Sample size calculation

A High-level Overview

Márton Kiss MD

2024-11-14

Introduction

This presentation provides a high-level overview of sample size calculation in clinical trials.

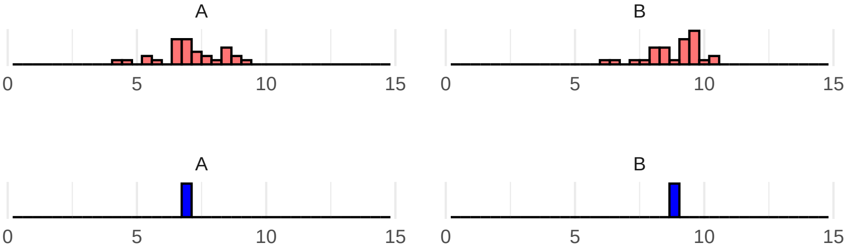
We will discuss the key concepts and considerations for estimating the sample size for clinical trials.

The goal of this presentation however is not to provide a comprehensive guide to sample size calculation, but rather to introduce the main concepts and considerations involved in this critical aspect of clinical trial design.

Example design

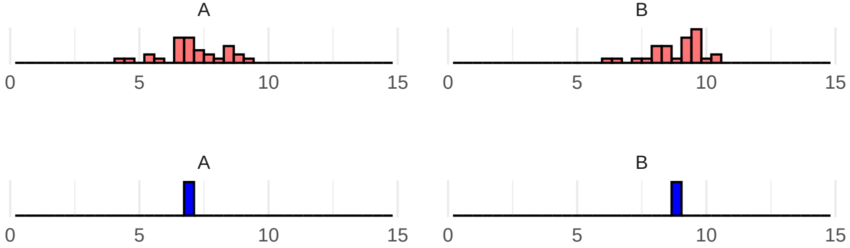
Let's consider a hypothetical clinical trial to illustrate the sample size calculation process.

- Two arms (types of treatment received by subjects): A and B
- Suppose the sample size is 30



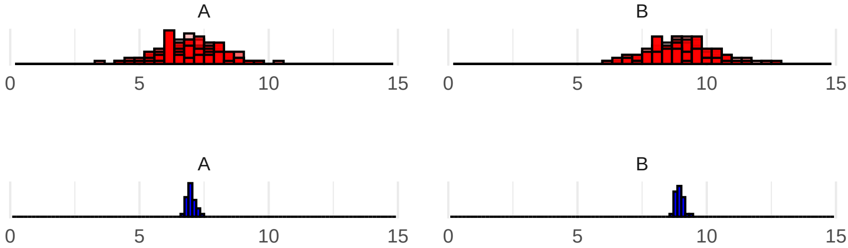
Example design (cont.)

The below plot shows the distribution of the outcome variable in the two treatment groups (in red) and the mean values (in blue) for a randomly generated study.
Parameters used for generation: $\mu_A = 7$, $\mu_B = 9$, SD=1 for both groups.



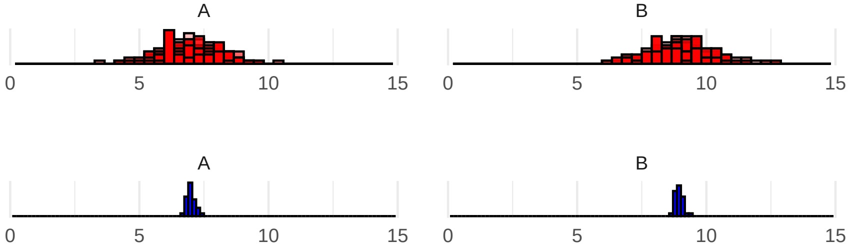
Example design, multiple studies

- Now we simulate 30 similar studies. (The red histograms are severely overplotted, ie. they are like 30 histograms put together.)
- The mean values are shown in blue. They are much more tightly distributed than the individual observations.



Example design, multiple studies (contd.)

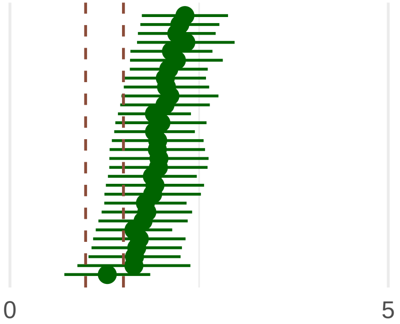
- The mean values per group (blue) have a SD of $\frac{1}{\sqrt{30}}$ being much more precise than the individual observations (red).
- This is the principle of the Central Limit Theorem. If **N** is *"big enough"* the mean of the observations will be normally distributed.



Next, we consider the mean difference between the **A** and **B** groups.

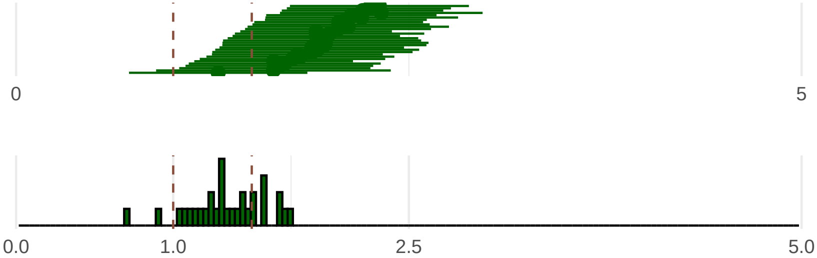
Example design, A-B difference

- These are the estimates **per simulated study** for the difference between the **A** and the **B** group means (point estimate + *97.5% CI*).
- If we tested the **difference** to be greater than a value below the smallest CI limit, we would have **rejected the null hypothesis** in **each case**.



Example design, A-B difference (contd.)

- The distribution of the **lower ends** of the CI's is of interest (2nd plot).
- If we tested the **difference** to be greater than 1, we would have been right 28/30 ~ 93% of cases.
- If we tested the **difference** to be greater than 1.5, we would have been right 9/30=30% of cases.



Defing Power

- If we tested the **difference** to be greater than 1, we would have been right 100% of cases.
- If we tested the **difference** to be greater than 1.5, we would have been right 9/30=30% of cases.

The (estimated) chance for rejecting the null hypothesis for a study is called the *Power* of the study.

A clinical trial should be powered between **80%** and **90%**.

- If the Power is less than 80%, the **underpowered** study is unethical, as the risk/suffering to the participants has to be justified by a high chance of meaningful results
- if the Power is more than 90%, the study is **overpowered** and unethical because *more* subjects are subjected to risk & undue suffering than necessary (by general consensus).

Calculating Power

- If we tested the **difference** to be greater than 1, we would have been right 100% of cases.
- If we tested the **difference** to be greater than 1.5, we would have been right 9/30=30% of cases.

The Power of the study was **simulated** in this example.

It is a flexible approach and can be useful in non-standard design or situations (eg. the variable of interest has a peculiar distribution, the primary endpoint is a composite of not independent variables, etc.)

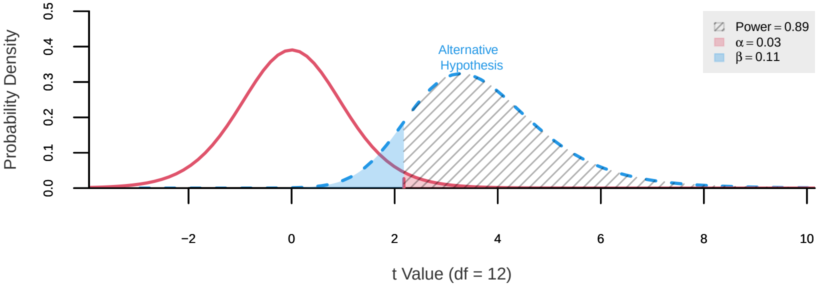
For standard designs, closed formulas exist for calculating the Power.

Distribution of the lower CI limits

- In general what we are looking for, is the distribution of the **lower CI limits** of the difference between the groups (or the critical value of the relevant statistic to be tested).
- We could calculate the proportion of cases where the difference is greater than a certain value and could establish a threshold above/below the chance of rejecting the null hypothesis is at our desired level for Power.
- The distribution of the lower CI limits can either be assumed to have a **normal distribution** or a **t-distribution** (this is better).
- The precise distribution is given by the **non-centrality parameter** $\lambda = \frac{\delta\sqrt{n/2}}{\sigma}$ (along with the *degrees of freedom* if a t-distribution is used).
- λ has a different form for different designs and it represents the effect size, as in the size of the difference relative to its variability.

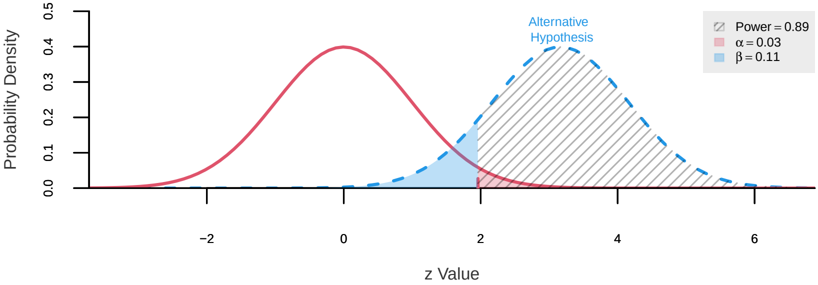
Power calculation based on the Noncentrality Parameter

Below is the Power calculation for the example study based on the closed form solution using the t-distribution. The results show that for a sample size of 6 subjects per arm, the Power would be ~89%.



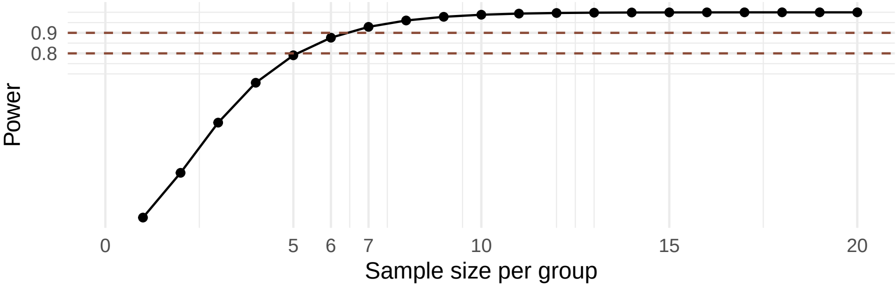
Power calculation based on the Noncentrality Parameter

Below is the Power calculation for the example study based on the closed form solution for the normal distribution. The results show that for a sample size of **5** subjects per arm, the Power would be ~89%.



Relationship between Power and Sample size

- Seen below is a **Power curve** for the study (using the t-distribution)
- For our example, the Power vs. the Sample size is shown below based on a closed formula
- The researcher may choose either **6** or **7** subjects **per group** to achieve a Power of at least 80% or at least 90%.



Corrections for dropouts

Assume 25% dropout rate. For 90% Power, the final sample size should be: $\frac{7}{100\%-25\%} = 9.33 \approx 10$ subjects per group.

For a total of $10 \times 2 = 20$ subjects

- 7 subjects per group after 25% dropouts would result in a total of 20 subjects for the study.
- A **correction** for dropouts is necessary because the Power calculation is based on the number of subjects that **finish** the study.
- Avoid an incorrect way of accounting for dropouts by multiplying $(7 \times 125\% = 8.75)$
- It may be necessary for logistical purposes to calculate the **screening failures** as well. A proportion of participants will fail screening (eg. by having out-of-range lab results when the study calls for healthy participants)

Sensitivity analysis for Sample Size estimation

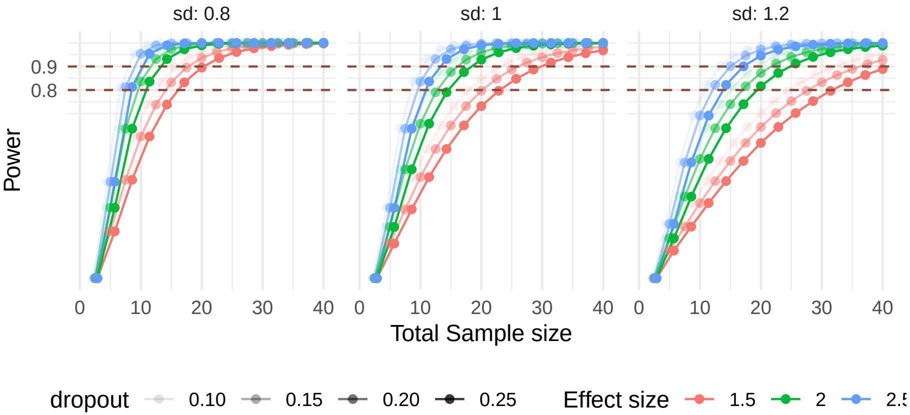
In our example we have arrived at 20 subjects for the sample size, but have used the following suppositions:

- **The standard deviation of the outcome variable is 1 in both groups**
- **The difference between the groups is 2**
- **The dropout rate is 25%**
- The Power is 90%
- The null hypothesis is that the difference is 0
- (The alpha level is 0.05 for two-sided testing)

The first three items are **estimated**, and therefore we invite uncertainty. Moreover, the errors in the estimates may compound one another.

Sensitivity analysis for SS estimation (contd.)

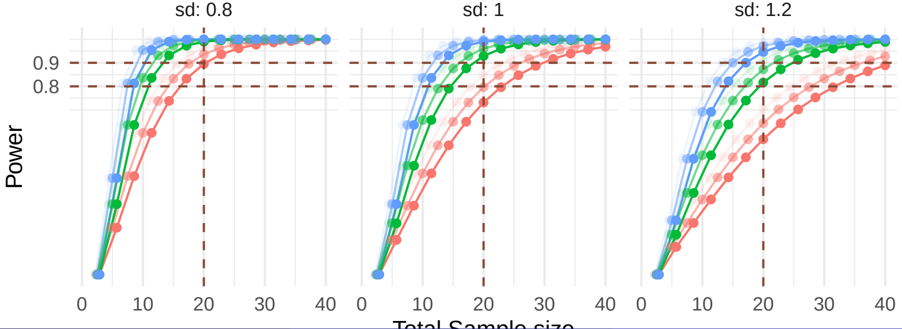
Unfortunately seldom seen in practice, but the interaction of these parameters should be investigated in a **sensitivity analysis** before the start of the study. This brings us to another line of thinking about the sample size.



Sensitivity analysis for SS estimation (contd..)

When a given sample size is considered (here: **20 subjects** total), we could investigate under what circumstances would our study be acceptable. Here the study would be acceptable if:

- The effect size were lower than anticipated (1.5) with a SD of 0.8 even with a dropout rate of 25%
- The SD was higher than anticipated (1.2) with a difference of 2 and a dropout rate of 25% *but not when the difference was also lower (1.5)*
- The SD was unchanged, but the difference was lower (1.5) with a dropout rate of 25%



Special cases

- In certain cases (survival analysis) it is not the *sample size* what matters, but the *number of events*. This introduces another parameters (ratio of patients experiencing an event) which adds another layer of uncertainty.
- We are not covering **adpative designs**. These are studies in which the sample size may be readjusted based on *"interim analyses"* along with the stopping the study or certain arms of the study for futility/early success.
- In general using an **adaptive design** may be more efficient, but adds additional complexity and uncertainty for the trial.
- In certain cases, the analysis method used are different (more elaborate) compared to the method used to calculate the sample size (eg. analysis with baseline correction for the investigation of change and using a mixed model while the sample size calculation was based on a simple t-test).
- The precise methodology for determining the Sample Size often (should) involve a statistician in these special cases.

Importance of design

- **Before** a valid csample size calculation could be performed, the study design has to be defined.
- Of particular interest is the **primary endpoint** of the study (along with its scale) and its method of evaluation (in broad strokes).
- Often Investigators are reluctant to give the relevant **parameters** for variability or effect size. This is perhaps due to the high level of uncertainty regarding these numbers, but its the joined task of the statistician and the investigator to come up with a reasonable estimate if necessary, through several iterative steps.
- While it is possible to define co-primary endpoints, or a joined hypothesis for the primary endpoint, the level of significance has to be adjusted in these cases, (often making the study unfeasible).
- In case multiple variables are to be investigated (which is in most cases), most investigations may be designated as *secondary endpoints* for which no power calculation is required (though the results of these endpoints therefore cannot be guaranteed and they are a bit less persuasive).

A remark about post hoc power.

Two friends are playing golf. The first one takes a swing and the ball lands on a particular spot. The second one says: “Wow! What were the odds of that ball landing exactly on that spot out of all other spots on the field?”

After a study, sometimes the **Post hoc** power is reported. This is done by taking the *observed* parameters and repeating the power calculation with them instead of the original assumptions.

This is problematic on a number of levels, one of which is the direct link between the observed p-value of the difference and post-hoc Power, so using the post-hoc Power to provide additional confirmation or to negate the results is circular reasoning.

Specific implementations

- A number of software has implementations for calculating Power in several situations where a closed-form solution is available.
- A common software is **GPower**, and implementations in SAS along with most modern statistical software are also available.
- The author's favorite implementation is the **pwrss** R package. For adaptive designs, the **rpact** package is recommended.
- To use any of these implementation requires the user to be informed about the specific distributions of their outcome variable and study design.