Tutorial 1: Power and Sample Size for the One-sample t-test

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.

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. The investigator usually hopes that the evidence in the data will disprove the null hypothesis. For a one-sample test of the mean, the null hypothesis $H_0$ is a simple statement about the expected value of the mean, if the sample is representative of a given reference population.

As an example, let us say that for a sample of early and late stage head and neck cancer patients, Ki-67 values (a measure of cell proliferation) are obtained from tissue biopsies. The goal is to compare this sample to a reference population, one with a mean Ki-67, $\mu_{Ki-67}$, of 42% (e.g. Seoane et al., 2010). The null hypothesis is that the sample came from that population, i.e. $H_0: \mu_{Ki-67} = 42\%$. An opposing, or alternative hypothesis $H_1$ is then stated, contradicting $H_0$, e.g. $H_1: \mu_{Ki-67} \neq 42\%$, $H_1: \mu_{Ki-67} < 42\%$, or $H_1: \mu_{Ki-67} > 42\%$. If the mean from the sample is sufficiently different from 42% a decision is made to reject the null in favor of the specific alternative hypothesis. If the sample and reference means are similar, 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 $\alpha$. The probability of failing to reject a
false null hypothesis, called Type II Error, is denoted by $\beta$. The possible outcomes of a hypothesis test are summarized in the table below along with the conditional probabilities of their occurrence.

<table>
<thead>
<tr>
<th>Hypothesis Testing Decision</th>
<th>State of Nature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$ True</td>
</tr>
<tr>
<td><strong>Fail to reject $H_0$</strong></td>
<td>True negative</td>
</tr>
<tr>
<td></td>
<td>A correct decision</td>
</tr>
<tr>
<td></td>
<td>$\text{Prob(True negative)} = 1-\alpha$</td>
</tr>
<tr>
<td><strong>Reject $H_0$</strong></td>
<td>False positive</td>
</tr>
<tr>
<td></td>
<td>A Type I Error</td>
</tr>
<tr>
<td></td>
<td>$\text{Prob(False positive)} = \alpha$</td>
</tr>
</tbody>
</table>

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 ($\alpha$, 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 include variability as summarized by $\sigma^2$, where $\sigma$ is the standard deviation of the population; and the difference between the mean under the null hypothesis and the mean under a specific alternative, i.e. $\mu_0 - \mu_1$. In practice, $\sigma$ 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 difference you wish to detect between the null and alternative hypothesis means 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 H₀ and H₁, \( |μ₀ - μ₁| \); Level of significance, \( α \);
- Decrease: Variability, \( σ₁^2 \) and \( σ₂^2 \)

To reduce required sample size, \( n = n₁ + n₂ \) -
- Increase: Detectable Difference, between the mean under H₀ and H₁, \( |μ₀ - μ₁| \); Level of significance, \( α \)
- Decrease: Variability, \( σ₁^2 \) and \( σ₂^2 \); power, 1-β

The applet available through the link below illustrates these relationships:

http://wise.cgu.edu/power_applet/power.asp

An additional consideration for power analysis and sample size estimation is the direction of the alternative hypothesis. A one-sided alternative hypothesis, e.g. \( H₁: μ_{Kί-67} > 42\% \), leads to greater achievable power than a two-sided alternative for a fixed sample size, or a lower required sample size for a fixed power. The use of one-sided tests is controversial and should be well justified based on subject matter and context.
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 0: Uncertainty in Power and Sample Size Estimation*. 
Content A: Features and Assumptions

Study design description: Ki-67 example

In the Ki-67 example, the inference relates to a single population. The outcome variable is continuous (% staining) with an unknown, but finite variance, $\sigma^2$. Since we base the % staining on a large number of tumor cells, we may reasonably assume a normal distribution value for that Ki-67 value. The \textit{a priori} stated hypothesis regards the mean of Ki-67 and whether the mean is the same as that of a reference population, in this case the population studied by Seoane et al (2010). These are the features that make the example suitable for a one-sample t-test.

In carrying out a one-sample t-test we make the assumption that the observations are randomly sampled from a well-characterized population and are independent of each other (i.e., that there is no clustering). In most cases, we can easily verify this assumption. In addition, we assume an approximately normal underlying distribution of change in Ki-67 with resveratrol. For large samples the t-distribution will be a good approximation to the t-statistic for the data, but in small to moderate sized samples the approximation will not be as good.

One-sample t-test: inference about the mean $\mu$ when the variance $\sigma^2$ is unknown

To test the hypothesis about the mean Ki-67, the standardized value of the sample mean, $\bar{X}_{Ki-67}$, becomes the test statistic, $t$. We assumed the $t$ value to follow a t distribution with n-1 degrees of freedom, i.e. the sample size minus 1:

$$t = \frac{\bar{X}_{Ki-67} - \mu_0}{s_{Ki-67}/\sqrt{n}} \sim t_{n-1}$$

where $s_{Ki-67}$ is the standard deviation of Ki-67 estimated in the sample. To test the hypothesis $H_0$: $\mu_{Ki-67} = 42\%$ vs. $H_1$: $\mu_{Ki-67} \neq 42\%$, we reject $H_0$ if:

$$\bar{X}_{Ki-67} > 42\% + t_{n-1,1-\alpha/2} s_{Ki-67}/\sqrt{n} \quad \text{or} \quad \bar{X}_{Ki-67} < 42\% - t_{n-1,1-\alpha/2} s_{Ki-67}/\sqrt{n},$$

where $t_{n-1, 1-\alpha/2}$ is the $(1-\alpha/2) \times 100$th percentile of the t distribution with n-1 degrees of freedom.
The t-test as a General Linear Model (GLM)

A more general characterization of the t-test can be made using the general linear model (GLM),  \( Y = X \beta + \varepsilon \). For the test of the mean Ki-67, \( Y \) is an \( n \times 1 \) matrix of Ki-67 values, \( X \) is an \( n \times 1 \) matrix of 1s, \( \beta \) is a 1x1 matrix of the unknown regression coefficient (in this case the estimated mean Ki-67), and \( \varepsilon \) is an \( n \times 1 \) matrix that represents the random deviations of single Ki-67 observations from their mean. The \( Y \) and \( X \) matrices are shown below:

<table>
<thead>
<tr>
<th>( Y )</th>
<th>( X )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_1 )</td>
<td>1</td>
</tr>
<tr>
<td>( Y_2 )</td>
<td>1</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( Y_n )</td>
<td>1</td>
</tr>
</tbody>
</table>

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,

\[
\sigma^2_{Y|X} = \sigma^2_{Y|X=1} = \sigma^2_{Y|X=2} = \ldots = \sigma^2_{Y|X=x}
\]

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

We obtain the estimate of \( \beta \), \( \hat{\beta} \), using the method of least squares. To test the hypothesis \( H_0: \mu_{\text{Ki-67}} = 42\% \), we use the standardized value of \( \hat{\beta} \), which is distributed as a t with \( n-1 \) degrees of freedom:  

\[
t = \frac{\hat{\beta} - 42\%}{s_{y|x}/\sqrt{n}} \sim t_{n-1}
\]

where \( s_{y|x} \) is the square root of the residual variance from the GLM.
Content B: How to obtain inputs for power analysis and sample size estimation

B.1 Inputs for Power analysis

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 $\alpha$-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 $\alpha$ can be selected to illustrate the sensitivity of the results to conservative vs. liberal choices for the Type I error rate.

As noted for the Ki-67 example, data are available on the mean and standard deviation in Seoane et al. (2010). Below is an excerpt of Table 1 from that publication.

| Table 1. Characteristics of cases and interval from first symptoms to diagnosis. |
|-------------------------------|----------------|-----------------|----------------|----------------|----------------|
| Variables         | Mean    | SD     | Median | Minimum | Maximum |
| 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. 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. Rarely is the variance reported in a publication. If reported, the researchers usually obtain the standard deviation as the square root of the variance.

Even when using information from large 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. Due to this uncertainty, when performing power analysis one must consider ranges of values for each of the inputs. For example, the standard deviation should usually range from 0.5 times the available estimate up to twice the estimate.
Specification of the difference between means, $\mu_0 - \mu_1$

As noted above, subject matter considerations dictate the choice of a value for the difference between the null and alternative means. The goal is to specify the smallest difference that would be considered scientifically important. We consider a difference of 10% biologically important for the one-sample Ki-67 example.

B.2 Inputs for sample size estimation

Sample size estimation for a single mean consists of determining the required sample size to achieve a specified level of power for a given difference between the mean under the stated $H_0$ and under the stated $H_1$, standard deviation and $\alpha$-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 estimates 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 knowledge of 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 $\alpha$ 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.

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.

Since there is only one sample, select *One group* on the *Study Groups* screen. Click the forward arrow to move to the next screen.

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 value for your primary hypothesis. Enter 42, then click the forward arrow to move to the next screen.

![Hypothesis screen](image)

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-67 example, enter 52%, or 52 in the screen below.

![Means screen](image)

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 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 on 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 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.

After you have made your inputs, click *Add* to populate the screen, then click the forward arrow to see your results.
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 standard deviations of Ki-67 with an $\alpha$-level of 0.05 (two-sided) and a sample size of 30. It can be seen that power is greatly reduced with increasing variability in Ki-67.

It can be observed in the graph below that above a certain amount 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.

![Power Curve](Image)
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.
List of matrices used in calculation:

\[ E_s(X) = \begin{bmatrix} 1.0000 \end{bmatrix} \]

\[ B = \begin{bmatrix} 52.0000 \end{bmatrix} \]

\[ C = \begin{bmatrix} 1.0000 \end{bmatrix} \]

\[ U = \begin{bmatrix} 1.0000 \end{bmatrix} \]

\[ \Theta_0 = \begin{bmatrix} 42.0000 \end{bmatrix} \]

\[ \Sigma_e = \begin{bmatrix} 625.0000 \end{bmatrix} \]

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 \( \Theta_0 \) matrix represents the mean under \( H_0 \) and the \( \Sigma_e \) matrix represents the between-subject variance. The remaining matrices \( E_s(X) \), \( C \) and \( U \) are scalars with a value of 1.

References


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

Copyright/LicenseRequest permission to reuse

Normal curves for power analysis and sample size estimation

Consider the distribution of $\bar{X}$ if $\mu = \mu_0$ and if $\mu = \mu_1$

$\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$ \hspace{1cm} $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$

Area = $\alpha/2$ \hspace{1cm} 1-$\beta$ \hspace{1cm} Area = $\alpha/2$

Lower Rejection Region \hspace{1cm} Acceptance Region \hspace{1cm} Upper Rejection Region