Tutorial 5: Power and Sample Size for One-way Analysis of Variance (ANOVA) with Equal Variances Across Groups

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. In order to understand our terms, we will review these key components of hypothesis testing before embarking on a numeric example of power and sample size estimation for a one-way ANOVA.

Overview of Power Analysis and Sample Size Estimation

A *hypothesis* is a claim or statement about one or more population parameters, e.g. a mean or a proportion. A *hypothesis test* is a statistical method of using data to quantify evidence in order to reach 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. H_0 is where we place the burden of proof for the data; it is usually what an investigator hopes to *disprove* if the evidence in the data is strong enough. In the one-way analysis of variance, the goal is to determine if the samples have the same average. The null hypothesis H_0 is then a simple statement about the means being equal.

As an example, let us say that a randomized three-arm trial in post-surgery/radiation head and neck cancer patients is planned, with a 1:1:1 ratio of resveratrol (an antioxidant found in the skins of grapes and other fruits): another supplement: placebo. The primary endpoint is change in Ki-67, a measure of cell proliferation (% of cells staining for Ki-67). Ki-67 is measured in tumors after surgery and before the administration of resveratrol, the other supplement, or placebo. Measurements are obtained from biopsies after 6 months on treatment. The goal is to compare mean change in Ki-67 across treatments, i.e. $\mu_{\Delta\text{Ki-67, resveratrol}}$ vs. $\mu_{\Delta\text{Ki-67, other supplement}}$ vs. $\mu_{\Delta\text{Ki-67, placebo}}$. The null hypothesis is that there is no difference in mean change in Ki-67, i.e. H_0: $\mu_{\Delta\text{Ki-67, resveratrol}} = \mu_{\Delta\text{Ki-67, placebo}} = \mu_{\Delta\text{Ki-67, placebo}}$. An opposing, or alternative hypothesis H_1 is then stated, contradicting H_0, e.g. H_1: at least one of the mean changes is different from the other two. If the mean changes from the three samples are sufficiently different a decision is made to *reject the null* in favor of the alternative hypothesis. If the sample mean changes are similar to each other, the null hypothesis

cannot be rejected, but neither can a claim be made that the hypothesis is unequivocally true.

Because of sampling there is inherent uncertainty in the conclusion drawn from a hypothesis test. Either a correct or an incorrect decision will be made, and the goal is to minimize the chances of making an incorrect one. The probability of rejecting a true null hypothesis, called Type I Error, is denoted by α.The probability of failing to reject a false null hypothesis, called Type II Error, is denoted by β. The possible outcomes of a hypothesis test are summarized in the table below 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 Prob(True negative) = $1-\alpha$	False negative A Type II Error Prob(False negative) = β
Reject H_0	False positive A Type I Error Prob(False positive) = α	True Positive A correct decision Prob(True positive) = $1-\beta$

A Type I error occurs when H_0 is incorrectly rejected in favor of the alternative (i.e. H_0 is true). A Type II error occurs when H_0 is not rejected but should have been (i.e. H_1 is true). To keep the chances of making a correct decision high, the probability of a Type I error (α, the level of significance of a hypothesis test) is kept low, usually 0.05 or less, and the power of the test ($1-\beta$, the probability of rejecting H_0 when H_1 is true) is kept 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.

Other factors that impact power and sample size determinations for a one-way ANOVA include variability as summarized by σ_1^2, σ_2^2 and σ_3^2 where σ_1, σ_2 and σ_3 are the standard deviations in the three populations; and the differences among the population means under the null hypothesis and under the alternative. In practice, σ_1^2, σ_2^2 and σ_3^2 can be difficult to specify. Previous work reported in the literature is a good source of information about variability, but sometimes a pilot study needs to be carried out in order to obtain a reasonable estimate. Specifying the magnitude of the differences you wish to detect between the null and alternative hypothesis varies by consideration of subject matter. The importance of a difference varies by discipline and context.

The following summarizes the interrelationships of power, sample size, level of significance, variability and detectable difference:

To increase power, 1-β:

- Increase: Required Sample Size, n; Detectable Difference, between the mean under H_0 and H_1 ; Level of significance, α;
- Decrease: Variability, σ_1^2 , σ_2^2 , σ_3^2

To reduce required sample size, $n = n_1 + n_2$ -

- Increase: Detectable Difference, between the mean under H_0 and H_1 ; Level of significance, α
- Decrease: Variability, σ_1^2 , σ_2^2 , σ_3^2 ; power, 1-β

Uncertainty in specifications of mean difference, variability

For a discussion of the role of uncertainty regarding inputs to power analysis and sample size estimation, refer to the tutorial in this series on *Uncertainty in Power and Sample Size Estimation*.

Content A: Features and Assumptions
Study design description: Ki-67 and resveratrol example

In the Ki-67 example, the inference relates to the differences in mean change in Ki-67 among three populations. The outcome variable for each sample is continuous (change in % staining) with an unknown, but finite variance of the change in each population, σ_1^2, σ_2^2 and σ_3^2. Since we base the % staining on a large number of tumor cells, we may reasonably to assume a normal distribution value for Ki-67 value. As linear functions of normally distributed quantities are also normally distributed, the change in Ki-67 from before treatment to after treatment is normally distributed. Finally, the *a priori* stated hypothesis regards the mean differences in change of Ki-67 and whether the mean differences are 0%. These are the features that make the example suitable for a one-way ANOVA.

In conducting a one-way ANOVA we make the assumption that the individual change values are randomly sampled from one of three well-characterized populations and that the observations within a sample are independent of each other, i.e. that there is no clustering between subjects or units of observation. In most cases, we can easily verify this assumption. As noted above, we assume an approximately normal underlying distribution of change in Ki-67 with resveratrol, and normal or approximately so for the other supplement or placebo . Finally, we make the assumption that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$.

The one-way ANOVA as a General Linear Model (GLM)

A general characterization of the one-way ANOVA can be made using the general linear model (GLM), $Y = X\beta + \varepsilon$. For the test of the mean change in Ki-67, *Y* is an $(n_1 + n_2 + n_3)$ x 1 matrix of Ki-67 change values, *X* is an $(n_1 + n_2 + n_3)$ x 3 matrix of 1s and 0s, β is a 1 x 3 matrix of the unknown regression coefficients, β_1 , β_2 , and β_3 – in this case the mean changes in Ki-67 in the resveratrol, other supplement and placebo groups, and ε is an $(n_1 + n_2 + n_3)$ x 1 matrix that represents the random deviation of a single Ki-67 change value from its group mean change. The Y and X matrices are shown below:

Y_Δ	X_1	X_2	X_3
$Y_{1,1Pre}$-$Y_{1,1Post}$	1	0	0
$Y_{1,2Pre}$-$Y_{1,2Post}$	1	0	0

.	.	.	.
.	.	.	.
$Y_{1,n1Pre}$-$Y_{1,n1Post}$	1	0	0
$Y_{2,1Pre}$-$Y_{2,1Post}$	0	1	0
$Y_{2,2Pre}$-$Y_{2,2Post}$	0	1	0
.	.	.	.
.	.	.	.
$Y_{2,n2Pre}$-$Y_{2,n2Post}$	0	1	0
$Y_{3,1Pre}$-$Y_{3,1Post}$	0	0	1
$Y_{3,2Pre}$-$Y_{3,2Post}$	0	0	1
.	.	.	.
.	.	.	.
$Y_{3,n3Pre}$-$Y_{3,n3Post}$	0	0	1

The assumptions of the GLM are:

- **Existence**: For any fixed value of the variable *X*, *Y* is a random variable with a certain probability distribution having finite mean and variance.
- **Independence**: The *Y*-values are statistically independent of one another.
- **Linearity**: The mean value of *Y* is a straight-line function of *X*.
- **Homoscedasticity (equal variances)**: The variance of *Y* is the same for any value of *X*. That is,

$$\sigma^2_{Y|X} = \sigma^2_{Y|X=1} = \sigma^2_{Y|X=2} = ... = \sigma^2_{Y|X=x}$$

- **Normal Distribution**: For any fixed value of *X*, *Y* has a normal distribution. Note this assumption does ***not*** claim normal distribution for *Y* .

To test the hypothesis H_0: $\mu_{\Delta Ki\text{-}67,\ resveratrol} = \mu_{\Delta Ki\text{-}67,\ other\ supplement} = \mu_{\Delta Ki\text{-}67,\ placebo}$ we determine if the between-group variation is significantly larger than the within-group variation. To perform this test, we calculate the Mean Square Between ($MS_{Between}$) and the Mean Square Within (MS_{Within}) by dividing the $SS_{Between}$ and SS_{Within} by their respective degrees of freedom:

$$MS_{Between} = SS_{Between} / (k\text{-}1)$$

$$MS_{Within} = Wi_{thin} / (n\text{-}k),$$

where k = 3, $SS_{Between} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_i})^2$, SSWithin = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_i} - \overline{\overline{Y}})^2$, Y_{ij} is the ith observation in group j, $\overline{Y_i}$ is the mean of group i, and $\overline{\overline{Y}}$ is the overall mean. We summarize this information in what is called an Analysis of Variance (ANOVA) table:

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Variance Ratio (F)**	**p-value**
Between	$SS_{Between}$	*k-1*	$MS_{Between}$	*F*	Pr($F_{k-1,n-k}$>*F*)
Within	SS_{Within}	*n-k*	$MS_{Within} = \sigma^2_{Y\|X}$		
Total	SS_{Total}	*n*-1			

Under the null hypothesis, the *F* statistic will have an *F* distribution with *k*-1 and *n-k* degrees of freedom. If $F > F_{k-1,n-k,1-\alpha}$ then we reject H_0.

B.1 Inputs for Power analysis

A power analysis for the differences in mean change consists of determining the achievable power for a specified: difference between means under the stated H_0 and under the stated H_1, sample size, standard deviation of the difference, and α-level. By varying these four quantities a set of power curves can be obtained that show the tradeoffs.

Information about mean change values and variability of change 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.

For the change in Ki-67 expected with resveratrol, the other supplement and placebo, we look to a study examining change in Ki67. A standard deviation of 14% for change in Ki67 with resveratrol was reported.[1] For this example we will assume that the variances in the three groups are all equal to $(14\%)^2$, or $196\%^2$.

When using information from studies in the literature, it can't be assumed that the means and standard deviations are known quantities. They are estimates and as such lead to uncertainty in estimated power. For further discussion about uncertainty refer to the *Tutorial 0: Uncertainty in Power and Sample Size Estimation*.

Specification of the differences between means for change, $\mu_{\Delta i} - \mu_{\Delta j}$

As noted above, subject matter considerations dictate the choice of a value for the difference between the null and alternative mean change . The goal is to specify the smallest difference that would be considered scientifically important. For the paired-sample Ki-67 example we consider a 20% difference between resveratrol and placebo,

[1] It should be noted that in some publications, standard error values are reported. Standard errors refer to the precision of a mean, not to the variability in a single sample. To convert a standard error to a standard deviation, multiply the standard error by the square root of the sample size. Often separate standard deviations of the pre- and post- measures are reported instead of the standard deviation of the change. In this case, assumptions must be made about the correlation, ρ, between the pre- and post- measures in order to obtain an estimate for the standard deviation of change. A value of 0 for ρ means the pre- and post- measures are uncorrelated while a value of 1 means they are perfectly correlated. Low values of ρ would lead to conservative power and sample size estimates while large values of ρ would lead to liberal ones. For a specific ρ value, the standard deviation of change is obtained as: $s_\Delta = \sqrt{s^2_{pre} + s^2_{post} - 2\rho s_{pre} s_{post}}$, where s^2 and s are the sample variance and standard deviation, respectively. Rarely is the variance reported in a publication. If it is, the standard deviation is obtained as the square root of the variance.

and a 20% difference between the other supplement and placebo, as biologically important.

B.2 Inputs for sample size estimation

Sample size estimation for multiple means from paired observations consists of determining the required sample size to achieve a specified level of power for a given difference between the mean changes under the stated H_0 and under the stated H_1, the common standard deviation of the change and α-level. By varying these four quantities we can obtain a set of sample size curves that show the tradeoffs.

As with power analysis, information about mean values and variability can be obtained from the published literature. Uncertainty in these values should be incorporated into the sample size estimation by either allowing those values to vary over a reasonable range, e.g. from 0.5 times up to twice the standard deviation, or by using sampling variability to obtain confidence intervals for power or sample size. Power can be varied over a desired 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 and Type II error rates.

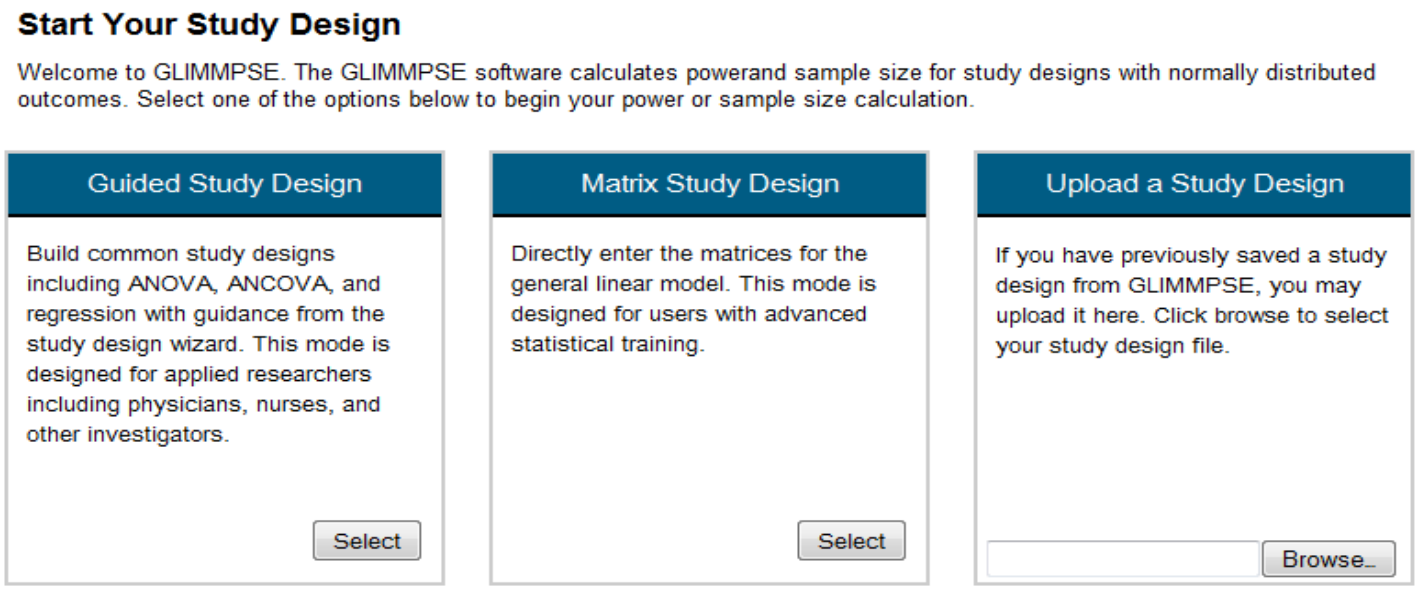
Content C: How to use the software

How to perform the analysis in GLIMMPSE

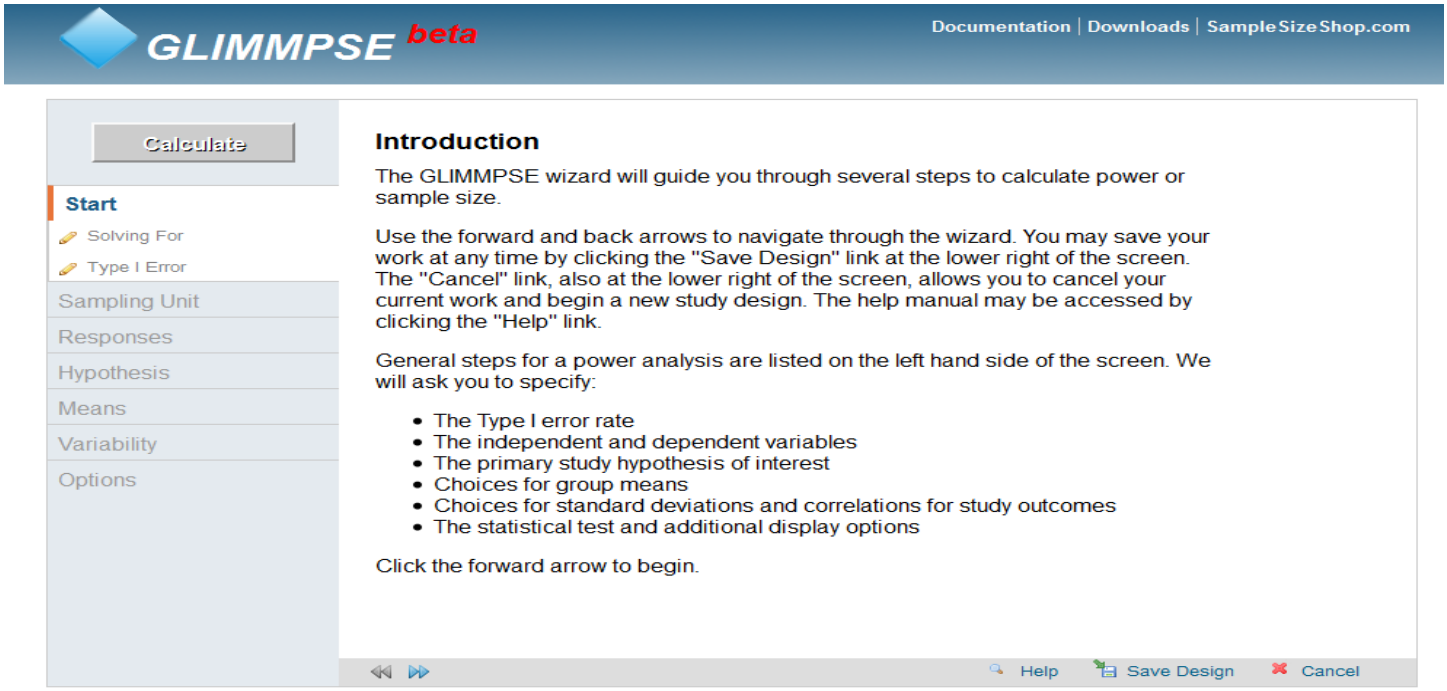
To start GLIMMPSE 2.0 beta, type http://samplesizeshop.com/calculate-power-and-sample-size-now/ in your browser window or visit www.SampleSizeShop.com and click on the GLIMMPSE tab, then on GLIMMPSE 2.0.0 Beta is Here! Google Chrome is the suggested browser for this application.

C.1 Power analysis using Guided Study Design mode

Guided Study Design mode is suggested for most users. For those familiar with Muller, LaVange, Ramey and Ramey (1992) we recommend using Matrix Study Design mode, which is illustrated in a separate tutorial.



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

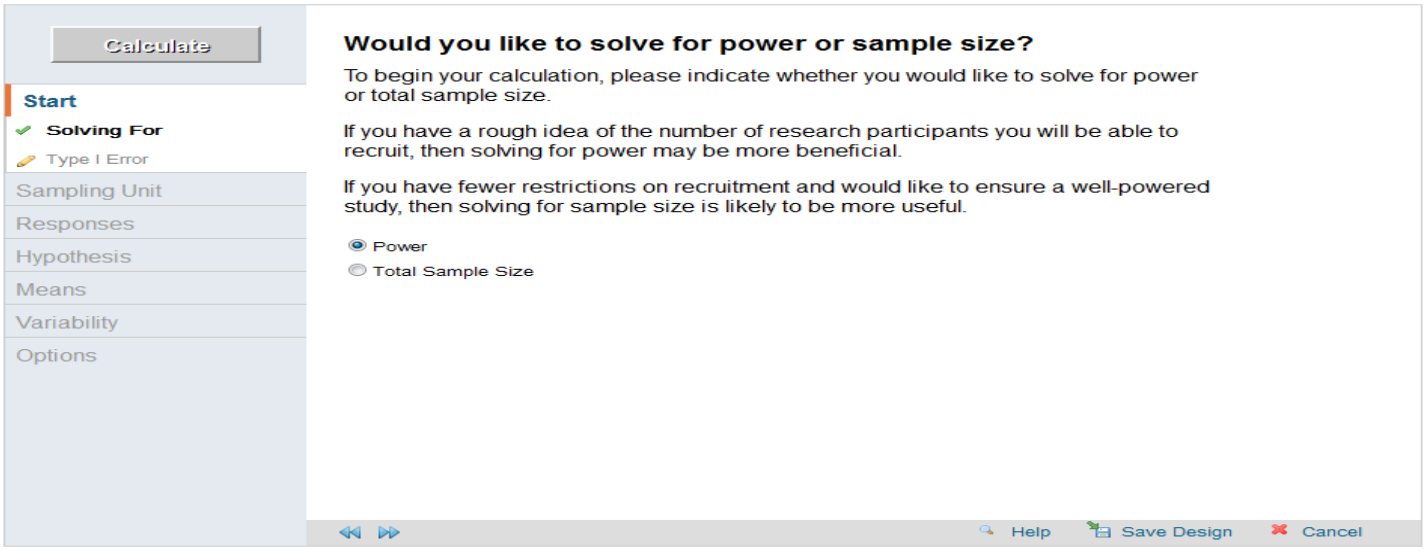


Throughout your use of the GLIMMPSE website, use the blue arrows at the bottom of the screen to move forward and backwards as you enter you values. The pencil symbols beside the screen names on the left indicate required information that has not yet been entered. In the above picture, the pencils appear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

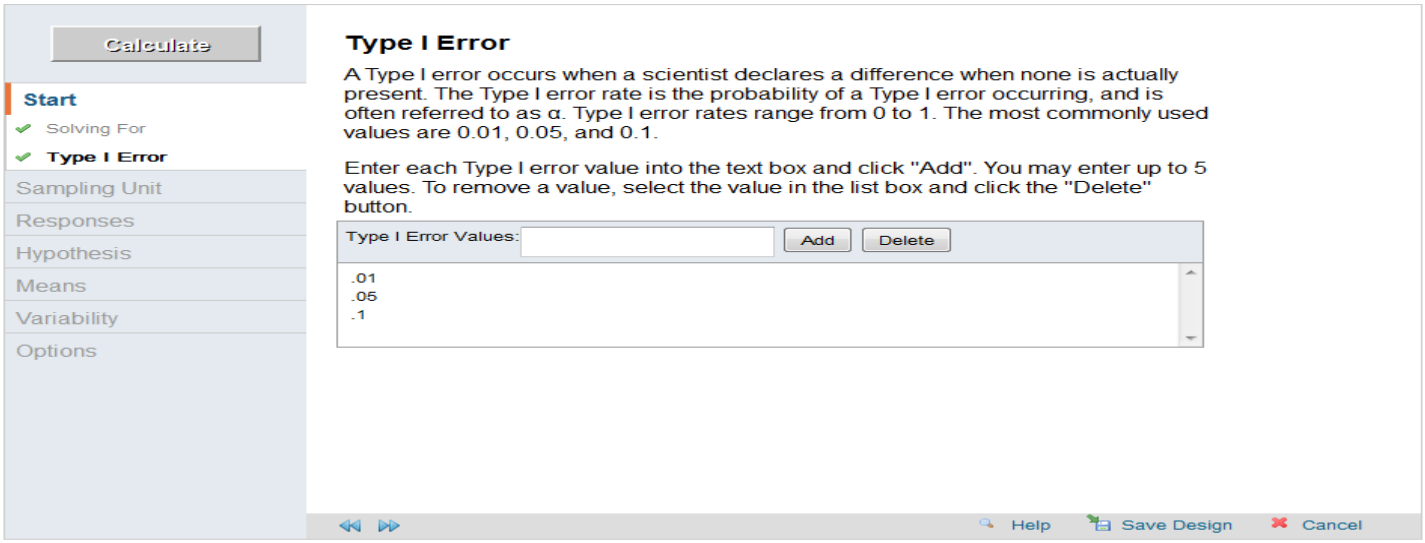
Once the required information is entered, the pencils become green check marks. A red circle with a slash through it indicates that a previous screen needs to be filled out before the screen with the red circle can be accessed. Search for a pencil beside a screen name in the previous tabs to find the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

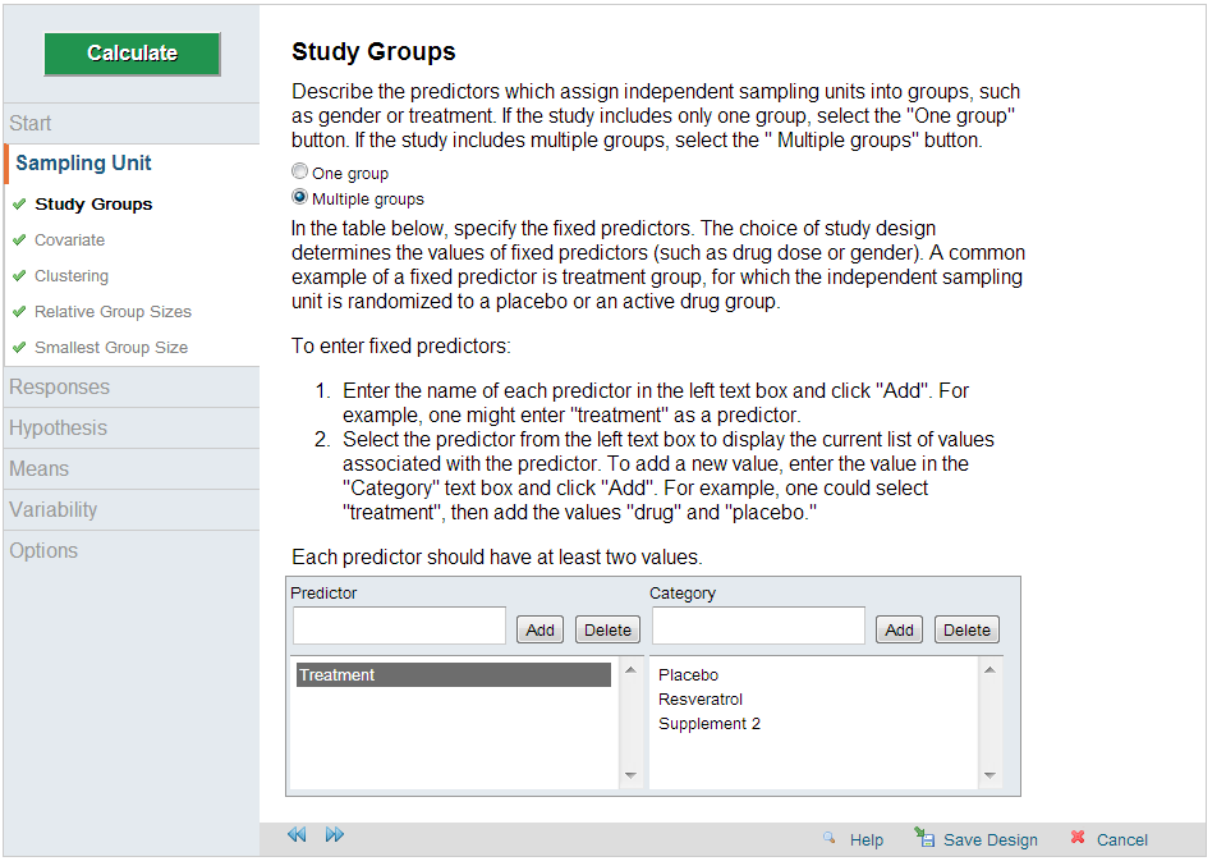


The *Type I Error* screen allows you to specify the fixed levels of significance for the hypothesis to be tested. Once you have entered your values, click the forward arrow to move to the next screen.

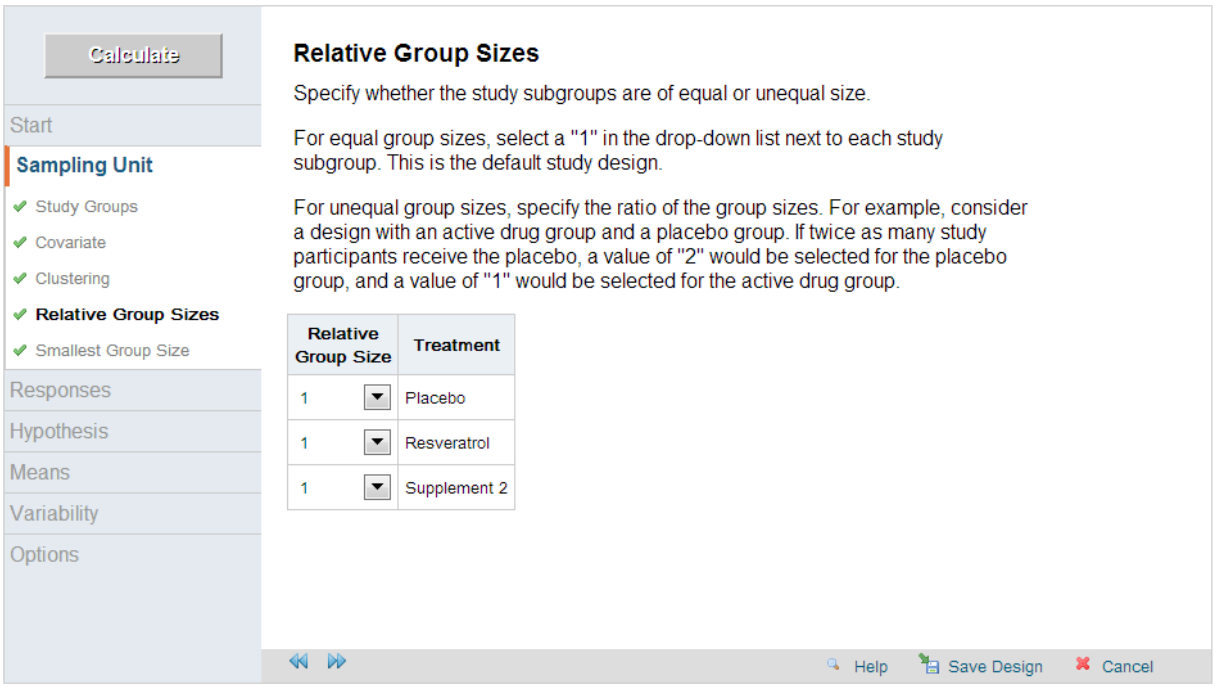


Read the information on the *Sampling Unit: Introduction* screen and click next to move to the next screen.

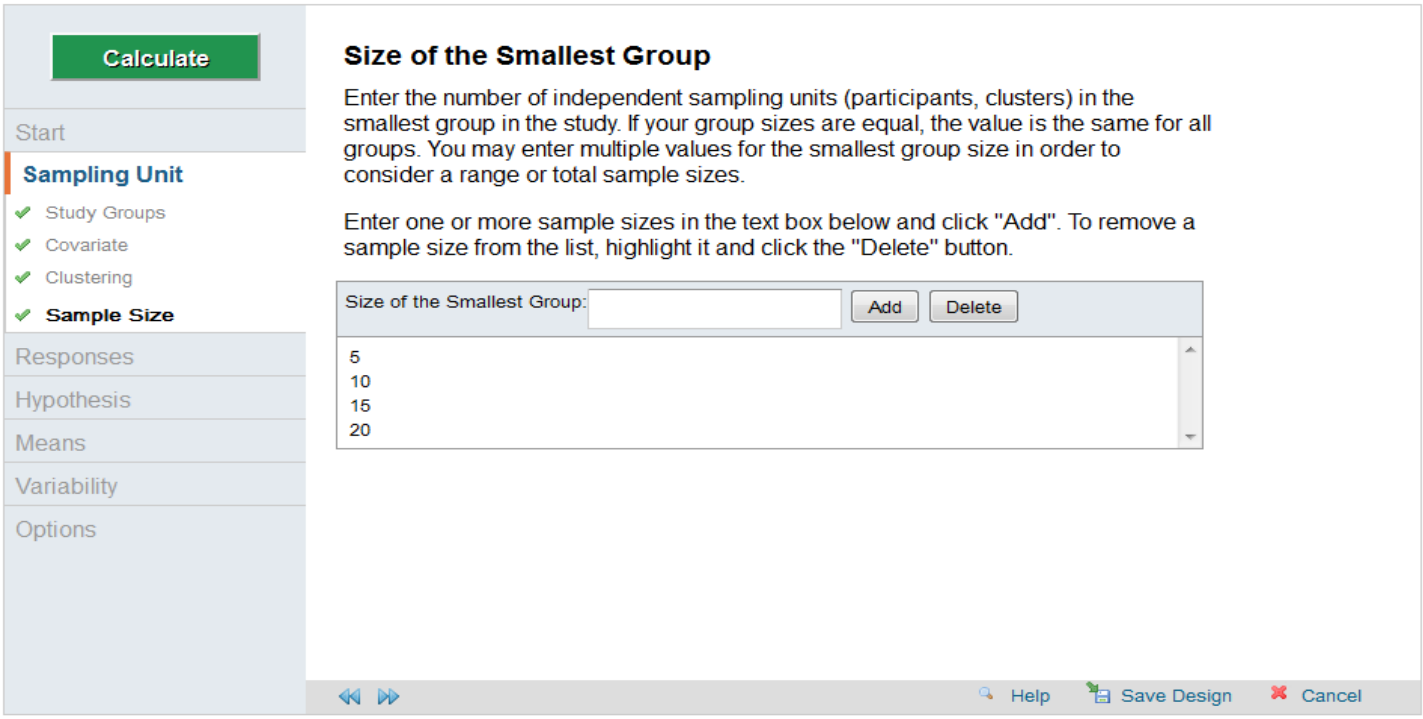
On the *Study Groups* screen, click on *Multiple Groups*. In the drop down menu that appears, enter *Treatment* as your predictor and *Resveratrol*, *Supplement 2* and *Placebo* as your categories.



The *Relative Group Sizes* screen allows you enter the ratio between the three groups. For this example, assume the groups are of equal size and leave all of the numbers as 1 .

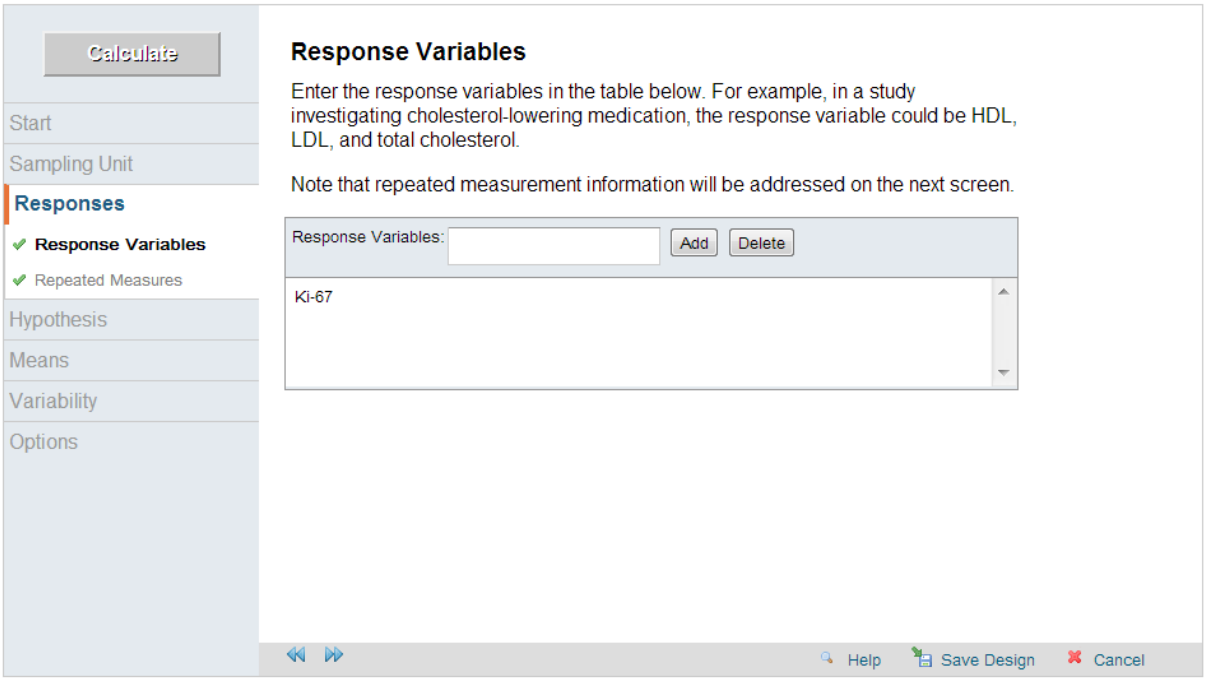


Use the *Sample Size* screen to specify the size of the smallest group for your sample size(s). Although only one sample is being used for this example, entering multiple values for the smallest group size allows you to consider a range of total sample sizes.



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

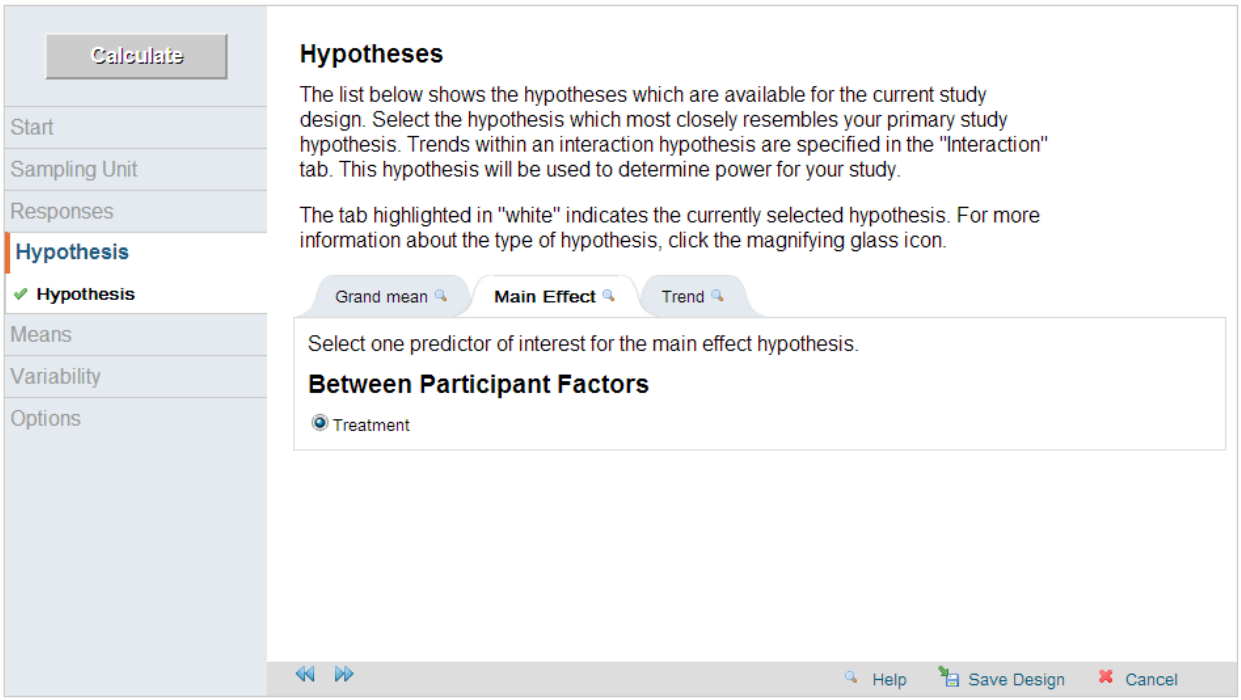
The *Response Variable* screen allows you to enter the response variable(s) of interest. Enter the three response variable, Ki-67.



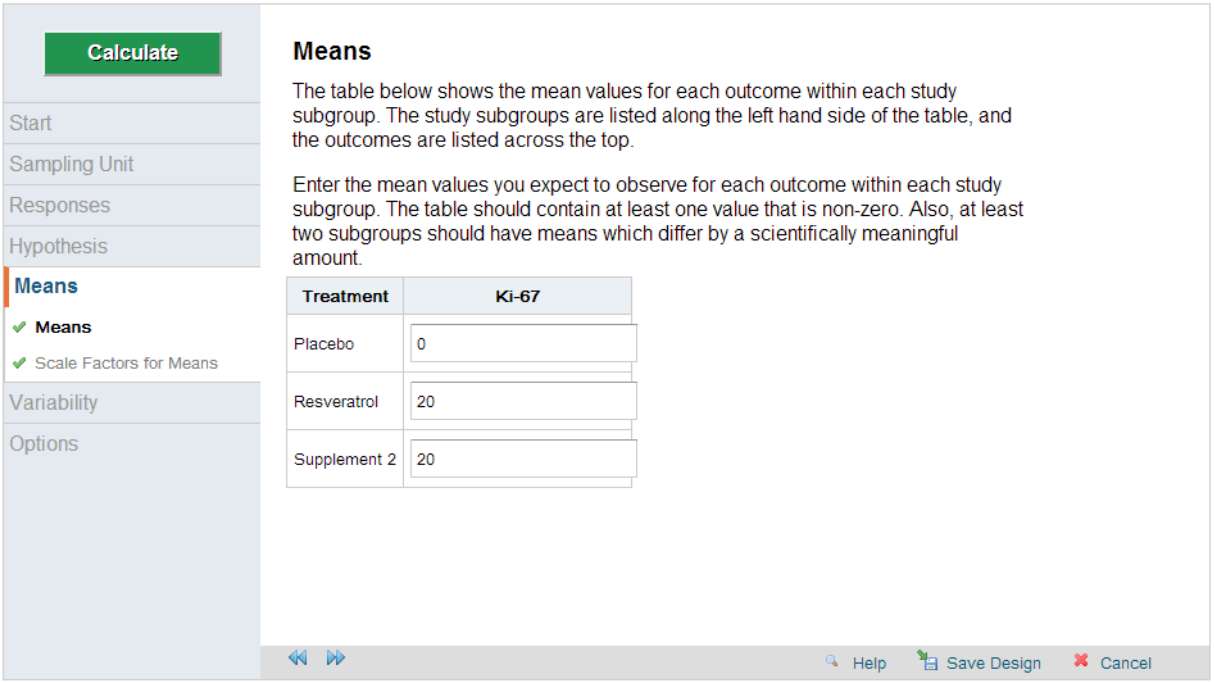
The *Repeated Measures* screen should be skipped, as there are no repeated measures for this example.

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

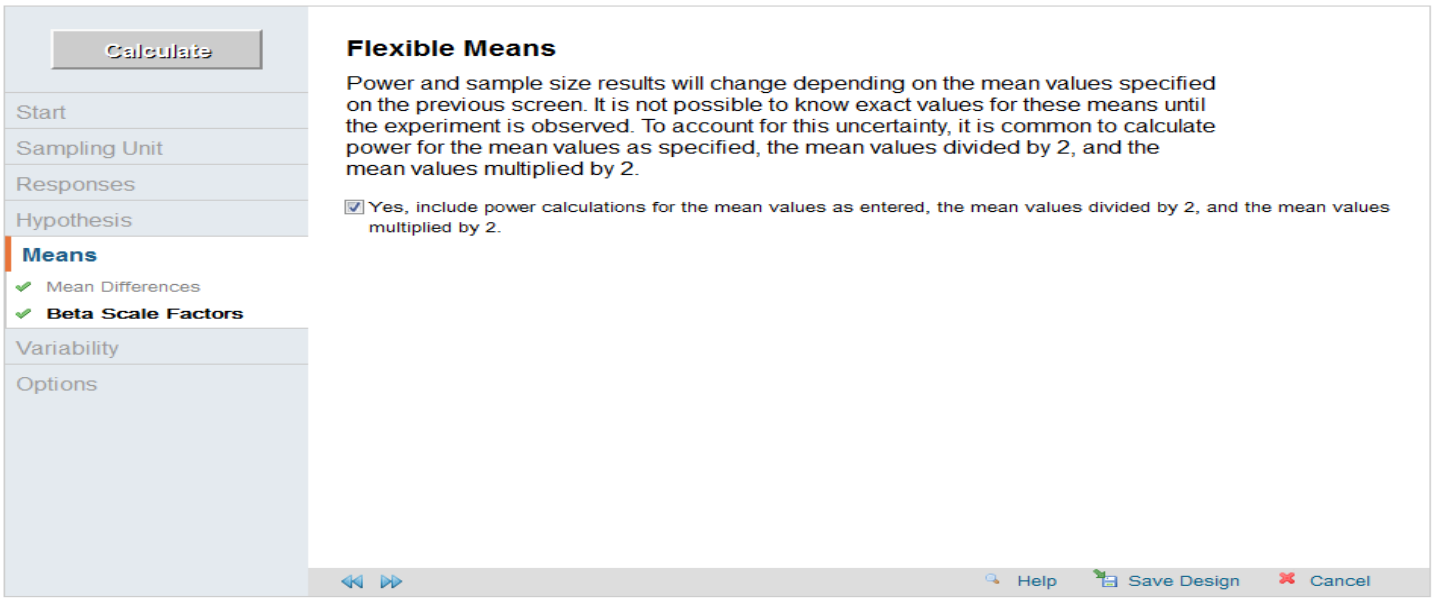
The *Hypotheses* screen allows you to specify the hypothesis which most closely resembles your primary study hypothesis, and to enter the known mean values for each hypothesis. After entering your hypothesis, click the forward arrow.



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 this example, enter 20%, or 20, in the first box, 20 in the second, and 0 in the last box.

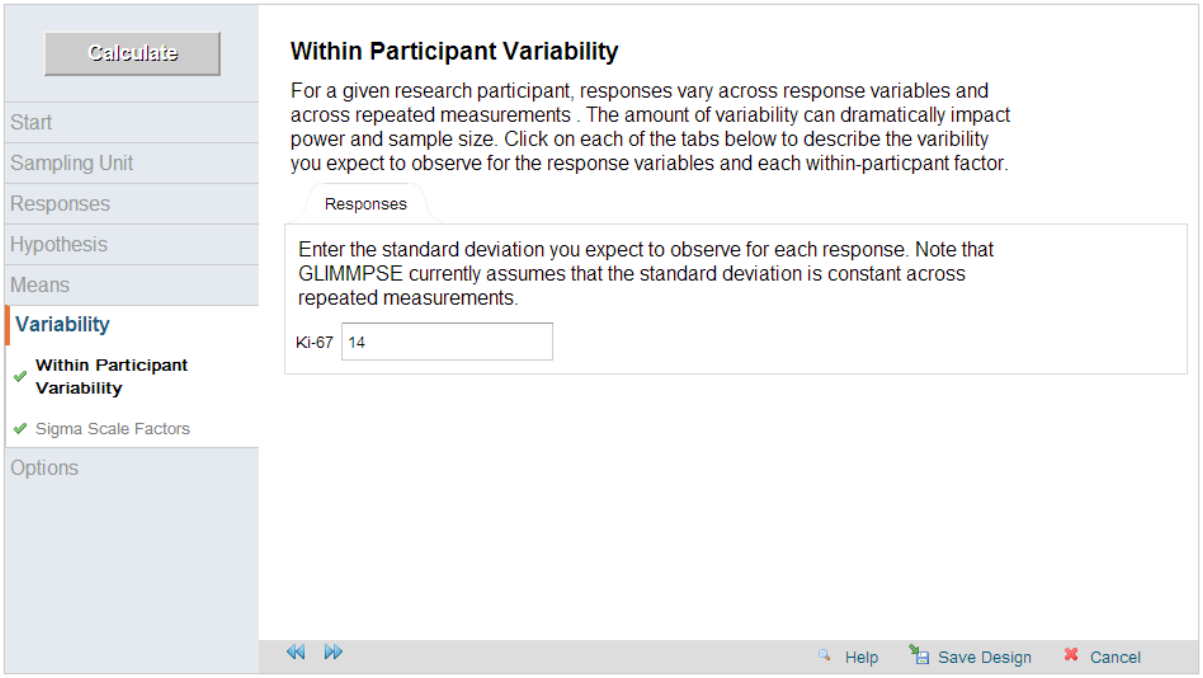


The true mean difference is uncertain. The *Beta Scale Factors* screen allows you to see how power varies with the assumed difference, so that you can allow it to vary over a reasonable range. GLIMMPSE allows this to be from 0.5x to 2x the stated difference, e.g. from 10% to 20% to 40%. Click *Yes*, then click on the forward arrow to continue.

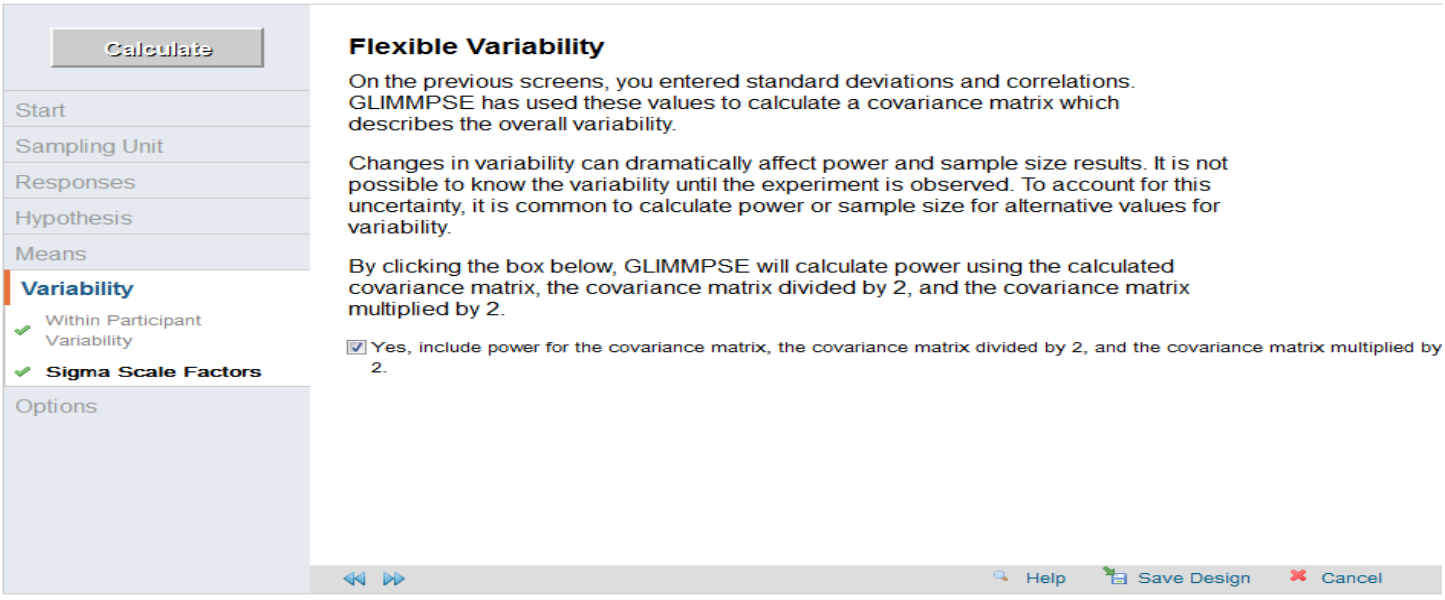


Read the information on the *Variability: Introduction* screen, then click the forward arrow to continue.

The *Within Participant Variability* screen allows you to specify the expected variability in terms of standard deviation of the outcome variable. For this example, the standard deviation is 14%. Enter 14, and the other boxes will self-populate. Click on the forward arrow to move to the next screen.

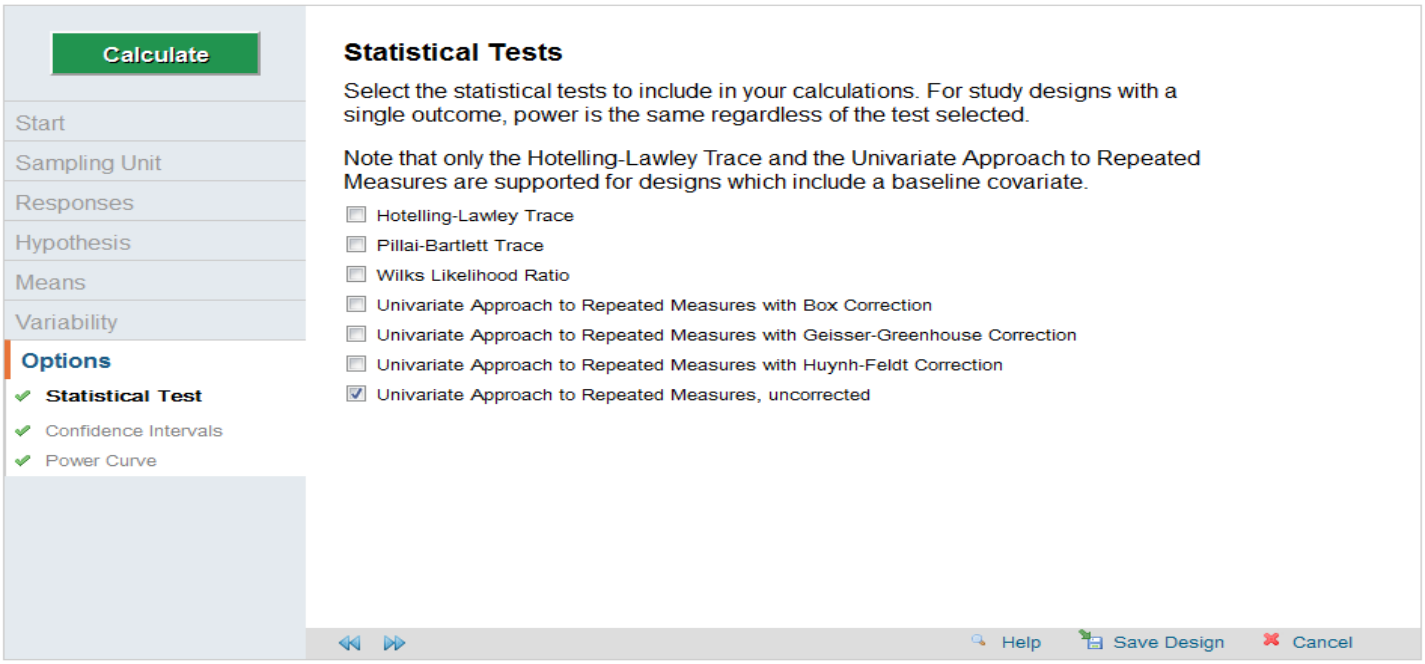


The true variability in Ki-67 is also uncertain. To see how power varies with the assumed standard deviation, you can allow the standard deviation to vary over a reasonable range. GLIMMPSE allows this to be from 0.5x to 2x the stated standard deviation, e.g. from 7% to 14% to 28%. Click *Yes*, then click the forward arrow to continue.



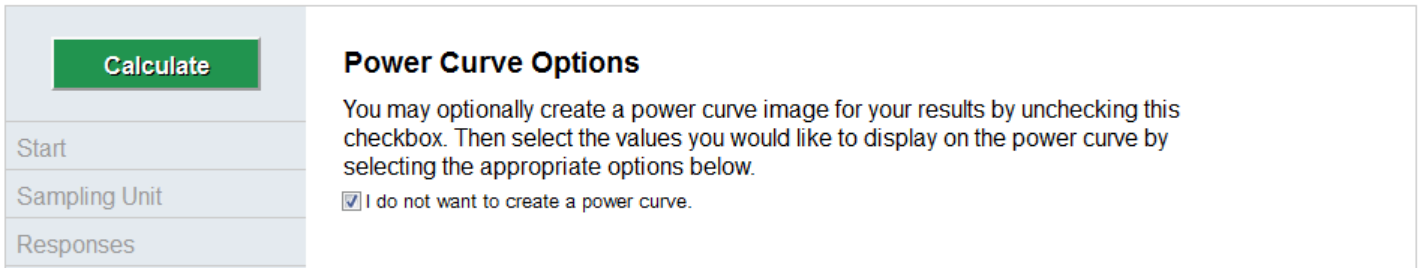
Read the information on the *Options* screen, then click the forward arrow to continue.

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



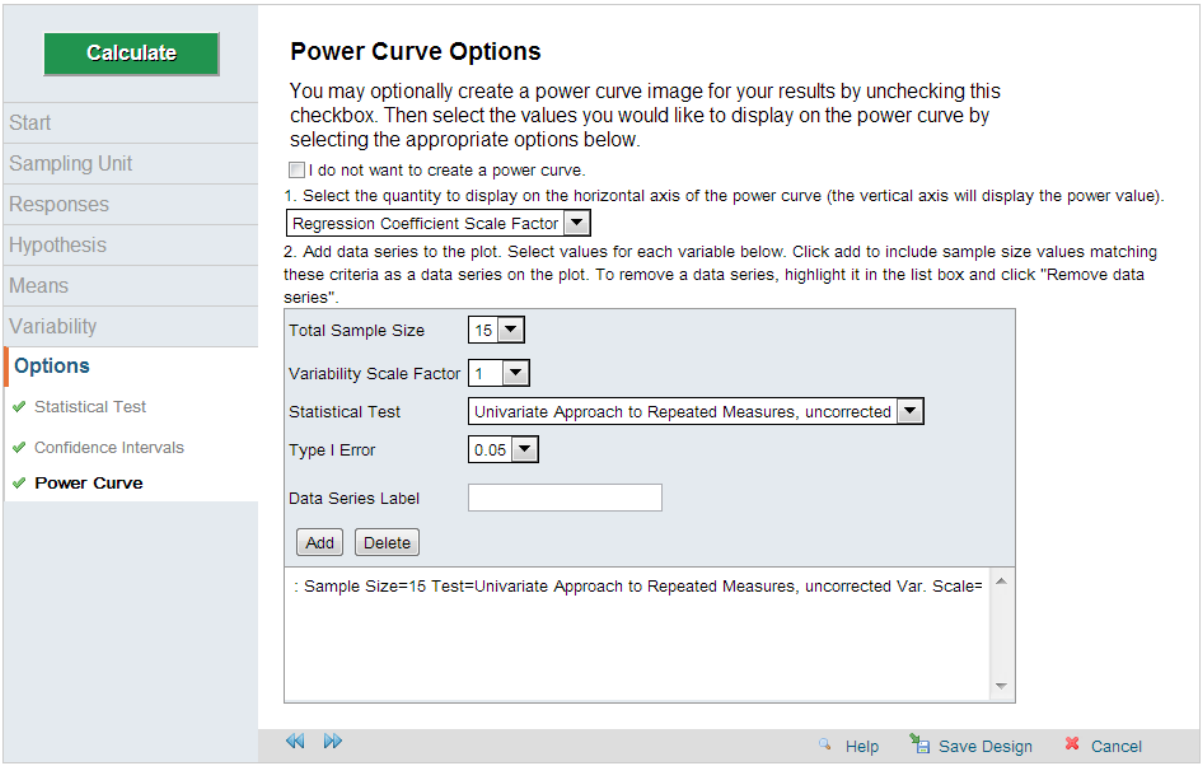
Leave the box checked in the *Confidence Interval Options* screen, and click forward to continue to the next screen.

Power analysis results are best displayed on a graph. To obtain a plot, first uncheck No thanks 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

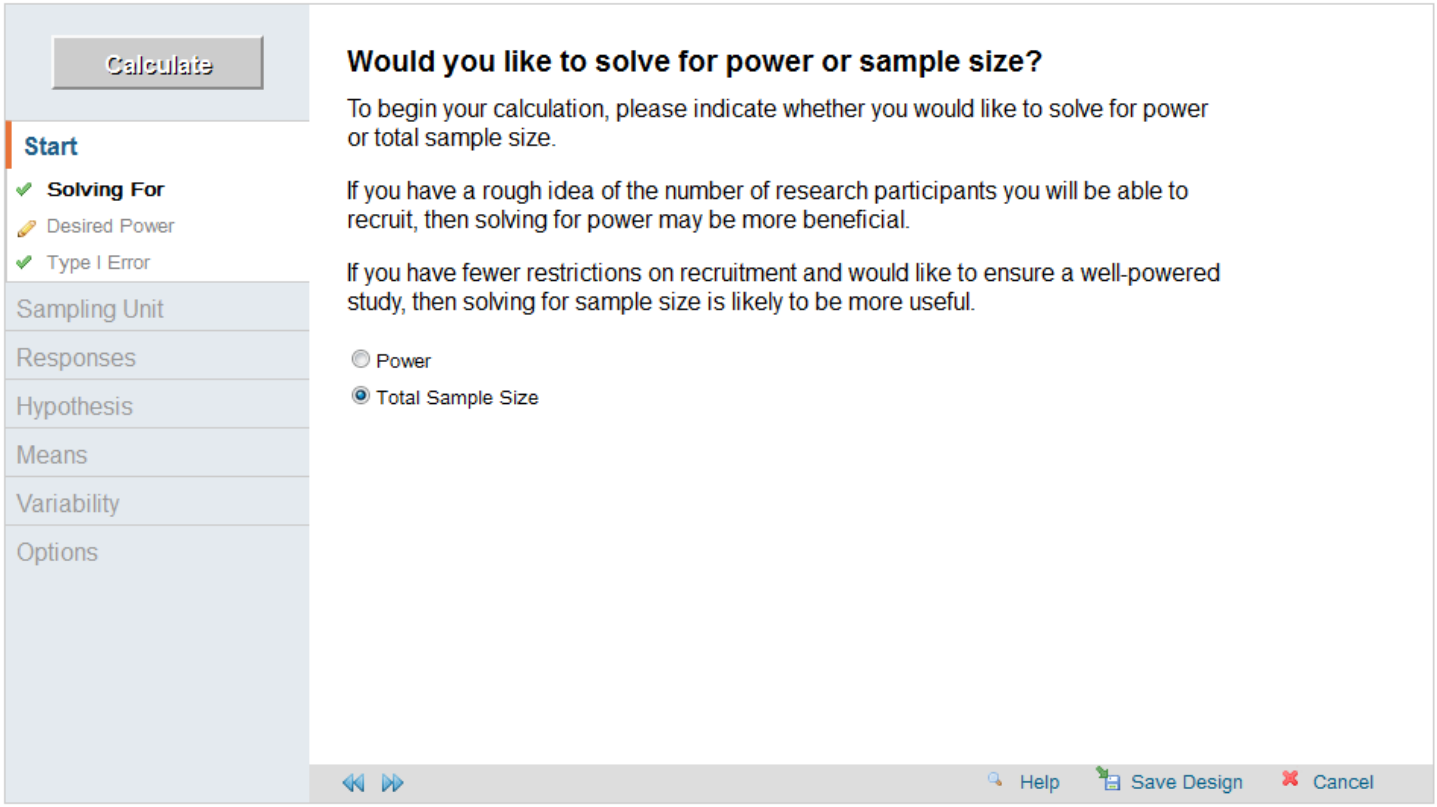
variables that you specify on this page. If you specified more than one value for an input variable, choose the specific level you want GLIMMPSE to use to plot the power curve.



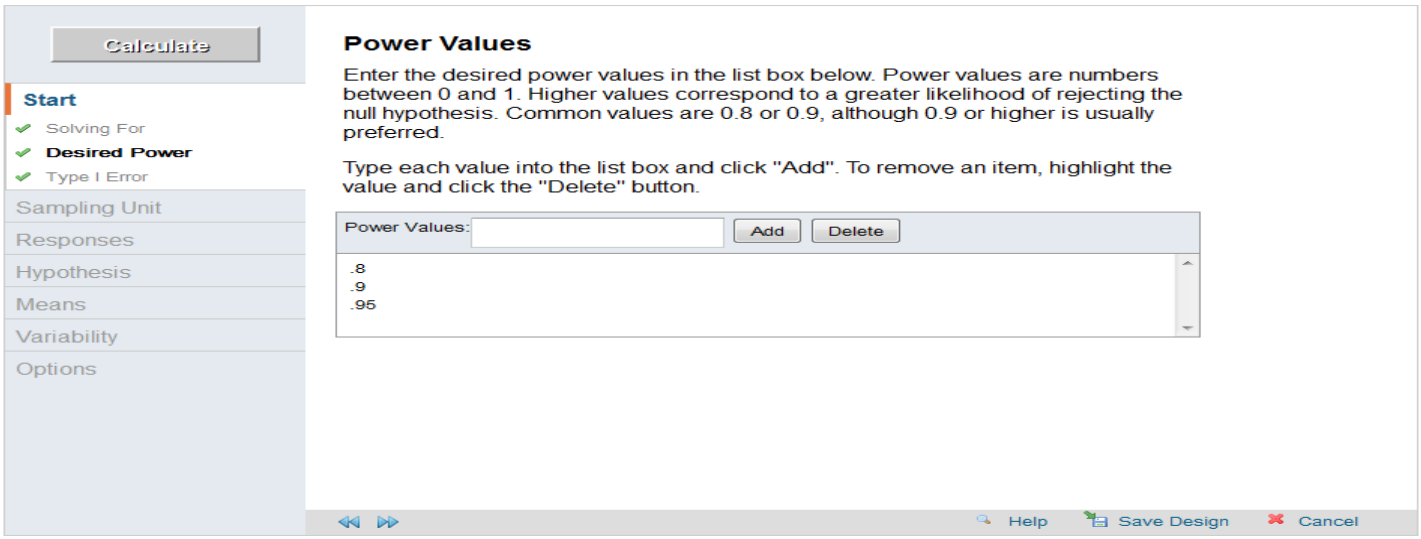
The power curve and table results for these inputs are shown in Section D below.

C.2 Sample size estimation with GLIMMPSE

To start, click on *Total Sample Size* in the *Solving For* screen. Click on the forward arrow to move to the next screen.



Enter the values for desired power, then click the forward arrow to move to the next screen.



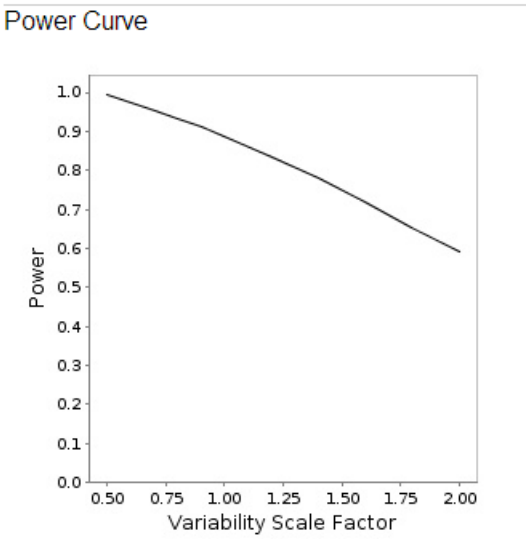
The remaining screens are filled in as above for power.

Content D: Interpret the results

Power analysis interpretation

For the power analysis inputs in Guided Study Design Mode (Section C.1), GLIMMPSE produces a curve showing the relationship between achievable power, mean difference (beta scale), standard deviation (variability scale), and fixed values of total sample size and level of significance. A complete downloadable table of results in Excel.csv format is also produced. The plot below shows achievable power over a range of differences in mean change in Ki-67 with an α-level of 0.05 (two-sided) and a sample size of 15 per group.

It can be observed in the graph below that above a certain level of variability, power is much less sensitive to changes in specified values. The sensitivity of power to assumptions within a certain range of values has been called the slippery slope of power. In order to avoid the slippery slope, it is best to assume a large amount of variability and/or a large difference between null and alternative means.



Sample size estimation interpretation

For the sample size estimation inputs in Guided Study Design mode above, GLIMMPSE produces a downloadable table showing the relationship between required sample size, desired power, level of significance, variability, and difference between null and alternative hypothesis means. This can be saved in Excel.csv format and plotted.

Content E: References cited

List of matrices used in calculation:

$$\mathbf{Es(X)} = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0.0000 \\ 20.0000 \\ 20.0000 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 1.0000 & -1.0000 & -0.0000 \\ 1.0000 & -0.0000 & -1.0000 \end{bmatrix}$$

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

$$\mathbf{\Theta_0} = \begin{bmatrix} 0.0000 \\ 0.0000 \end{bmatrix}$$

$$\mathbf{\Sigma_E} = [196.0000]$$

For the one-way ANOVA, GLIMMPSE works with the matrices listed above in making the computations. Since three means are being tested the essence matrix Es(**X**) is of dimension 3 x 3. The $\mathbf{\Theta_0}$ matrix represents the mean under H_0 and the $\mathbf{\Sigma_e}$ matrix represents the between-subject variance. The **B** is a vector of regression coefficients and **C** is the 2 x 3 contrast matrix for **B**.

References

Muller KE, LaVange LM, Ramey SL, Ramey CT (1992). Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications. *Journal of the American Statistical Association*, 87:1209-1226.

Content F: Exhibits, tables, figures

Ki-67 staining

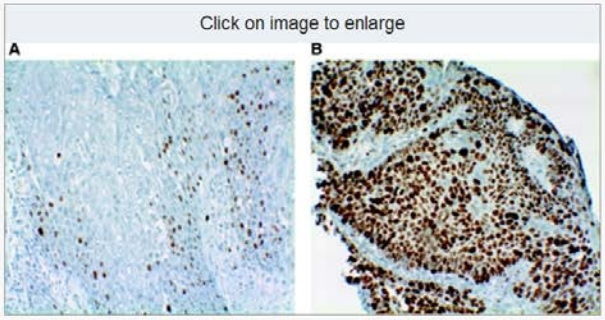


Figure 2

Immunohistochemical detection in primary pretreatment biopsies of: (**A**) a low percentage (17%) of Ki-67-positive nuclear staining; (**B**) a high percentage (93%) of Ki-67-positive nuclear staining.

From: Br J Cancer. 2008 October 7; 99(7): 1121–1128. Published online 2008 September 2. doi: 10.1038/sj.bjc.6604633