Tutorial 0: Uncertainty in Power and Sample Size Estimation

Anna E. Barón, Keith E. Muller, Sarah M. Kreidler, and Deborah H. Glueck

Acknowledgements:

The project was supported in large part by the National Institute of Dental and Craniofacial Research under award NIDCR 1 R01 DE020832-01A1. The content is solely the responsibility of the authors, and does not necessarily represent the official views of the National Cancer Institute, the National Institute of Dental and Craniofacial Research, nor the National Institutes of Health.

Tutorial 0: Uncertainty in Power and Sample Size Estimation

Preface

Power is the probability that a study will reject the null hypothesis. The estimated probability is a function of sample size, variability, level of significance, and the difference between the null and alternative hypotheses. Similarly, the sample size required to ensure a pre-specified power for a hypothesis test depends on variability, level of significance, and the null vs. alternative difference.

Power analysis consists of determining the achievable power for the specified null vs. alternative hypotheses. For example, varying the inputs, the standard deviation, or the detectable mean difference will show the power tradeoffs.

Likewise, sample size estimation consists of determining the required sample size for the null vs. alternative hypotheses. For example, varying the inputs, standard deviation and detectable mean difference, will show the sample size tradeoffs.

Before embarking on examples of power and sample size estimation for specific designs in the other tutorials, it is important to review the impact of uncertainty in the inputs on estimates of power and sample size.

Two types of uncertainty

Information about expected mean differences and variability can be obtained from a variety of sources. Often one's own previous research, either pilot or demonstration studies, will serve as a good starting point for power and sample size estimation. In other cases, the published literature will provide the needed information. In the absence of such information, mean values and standard deviations will simply be the best educated guesses, as in the case of a brand new exploration in a field of research.

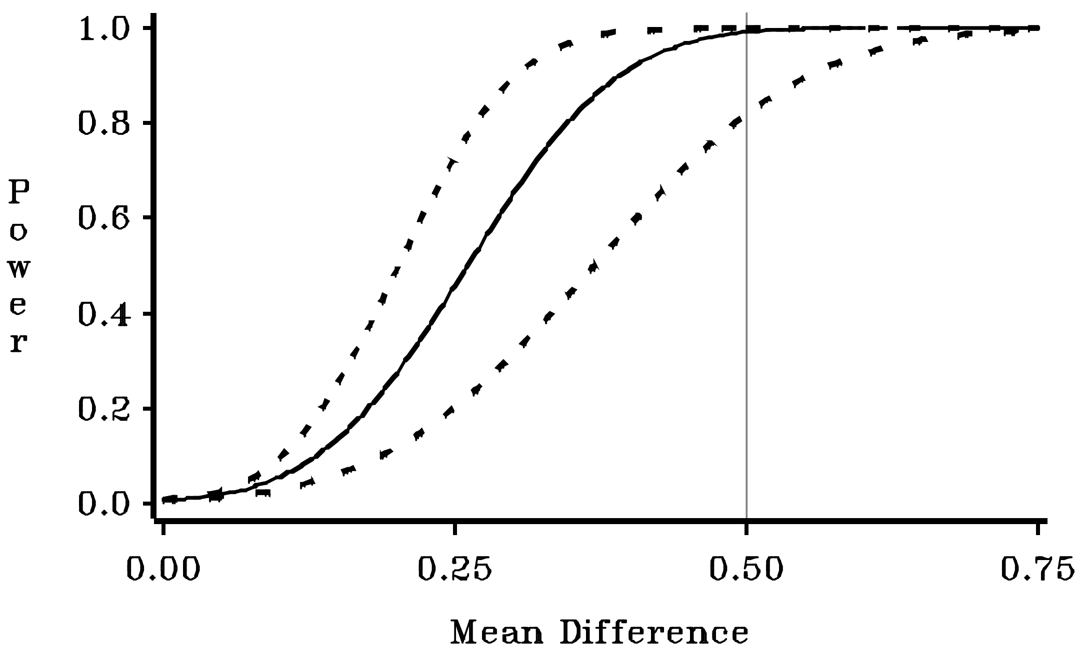
There are two ways in which uncertainty in power or sample size can be conveyed:

1) <u>Sampling uncertainty</u>

   Even when using information from large studies in the literature, it can't be assumed that means and standard deviations are known quantities. They are *estimates* and as such lead to uncertainty in estimated power.

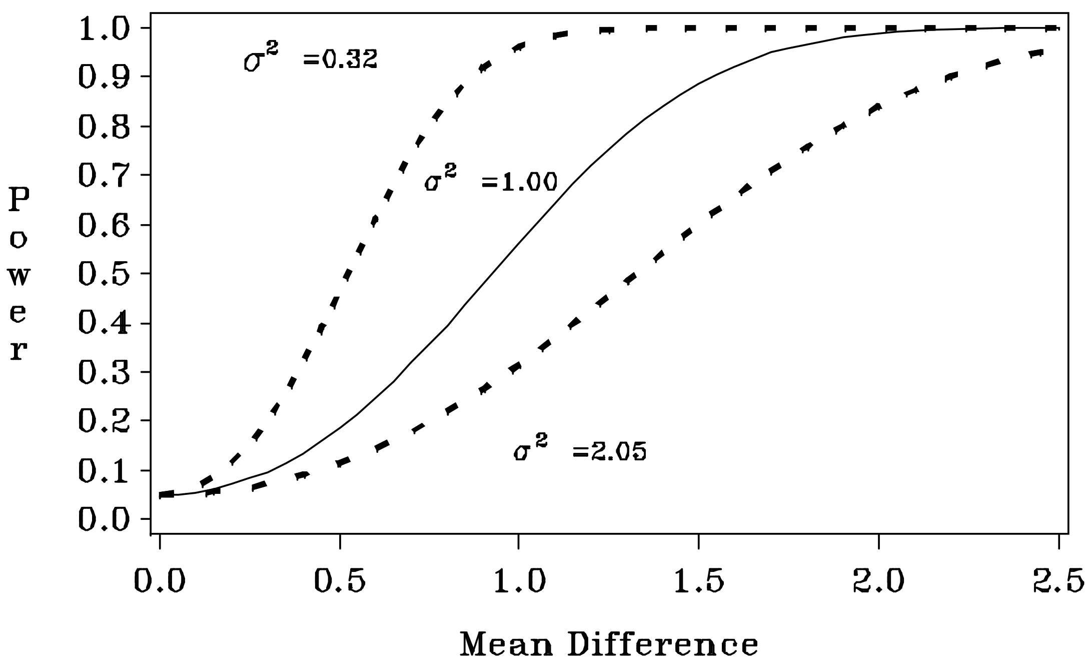
Thus, estimated power for a fixed sample size and estimated sample size for a fixed power are *random variables* with sampling variability, similar to a mean or a proportion (Taylor and Muller, 1995).

To convey uncertainty in power estimates due to sampling variability, study design information is needed from the literature or one's own work. Specifically, the sample size and the particular design (e.g. one group, two-group, multi-group) that gave rise to the standard deviation and/or detectable mean difference estimates can be used to obtain simultaneous confidence bands for the estimated power. Below is an example of a power curve with 95% two-sided simultaneous confidence bands for a two-sample t-test with a sample size of 15 in each group. To obtain these bands, the estimated standard deviation, 0.26, was used from a study with a *total sample size of 24* and *two independent groups of observations*. The tradeoff shown in the graph is power vs. mean difference with α = 0.01 (two-sided), where the mean difference was fixed over a reasonable range (see "*Sensitivity analysis …*" below). By focusing on a scientifically meaningful difference between means, e.g. 0.5, we can be 95% confident that the power to detect that difference as significant at the 1% level of significance with a total sample size of 30 is between about 80% and 100%.



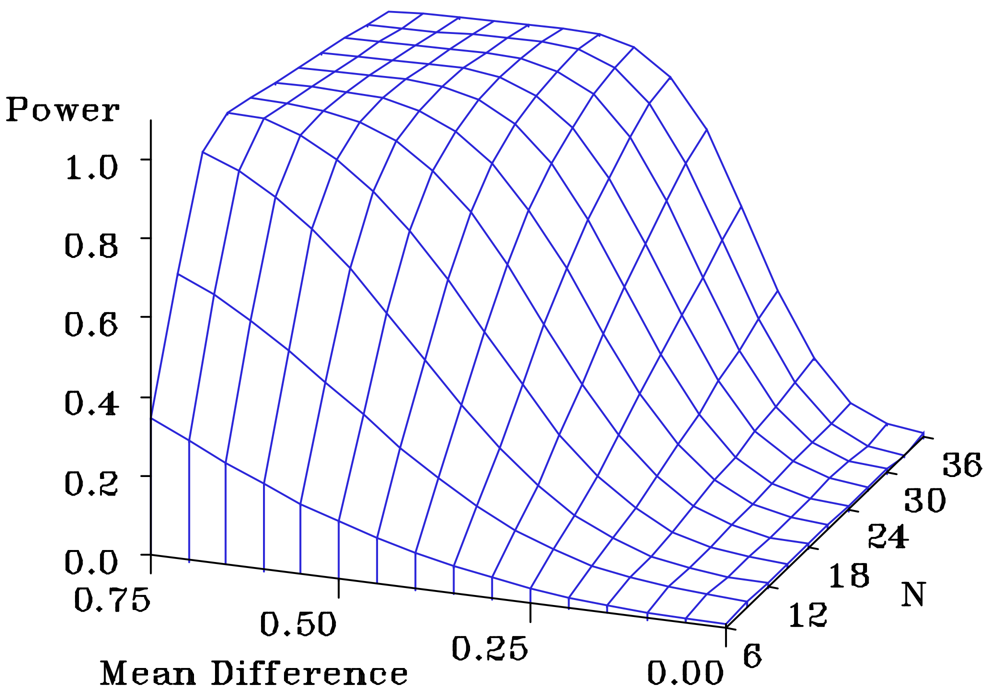
2) <u>Sensitivity analysis using speculation or an educated guess</u>

To accomplish this analysis we allow input values, e.g. the standard deviation or the detectable mean difference, or both, to vary over a *fixed*, reasonable range. The fixed, reasonable range can be, for example, from 0.5 times to twice the standard deviation and/or the detectable mean difference. Thus, sampling variability is not incorporated into the calculations, but by considering a range of fixed input values for the standard deviation and/or mean difference we indicate that these are unknown, yet plausible values. The figure below illustrates how power varies with fixed values of the mean difference when considering three different possible fixed values of the variance for a two-sample t-test with sample sizes of 10 in each group and α = 0.05 (two-sided) (Muller and Benignus, 1992). From the plot, it can clearly be seen that increased variability leads to a notable loss of power for most values of the mean difference considered.



Avoiding the Slippery Slope of Power

From the previous power curves we can see that, below certain levels, power is sensitive to the magnitude of the mean difference. Variance also affects power's threshold of sensitivity. We call this the "slippery slope of power." In practice, to assure that sample size is large enough to achieve adequate power, even when the inputs have been underestimated (in the case of standard deviation) or overestimated (in the case of mean difference), it is recommended, where feasible, that sample size be chosen to correspond to very high levels of power, e.g. 90% or even 95%. The plot below illustrates the power—sample size—mean difference tradeoffs for a two-sample t-test with standard deviation of 0.26 and α=0.01 (two-sided). It can be seen that a total sample size (N) of 36 (or 18 per group) assures 90% power or greater even if the mean difference is as low as 0.4.



Content A: Uncertainty in the inputs for power analysis

A.1 Inputs for Power analysis

We now illustrate uncertainty in a power analysis using the one-sample t-test (see also *Tutorial 1: Power and Sample Size for the One-sample t-test*). A power analysis for a single mean consists of determining the achievable power for a specified difference between the mean under the stated H_0 and under the stated H_1, sample size, standard deviation, and α-level. By varying these four quantities a set of power curves can be obtained that show the tradeoffs.

Information about mean values and variability can be obtained from the published literature. Sample size can be varied over a feasible range of values, and various values of α can be selected to illustrate the sensitivity of the results to conservative vs. liberal choices for the Type I error rate.

In Tutorial 1 we use data on a cell proliferation marker used in chemoprevention research on head and neck cancer, Ki-67. Data are available on the mean and standard deviation of Ki-67 in Seoane et al. (2010). Below is an excerpt of Table 1 from that publication. The data are based on 63 incident cases of oral cancer.

Table 1. Characteristics of cases and interval from first symptoms to diagnosis.

Variables	Mean	SD	Median	Mínimum	Maximun
Age	61.8	12.2	61.5	42	89
Follow-up, mo	38.58	24.02	38.23	1.3	86.59
Ki-67, %	41.6	24.8	39	6	93

The mean Ki-67 of the sample is 41.6% and the standard deviation is 24.8%. For simplicity, these values will be taken to be 42% and 25%, respectively. The sample size used was 63 and there was only one group of patients used to estimate the standard deviation. For the one-sample Ki-67 example, a 10% deviation from 42% is considered to be biologically important.

When performing a power analysis using these data we have two choices for incorporating uncertainty: 1) sampling variability, and 2) sensitivity analysis. For sampling uncertainty, we can obtain a confidence interval for the power curve using the sample size of 63 and the fact that the standard deviation was estimated on one group of patients. To accomplish a sensitivity analysis, we consider, instead, a fixed range of values for the standard deviation and/or the mean difference under the null vs.

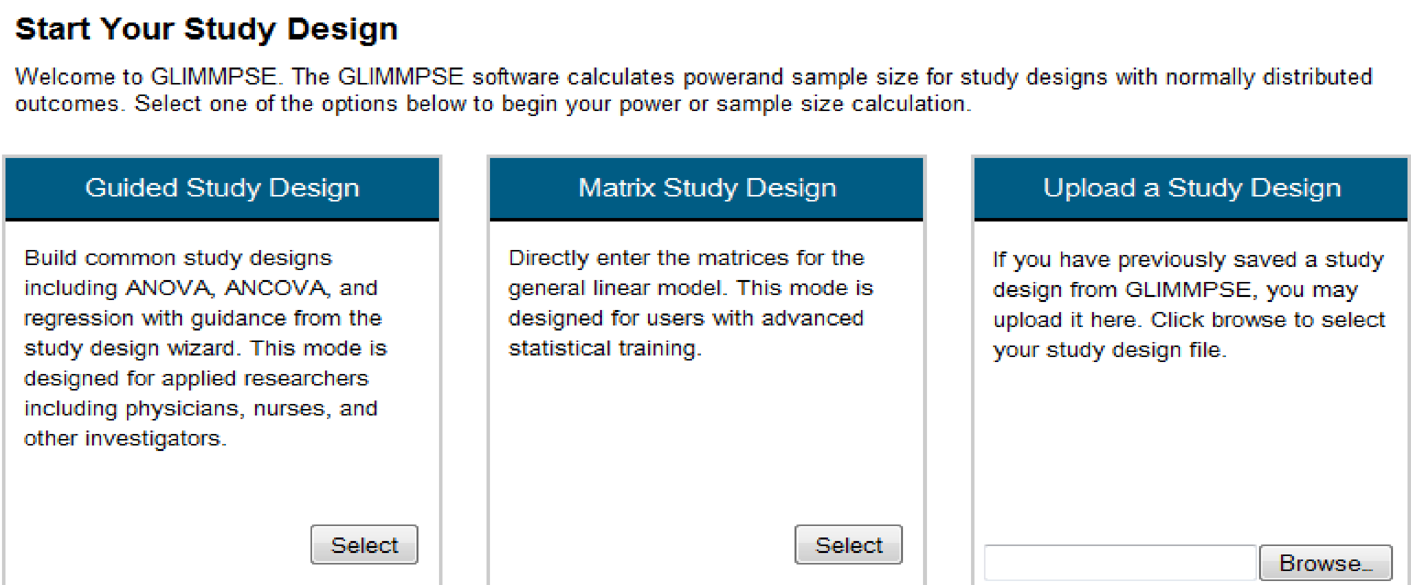
alternative hypotheses. For example, we can allow the standard deviation and/or mean difference to range from 0.5 times the speculated values up to twice the speculated values.

Content B: How to use the software to incorporate uncertainty
How to perform the analysis in GLIMMPSE

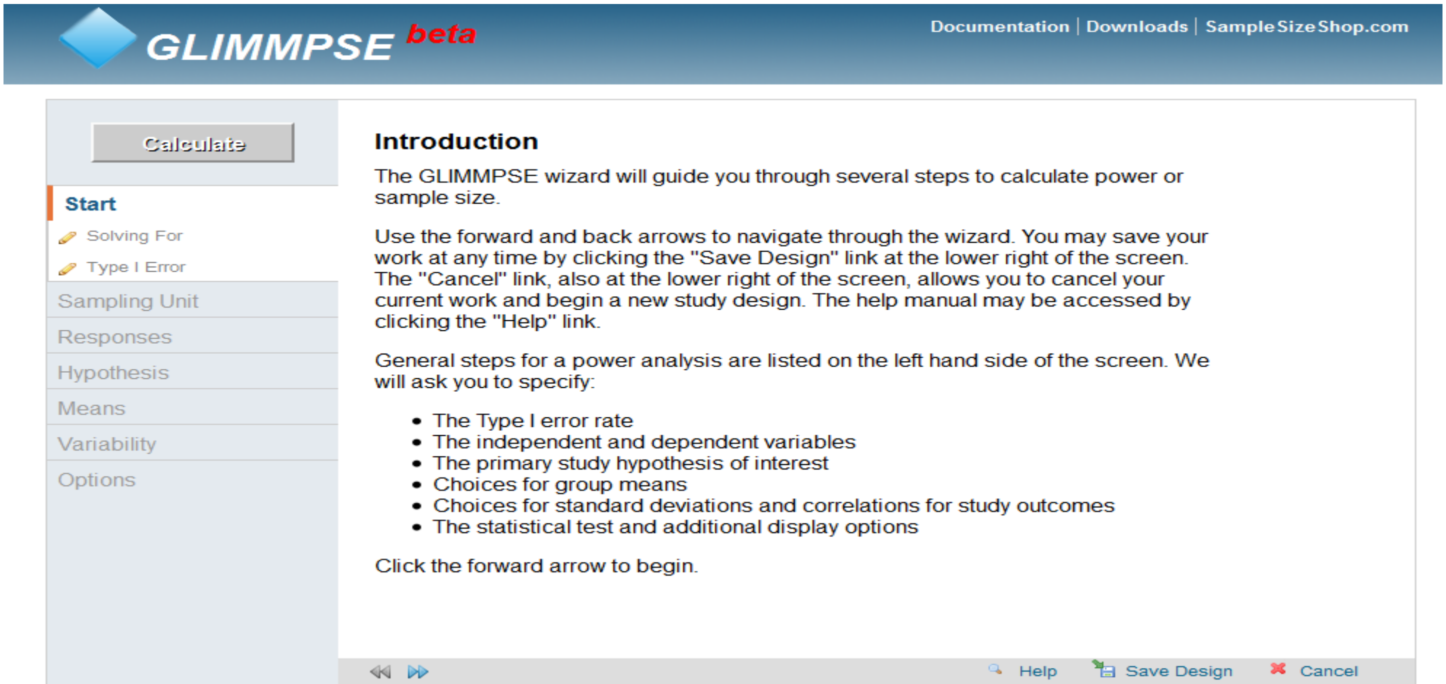
To start GLIMMPSE 2.0 beta, type http://samplesizeshop.com/calculate-power-and-sample-size-now/ in your browser window or visit www.SampleSizeShop.com and click on the GLIMMPSE tab, then on GLIMMPSE 2.0.0 Beta is Here! Google Chrome is the suggested browser for this application.

B.1 Uncertainty in Power analysis using Guided Study Design mode

Guided Study Design mode is suggested for most users.



Selecting *Guided Study Design* takes you to the *Introduction* page:

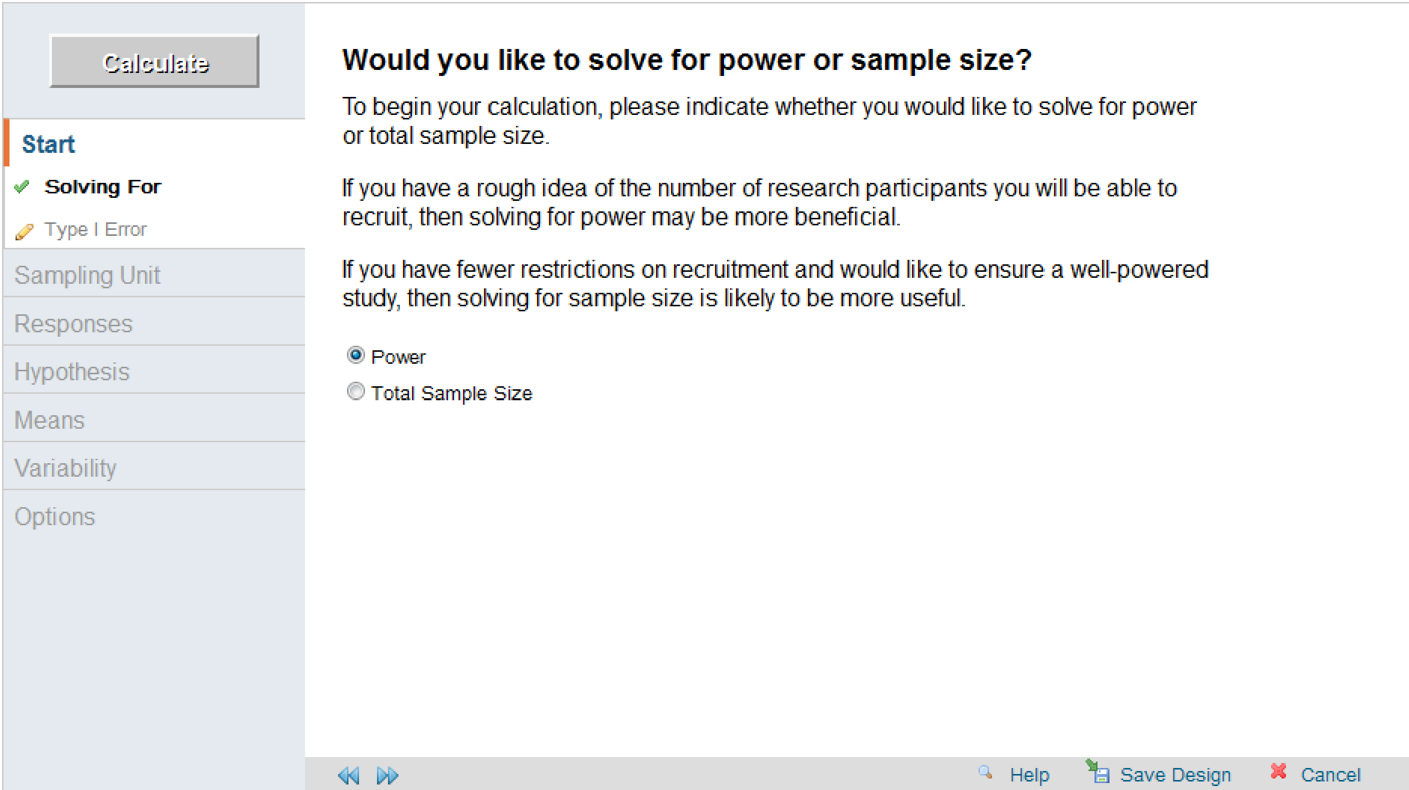


Throughout your use of the GLIMMPSE website, use the blue arrows at the bottom of the screen to move forward and backwards. The pencil symbols beside the screen names on the left indicate required information that has not yet been entered. In the above picture, the pencils appear next to the *Solving For* and the *Type I Error* screen names.

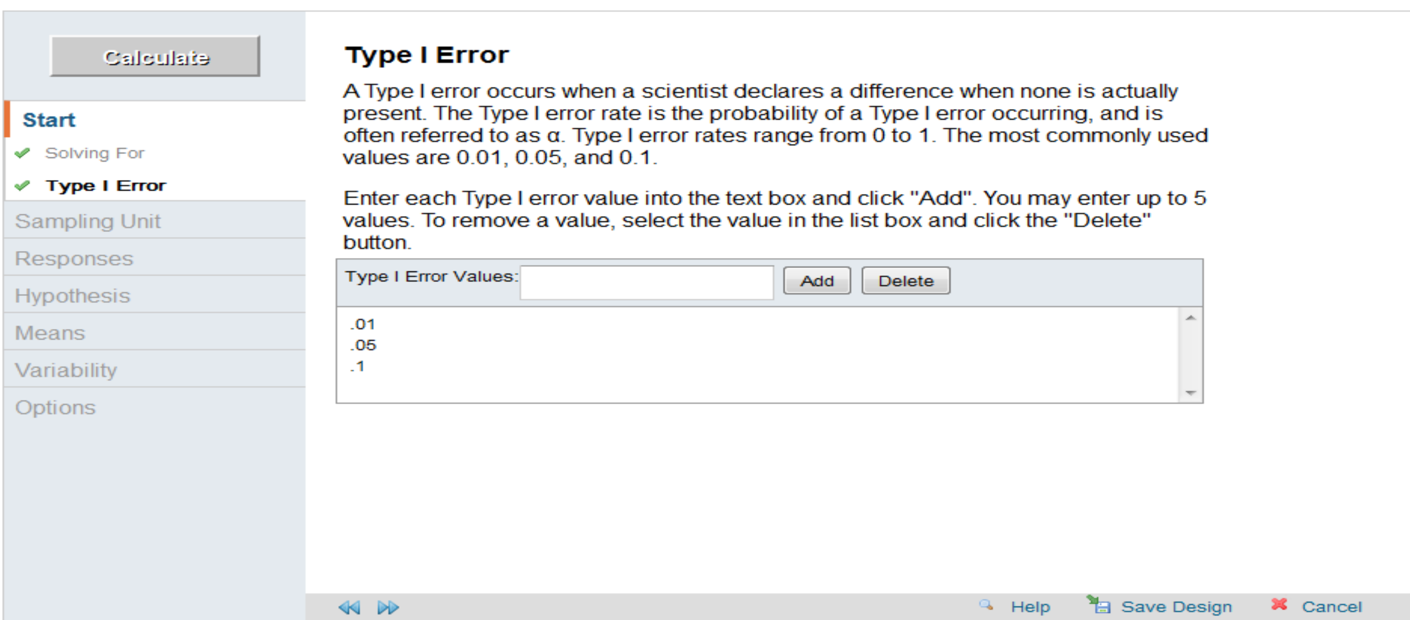
Once the required information is entered, the pencils become green check marks. A red circle with a slash through it indicates that a previous screen needs to be filled out before the screen with the red circle can be accessed. Search for a pencil beside a screen name in the previous tabs to find the screen with missing entries.

Once you have read the information on the *Introduction* screen, click the forward arrow to move to the next screen.

The *Solving For* screen allows you to select Power or Sample Size for your study design. For the purposes of this tutorial, click *Power*. Click the forward arrow to move to the next screen.

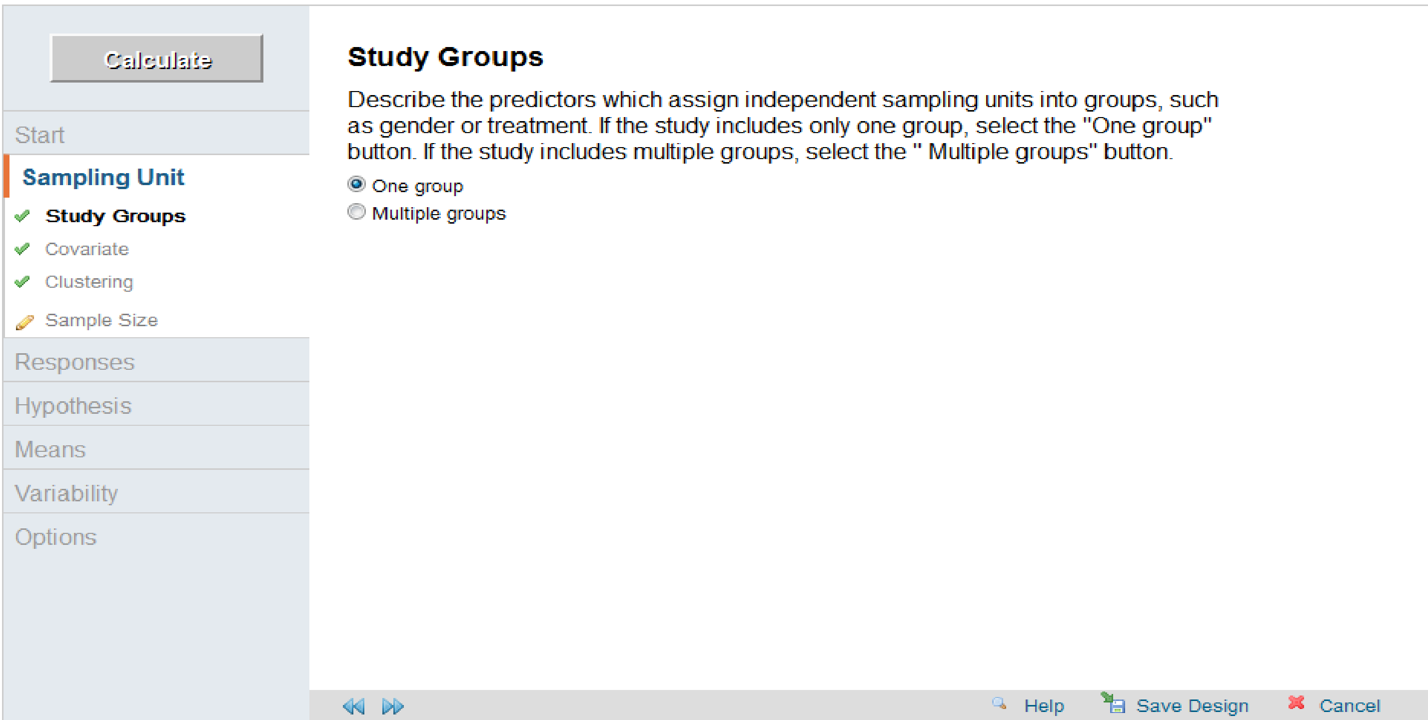


The *Type I Error* screen allows you to specify the fixed levels of significance for the hypothesis to be tested. Once you have entered your values, click the forward arrow to move to the next screen.



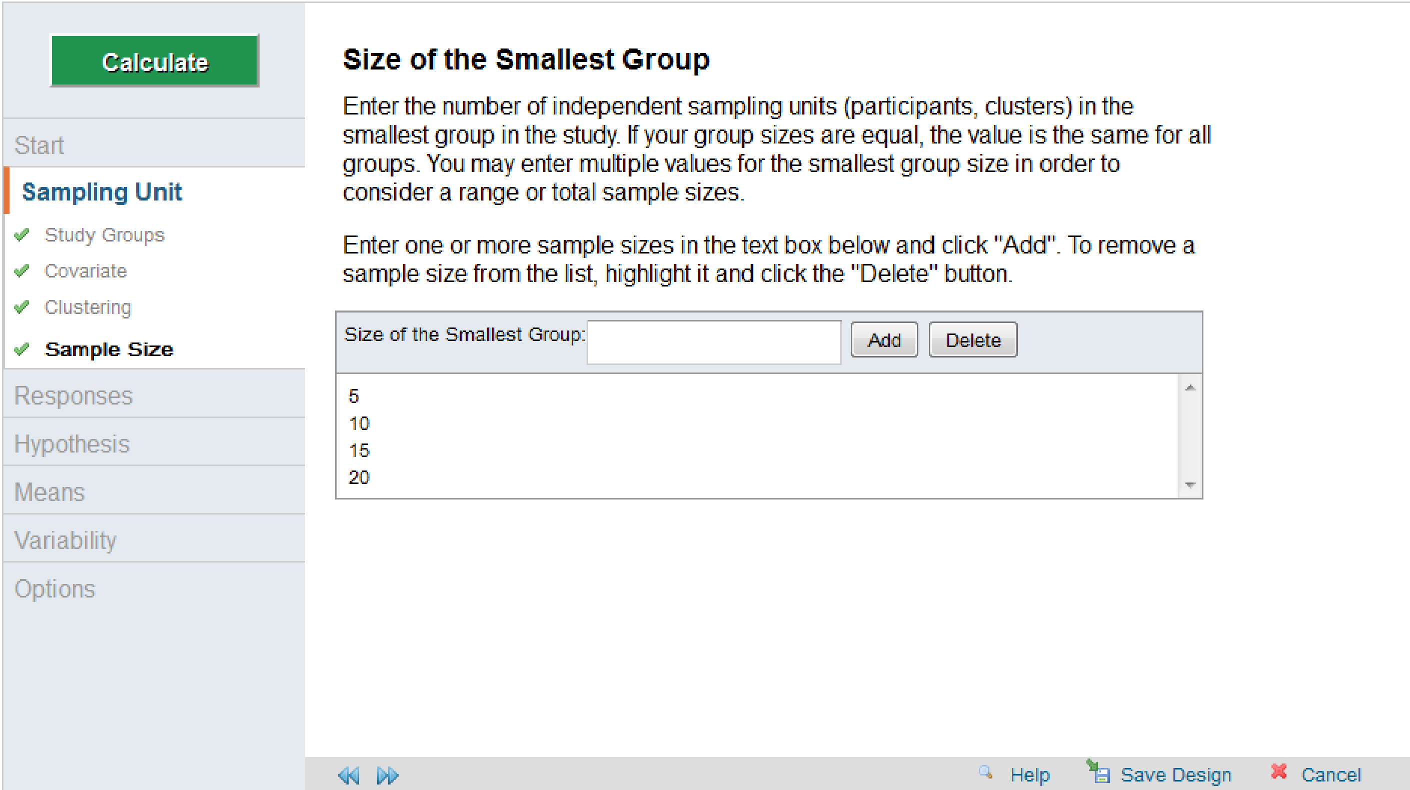
Read the information on the *Sampling Unit: Introduction* screen and click next to move to the next screen.

For a one-sample test, there is one fixed predictor with one level. On the *Study Groups* screen, select *One group* and click the forward arrow to move to the next screen.



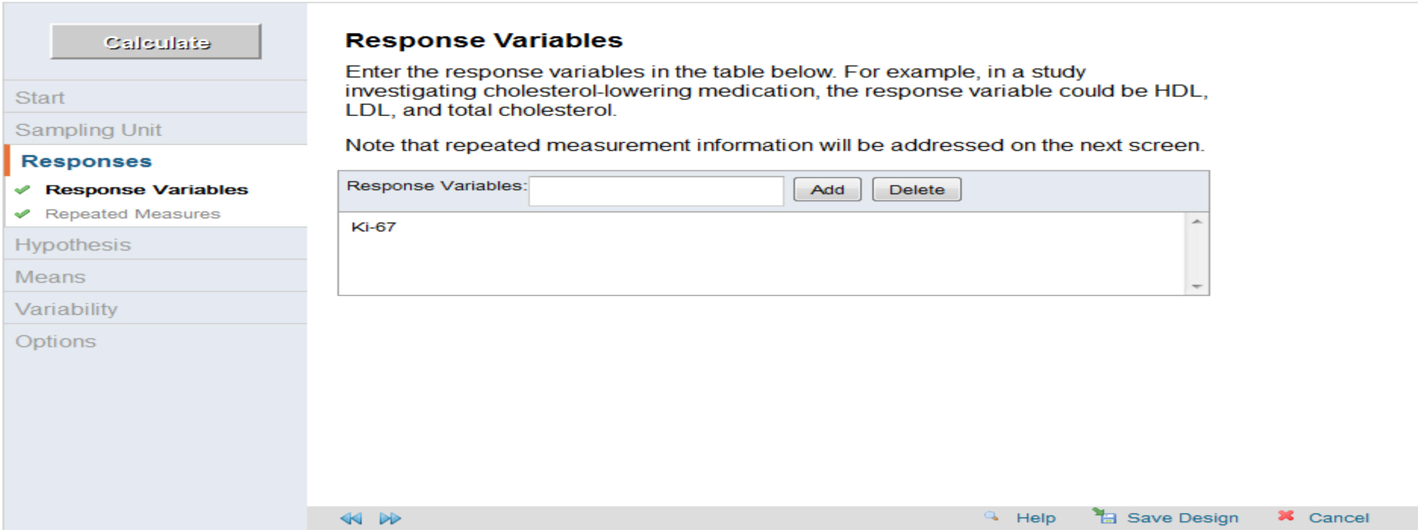
Since no covariate and no clustering will be used, skip the *Covariate* and the *Clustering* screens.

Use the *Sample Size* screen to specify the size of the smallest group for your sample size(s). Although only one sample is being used for this example, entering multiple values for the smallest group size allows you to consider a range of total sample sizes.



Read the *Responses: Introduction* screen, and click the forward arrow when you have finished.

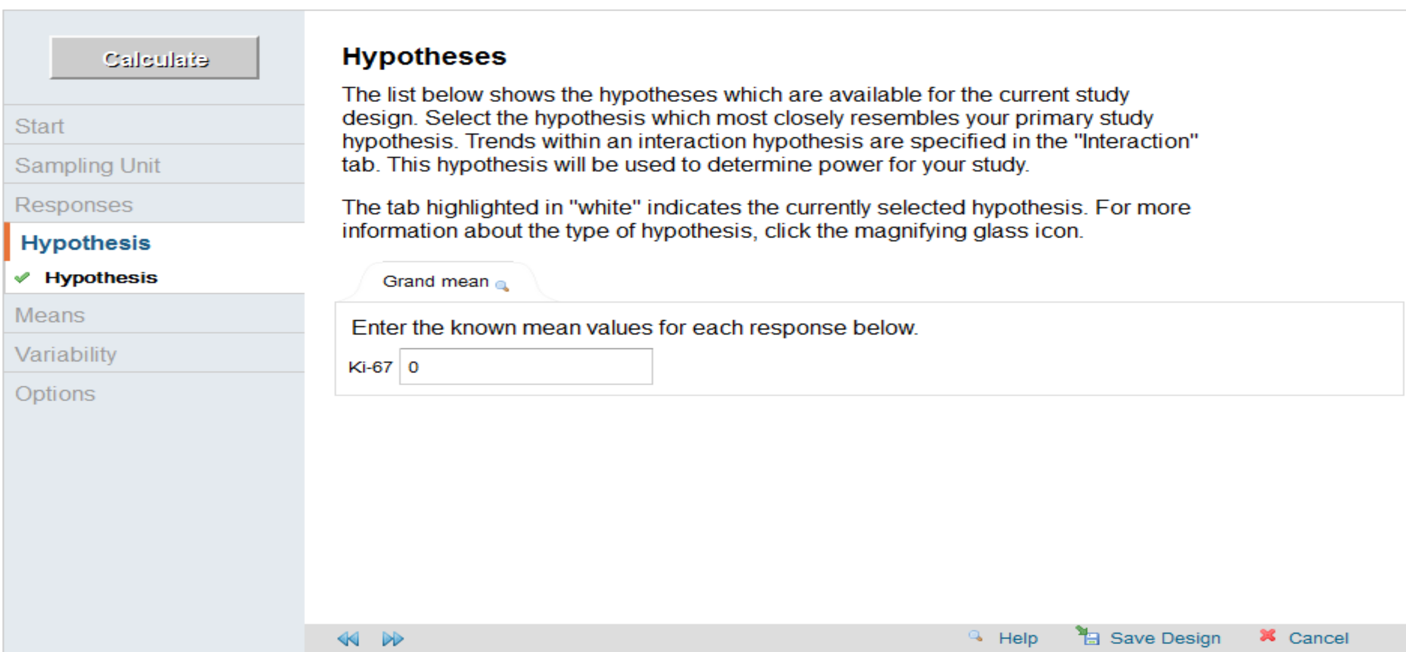
The *Response Variable* screen allows you to enter the response variable(s) of interest. In this example, the response variable is Ki-67. Click the forward arrow when you have finished entering your value(s).



The *Repeated Measures* screen should be skipped, as there are no repeated measures for this example.

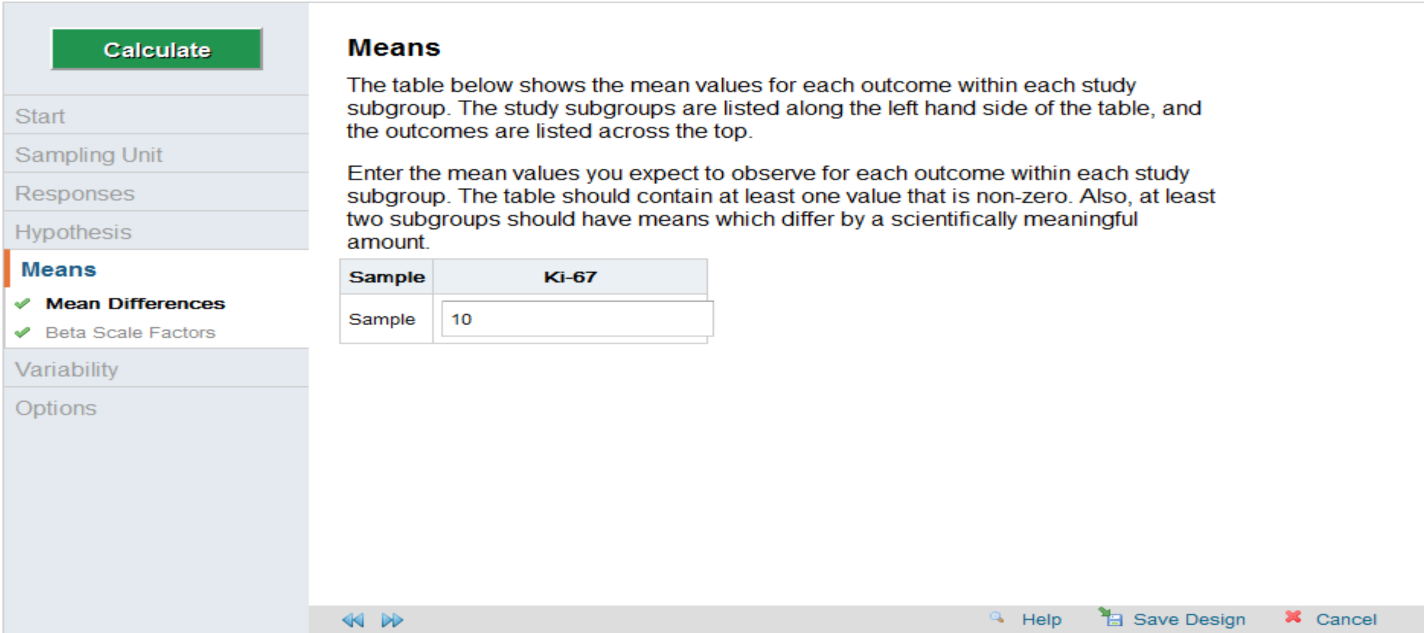
Read the *Hypothesis: Introduction* screen, then click the forward arrow to move to the *Hypothesis* screen.

The *Hypotheses* screen allows you to enter the known mean values for your primary hypothesis. For this example, enter 0.

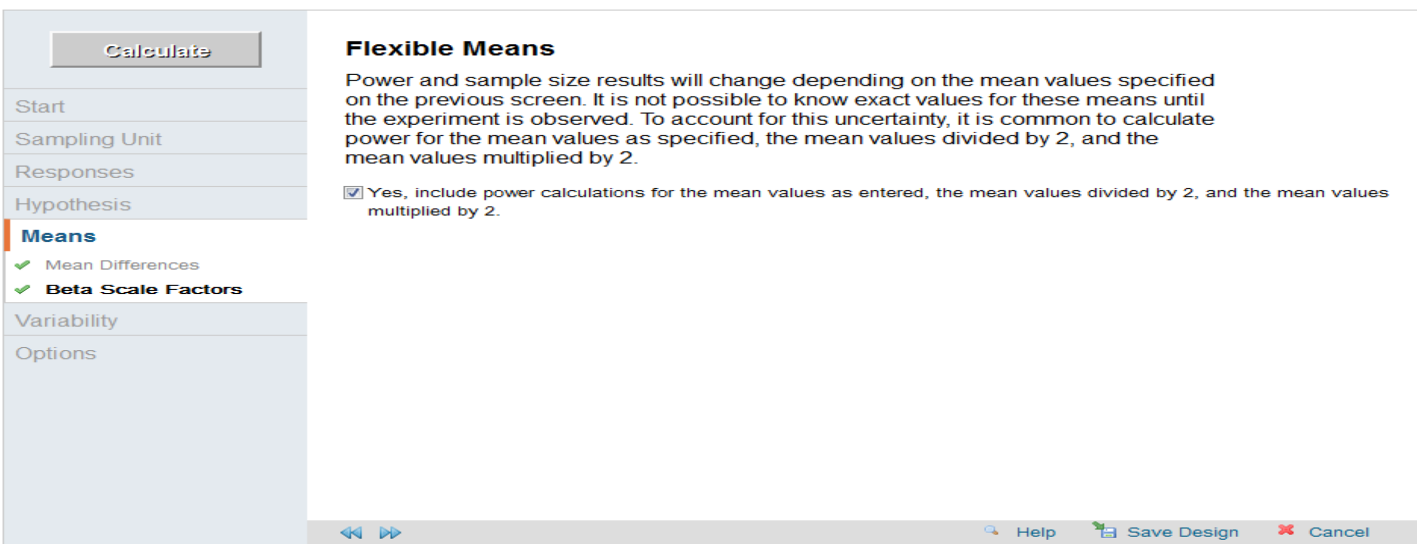


Read the information on the *Means: Introduction* screen, then click the forward arrow.

The *Mean Differences* screen allows you to specify the difference between the null and alternative hypothesis means. For the Ki-67example, this is 52% vs. 42% or a mean difference of 10%. Enter the difference of 10 and click the forward arrow to continue.

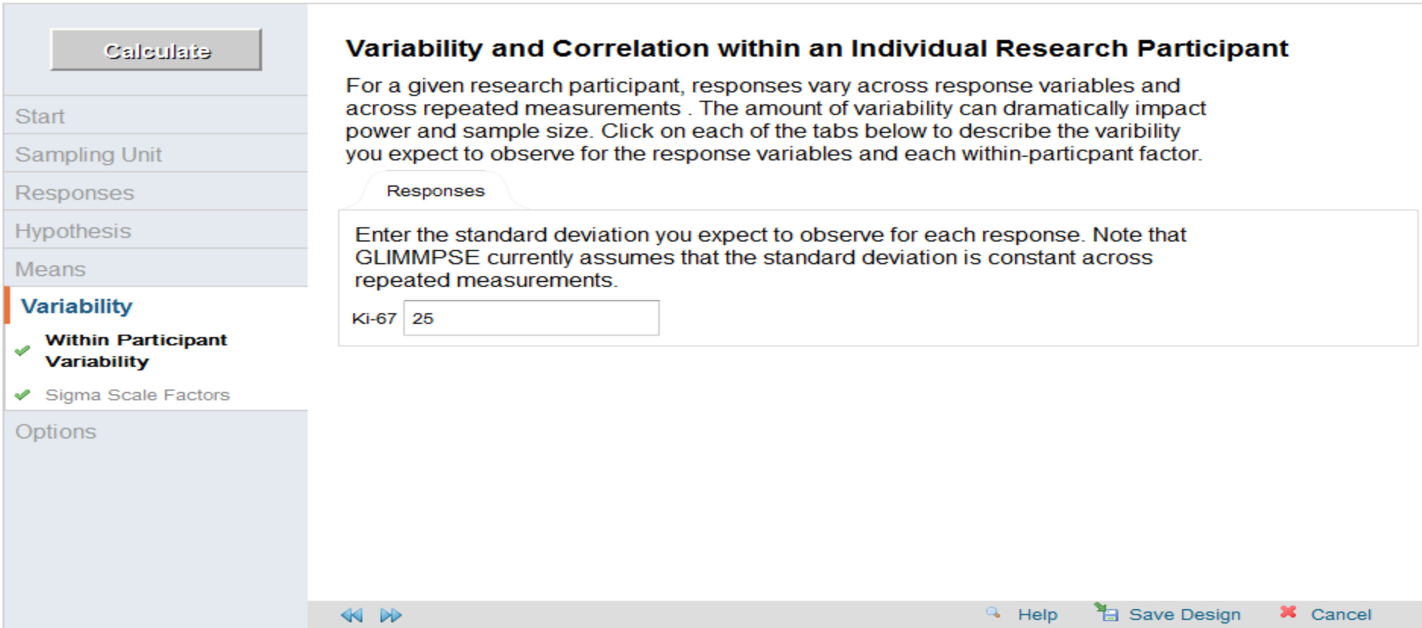


The *Beta Scale Factors* screen allows you to see how power varies with the assumed difference, so that you can allow it to vary over a reasonable range. GLIMMPSE allows this to be from 0.5x to 2x the stated difference, e.g. from 5% to 10% to 20%. Click *Yes*, then click on the forward arrow to continue.

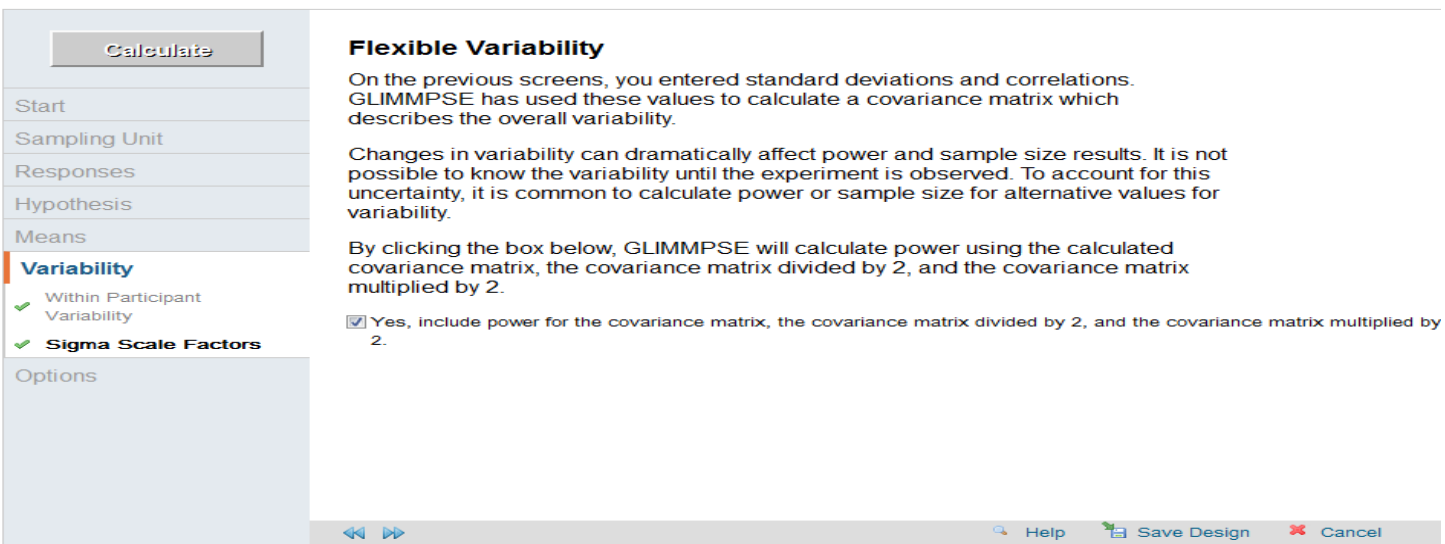


Read the information on the *Variability: Introduction* screen, then click the forward arrow to continue.

The *Within Participant Variability* screen allows you to specify the expected variability in terms of standard deviation of the outcome variable. Using data from Seoane et al, 2010, the standard deviation of Ki-67 in a group of early and late stage head and neck cancer patients was estimated to be 25%. Enter 25 and click the forward arrow to continue.

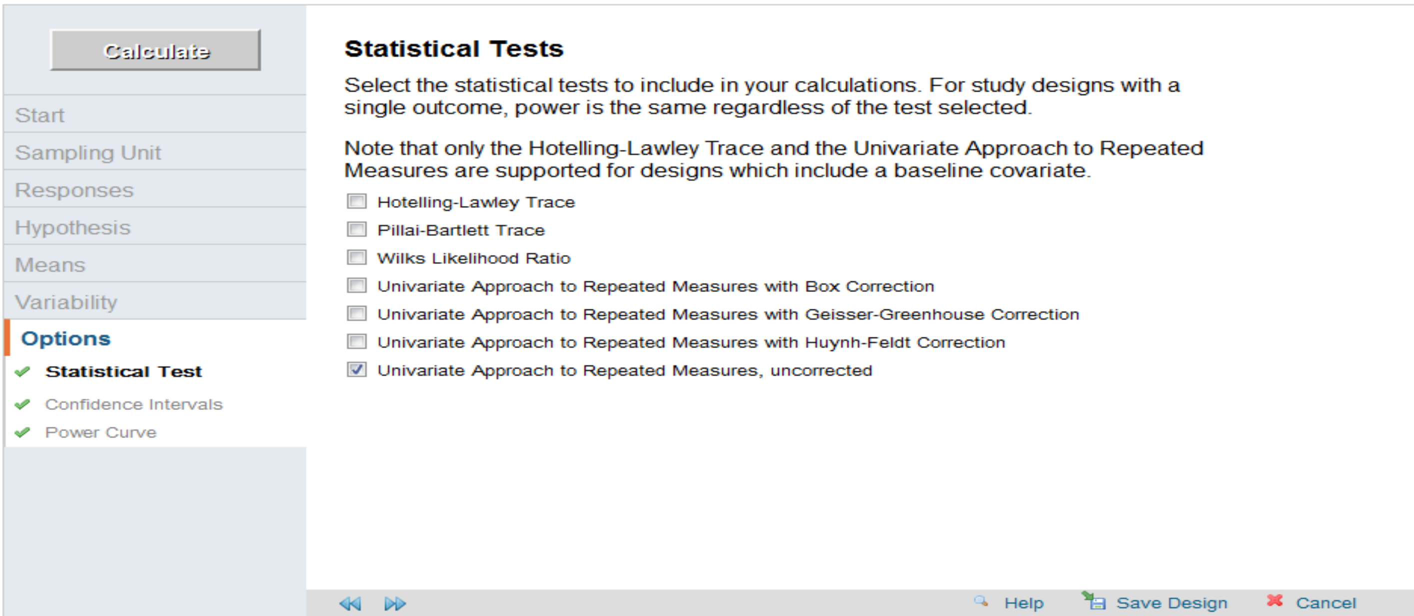


The true variability in Ki-67 is also uncertain. To see how power varies with the assumed standard deviation, the *Flexible Variability* screen allows the standard deviation to vary over a reasonable range. GLIMMPSE allows this to be from 0.5x to 2x the stated standard deviation, e.g. from 12.5% to 25% to 50%. Click *Yes*, then click the forward arrow to continue.



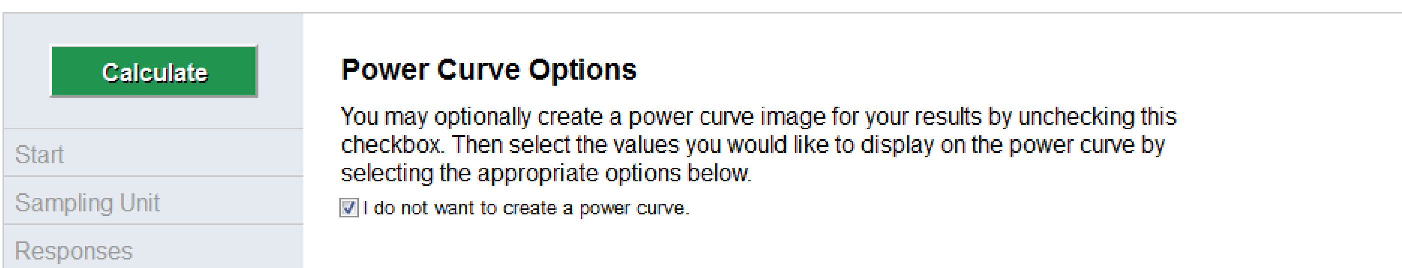
Read the information on the *Options* screen, then click the forward arrow to continue.

For the one-sample test of a single mean, the available tests in GLIMMPSE yield equivalent results. More information on choosing a test can be found in the Tutorial on Selecting a Test. Click on any one of the tests in the *Statistical Test* screen and then click the forward arrow to continue.

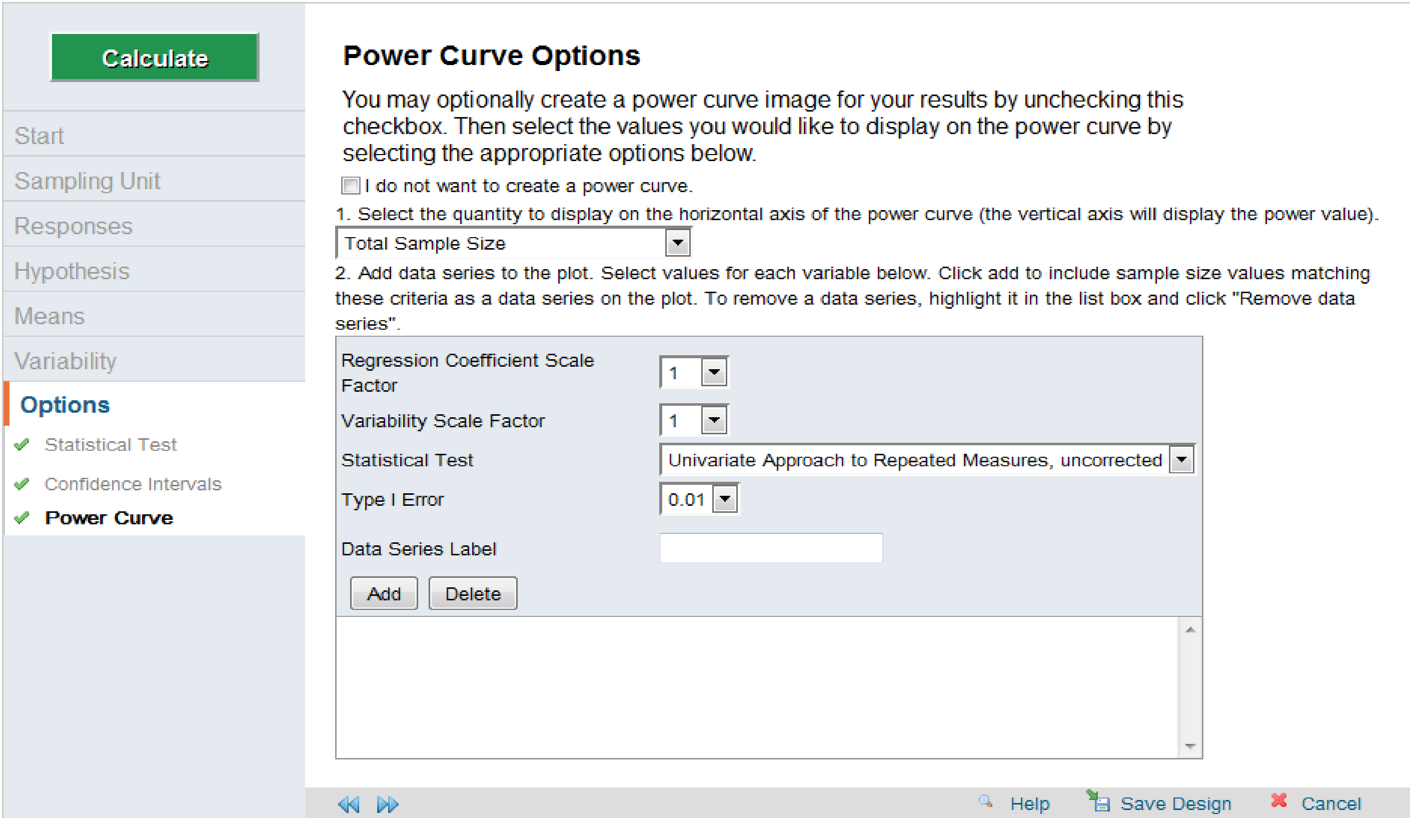


Leave the box checked in the *Confidence Interval Options* screen, and click forward to continue to the next screen.

Power analysis results are best displayed on a graph. To obtain a plot, first uncheck the box on the *Power Curve Options* screen.



From the pull down menu that appears once you have unchecked the box, select the variable to be used as the horizontal axis (e.g. *Total sample size* or *Variability Scale Factor*). GLIMMPSE will produce one power plot based on specific levels of the input variables that you specify on this page. If you specified more than one value for an input variable, choose the specific level you want GLIMMPSE to use to plot the power curve.



Content C: Interpret the uncertainty results

Interpretation of Uncertainty in Power Analysis

For the power analysis inputs in Guided Study Design Mode (Section B.1), GLIMMPSE produces a curve showing the relationship between achievable power, and fixed values of mean difference (beta scale), standard deviation (variability scale), total sample size and level of significance. A complete downloadable table of results in Excel.csv format is also produced. For the one-sample Ki-67 example, the plot shows achievable power over a range of differences in mean Ki-67 and standard deviations of Ki-67 with an α-level of 0.05 (two-sided) and a sample size of 30.

When confidence intervals are requested, additional inputs - sample size of the source study, and number of groups giving rise to the estimates of mean difference and/or standard deviation from the source study - must be provided, leading to a simultaneous confidence interval which is overlaid on the power curve. This plot shows the uncertainty in power due to sampling variability associated with available estimates of both the mean difference and the standard deviation used in the calculations.

Content D: References cited

List of matrices used in calculation:

$$\mathbf{Es(X)} = [1.0000]$$

$$\mathbf{B} = [10.0000]$$

$$\mathbf{C} = [1.0000]$$

$$\mathbf{U} = [1.0000]$$

$$\mathbf{\Theta}_0 = [0.0000]$$

$$\mathbf{\Sigma}_E = [625.0000]$$

For the one-sample t-test, GLIMMPSE works with the matrices listed above in making the computations. Since only a single mean is being tested the matrices are all of dimension 1 x 1. The $\mathbf{\Theta_0}$ matrix represents the mean under H_0 and the $\mathbf{\Sigma_e}$ matrix represents the between-subject variance. The remaining matrices Es(**X**), **C** and **U** are scalars with a value of 1.

References

Muller KE and Benignus VA (1992). Increasing Scientific Power with Statistical Power. *Neurotoxicology and Teratology*, 14: 211-219.

Seoane J, Pita-Fernandez S, Gomez I et al (2010). Proliferative activity and diagnostic delay in oral cancer. *Head and Neck*, 32:1377–1384.

Taylor DJ, Muller KE (1995). Computing Confidence Bounds for Power and Sample Size of the General Linear Univariate Model. *The American Statistician*, 49(1), 43-47.