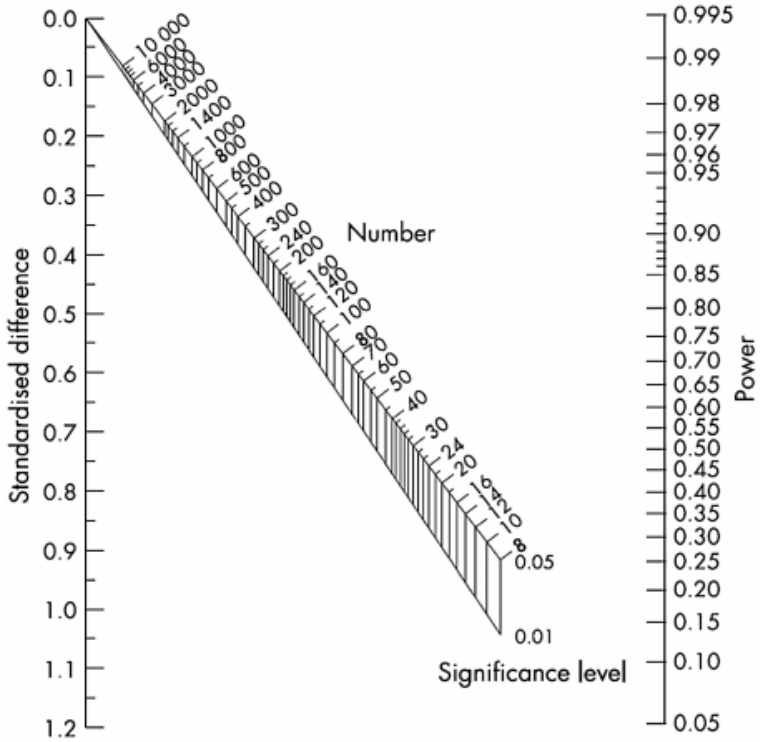
Formula $N' = \frac{N}{1-q}$ will not give an exact result in the example above. It will be necessary to take into account and estimate the number of non-compliant patients in each group: let μ_t and μ_c – mean indicators of the observed treatment effect in both treatment and control group. The difference in the effects can be expressed as the difference between means: $\mu_t - \mu_c$. Let P_t and P_c – proportions of non-compliant patients in each group. If $P_t = 0$ – all the patients in treatment group are expected as fully compliant, so μ_t will not change. Conversely if $P_t = 1$, then all the patients in treatment group are expected to be non-compliant, hence μ_t will be equal to μ_c. Hence, the recalculated expected difference of mean values will be: $\mu_t' - \mu_c' = (\mu_t - \mu_c)(1 - P_c - P_t)$.

It should be noted that the difference between the means is included in the denominator of the final formulas in the square, so to apply this correction, the calculated sample size must be multiplied by $1/(1 - P_c - P_t)^2$. Ignoring these calculations in practice might lead to a significant loss of the power of the criterion. Exclusion from consideration of such patients may violate the principle of "intent to treat".

GENERAL ALGORITHM

Before speaking of more complex formulas, a simple method of how to calculate the sample size will be reviewed. It can give an idea of the required sample size without applying formulas or calculations. This method is suitable for simple clinical trials where it is sufficient to know the level of significance, required power and value of the expected standardized difference. Basically it could be used when the hypotheses about the means and proportions are tested. On the graph below, it is needed just to connect the power and the deviation with a line, select the significance level, and then find in the nomogram what sample size corresponds to the choice.

Altman's nomogram:



Further in the paper the particular and most popular cases of calculating the sample size are described. They are used for different trial designs and for different types of treatment effect. For each of them specific statistical criteria and hypotheses are used to calculate the sample size. Statistical significance is achieved through the determination of power and level of significance.

In this section a general approach to estimate sample size and achieve statistical significance is described. Suppose that the test statistics (a function of the sample) T is used to test the hypothesis. Ω_α is a critical region of the criterion, that depends on the level of significance α. Then if H_0 hypothesis is true, from the definition comes:

$$P\{T \in \Omega_\alpha | H_0\} = \alpha$$

Consider Ω_β to be a critical region, dependent from error type II β. Then if H_0 hypothesis is false, and an alternative hypothesis is true, comes:

$$P\{T \notin \Omega_\beta | H_a\} = 1 - \beta$$

These formulas will be used to estimate the sample size in the case when the difference between the mean values of a parameter in both treatment and control groups is observed. The test statistics T in this case is the standardized difference between the means:

$$T = \frac{\mu_t - \mu_c}{\sigma\sqrt{2/n}}$$

T has a standard normal distribution. μ_t and μ_c are the means in treatment and control groups respectively. σ is standard deviation and n is the number of patients in each group. H_0 hypothesis states that means in groups are equal: $\mu_t = \mu_c$. Using the two-tailed test and the definition of type II error makes it possible to derive a probability to reject H_0 when it's not true:

$$P\left\{\frac{|\mu_t - \mu_c|}{\sigma\sqrt{2/n}} > Z_{1-\frac{\alpha}{2}} \middle| H_a\right\} = 1 - \beta$$

On the other hand, an alternative hypothesis states that $\mu_t \neq \mu_c$, so $\delta = \mu_t - \mu_c$ is the non-zero difference between the means. Hence the probability to accept H_a, when it's true is:

$$P\left\{\frac{|\mu_t - \mu_c| - \delta}{\sigma\sqrt{2/n}} > Z_\beta \middle| H_a\right\} = 1 - \beta$$

After some conversion:

$$P\left\{\frac{|\mu_t - \mu_c|}{\sigma\sqrt{2/n}} > Z_\beta + \frac{\delta}{\sigma\sqrt{2/n}} \middle| H_a\right\} = 1 - \beta$$

Combining with the previous formula:

$$Z_\beta + \frac{\delta}{\sigma\sqrt{2/n}} = Z_{1-\frac{\alpha}{2}}$$

whence

$$\frac{\sqrt{n}\delta}{\sigma\sqrt{2}} = Z_{1-\frac{\alpha}{2}} - Z_\beta$$

and

$$N = \frac{2\sigma^2(Z_{1-\frac{\alpha}{2}} - Z_\beta)^2}{\delta^2}$$

From the symmetry of standard normal distribution follows: $Z_\beta = -Z_{1-\beta}$, then

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

EXAMPLES OF FORMULAS

Below there are three examples of most popular cases of the sample size calculations.

SAMPLE SIZE ESTIMATION FOR PROPORTIONS

This section concerns the cases when one of the endpoints of a clinical trial is the analysis of proportion (or percentages) of patients with some characteristics or condition. This means that for each patient, the result of observation will be either "success" when the characteristic is present, or "failure" when it is not. Usually, a comparison is made between the two groups of patients, when the presence or absence of some treatment effect is examined.

Required assumptions to use the following formula are that patient's responses must be independent and variance must be the same in both groups. Then the sample size in each group can be estimated using formula:

$$N = \frac{(\pi_t(1-\pi_t) + \pi_c(1-\pi_c))(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{(\pi_c - \pi_t)^2}$$

where:
π_t is the probability of observing "success" in the treatment group.
π_c is the probability of observing "success" in the control group.

The advantage of this formula is that it is not necessary to know the exact value of variance or the standard deviation. However, unlike the previous example, the variance will not always be equal for both groups. In this case, the sample size for each of the groups can be calculated by the formula:

$$N = \frac{(Z_{1-\frac{\alpha}{2}}\sqrt{2\pi(1-\pi)} + Z_{1-\beta}\sqrt{\pi_t(1-\pi_t) + \pi_c(1-\pi_c)})^2}{(\pi_c - \pi_t)^2}$$

where:
$\pi = \frac{\pi_c + \pi_t}{2}$ – the mean frequency of the observed treatment effect in groups.
Both formulas can be derived on the basis of the chi-square (Pearson) test. Fisher's exact test can be used depending on the input parameters.

Example: The trial compares the efficacy of corticosteroid injections to physiotherapy in the treatment of a painful, rigid shoulder. The treatment is considered successful after 7 weeks if a patient considers himself fully recovered or he has improvement (by Likert scale). Success means the presence of this indicator in 40% of patients in the group where treatment is less effective.
Desired power is 80%, the level of significance is 5%. The main question regarding the sample size is how many patients should be included to find a clinically important difference of 25% between the groups.

Input data looks as follows:

$$\pi_c = 0.4, \pi_t = 0.4 + 0.25 = 0.65, \pi = \frac{0.4 + 0.65}{2} = 0.525,\ Z_{1-\frac{\alpha}{2}} = 1.96,\ Z_{1-\beta} = 0.84$$

Then

$$N = \frac{(Z_{1-\frac{\alpha}{2}}\sqrt{2\pi(1-\pi)} + Z_{1-\beta}\sqrt{\pi_t(1-\pi_t) + \pi_c(1-\pi_c)})^2}{(\pi_c - \pi_t)^2} =$$

$$= \frac{(1.96(\sqrt{2 \cdot 0.525 \cdot 0.475} + 0.84\sqrt{0.65 \cdot 0.35 + 0.4 \cdot 0.6}))^2}{(0.4 - 0.65)^2} = \frac{(1.384 + 0.574)^2}{(-0.25)^2} =$$

$$= \frac{3.834}{0.0625} \cong 62$$

The number of patients required in each treatment group is 62. Taking into account losses to follow-up, and correcting the sample size by 10% will result in:

$$N' = \frac{N}{1-q} = \frac{62}{1-0.1} = \frac{62}{0.9} \cong 69$$

Here and further in the paper, the sample sizes were derived on the basis of direct substitution of the initial data into the indicated formulas.

SAMPLE SIZE ESTIMATION FOR TWO MEANS

Here a clinical trial includes two groups of patients – not necessarily of the same size. The purpose of the trial is to identify significant differences in some parameter of interest. This paragraph considers the difference between the means, however odds ratios, hazard ratios, or other statistics can also be analyzed.
Several assumptions should be made to use this method. The difference between sample data should have normal distribution. If it doesn't, then some arithmetic transformation may be applied (for example, logarithmic transformation). Also the patient's responses should be independent and the variance should be the same in both groups. Under these assumptions the formula for estimation of the sample size will be:

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

where δ is the supposed difference between the means to be discovered.
In practice, the variance will not be known, it will have to be estimated on the basis of either past experience or data from the pilot study. If the number of patients in the groups is different, the sample size for patients in the control group will be:

$$N_c = \left(1 + \frac{1}{k}\right)\frac{\sigma^2(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{\delta^2}$$

where $k = \frac{N_t}{N_c}$ – the ratio of the number of patients in the treatment group to the control group.
Both formulas can be derived on the basis of the hypothesis of the equality of two means. Null hypothesis states that the mean values in the two groups are equal.

Example: the effect of Vitamin D on the prevention of neonatal hypocalcaemia is examined on pregnant women compared to placebo. The main efficacy parameter is the infant's serum calcium level one week after the birth. Previous research shows that the mean serum calcium level is 9.0mg/100ml and standard deviation is 1.8 mg/100ml. An increase of 0.5 mg/100ml is considered to be clinically relevant and the significance level is 5% and power is 95%. According to the formula above, 337 patients should be included in each group.

SAMPLE SIZE ESTIMATION FOR TIME-TO-EVENT DATA USING LOG-RANK TEST

In this type of clinical trials the time to event of interest is observed. This event could be death, disease progression, tumor recurrence, etc. In this example it is assumed that groups of patients should be of equal size and the probability of event for each patient within a group is the same. Under these assumptions the formula for estimation of the sample size will be:

$$N = \frac{1}{\pi_c + \pi_t}\left(\frac{\theta + 1}{\theta - 1}\right)^2 (Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2$$

where
π_t is the probability of the event to happen of a patient in the treatment group during a trial.
π_c is the probability of the event to happen of a patient in the control group during a trial.
$\theta = \frac{\log(1-\pi_c)}{\log(1-\pi_t)}$ is a relative risk of the event occurrence, or hazard ratio.
This formula can be simplified under an additional assumption that the time to event is subjected to the law of exponential distribution. On the basis of this assumption, the above formula will have the following form:

$$N = \frac{4}{(\pi_c + \pi_t)} \cdot \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{(\log\theta)^2}$$

If the numbers of patients in the groups are different, then the number of patients in the control group will be:

$$N_c = \frac{(k+1)^2}{k(\pi_c + \pi_t)} \cdot \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})^2}{(\log\theta)^2}$$

where $k = \frac{N_t}{N_c}$ – the ratio of number of patients in treatment group to control group.
Log-rank test is used as a basis to derive these formulas. The test hypothesis statement is that the survival curves in the two groups are the same. In other words, the probability of occurrence of an event at any given time point in a group is the same. In practice it is difficult to estimate the time to event, and the number of patients, who can experience the event in advance. However, some assumptions could be made about the distribution of the event occurrence.

Example: The Myocardial Infarction Prevention Trial. Influence of HDL-Plus on the risk of a heart attack is studied. 20% of patients are expected to suffer a heart attack over the course of the 5 years of follow-up period. Those taking HDL-Plus can expect their risk to decrease to approximately 15%.
Given this, $\theta = 1.373$. Using the formula for 80% power and the significance level being 5% it follows that 907 patients should be included in each group.

Sometimes, to simplify the situation, it is possible to switch the model and determine the sample size using proportions. To do this, the endpoint should be re-formulated as, for example, "success" - the survival rate for one year, or the non-occurrence of the disease progression. As this approach simplifies the model, then some design details may not be taken into account. That might lead to an overassessment of the sample size. For example, a more general question can be asked in the previous case: how many patients are needed to detect the difference of 50% between the patient rates in both groups for the significance level of 5% and the power 80%. It is expected that within 3 months 10% of the patients in the control group and 20% of patients in the treatment group will have HDL Cholesterol level above 45 mg / dl.
Using the formula to estimate sample size for proportions, it turns out that 199 patients in each group are needed, which is much less than 907 patients when giving an answer to the previous, more particular question about the treatment.

These formulas describe most common cases of the sample size calculation and can be applied before the research with a number of simplifications and assumptions. In practice, more complex models and formulas can be used, for example in time to event trial that involves the length of accrual period and the follow-up period.

IMPLEMENTATION IN SAS AND R

The power and the sample size calculations can be done in SAS in various ways. One of them is PROC POWER, that is used for evaluating power and sample size for a variety of mostly basic statistical analyses such as described in the examples and also the equivalence test, multiple regression, logistic regression, and some nonparametric

tests. PROC GLMPOWER can be used for more complex designs, such as factorial design, randomized complete block design; for analysis of covariance etc. However, an example of use of PROC GLMPOWER for basic analysis will be provided.

PROC GLMPOWER: ESTIMATING THE SAMPLE SIZE FOR THE DIFFERENCE BETWEEN MEANS

Values are taken from the example of the estimation of sample size for two means. Here `a1` and `a2` are the anticipated mean values in the control and treatment groups respectively.

```
data oneway;
    level = "a1"; meanest = 9; output;
    level = "a2"; meanest = 9.5; output;
run;
proc glmpower data=oneway;
    class level;
    model meanest = level;
    power
    stddev = 1.8
    alpha = 0.05
    ntotal = .
    power = 0.95;
run;
```

Result:

The SAS System
The GLMPOWER Procedure

Fixed Scenario Elements	
Dependent Variable	meanest
Source	level
Alpha	0.05
Error Standard Deviation	1.8
Nominal Power	0.95
Test Degrees of Freedom	1

Computed N Total		
Error DF	**Actual Power**	**N Total**
674	0.950	676

PROC POWER: ESTIMATING THE SAMPLE SIZE FOR TIME-TO-EVENT DATA: LOG-RANK TEST

Values are taken from the example of the estimation of sample size for time-to-event data using log-rank test. Accrual time is accepted as half of the total study time, which is equal to 5 years.

```
proc power;
    twosamplesurvival
     test=logrank
     curve("Control")   = (0 5):(1 0.8)
     curve("Treatment") = (0 5):(1 0.85)
     refsurvival = "Control"
     accrualtime = 2.5
     followuptime = 2.5
     hazardratio = 1.373
     alpha = 0.05
     sides = 2
     ntotal = .
     power = 0.8;
run;
```

Note, while setting power to missing and specifying `ntotal` (or `npergroup`) parameter it is possible to estimate power for a given sample size. Also a graphic illustration of the results is available by using `plot` parameter. Changing `curve` parameter to `curve("Control") = (5):(0.8)` will give the exponential distribution of the event. The same change should be applied to treatment curve.

Result:

The SAS System
The POWER Procedure
Log-Rank Test for Two Survival Curves

Fixed Scenario Elements	
Method	Lakatos normal approximation
Number of Sides	2
Accrual Time	2.5
Follow-up Time	2.5
Alpha	0.05
Reference Survival Curve	Control
Form of Survival Curve 1	Piecewise Linear
Form of Survival Curve 2	Proportional Hazards
Hazard Ratio	1.373
Nominal Power	0.8
Number of Time Sub-Intervals	12
Group 1 Loss Exponential Hazard	0
Group 2 Loss Exponential Hazard	0
Group 1 Weight	1
Group 2 Weight	1

Computed N Total	
Actual Power	**N Total**
0.800	1772

The same calculations in R can be performed using functions either from the PWR package, POWERSURVEPI package (for time to event analyses) or from the built-in R functions.

PWR PACKAGE IN R: ESTIMATING THE SAMPLE SIZE FOR TWO PROPORTIONS
Values are taken from the example of the estimation of sample size for two proportions.

```
install.packages("pwr")

library(pwr)

effect <- ES.h(0.4, 0.65)

pwr.2p.test(h=effect, n=NULL,
            sig.level=0.05, power=0.8,
            alternative = "two.sided")
```

Result:

```
##
## Difference of proportion power calculation for binomial distribution (arcsine
## transformation)
##
##              h = 0.5060506
##              n = 61.29835
##      sig.level = 0.05
##          power = 0.8
##    alternative = two.sided
##
## NOTE: same sample sizes
```

In this example the PWR package should be downloaded and installed first using `install.packages` command. Note that `ES.h` function should be applied to calculate the value of effect size, depending on proportions. The same calculations can be made using `power.prop.test` built-in function.

COMPARISON OF THE RESULTS

In the example of estimating the sample size for two proportions PWR package and the given formulas were used. The same data after being input into the PROC POWER gives the same result – 62 patients per group after rounding up.

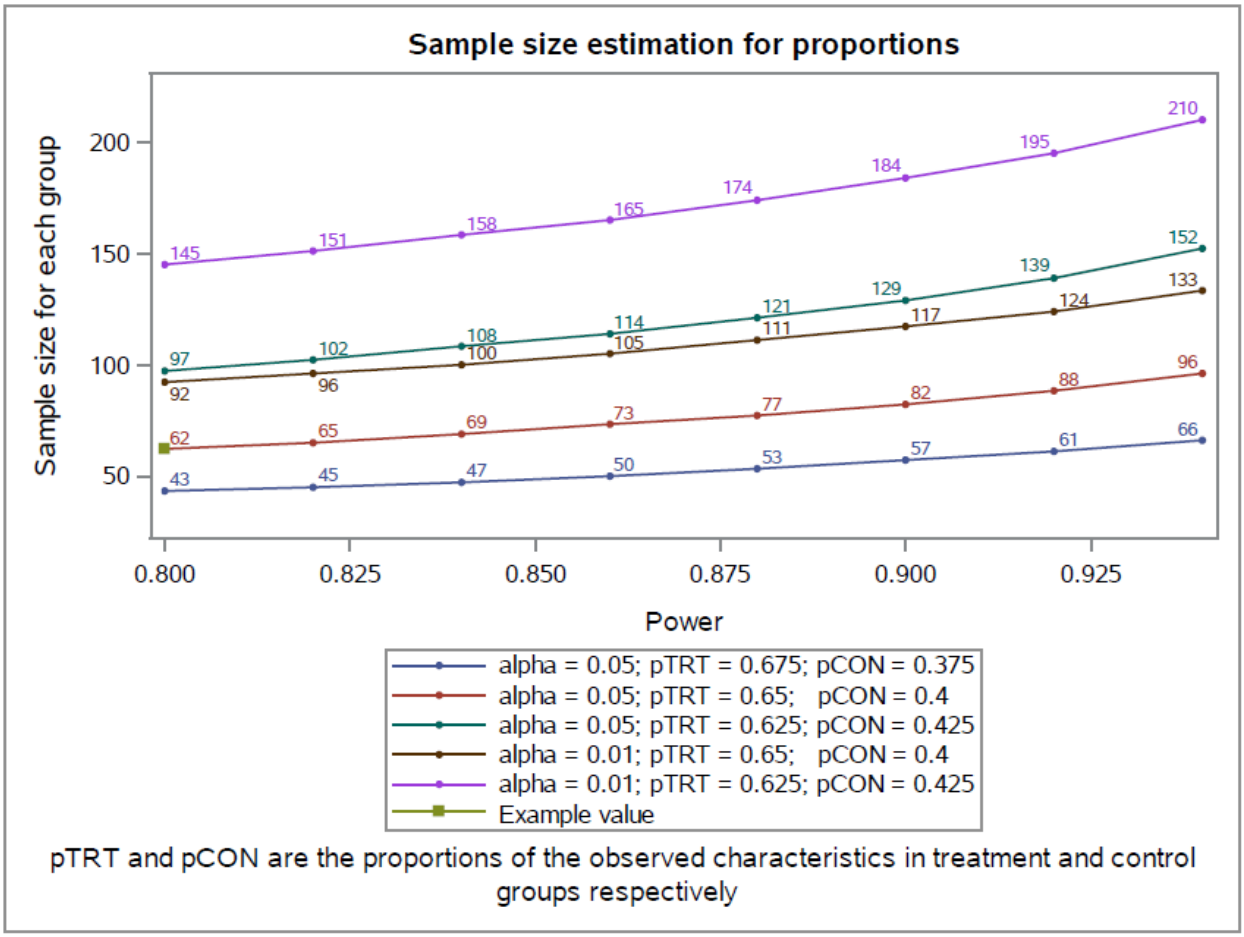
The results of PROC GLMPOWER, PROC POWER and the PWR package in case of estimating the sample size for two means are the same – 676 patients in total are required (338 per each group). Using the given formulas gives the similar result: 337 patients per each group (674 in total).

Estimation of the sample size for time to event data is not available in the PWR package. However, there are multiple power and sample size functions in POWERSURVEPI package that can be used for various models of time to event analysis. `powerCT.default` function produces the same result as the given formula in case when the reference survival curve is piecewise linear and when the hazard ratio is given. There are 1816 patients (908 per group) required in total for the both cases. Using PROC POWER for the same input data gives a result of 1772 patients. The result is a bit different from the above, as PROC POWER should include the information about the accrual period and the length of the trial in the input parameters. In summary, a researcher has a wide range of functions and procedures for estimating the sample size.

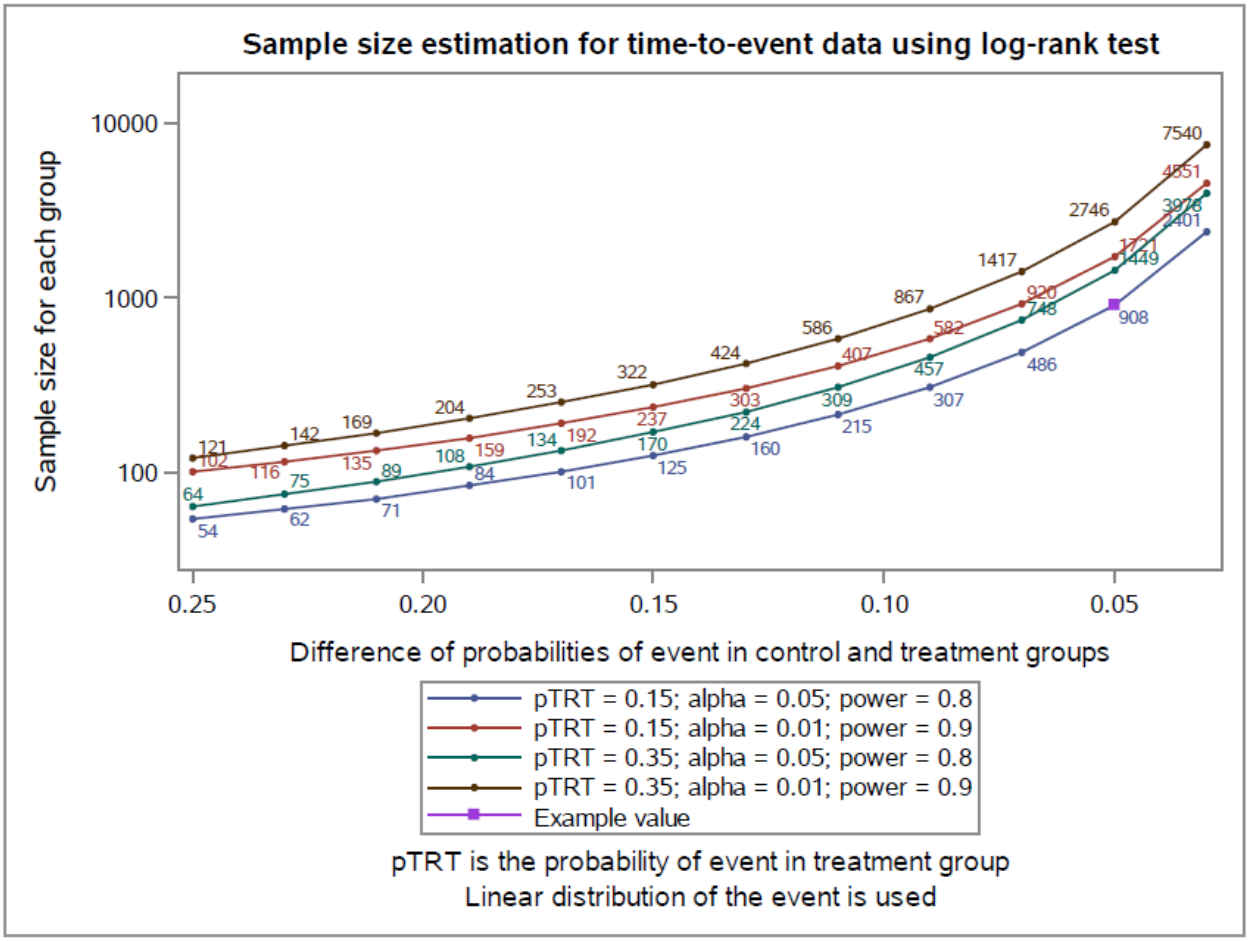
WHEN PARAMETERS CHANGE

As the exact values of such parameters as expected means, standard deviations or event rates remain unknown at the beginning of a clinical trial, it's interesting to see how the change of these parameters affects the power and the sample size. The graphs illustrating two given examples are shown below. The input parameters were taken from the examples above.

First graph illustrates the dependency of the sample size on power, proportions, and level of significance. The results show that when the level of significance is fixed then the sample size mostly depends on the difference between proportions – the lesser it is, the bigger sample size is required. Reducing the difference from 3 to 2 entails an increase of more than 2 times in the sample size.



Second graph illustrates the dependency of sample size on power, event rates in treatment groups, and level of significance. Largest effect on sample size makes the level of significance and the difference between groups. The lesser difference is desirable to be found the larger sample size is needed. This dependency is non-linear and when the difference is close to 0, sample size sharply increases and can reach over 7500 patients. Changing the statistical significance of the test, i.e. moving from the level of significance of 0.05 and power of 0.8 to 0.01 and 0.9 respectively almost doubles the required sample size.



GENERAL THOUGHTS

Changing statistical parameters directly affects the required sample size. But this process needs to be strictly controlled, without making the only purpose to reduce the number of patients. For example, consider the study of the effect of Anturan compared to placebo in the patients who have suffered myocardial infarction. The main criterion of interest is death for any reason during the first year after taking the treatment.

Assume that it is required to detect a decrease in mortality by 20% (from 10 to 8%) with the power of 95% and the level of significance 1%. The input data then has the following form: $\pi_c = 0.1,\ \pi_t = 0.08, \alpha = 0.01, \beta = 0.05$. To fulfil these requirements, a sample of 7292 patients would be needed for each group (14584 in total). Compliance with these conditions is very expensive and is almost impossible in real life.

In practice, milder conditions were applied, namely, the difference between the ratios was 5% (i.e. a decrease in mortality by 50%), the significance level was 5% with the power being 90%, which resulted in the sample size of 582 patients in each group (1164 in total). Total number of patients was increased to 1500 taking into account non-compliance and losses to follow-up.

On the other hand, it is possible to set the power to 50% to detect a reduction in mortality of 80% at the significance level of 10%. Then the input data will be as follows: $\pi_c = 0.1,\ \pi_t = 0.02, \alpha = 0.1, \beta = 0.5$. Such conditions will give a sample of only 48 patients for each group. However, a study with a sample of this size would be absurd, since probability of making type II error is too high and previous studies of myocardial infarction indicate that it is impossible to detect such significant reduction in mortality.

From the foregoing it follows directly that it is necessary to choose reasonable power and significance level, as well as make realistic estimates of characteristics (standard deviation, difference of means, difference in frequencies,

etc.). Similarly, several treatment groups, several endpoints, the presence of interim analysis can affect specifically the level of significance and the sample size as a whole.

From the ethical point of view, it is irresponsible to start a trial, which has only 40% probability of detecting the treatment effect. An experiment with an insufficient size subjects patients to potentially harmful treatments without increasing knowledge about the treatment. In the experiment with an excessive size, an unnecessary number of patients undergo a potentially dangerous treatment or at the same time a larger number of patients do not receive a potentially beneficial treatment (in the case of the experiment with a control group of patients).

CONCLUSION

There is no simple answer to the question "what size of the sample should be", it is easier to answer the reverse question - what it should not be, and when clinical trial should not start at all. Estimating the sample size with the usage of the formulas shown above can help to make the most optimal decision. Many factors need to be taken into account in every case before making a decision. The clinical trial budget, design, inclusion criteria, availability of data from previous trials or literature – all of these influences the sample size.

It is very important to assess the sample size before a clinical trial starts, because, for example, if the trial budget is limited and the estimated sample size requires significantly bigger budget than planned, then such trial should not begin – at least under the present conditions. If something similar happens, then one of possible solutions can be to reduce the requirements for the study, increase the magnitude of the treatment effect to be detected, reduce the power, or even change the design, however, that option may not always be acceptable. Otherwise, conducting such trial can lead to wasting of resources for achieving a statistically insignificant result, as well as causing inconvenience to patients. It is also worth noting that not all the patients will simultaneously join the trial, and although it is a rear practice, it could make sense to stop the trial earlier for ethical and economic reasons if patients start to show a significant difference in the parameters of interest between the groups. A very large sample size is also a bad option, as this leads to a useless waste of resources, and exposes a number of patients (beyond what is necessary) to a potentially dangerous treatment, or vice versa - does not allow patients to receive a potentially effective treatment.

Determining the sample size is often a difficult task, and is very important to neglect while planning a clinical trial. This problem can be partially solved using formulas, but the result can still be statistically inaccurate. Therefore, it is necessary to take into account all the aspects of the trial, and use the model which is closest to the trial design.

REFERENCES

Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002;24:39–53.

Aviva Petrie, Caroline Sabin. Medical Statistics at a Glance, 3rd Edition Jul 2009, Wiley-Blackwell pp. 50-52, 108-112, 149.

Pocock SJ: Clinical trials. pp.123-130. Wiley, 1983.

Shuster JJ: Handbook of Sample Size Guidelines for Clinical trials. pp. 51-130. CRC Press, 1990.

Russell V Lenth (2001) Some Practical Guidelines for Effective Sample Size Determination, The American Statistician, 55:3, 187-193.

Conroy, Ronán. (2018). The RCSI Sample size handbook. 10.13140/RG.2.2.30497.51043.

M Castelloe, John. (2000). Sample size computations and power analysis with the SAS system. 265-25.

Suresh KP, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. J Hum Reprod Sci 2012;5:7-13

SAS Institute Inc. 2013. SAS/STAT® 13.1 User's Guide. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Andrii Artemchuk
Company: Experis Clinical Solutions
Address: 43/2 Gagarin Avenue
City / Postcode: Kharkov 61001, Ukraine
Work Phone: +38 057 755 7020 ext. 2425
Fax: +38 057 728 30 35
Email: andrii.artemchuk@intego-group.com
Web: https://intego-group.com