Grant Writing Seminar Series
SESSION 5:

Hypothesis & Statistical Testing, Sample Size, Power, & Detectable Effect Size

Betsy Tolley, PhD

OBJECTIVES

- **Provide a concise description of how the process of sample size estimation and power calculation works**
- **Encourage young investigators to participate actively in considering the assumptions and estimates (or informed guesses) involved in sample size planning**

- Formulating a specific, simple research hypothesis at the start of the study design process.
- Identifying single primary null and alternative hypotheses.
- Specifying only one major predictor variable and only one clinically important outcome variable.
- Choosing the “best” study design.
- Identifying the “best” variable to measure the outcome.
- Considering potential problems before it is “too late”.

Question:

- **What type of issue is sample size determination and why is it important to specify the appropriate number of participants needed to complete a study?**

  - A scientific issue.

  - A statistical issue.

  - A practical issue.

  - An ethical issue.

.

Answer

- **All of the above.**

The overall goal of this lecture is to provide a framework for considering the entire study rather than focusing exclusively on one aspect—the number of participants completing the study.

- **The goal of sample size planning is estimating the appropriate number of participants for a chosen study design.**
  - Too small a sample size leaves the research question unanswered.
  - Too large a sample size is more complicated and costly than necessary.
- **Sample size calculations are useful and informed guesses.**

These calculations are only as accurate as the data and estimates on which they are based. Sample size planning can reveal (1) when a study is feasible or not and (2) when another predictor or outcome variable is needed.

Hypotheses

- **Sample size planning begins with restating the research question as a research hypothesis.**

  The research hypothesis summarizes the main elements of the study—the design, the characteristics of the sample, and the predictor and outcome variables.

- **Hypotheses are needed for studies that will use statistical tests to compare findings among groups suggested by the following phrases:**
  - "greater than" or "less than".
  - "more likely than".
  - "associated with", "compared with", "related to", "similar to", or "correlated with".
  - "causes" or "leads to".

- **Hypotheses are not needed for descriptive studies, such as prevalence.**

  Sample size planning still should be done prior to starting these studies.

Characteristics of a Good Research Hypothesis

- **Simple versus Complex**

A simple hypothesis contains only one predictor and one outcome variable. The predictor or outcome variable can be a combined variable, such as complications.

- **Specific versus Vague**

A specific hypothesis leaves no ambiguity about the groups, the predictor and outcome variables, or the statistical test that will be used.

- **In-Advance versus After-the-Fact**

A single research hypothesis stated at the start of the study design process keeps the research focused on the primary objective, creates a stronger basis for interpreting the results, and prevents over-interpretation of the importance of the findings.

.

Example

Overall Goal: Use a pragmatic randomized trial to determine the comparative effectiveness of patient-driven text-messaging (TM) versus health-coaching (HC) versus enhanced usual care (EC) for African-American adults with uncontrolled DM and multiple chronic conditions (MCC) in medically under-served areas (MUA) with an emphasis on identifying and quantifying important interactions between key baseline characteristics and treatment arm.

Aim 1: Quantify the effectiveness of TM, HC, and EC in improving the primary outcome measures: diabetes self-care activities related to general diet, exercise and medication adherence.

Hypotheses: We predict that TM and HC will provide significant and equivalent improvements in primary and secondary outcomes overall versus EC.

Null and Alternative Hypotheses

- **The Null Hypothesis**

The process begins with restating the research hypothesis in a statement that proposes no difference between the groups or no association. This hypothesis is what the researcher wants to reject.

- **The Alternative Hypothesis**

The alternative hypothesis should not specify the expected direction, but state that there is a difference or an association. This is called a two-sided alternative hypothesis. The alternative hypothesis is accepted, if, and when, the null hypothesis is rejected.

- **What about a one-sided alternative?**

Most grant and manuscript reviewers expect two-sided alternative hypotheses and are very critical of one-sided alternative hypotheses.

If a one-sided alternative is specified and the result lies in the opposite direction, then the null is automatically accepted.

Example

H_0: The average improvement in the number of days that African-American adults with DM and MCC randomized to the TM group ate healthy meals is the same as the average improvement in the number of days that those randomized to the EC group did.

H_1: The average improvement in the number of days that African-American adults with DM and MCC randomized to the TM group ate healthy meals is not the same as the average improvement in the number of days participants randomized to the EC group did.

Underlying Statistical Principles

- **The Study Sample versus the Reference Population**

Statistics is the science that allows researchers to develop probability models.

These probability models allow inferences to be made about a specific random phenomenon or characteristic of a reference population based on relatively limited observations from a sample.

Effect Size

- **The ability to detect an association between a predictor and an outcome variable or a difference between groups depends on the size of the association or the difference.**

  Large associations or differences are easier to detect than small ones.

  Laboratory studies detect large associations or differences with very small samples because the environment and extraneous variables are eliminated or controlled.

  Clinical studies typically are designed to detect small to medium associations or differences and require much larger samples.

- **The effect size is the size of the association between the predictor and the outcome or the difference between the groups as measured by the outcome variable.**

  Choosing an appropriate effect size is notoriously difficult!

Determining an Appropriate Effect Size

- **Locate published data from previous studies using similar reference populations.**
  - Means and standard deviations or standard errors.
  - Proportions or odds ratios.
  - Correlation coefficients or regression coefficients.
- **Choose the smallest effect size that would be clinically important.**
- **Conduct a small pilot study or chart review.**
- **Alternatively, start with the number of participants available and estimate the effect size that can be detected.**

Example

Using mean changes and standard deviations for the three primary outcomes as reported in Rosenberg et al., Thom et al., or Arora et al., we expect improvements in both active intervention arms of 1.6 ± 2.2 day, 1.1 ± 2.4 day, and 0.9 ± 2.5 day for the three primary outcomes, respectively.

Type I and Type II Errors

	True State of Nature	
	Null hypothesis is true.	Alternative hypothesis is true.
Decision in Favor of		
Null Hypothesis	Correct	Type II Error (β)
Alternative Hypothesis	Type I Error (α)	Correct

- Type I Error (false positive) occurs when the investigator rejects the null, but the null is true.
- Type II Error (false negative) occurs when the investigator fails to reject the null, but the alternative is true.
- Increasing sample size reduces the likelihood of making these errors, making the sample more likely to represent the reference population.

α, β, and Power

- **α is the maximum probability of making a Type I error (rejecting the null hypothesis when the null is true) and is called the level of statistical significance.**

  By convention, α is usually no larger than 0.05, but can range from 0.01 to 0.10.

- **β is the probability of making a Type II error (failing to reject the null hypothesis when the alternative is true).**

  β must be smaller than 0.50. Typical values of β are 0.20, 0.15, 0.10, 0.05, and 0.01.

- **1-β is called power and it is the probability of correctly rejecting the null hypothesis in the sample, if the actual effect in the reference population is equal to or greater than the specified effect size.**

  If power is 0.80, then the investigator is willing to take a 20% chance of missing a real association or difference of the specified effect size, if it exists.

α, β, and Power

- **Ideally, investigators want α and β as close to zero as possible.**

  However, reducing them both requires increasing sample size.

- **Sample size planning aims to choose a sufficient number of participants who will complete the study**

  To keep α and β at acceptably low levels

  Without requiring unnecessary expense and difficulty.

- **Choice of the "best" levels of α and β depend on the research question and the risks associated with making each type of error.**

  Testing the efficacy of a potentially harmful medication requires avoiding a Type I error or a false-positive result.

  Reassuring the public that living near an ore smelter is safe requires avoiding a Type II error or a false-negative result.

What about the p-value?

- **The null hypothesis is assumed to be true so that it can be rejected as false with a statistical test.**

- **When data are analyzed, a statistical test is used to determine the p-value.**

The p-value is the probability of getting an effect as big or bigger than the one found in the study, if the null hypothesis were actually true.

Importantly, if the null hypothesis is true, and there is really no difference between the groups or no association, then the only way the study could have produced a difference is by chance alone (i.e., just plain “bad” luck).

If that chance is “small” or less than α, then the null hypothesis can be rejected in favor of the alternative, and the inference is that an association or a difference has been detected.

What about non-significant results?

- **Getting a p-value greater than α (but often less than 0.10) does not mean that there is no association or difference between the groups.**

- **Non-significant results may indicate that the result observed in the sample (the observed effect) is small compared with what could have occurred (based on the proposed or expected effect) with a given sample size at α of 0.05.**

- **A reasonable alternative is reporting a 95% confidence interval and state that the results, although suggestive of an association or a difference, did not achieve statistical significance (p = 0.xx).**

This alternative approach preserves the integrity of the original hypothesis design but acknowledges that statistical significance is not a yes-or-no situation, and the reported results can be used for planning future studies.

Types of Statistical Tests & Sample Size

- Each type of statistical test is based on a set of mathematical assumptions.

- The choice of statistical test primarily depends on the predictor and the outcome variables and the study design.

- Each type of statistical test requires a different formula to determine sample size.

Simple Statistical Tests for Use in Estimating Sample Size

Predictor Variable	Outcome Variable	
	Dichotomous	Continuous
Dichotomous	Chi-square test for a contingency table	t-test
Continuous	t-test	Regression or Correlation Coefficient

Example

Power calculations are based on two-sided t-tests with type-I error rates of 0.05.

A standardized effect size is the effect divided by the standard deviation.

Standardized effect sizes will range from small (standardized difference = 0.36) to medium (= 0.73).

Variability of the Outcome Variable

- **The larger the variability or spread of the outcome variable among subjects, the more difficult it will be to demonstrate a difference between groups.**

The difference between the means of the groups (i.e., the effect size) may be moderate, but a large between-subject variability can cause the values in the groups to overlap. Distinguishing between the means with a statistical test will require a larger sample size.

- **Measurement error due to less precise measurements increases the overall variability of the outcome variable.**

- **As variability increases, sample size requirements increase.**

Steps for Estimating Sample Size and Power

- State the primary research hypothesis with one predictor and one outcome variable, being specific about the groups and the reference population.

- State the null hypothesis and two-sided alternative hypothesis

- Identify the types of predictor and outcome variable

- Choose a reasonable effect size and variability, if necessary. Sample size estimates for t-tests require prior estimates of variability, but those for proportions and odds ratios do not.

Other Considerations

- **Each participant in the sample must be available for analysis.**

  Participants without outcomes (dropouts) cannot be analyzed and do not count in the sample size computations.

  The investigator should estimate the anticipated proportion of dropouts and increase the number of participants enrolled to offset the number who will potentially drop out.

- **Ordinal Variables**

  Ordinal variables can be treated as continuous variables, if there are 6 or more categories and/or averaging of the values make sense.

  Otherwise, the best strategy may be to change the the scale to a dichotomous one.

- **Survival Analysis**

  A reasonable strategy can be simply estimating the proportion of each group expected to ever have the outcome.

Other Considerations

- **Matching**

  Matching can be used to substantially reduce the total number of participants required.

  Participants can serve as their own controls.

  Alternatively, participants can be matched on a few important baseline characteristics.

Sample Size Techniques for Descriptive Studies

- Concepts of the null and alternative hypothesis and power do not apply.
- There are no predictor and outcome variables.
- There are no comparisons between groups.
- Descriptive studies frequently report confidence intervals.
- Sample size for descriptive studies can be determined with techniques that are counterparts to those used for hypothesis testing of dichotomous and continuous variables.

What to Do When the Sample Size Is Fixed

- For secondary data analyses and limited numbers of available participants, the sample size is fixed.

- The approach is to work "backwards" from the fixed number of participants to determine the size of the effect that can be detected with the available sample size.

- Generally, the study should have a power of 80% to detect a "reasonable" effect size.

Strategies to Minimize Sample Size & Maximize Power

- Use continuous variables
- Use paired measurements
- Use more precise variables
- Use a more common outcome variable

Common Errors to Avoid

- Estimating the sample size late during the design of the study.
- Misinterpreting proportions expressed as percentages as continuous variables.
- Not taking missing subjects (dropouts) and missing data into account.
- Not considering that the potential group sizes may not the same.
- Not addressing paired data appropriately.

Assistance with Sample Size Determinations

BERD Unit

https://berd.uthsc.edu/

- On the Biostatistics, Epidemiology, and Research Design (BERD) page, simply use the "REQUEST ASSISTANCE" button

BERD Clinic for Sample Size Determinations

- Be prepared to discuss the study design, the main research hypothesis, the predictor and outcome variables, and if known, the effect size and the variability, if necessary.

- BERD faculty and staff will make the sample size computations in consultation with the investigators. They may offer other helpful suggestions, such as study design.

Questions?