Colorado School of PUBLIC HEALTH

UF Health Outcomes & Policy College of Medicine

Tutorial: Selecting a Test

Authors:
Sarah M. Kreidler, Anna E. Barón, and Deborah H. Glueck

December 10, 2012

This project is funded by NIDCR 1 R01 DE020832-01A1 to the University of Florida (Keith E. Muller, PI; Deborah Glueck, University of Colorado site PI). The associated GLIMMPSE software is released under the GNU Public License version 2. Previous funding for the GLIMMPSE software was received from an American Recovery and Re-investment Act supplement (3K07CA088811-06S) for NCI grant K07CA088811.

Contents

1. Introduction

When designing a research study, scientists must determine the number of participants they need to recruit. Researchers perform power and sample size calculations to select a sample size. The sample size needs to be large enough to answer the scientific question of interest, but not so large that participants are unnecessarily exposed to risk. To perform a power calculation, the scientist must specify several input values. The values include choices for means and variance, the Type I error rate, and the statistical test.

In this tutorial, we provide guidelines for selecting an appropriate test to calculate power and sample size for study designs with normally distributed outcomes. In Section 2, we review basic concepts of hypothesis testing and power and sample size analysis. In Section 3, we review criteria for the comparison of tests. In Section 4, we provide guidelines for selecting an appropriate test. In Section 5, we describe the user interface for selecting a test when conducting a power or sample size analysis with the GLIMMPSE software.

2. Overview of Hypothesis Testing, Power, and Sample Size

A hypothesis is a claim or statement about one or more population parameters, such as a mean or a proportion. A hypothesis test uses data to provide evidence for a decision about a hypothesis. We begin by stating a null hypothesis, H_0, a claim about a population parameter (for example, the mean). We initially assume the null hypothesis to be true. The investigator usually hopes that the evidence in the data will disprove the null hypothesis.

Because of sampling, there is inherent uncertainty in the conclusion drawn from a hypothesis test. The scientist will either make a correct decision or make an error. Since perfect certainty cannot be achieved, the scientists attempts to minimize the chances of making an incorrect decision. The probability of rejecting a true null hypothesis is denoted by α, and the probability of failing to reject a false null hypothesis is denoted by β. The possible outcomes of a hypothesis test are summarized in Table 1 along with the conditional probabilities of their occurrence.

	State of Nature	
Hypothesis Testing Decision	H_0 True	H_1 True
Fail to reject H_0	True negative A correct decision Pr(True Negative) $= 1 - \alpha$	False negative A Type II error Pr(False Negative) $= \beta$
Reject H_0	False positive A Type I error Pr(False positive) $= \alpha$	True positive A correct decision Pr(True positive) $= 1 - \beta$

Table 1: Outcomes of Hypothesis Testing

A scientist makes a Type I error when she rejects H_0, when in fact, H_0 is true. Similarly, a scientist makes a Type II error when she fails to reject H_0, when in fact, H_1 is true and H_0 is false. To keep the chances of making a correct decision high, α, the probability of a Type I error, is usually chosen to be 0.05 or less. Similarly, the sample size is chosen so that $1 - \beta$, the conditional power of the test, is high, usually 0.8 or more. When H_0 is true, the power of the test is equal to the level of significance. For a fixed sample size, the probability of making a Type II error is inversely related to the probability of making a Type I error. Thus, in order to achieve a desirable power for a fixed level of significance, the sample size will generally need to increase.

2.1. Types of Hypotheses

To begin our discussion of hypotheses, we first need to define the concept of the *independent sampling unit* (Muller and Stewart 2006). Independent sampling units may be people, rats, or groups of participants such as schools

or neighborhoods. Observations on one independent sampling unit are statistically independent of observations from another independent sampling unit. However, observations within the same independent sampling unit may be correlated. For example, test results for two students in the same school are often correlated, because the students share the same learning environment. Test results from different schools, however, may be assumed to be independent. Similarly, Ki-67 levels for two oral lesions excised from the same mouth are typically correlated, because they are taken from the same individual, but Ki-67 levels from two different patients are independent.

Between-participant hypotheses concern treatments applied to different independent sampling units, or characteristics that distinguish different independent sampling units. In randomized controlled trials, or laboratory experiments, scientists apply treatments to different independent sampling units. For example, a scientist may conduct a study in which participants with pre-malignant oral lesions are randomized to receive either a smoking cessation program or the standard of care. In an observational study, independent sampling units may be classified into groups by characteristics such as gender or smoking habits. For example, we might want to compare cotinine levels between men and women in a smoking cessation program.

Alternatively, experimenters may want to compare repeated measurements within independent sampling units. Hypotheses that compare outcomes within an independent sampling unit are called *within-participant* hypotheses. For example, scientists might examine repeated cotinine levels across time in participants in a smoking cessation program. An observational study might compare gene expression levels in normal, pre-malignant, and oral cancer tissues from the same study participant.

We note that in older experimental design books, these classifications are called *between-subject* hypotheses, and *within-subject* hypotheses. We adopt the word *participant* instead of *subject* to incorporate the spirit of the Helsinki Proclamation, that human participants in research are independent actors who make a contribution to society by autonomously agreeing to be in an experiment.

We can further classify hypotheses as *simple* or *complex*. We define a *simple between* hypothesis as a hypothesis that compares at most two different groups of independent sampling units. A *complex between* hypothesis compares three or more different groups of independent sampling units. Similarly, a *simple within* hypothesis is a hypothesis that compares at most two different measurements within independent sampling units. A *complex within* hypothesis compares three or more measurements within independent sampling units. We list examples of common designs with simple and complex hypotheses below.

2.2. Designs with *simple between* and *simple within* hypotheses

For designs with both *simple between* and *simple within* hypotheses, the hypothesis of interest can compare at most two different groups of independent sampling units, and at most two repeated measures. The following designs and hypotheses have this form.

- Participants with oral cancer are randomized to receive either chemotherapy or radiation. The researchers wish to test if the Ki-67 levels differ between the two groups at one month post randomization. This design is typically analyzed with a two-sample t-test.
- Ki-67 levels are measured in participants with oral cancer before and after treatment with the antioxidant resveratrol. The researchers wish to test if the Ki-67 levels differ between the pre- and post-treatment observations. This design is typically analyzed with a paired t-test.
- Participants with oral cancer are randomized to receive either chemotherapy or radiation. Researchers measure the participants' Ki-67 levels at one month and six months post randomization. The researchers wish to test if the change in Ki-67 levels from one to six months differs between the chemotherapy and radiation groups. That is, the researchers wish to test the interaction between time and treatment.

2.3. Designs with a *simple between* hypothesis and a *complex within* hypothesis

Designs with a *simple between* hypothesis but a *complex within* hypothesis can compare at most two groups of independent sampling units, but may compare three or more repeated measurements within an independent sampling unit. The following designs and hypotheses have this form.

- Participants with oral cancer are randomized to receive either chemotherapy or radiation. Researchers measure the participants' Ki-67 levels at one month, three months, six months, and twelve months post randomization. The researchers wish to test if the trend in Ki-67 levels over time differs between the chemotherapy and radiation groups. In this case, the researchers are comparing the two between participant groups, and analyzing the linear and quadratic trends over time within each participant. The hypothesis tests for a time by treatment interaction.
- Participants with oral cancer are randomized to receive either chemotherapy or radiation. Researchers measure the participants' Ki-67 levels at one month, three months, six months, and twelve months post randomization. The researchers wish to test if the Ki-67 levels differ among any of the follow-up times, when averaged across the two treatment groups. This hypothesis tests for the main effect of time on Ki-67 levels.

2.4. Designs with a *complex between* hypothesis and a *simple within* hypothesis

Designs with a *complex between* hypothesis but a *simple within* hypothesis compare three or more groups of independent sampling units, but may only compare at most two repeated measurements within an independent sampling unit. The following designs and hypotheses have this form.

- Participants with oral cancer are randomized to receive either low, medium, or high dose chemotherapy. Researchers measure the participants' Ki-67 levels at one month post randomization. The researchers wish to test if Ki-67 levels differ among the three chemotherapy doses. This design is commonly analyzed with a one-way analysis of variance.
- Participants with oral cancer are randomized to receive either 10mg, 50mg, 100mg, or 200mg of resveratrol. Researchers measure the participants' Ki-67 levels at one month post randomization. The researchers wish to test if there is a dose-response relationship in the effect of resveratrol on Ki-67 levels. This hypothesis tests for linear, quadratic, and cubic trends of the response as a function of dose.

2.5. Designs with a *complex between* and a *complex within* hypotheses

Designs with *complex between* and *complex within* hypotheses can compare three or more groups of independent sampling units, and three or more repeated measurements within an independent sampling unit. The following designs and hypotheses have this form.

- Participants with oral cancer are randomized to receive either 10mg, 50mg, 100mg, or 200mg of resveratrol. Researchers measure the participants' Ki-67 levels at one month, three months, six months, and twelve months post randomization. Researchers wish to test the time trend by treatment interaction. That is, they wish to examine if the dose of resveratrol modifies the change in response curve over time.
- Participants with oral cancer are randomized to receive either 10mg, 50mg, 100mg, or 200mg of resveratrol. Researchers measure the participants' Ki-67 levels at one month, three months, six months, and twelve months post randomization. Researchers wish to test if the the Ki-67 values differ among the dose groups at any of the follow-up times.

3. Criteria for Evaluating Tests

An optimal statistical test will minimize both the Type I and Type II error rates. Ideally, we select a test with high power and a fixed Type I error rate for the hypothesis of interest. A test is said to be uniformly most powerful if it has the greatest power among all possible tests of the same size. Uniformly most powerful tests exist only for certain special cases.

4. Guidelines for Choosing a Test for Power and Sample Size Analysis

We provide guidelines for choosing tests for power analysis for the general linear multivariate model with normally distributed outcomes (Muller and Stewart 2006) and for the general linear mixed model (Laird and Ware 1982). While there are many hypothesis tests available for the mixed model, the Wald test with Kenward-Roger degrees of freedom (Kenward and Roger 1997) controls the Type I error rate, even in small samples. For the general linear multivariate model, there are at least seven tests, listed below:

- Hotelling-Lawley Trace
- Pillai-Bartlett Trace
- Wilks' Lambda
- Univariate approach to repeated measures (uncorrected)
- Univariate approach to repeated measures with Box correction
- Univariate approach to repeated measures with Geisser-Greenhouse correction
- Univariate approach to repeated measures with Huynh-Feldt correction

The Hotelling-Lawley Trace, the Pillai-Bartlett Trace, and the Wilks' Lambda test are collectively referred to as the multivariate approach to repeated measures, or MULTIREP, tests. The remaining tests are different forms of the univariate approach to repeated measures, or UNIREP tests (Muller and Stewart 2006). None of the tests is uniformly most powerful for all study designs and hypotheses. Therefore, the optimal test choice depends on the specific study design and hypothesis of interest.

Often, the Hotelling-Lawley trace test for the general linear multivariate model coincides with the Wald test for the general linear mixed model with Kenward-Roger degrees of freedom (Kenward and Roger 1997). The two tests coincide when there are no missing observations, and each independent sampling unit has the same number of observations. The coincidence allows scientists to perform power analysis for the mixed model using well-studied methods in the general linear multivariate model.

For complicated designs and hypotheses, the MULTIREP and UNIREP may yield different hypothesis test results, and require different sample sizes. A complete discussion of test choice for complex designs and hypotheses is presented in Muller and Stewart (2006), Section 3.7. The discussion in Muller and Stewart (2006) is aimed at professional statisticians. Here, we present a more informal discussion, and a heuristic for test choice aimed at scientists. We give simple rules below, which usually work, but may not include every special case.

In Table 2, we recommend appropriate tests for power and sample size analysis based on the type of hypothesis. You may notice that in many cases, some or all of the tests coincide. When tests coincide, they give exactly the same p-values and decisions for data analysis, and provide exactly the same power and sample size when designing a study. When the choice of test makes no difference, we suggest selecting only the Hotelling-Lawley trace for power and sample size calculations. In general, no matter what hypothesis test you are considering, if you plan to

use the mixed model for your data analysis, and are going to use the Wald test with Kenward-Roger degrees of freedom, you should perform your power and sample size analysis using the Hotelling-Lawley test for the general linear multivariate model.

Within Participant Hypothesis	**Between Participant Hypothesis**	**Recommended Test**
simple	simple	All tests coincide, use Hotelling-Lawley Trace
simple	complex	All tests coincide, use Hotelling-Lawley Trace
complex	simple	MULTIREP tests coincide, use Hotelling-Lawley Trace
complex	complex	Hotelling-Lawley Trace

Table 2: Heuristics for Selection

It may seem odd that we provide power for six MULTIREP and UNIREP in the GLIMMPSE software when we recommend the Hotelling-Lawley Trace test in most cases. However, we wanted to build power and sample size software for a wide variety of applications. Tests other than the Hotelling-Lawley trace may provide better power for some experimental designs and hypotheses. In addition, researchers in some fields may be accustomed to using other tests.

5. How to Select a Test in GLIMMPSE

The GLIMMPSE software calculates power and sample size for the general linear multivariate model and certain classes of mixed models. GLIMMPSE can be accessed through a standard web browser at `http://glimmpse.samplesizeshop.com`. Statistical tests are specified on the *Statistical Test* screen under the *Options* menu. The user may calculate power for one or more tests by clicking the appropriate boxes.

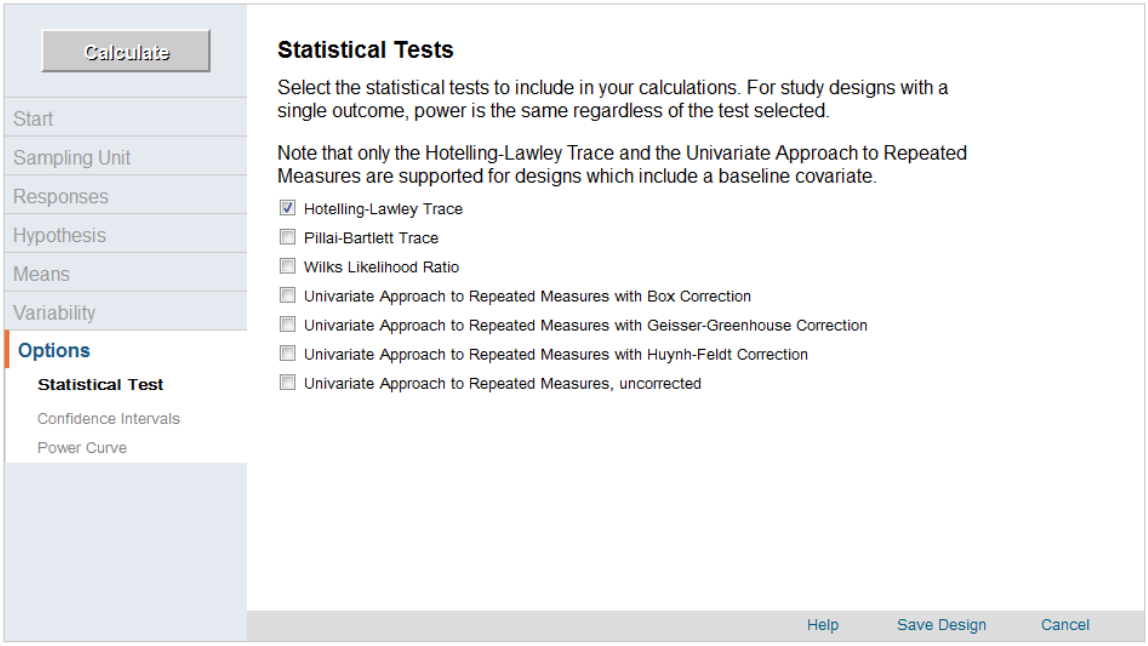

Figure 1: The *Statistical Test* screen

References

Kenward MG, Roger JH (1997). "Small sample inference for fixed effects from restricted maximum likelihood." *Biometrics*, **53**(3), 983–997.

Laird NM, Ware JH (1982). "Random-effects models for longitudinal data." *Biometrics*, **38**(4), 963–974. ISSN 0006-341X. PMID: 7168798, URL http://www.ncbi.nlm.nih.gov/pubmed/7168798.

Muller KE, Stewart PW (2006). *Linear Model Theory: Univariate, Multivariate, and Mixed Models*. John Wiley and Sons, Hoboken, New Jersey.