SAMPLE SIZE ESTIMATION IN CLINICAL TRIALS

Ana García Romero

Master in Pharmacological Research
Course in Data analysis
2023-2024

Importance of the sample size

Clinical trials need careful planning, which is summarized in a trial protocol. This includes details such as the hypothesis, the trial's objective, the design, data collection methods, criteria for selecting participants, scientifically justified sample size, procedures and intervention, data handling procedures, primary and secondary outcome definition, statistical analysis plan and assumptions. Among this information, we will focus on sample size estimation, which is one of the pivotal aspects in the design of a clinical study.

Despite awareness of the significance of biostatistics in preclinical and clinical research, many researchers have insufficient statistics knowledge. Consequently, unintentionally they often draw unsupported conclusions leading to numerous statistical errors that impact research outcomes. These errors range from flawed hypotheses and improper study design to issues like inadequate sample size, circular analysis, p-hacking and inappropriate presentation of results.

In clinical trials, it is neither reasonable nor possible to study the whole population, researchers typically select a sample, a smaller group of participants that is thought to be representative of the population. This sample is then used to draw inferences about the whole population. It is crucial to choose an appropriate sample size estimation to avoid over-estimation and under-estimation consequences on trial's outcomes. Any negligence in this calculation can result in different types of errors, including approval of false results (type I errors) or rejection of true findings (type II errors).

Too small sample sizes may make the results unrepresentative and not generalizable to the whole population, as well as may not allow to identify clinically significant differences when they exist. In this case, even though there is a therapeutic effect observed, it could be caused by random variations. Alternatively, too large sample sizes result in wastage of researcher's time and valuable resources, and the exposure of a greater portion of the population to the potential risks associated with the intervention, thereby raising ethical concerns. In addition, not meaningful changes could produce statistically significant differences.

From an economic point of view, both overestimating and underestimating the sample size can result in increased costs of the studies. Overestimation leads to higher sampling costs, as analyzing more samples implies additional expenses in terms of money, time and resources. On the other hand, underestimation contributes to increase costs arising from incorrect decisions, as fewer samples may lead to a higher likelihood of errors and their associated consequences. Considering all these aspects, in the approval or rejection of clinical trials the sample size plays a crucial role.

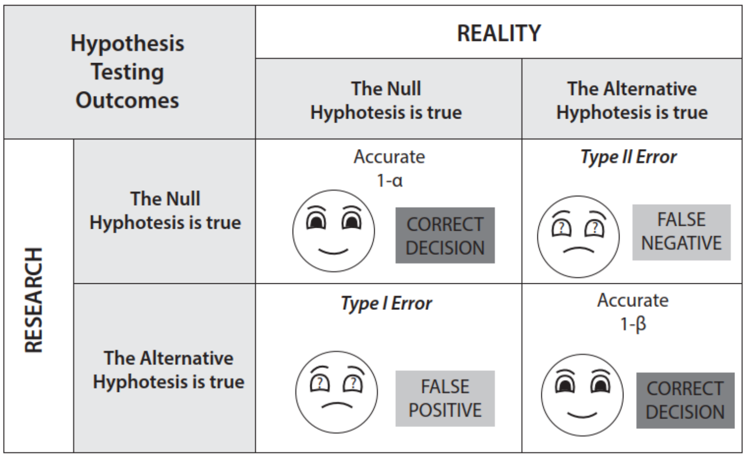
Factors influencing the sample size calculation

Main parameters that determine the required sample size

The calculation of sample size depends on different components, such as type I errors (p-value), type II errors (power), effect size and variability. Evaluating these errors involves considering the results in the context of its statistical initial hypothesis. In significance studies researchers compare two groups, usually the experimental group that receives the tested treatment and the control group that receives placebo, to try to find statistically significant differences associated to the treatment. Commonly used hypotheses are null hypothesis (H_0), suggesting no difference between groups, which means no therapeutical effect of the drug, and alternative hypothesis (H_1), expressing the prediction of the experimental group's outcome after the treatment.

Type I error occurs when a true null hypothesis is incorrectly rejected. It happens when the statistical test indicates a significant result (rejecting the null hypothesis) when, in reality, there is no true effect or difference in the population. The probability of committing such error is denoted as alpha (α). Type I errors are determined by the p-value, which represents the probability of obtaining the observed results (or more extreme) if the null hypothesis is true. On the other hand, the confidence level ($1 - \alpha$) represents the probability that the true value falls within the confidence interval.

Type II error occurs when a false null hypothesis is not rejected, meaning that the study fails to detect a true difference or effect that actually exists in the population, leading to a false negative result. The probability of committing such error is denoted as beta (β). On the other hand, power ($1 - \beta$) is the probability of correctly rejecting a false null hypothesis, that is, to detect a difference between two groups when it truly exists. A higher power reduces the risk of committing this type of error.



Hypothesis Testing Outcomes		REALITY	
		The Null Hyphotesis is true	The Alternative Hyphotesis is true
RESEARCH	The Null Hyphotesis is true	Accurate 1-α CORRECT DECISION	*Type II Error* FALSE NEGATIVE
	The Alternative Hyphotesis is true	*Type I Error* FALSE POSITIVE	Accurate 1-β CORRECT DECISION

Figure 1: Illustration of Type I and Type II errors [4]

Before conducting the study, researchers must establish a balance between type I and II errors. To this end, they must set the acceptable limit for p-value, (α, the level of significance), and for the false negative rate (β), to determine the threshold to reject the null hypothesis. A common α level is 0.05, indicating a 5% chance of such an error. However, α levels can vary based on study goals. Lower α levels minimize the risk of declaring an ineffective treatment as effective, decreasing the chance of making a Type I error but may increase the likelihood of Type II errors. Clinical trials often aim for a power of 80%, meaning there is a 20% chance of missing a real difference (the maximum acceptable value for β is often set at 0.20).

Both types of errors are highly dependent on the sample size. First, type I and II error probability is inversely proportional to the sample size, so larger sample sizes generally enhance the power of study and reduce the likelihood of false negatives and positives. On the other side, increasing the sample size may have ethical considerations, as well as rise the cost and time required for the study. For this reason, researchers need to consider all these practical and ethical constraints when determining the optimal sample size.

Variability refers to the dispersion or spread of data points in a sample, which is the extent to which individual data points in a dataset differ from the mean. Researchers need to anticipate the population variance of an outcome variable, typically estimated by the standard deviation (SD). As variance is usually unknown, an estimate must be employed instead. Investigators often rely on estimates from previous studies or from pilot studies in the population of interest. Variability depends on the homogeneity of the sample, as the more homogeneous the population is, the smaller its variance. This influences the required sample size required to achieve statistical significance, which is larger the higher the variance of the measurements.

Another aspect that influences the sample size is the effect size, a parameter that measures the minimal magnitude of the difference that investigators aim to detect between study groups, which is usually the minimal clinically important difference (MCID). To determine the most appropriate effect size there are different approaches. Some experts in the field are often consulted to determine the smallest difference that would be beneficial in view of its costs and risks. Effect size has a relevant statistical impact, as it is crucial for calculating the required sample size in study design. Sample size is inversely proportional to the square of the effect size, so even small changes in the expected difference have a significant effect on the estimated sample size. Larger effect sizes allow for the detection of effects with smaller sample size, as a fairly wide probability distribution may be acceptable. On the other hand, if small differences want to be detected, great precision and small probability distributions are required, which can be achieved with higher sample sizes.

In Figure 2 there is a practical example of the influence of the effect size in the sample size. For a constant pre-stablished power of 0.8, as the effect size decreases, the required sample size increases. Case 1 (where an effect size of 0.2 required a sample size of 778), is common in epidemiological or meta-analysis studies, where smaller effects are important and sample sizes are very large. On the other side, Case 2 (where an effect size of 1 required a sample size of 34), is common in clinical studies and Case 3 (where an effect size of 2.5 required a sample size of 8), is more common in pre-clinical studies with cell cultures or animals, where samples are usually small (5-10).

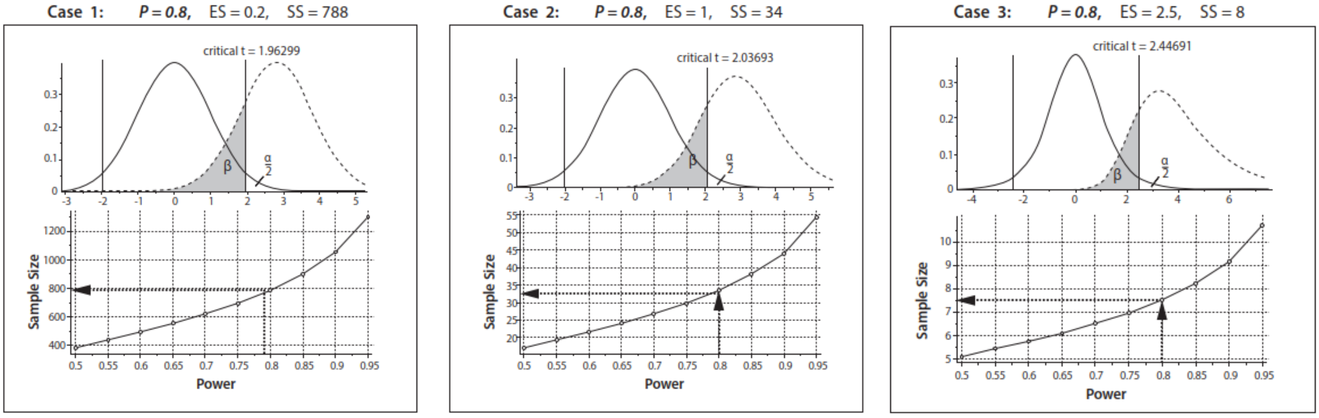


Figure 2: Relationship between effect size and sample size. P – power, ES – effect size, SS – sample size [4]

Other considerations

The calculation of sample size is also influenced by the drop-out rate, which refers to the estimated number of subjects who may exit the study for various reasons. While the sample size calculation provides the number of subjects needed to achieve a specified level of statistical significance for a given hypothesis, the reality of clinical practice may require enrolling additional subjects to compensate these potential drop-outs. The adjusted sample size (N_d) taking this into account can be calculated as:

$$N_d = \frac{n}{(1 - d)}$$

, where n represents the initially calculated sample size and d is the expected drop-out rate. For example, if the necessary sample size is n = 90 and the expected drop-out rate (d) is 0.1, then the sample size considering drop-out will be $N_d = \frac{90}{(1-0.1)} = 100$.

On the other hand, in some clinical trials, it is ethically desirable to perform an unequal treatment allocation, with more subjects in one arm compared to the other. This situation arises in placebo-controlled trials with very ill subjects, where it might be more ethical to assign more patients to the treatment group than to placebo group. The sample size for each arm (n_1 and n_2) can be calculated using the following formulas:

$$n_1 = 0.5 \cdot n \cdot (1 + k) \qquad n_2 = 0.5 \cdot n \cdot \left(1 + \frac{1}{k}\right)$$

,where n is sample size required for the clinical trial in each arm (if both arms were equal) and k is the desired ratio for the sample size of test group and the placebo group ($k = n_1/n_2$). For instance, if the necessary sample size is n = 80 and the desired ratio between test (n_1) and placebo (n_2) is 2:1 (k = 2), then $n_1 = 0.5 \cdot 140 \cdot (1+2) = 120$ and $n_2 = 0.5 \cdot 140 \cdot [1+(1/2)] = 60$.

Finally, the prevalence rate of the condition in the population is also a critical factor in the process of calculating the required sample size, because it influences the statistical power of the study. It is typically estimated from previous literature. However, researchers must be prepared to readjust the sample size if the observed event rate during the trial differs significantly from initial expectations. This adaptability is crucial for maintaining the robustness and reliability of the study.

How to calculate the sample size

There are different methods to determine the most adequate sample size depending on the specific study design, the hypothesis testing and the statistical analysis planned to use. In each case, different formulas are used. Two main examples are the comparison of two proportions and the comparison of two means.

Simple size calculation in the comparison of two means

For the comparison of two means, the formula is $n = \frac{(Z_{\alpha/2}+Z_{1-\beta})^2 2\sigma^2}{(\mu_1+\mu_2)^2}$, where

- n: sample size required in each group
- σ: standard deviation
- μ_1 and μ_2: mean of group 1 and 2
- $\mu_1 - \mu_2$: clinically significant difference of means of both groups (effect size)
- $Z_{1-\beta}$: Z-score for the desired power
- $Z_{\alpha/2}$: Z-score for the desired level of significance

Z values depend on the desired power and level of significance. Values for conventional values of α and β are shown in Table 1.

Table 1. Showing constant values for conventional values of α and β [3]

Z-values for conventional values of α		$Z_{\alpha/2}$
α	0.05	1.96
	0.01	2.58
Z-values for conventional values of β		$Z_{1-\beta}$
β	0.20	0.84
	0.1	1.28
	0.05	1.64
	0.01	2.33

For example, a placebo-controlled randomized trial aims to compare the effectiveness of Drug A in preventing the stress response to laryngoscopy, by studying the difference in mean systolic blood pressure between two groups. Considering that it is set a level of significance (α) of 5% and a power ($1-\beta$) of 90%, and that previous studies showed that during laryngoscopy there is an average rise of 20 mm Hg in systolic blood pressure, with a standard deviation of 15 mm Hg, sample size can be obtained by substituting in the formula.

- $\sigma = 15$ mm Hg
- $\mu_1 - \mu_2 = 20$ mm Hg
- $Z_{1-\beta} = 1.28$ ($1-\beta = 0.90$)
- $Z_{\alpha/2} = 1.96$ ($\alpha = 0.05$)

$$n = \frac{(1.96 + 1.28)^2 \cdot 2 \cdot 15^2}{20^2} = 11.81 \approx 12$$

In this case, a sample size of 24 individuals, 12 in each arm, is sufficient to detect a clinically relevant difference of 20 mm Hg between groups in systolic blood pressure with 90% power and a 5% level of significance.

Sample size calculation for the comparison of two proportions

For the comparison of two proportions, the formula is $n = \frac{(Z_{\alpha/2}+Z_{1-\beta})^2[p_1(1-p_1)+p_2(1-p_2)^2]}{(p_1-p_2)^2}$, where

- n: sample size required in each group
- $Z_{1-\beta}$: Z-score for the desired power
- $Z_{\alpha/2}$: Z-score for the desired level of significance
- p_1 and p_2: proportion of event of interest (outcome) for group 1 and group 2
- $p_1 - p_2$: clinically significant difference of proportions of both groups

For example, a placebo-controlled randomized trial aims to compare the effectiveness of Drug A in curing infants suffering from sepsis, by studying the difference in the incidence between two groups. Considering that it is set a level of significance (α) of 5% and a power ($1-\beta$) of 80%, and that previous studies showed that Drug A could cure 50% of subjects, and a clinically important difference of 16% as compared to placebo is acceptable, sample size can be obtained by substituting in the formula.

- $p_1 = 50\% = 0.5$ (proportion of subjects cured in drug A group)
- $p_2 = (50 - 16)\ \% = 34\ \% = 0.34$ (proportion of subjects cured in placebo group)
- $p_1 - p_2 = 16\% = 0.16$ (effect size: clinically significant difference)

- $Z_{1\text{-}\beta} = 0.84$ ($1- \beta = 0.80$)
- $Z_{\alpha/2} = 1.96$ ($\alpha = 0.05$)

$$\mathrm{n} = \frac{(1.96 + 0.84)^2[0.5(1 - 0.5) + 0.34(1 - 0.34)^2]}{0.16^2} = 145.29 \approx 146$$

In this case, a sample size of 292 infants, 146 in each arm, is sufficient to detect a clinically relevant difference of 16% between groups in curing sepsis with 80% power and a 5% level of significance.

Nonograms for sample size calculation

Besides the formulas, there are available different nonograms of software to estimate the most adequate sample size. Figure 3 illustrates one of the most commonly used nonograms, developed by Gore and Altman, which allows to estimate the most adequate sample size initially selecting the effect size (standardized difference) and aimed power for the study. It assumes Gaussian distributions. To obtain the sample size, a straight line must be drawn connecting the two values, which will cross the significance level region. Choosing the desired significance level, the intercept allows to obtain the required sample size.

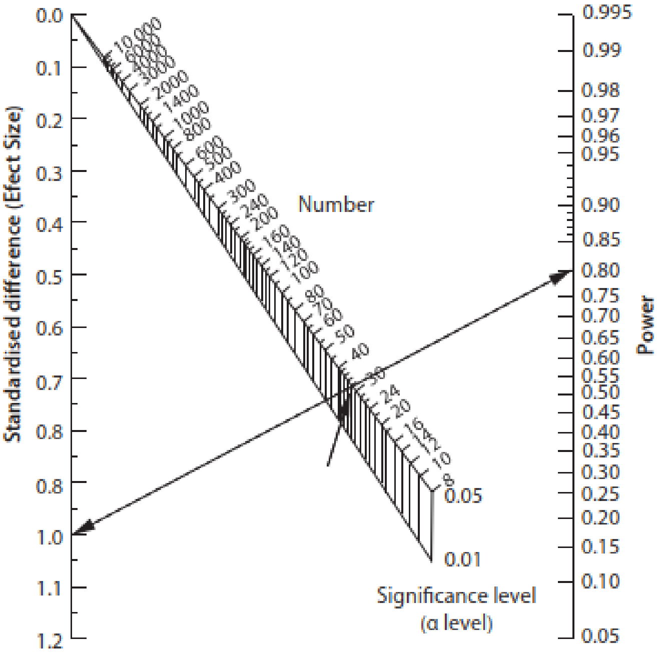


Figure 3: Nonogram for the calculation of sample size, in the comparison of two groups of equal size [4]

For example, for an effect size of 1, a power of 0.8 (80%) and a significance level of 0.05, the determined sample size is 30 (Figure 3).

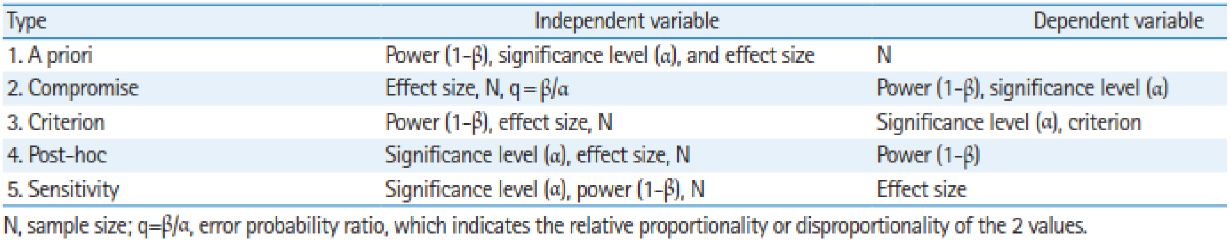
Software for the sample size calculation

In many cases, manual calculation of the sample size is too complex and difficult, specially for researchers that are not statisticians. For this reason, recently there has been a development of different software and websites that aim to effectively determine sample sizes for a variety of study types. Some tools that are free-to-use and don't require payment are G-Power, R and Piface. For example, R can be customized to meet individual statistical requirements, as it has specific program modules called packages that can be added to a base program.

On the other hand, Piface is a Java application designed specifically for sample size estimation and post-hoc power analysis. The most professional software is PASS (Power Analysis and Sample Size), which is not freely available, but offers the possibility to analyze approximately 200 different study types.

G-power, which can be downloaded for free at www.psycho.uni-duesseldorf.de/abteilun-gen/aap/gpower3, is capable of calculating statistical power and sample size for various types of statistical tests, including t-tests, F-tests, χ^2 tests, Z-tests and some exact tests. To this end, first researchers need to establish the objective and hypotheses of the study and choose the most appropriate statistical test. Then, they can select between five possible power analysis methods (summarized in Table 2), depending on the variables to be calculated and the given variables. *A priori* analysis are conducted before conducting the study, and the aim is to determine the required sample size for a specific desired effect size, confidence level and power. On the other hand, *post-hoc* analysis are performed after the completion of the study, and the aim is to calculate the power of the study for the sample size, effect size and confidence level used in the study. To obtain the desired variable, researchers need to input the required variables for analysis and select the "calculate" button.

Table 2. Power analysis methods in G-power [6]



Type	Independent variable	Dependent variable
1. A priori	Power (1–β), significance level (α), and effect size	N
2. Compromise	Effect size, N, q = β/α	Power (1–β), significance level (α)
3. Criterion	Power (1–β), effect size, N	Significance level (α), criterion
4. Post-hoc	Significance level (α), effect size, N	Power (1–β)
5. Sensitivity	Significance level (α), power (1–β), N	Effect size

N, sample size; q=β/α, error probability ratio, which indicates the relative proportionality or disproportionality of the 2 values.

Reporting of the sample size

As per the CONSORT statement guidelines, it is imperative for all published randomized controlled trials to report and justify their sample size calculations. This report has greatly increased in the past decades, but only about a third perform it adequately. These calculations are commonly reported in a deficient way, only mentioning the confidence level and the power, but not specifying essential parameters such as the effect of interest and variability. Despite the recommendation from the CONSORT group to disclose details about sample size determination, researchers and reviewers often do not take this seriously, so there is a need to enhance transparency in this aspect.

Bibliography

1. Sakpal TV. Sample size estimation in clinical trial. *Perspect Clin Res*. 2010; 1(2): 67-9
2. Kennedy-Saffer L, Hughes MD. Sample size estimation for stratified individual and cluster randomized trials with binary outcomes. *Stat Med*. 2020; 39(10): 1489–1513. DOI: 10.1002/sim.8492
3. Gupta KK, Attri JP, Singh A, Kaur H, Kaur G. Basic concepts for sample size calculation: Critical step for any clinical trials! *Saudi J Anesth*. 2016; 10(3):328-31. DOI: 10.4103/1658-354X.174918
4. Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochem Med (Zagreb)*, 2021; 31(1):010502. DOI: 10.11613/BM.2021.010502

5. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J*. 2003; 20(5):453-8. DOI: 10.1136/emj.20.5.453
6. Kang H. Sample size determination and power analysis using the G*Power software. *J Educ Eval Health Prof*. 2021; 18:17. DOI: 10.3352/jeehp.2021.18.17
7. Brasher PM, Brant RF. Sample size calculations in randomized trials: common pitfalls. *Can J Anaesth*. 2007; 54(2):103-6. DOI: 10.1007/BF03022005
8. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. *BMJ*. 2009; 338:b1732. DPI: 10.1136/bmj.b1732