Sample Size Justification

1

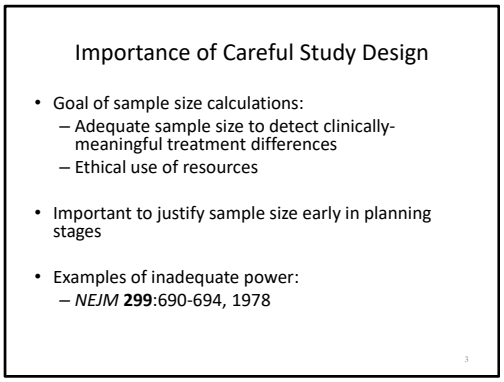
In this module, I will present a summary of sample size justification.

Outline

- Statistical Concepts: hypotheses and errors
- Effect size and variation
- Summary of factors influencing sample size and power

2

In this module I will briefly review information related to hypothesis testing, which will form the framework for our discussion of sample size requirements.
I will also present a summary of the primary factors that influence sample size estimates, namely, the significance level, power, effect size and variability. Keep in mind that additional study design factors, such as withdrawal and longitudinal repeated measures, also impact the sample size.



Let's begin by first noting the importance of careful planning before initiating a research project. After developing your question of interest, as you begin to think about the research design, you will also need to consider the number of participants or observations that are needed to make "confident" statements about associations or intervention effects based on your sample of observed data.

Sample size calculations are performed early in study planning and development to ensure that you will have sufficient information to detect clinically-meaningful treatment differences, if they exist. Furthermore, you want to use financial, staff, and patient resources ethically. If the study is poorly planned or implemented, no useful information will be gained and those resources will have been wasted. Therefore, it is important to justify the sample size early in the planning stages of the study.

Journal editors and grant funding agencies are more proactive in their requirement for evidence that the sample size has been appropriately chosen for a given project. This was not always the accepted standard. In a 1978 review of the medical literature, published in the New England Journal of Medicine, found that, of 71 "negative" trials, 94% had >10% chance of missing a true effect and 70% had >10% chance of missing a true 50% reduction. In general, the investigators found that TRIALS WERE UNDERPOWERED TO DETECT TRUE DIFFERENCES.

Example

- Double-blind randomized trial
- Compare inhaled corticosteroids with oral corticosteroids in the management of severe acute asthma in children
- 100 children were randomized
- Primary outcome: forced expiratory volume (as a percentage of the predicted value) 4 hours after treatment administration

  – Schuh et al., (2000) *NEJM.* **343**(10)689-694.

A double-blind randomized trial was conducted to determine how inhaled corticosteroids compare with oral corticosteroids in the management of severe acute asthma in children.

In the study, 100 children were randomized to receive one dose of either 2 mg of inhaled fluticasone or 2 mg of oral prednisone per kilogram of body weight.

The primary outcome was forced expiratory volume (as a percentage of the predicted value) 4 hours after treatment administration.

Example: Sample Size Justification

- In the article the authors state

"The sample size was based on an estimated standard deviation of 15 for the change in the percentage of the predicted FEV_1 in the Prednisone group. In order to allow detection of a 10 percentage point difference between the groups in the degree of improvement in FEV_1 (as a percentage of the predicted value) from base line to 240 minutes and to maintain an alpha (α) error of 0.05 and a beta (β) error of 0.10, the required size of the sample was 94 children.".

Reference: Schuh et al., (2000) *NEJM.* **343**(10)689-694

5

In the methods section of the manuscript the authors provided this justification for the sample size. After reviewing this series, you will be able to identify the factors and assumptions that impacted the required sample size; namely, the alpha level, the power, the effect size, and the variation.

Study Design and Primary Endpoint

- Sample size calculations depend on:
  - Study Design:
    - Number of groups being compared and prevalence of group status
    - Experimental or observational study (account for confounding factors)
    - Independent observations or cluster-correlated data
  - Type of response variable:
    - Continuous response: weight, blood pressure
    - Dichotomous response: presence/absence
    - Time to event: survival time, time to relapse

6

There are multiple design features that will impact the sample size justification.

First, the required sample size will differ depending on the study design. Are we comparing only two groups or are there more than two groups? Will the groups be equally sized? Will the assignment to the intervention or control be randomized? Do we need to account for confounding factors in our analysis? Are observations independent or do they cluster within families or social groups such as schools or households?

The sample size will also differ depending on the type of endpoint. Are we comparing means from continuous measures like weight? Are we comparing proportions between groups, such as the proportion of patients responding to treatment? Are we comparing the time to event distributions between groups, such as the time to death between treatment groups?

Answers to these questions will impact our approach to estimating the required sample size.

Statistical Concepts
Hypotheses

- Null hypothesis: H_0
  - Typically a statement of no treatment effect
  - Assumed true until evidence suggests otherwise
  - Example: H_0: Mean FEV_1 is same in treatment groups
- Alternative: H_A
  - Reject null hypothesis in favor of alternative hypothesis
  - Often two-sided
  - Example: H_A: Mean FEV_1 differs between treatment groups

7

In hypothesis testing, when we aim to detect intervention effects or associations between exposures and outcomes, meaning, superiority settings, we have two statistical hypotheses in mind, the Null Hypothesis and the Alternative Hypothesis. In a superiority setting, the Null Hypothesis is a statement of no treatment effect or no association and is assumed true until we find evidence to suggest otherwise. Relative to our study example, the Null Hypothesis will be that the mean FEV_1 is the same in the treatment groups.

The Alternative Hypothesis is the hypothesis that we hope to find evidence in favor of. In practice, we will reject the Null Hypothesis in favor of the Alternative Hypothesis. The Alternative Hypothesis is typically two-sided, we will discuss what this means on the next slide. Relative to our study example, the Alternative Hypothesis will be that the mean FEV_1 differs between the treatment groups.

Statistical Concepts
Hypotheses

- Alternative hypothesis may be one-sided or two-sided
  - Example:
    - Null hypothesis: Mean FEV_1 is same in patients receiving different treatments
    - Alternative hypothesis:
      - One-sided: Mean FEV_1 is lower in patients receiving treatment A
      - Two-sided: Mean FEV_1 is different in patients receiving treatment A relative to treatment B
- Choice of alternative does affect sample size calculations. Typically a two-sided test is recommended.

8

The Alternative Hypothesis may be one or two sided and the Null Hypothesis is the compliment of the Alternative Hypothesis. As an example, if the Null Hypothesis is that there is no difference in mean FEV_1 between the groups, a two-sided alternative would indicate that there is a difference in the mean FEV_1 measure and would not specify the direction of the difference. On the other hand, if the Alternative Hypothesis was one-sided, we would need to specify the direction of the difference, for example, indicating that mean FEV_1 was lower among patients who received the oral agent compared to the inhaled agent. The Null Hypothesis in this case would be that mean FEV_1 for patients receiving the oral agent was greater than or equal to that among patients who received the inhaled agent; the complement of the alternative hypothesis.

The choice of a one-sided or two-sided alternative hypothesis does impact the required sample size estimate. The required sample size will be larger for a two-sided alternative hypothesis. Therefore, alternative hypotheses are typically specified as 2-sided to be conservative.

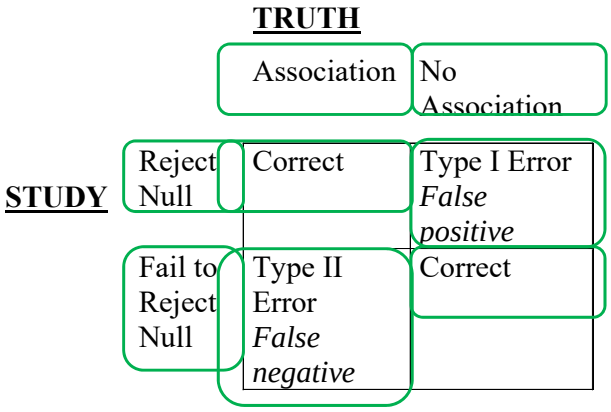
Historically, there are some very well known clinical trials that were designed to have a one-sided alternative hypothesis, which was incorrectly assumed.

The Cardiac Arrhythmia Suppression Trial addressed the question of whether the

suppression of asymptomatic or mildly symptomatic ventricular arrhythmias after MI would reduce the rate of death from arrhythmia. The investigators proposed as one-sided alternative stating that the mortality rate would be lower under the anti-arrhythmia treatment; however, they actually observed a higher mortality rate under the treated arm – opposite of their assumed one-sided alternative.

Statistical Concepts
Errors

- Errors associated with hypothesis testing



9

When conducting hypothesis tests, we collect a sample of data and then make a decision or form inferences based on the sample of data. When making a decision, we either make a correct decision or make an error relative to the true status. In truth, there either is an association or is no association between, say, treatment and outcome. Based on our data, we make a decision to either reject the null and conclude that there is an association or we fail to reject the null hypothesis and conclude that there is no significant indication of an association.

If there is an association between the intervention and response, and we reject the null, we have made a correct decision.
If there is no association between the intervention and response, and we fail to reject the null, we have made a correct decision.
In both other cases, we make an error. If there is no association between the intervention and response, and we reject the null, we have committed a false positive error, which we refer to as a Type I error. Alternatively, if there is an association between the intervention and response, and we fail to reject the null, we have failed to detect a true intervention effect and have committed a false negative error, which we refer to as a Type II error.

Statistical Concepts
Significance Level

- Significance level: alpha (α)
  - Probability of a Type I error
  - Probability of a false positive
  - Example: If the effect on FEV_1 of the treatments do not differ, what is the probability of incorrectly concluding that there is a difference between the treatments?
  - Typically chosen to be 5%, or 0.05

10

Let's review the two types of errors that we may commit in hypothesis testing.

If there is no association between the intervention and response, and we reject the null, we have committed a false positive error, which we refer to as a Type I error. We denote the probability of a type I error as alpha. In terms of our example study, the alpha level addresses the question, "If the effect on FEV_1 of the treatments do not differ, what is the probability of incorrectly concluding that there is a difference between the treatments?". In practice, the probability of a Type I error should be low and is typically set to be 5%. Meaning, we often test hypotheses using an alpha level of 0.05.

Statistical Concepts
Power

- Power: 1-beta (1-β)
  - Probability of detecting a true treatment effect
  - Power = (1- probability of a false negative)
    = (1-probability of Type II error)
    = (1-β) = probability of a true positive
  - Example: If the effects of the treatments do differ, what is the probability of detecting such a difference?
  - Typically chosen to be 80-99%

11

The other type of error that can be made in hypothesis testing is a Type II error. If there is an association between the intervention and response, and we fail to reject the null, we have failed to detect a true intervention effect and have committed a false negative error, which we refer to as a Type II error.

The probability of a Type II error is related to the Power of the study. The Power of the study is the probability of rejecting the null hypothesis when the alternative hypothesis is true. This is a correct decision; if there is a true treatment effect, Power is the probability of detecting that treatment effect. The Power can be calculated as one minus the probability of a Type II error, where the probability of a Type II error is denoted as beta.

In our example study, power answers the question "If the effects of the treatments do differ, what is the probability of detecting such a difference?".

In practice, we want the Power of a study to be high. We typically design our studies to have Power of at least 80%.

Treatment Effect

- What is the minimal, clinically significant difference in treatments we would like to detect?
- Pilot studies may indicate magnitude
- Example: The authors felt that a 10 percentage point difference in FEV_1 between the treatment groups was clinically significant
- Denoted by delta (δ)

12

In addition to the Type I and Type II error rates, the sample size will also be impacted by the intervention effect size and the amount of variability in the data.

The effect size is the minimal, clinically significant difference in treatments we would like to detect. The effect size reflects both the expected impact of the intervention (how large of an effect can be expected under the new intervention) as well as
the clinical threshold that defines the magnitude of the effect that is important. For example, a new intervention may only have a slight impact on response and this small effect would not be clinically meaningful.

As the effect size increases, it becomes easier to detect a difference between groups, or an association between exposure and outcome, and therefore, the required sample size that is needed, assuming all other factors remain fixed, is smaller.

Estimates of the effect size can be derived from pilot studies or related studies

reported in the literature.

In terms of our example, the authors felt that a 10 percentage point difference in FEV_1 between the treatment groups was clinically significant.

The effect size is typically denoted as delta.

Variability in Response

- To estimate sample size, we need an estimate of the variability of the response in the population
- Estimate variability from pilot or previous, related study
- Example: The authors estimate that the standard deviation of FEV_1 is 15 percentage points.
- Denoted by sigma (σ)

13

The required sample size is also impacted by the variability in the response.

As the variability, or noise in the data, decreases it becomes easier to detect a difference between groups, or an association between exposure and outcome, and therefore, the required sample size that is needed, assuming all other factors remain fixed, is smaller.

Estimates of the variance can be derived from pilot studies or related studies reported in the literature.

In terms of our example, the authors estimate that the standard deviation of FEV_1 is 15 percentage points.

The variability is typically denoted as sigma, corresponding to the standard deviation.

Sample Size Calculators

- PS: Power and Sample Size Calculation
  http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize

- Warning: very easy to generate sample size estimates
  - More difficult to generate appropriate sample size requirements

14

There are several online sample size calculators and programs. One that I find to be an accurate and broadly-applicable program is the PS Power and Sample Size software from the biostatistics department at Vanderbilt. The link is provided on this page.

Keep in mind that while it is very easy to generate required sample sizes using the online tools and programs, it is much more difficult to know if the assumptions that you have made and the study design parameters are appropriate and meaningful for the question and study at hand. Involvement of a statistician in study planning is recommended.

Factors Influencing Sample Size

Assuming all other factors fixed, sample size increases when the following changes occur:

↑ power ⇒ ↑ sample size
↓ significance level ⇒ ↑ sample size
↑ variability in response ⇒ ↑ sample size
↓ effect size ⇒ ↑ sample size

15

Now, we can consider the four primary factors that impact sample size and note some general results.

Assuming all other factors fixed, sample size increases when the following changes occur:

Increased power: if we want a greater probability of detecting a true difference, we need more information, and therefore, a larger sample size.

Lower significance level: if we want a smaller chance of incorrectly concluding there is a treatment effect (meaning, a lower false positive error), we will perform our hypothesis tests using a lower alpha level and therefore, it will be more difficult to reject the null hypothesis in favor of the alternative hypothesis. If it is more difficult to detect a significant difference, we need more information and therefore, a larger sample size.

Increased Variability: if the response is more variable, it is more difficult to detect signals from the data amidst the noise and therefore, we need more information and therefore, a larger sample size.

Decreased effect size: if the effect size is smaller, the groups are more similar, and therefore, it is more difficult to detect a significant difference and we need more information and hence, a larger sample size.

Factors Influencing Power

Assuming all other factors fixed, power <u>decreases</u> when the following changes occur:

$\downarrow$ significance level $\Rightarrow \downarrow$ power
$\downarrow$ effect size $\Rightarrow \downarrow$ power
$\uparrow$ variability in response $\Rightarrow \downarrow$ power
$\downarrow$ sample size $\Rightarrow \downarrow$ power

16

Assuming all other factors fixed, <u>power decreases</u> when the following changes occur:

Lower significance level: if we want a smaller chance of <u>incorrectly</u> concluding there is a treatment effect (meaning, a lower false positive error), we will perform our hypothesis tests using a lower alpha level and therefore, it will be more difficult to reject the null hypothesis in favor of the alternative hypothesis. If it is more difficult to detect a significant difference, we have lower power.

Decreased effect size: if the effect size is smaller, the groups are more similar, and therefore, it is more difficult to detect a significant difference and therefore, we have lower power.

Increased Variability: if the response is more variable, it is more difficult to detect signals from the data amidst the noise and therefore, we have lower power.

Decreased sample size: with a smaller sample size, we have less information and therefore, we are less likely to detect significant associations or differences and hence have lower power

Summary

- Sample size calculations are an important component of study design
- Want sufficient statistical power to detect clinically significant differences between groups when such differences exist
- Calculated sample sizes are estimates

17

In summary, we recognize the importance of sample size planning early in research project planning. We aim to have sufficient statistical power to detect clinically significant differences between groups when such differences exist. The required sample size is impacted by hypothesis testing error rates (related to the significance level and power), the effect size and variability. Pilot studies are useful for deriving estimates of the effect size and variability. It is important to keep in mind that the calculated sample sizes are estimates that reflect our best guess regarding the effect size and variability. In practice, we may specify more conservative estimates to ensure sufficient power.

References

Sample Size Justification

- Freiman, J. A. et al. "The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 72 "negative" trials. *N Engl J Med*. **299**:690-694, 1978.
- Friedman, L. M., Furberg, C. D., DeMets, D. L., *Fundamentals of Clinical Trials*, Springer-Verlag, 1998, Chapter 7.
- Lachin, J. M. "Introduction to sample size determination and power analysis for clinical trials". *Controlled Clinical Trials*. **2**:93-113. 1981.

18

This slide includes several references that provide a useful overview of sample size calculations and sample size justification.