UNDERSTANDING THE STATISTICAL POWER OF A TEST

Hun Myoung Park
Software Consultant
UITS Center for Statistical and Mathematical Computing

How powerful is my study (test)? How many observations do I need to have for what I want to get from the study? The statistical power analysis estimates the power of the test to detect a meaningful effect, given sample size, test size (significance level), and standardized effect size. Sample size analysis determines the sample size required to get a significant result, given statistical power, test size, and standardized effect size. These analyses examine the sensitivity of statistical power and sample size to other components, enabling researchers to efficiently use the research resources.

1. WHAT IS A HYPOTHESIS?

A hypothesis is a specific conjecture (statement) about a property of population. There is a null hypothesis and an alternative (or research) hypothesis. Researchers often expect that evidence supports the alternative hypothesis. The null hypothesis, a specific baseline statement to be tested, usually takes such forms as "no effect" or "no difference."[1] A hypothesis is either two-tailed (e.g., $H_0 : \mu = 0$) or one-tailed (e.g., $H_0 : \mu \geq 0$ or $H_0 : \mu \leq 0$).[2]

Three points deserve being taken into account in making a hypothesis. *A hypothesis should be specific* enough to be falsifiable; otherwise, the hypothesis cannot be tested successfully. Second, a *hypothesis is a conjecture about a population (parameter), not about a sample (statistic).* Thus, $H_0 : \bar{x} = 0$ is not valid because we can compute and know the sample mean $\bar{x}$ from a sample. Finally, a *valid hypothesis is not based on the sample to be used to test the hypothesis*. This tautological logic does not generate any productive information.[3]

2. SIZE AND POWER OF A TEST

The *size of a test*, often called *significance level*, is the probability of Type I error. The Type I error occurs when a null hypothesis is rejected when it is true (Table 1). This test size is denoted by α (*alpha*). The 1- α is called the *confidence level*.

[1] Because it is easy to calculate test statistics (standardized effect sizes) and interpret the test results (Murphy 1998).
[2] μ (Mu) represents population mean, while $\bar{x}$ denotes sample mean.
[3] This behavior, often called "data fishing," just hunts a model that best fits the sample, not the population.

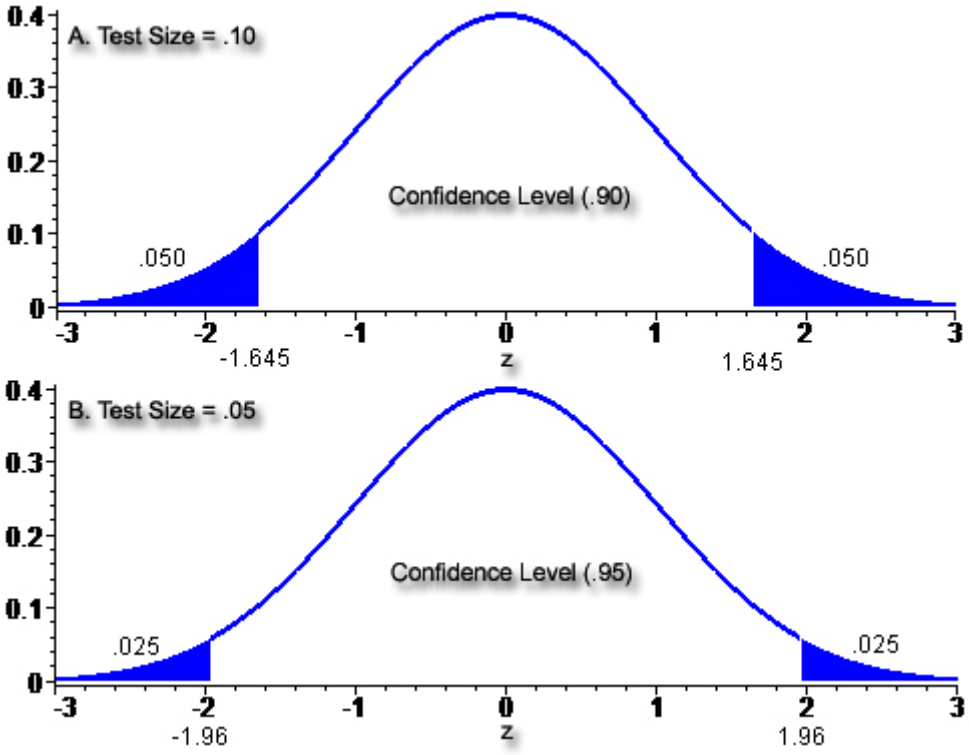
In a two-tailed test, the test size (significance level) is the sum of the two symmetric areas at the tails of a probability distribution. See the shaded areas of two standard normal distributions in Figure 1. These areas are called null hypothesis *rejection regions* in the sense that we reject the null hypothesis if a test statistic falls into these regions. The test size is a subjective criterion, although the .10, .05, and .01 levels are conventionally used.

Table 1. Size and Power of a Test

	Do not reject H_0	*Reject H_0*
H_0 is true	Correct Decision 1-α: Confidence level	Type I Error Size of a test α: Significance level
H_0 is false	Type II Error β	Correct Decision 1-β: Power of a test

Think about the .05 test size (see B in Figure 1). We need to know a particular value from which the sum of the areas up to the both infinities is .05. The value is called the *critical value* of the significance level. Critical values depend on test statistics, probability distributions, and test types (one-tailed versus two-tailed).

Figure 1. Test Size (Significance Level) and Critical Value



For example, the 1.96 (-1.96) in the standard normal distribution is the critical value of the two-tailed test at the .05 test size (significance level). Thus, the test size is the sum of probabilities that a sample statistic goes beyond the critical value (larger than 1.96 and less than -1.96). In the standard normal distribution, the critical value of a two-tailed test at the .10 significance level is 1.645 (see A in Figure 1) and 2.58 for the test size .01. As

test size (significance level) decreases, the critical value is shifted to the extremes; the rejection areas become smaller; it is less likely to reject the null hypothesis.

How do we substantively understand the test size or significance level? It is the extent that we are willing to take a risk of making wrong conclusion. The .05 level means that we are taking a risk of being wrong five times per 100 trials. A hypothesis test using a lenient test size like the .10 is more likely to reject the null hypothesis, but its conclusion is less convincing. In contrast, a stringent test size like .01 reports significant effects only when the effect size (deviation from the baseline) is large. Instead, the conclusion is more convincing (less risky). For example, the 1.90 in Figure 1 is considered exceptional at the .10 level (A), but not statistically discernable at the .05 level (B).

What is the *power of a test*? The power of a statistical test is the probability that it will correctly lead to the rejection of a false null hypothesis (Greene 2000). The statistical power is the ability of a test to detect an effect, if the effect actually exists (High 2000). Cohen (1988) says, it is the probability that it will result in the conclusion that the phenomenon exists (p.4). A statistical power analysis is either retrospective (post hoc) or prospective (a priori).

The statistical power is denoted by 1 – ß, where ß (*beta*) is the Type II error, the probability of failing to reject the null hypothesis when it is false (Table 1). See the shaded areas of B and B' in Figure 2. Conventionally a test with power greater than .8 level (or ß<=.2) is considered statistically powerful.

3. Components of the Statistical Power Analysis

What do we need to consider when conducting the statistical power analyses? There are six components:

1. Model (test)
2. Standardized effect size: (1) effect size and (2) variation (variability)
3. Sample size (n)
4. Test size (significance level α)
5. Power of the test (1–ß)

A research design contains specific models (tests) on which their statistical powers are based. Different models (tests) have different formulas to compute test statistics. For example, the T-test uses the T distribution to determine its statistical power, while ANOVA depends on the F distribution.

A *standardized effect size*, a test statistic (e.g., T and F scores) is computed by combining the effect size and variation.[4] An effect size in actual units of the response is the "degree to which the phenomenon exists" (Cohen 1988). Alternatively, an effect size is the deviation of hypothesized value in the alternative hypothesis from the baseline in the null

[4] In T-test, for example, the deviation (effect size) is divided by the standard error.

hypothesis. Variation (variability) is the standard deviation of population. Cohen (1988) calls it the reliability of sample results. This variation usually comes from previous research or pilot studies; otherwise, it needs to be estimated.

Sample size (N) is the number of observations (cases) in a sample. As mentioned in the previous section, the *test size* or *significance level* (α) is the probability of rejecting the null hypothesis that is true. The *power of the test* (1–ß) is the probability of correctly rejecting a false null hypothesis.

4. Power Analysis Using an Example

The computation of statistical power depends on specific a model (or test). The easiest model is the T-test with a relatively simple formula.

Imagine a random variable like the number of deaths per 100 thousand people from lung cancer. Suppose that the variable is known to be normally distributed with a mean of 20 and a standard deviation of 4 ($H_0 : \mu = 20$). See the probability distribution A in Figure 2.

Now, we believe that the mean is not 20, but 22 with the same standard deviation. See the probability distribution B in Figure 2. We also think that test size .05 will be fine. So, we are going to conduct a two-tailed T-test at the .05 significance level with the alternative hypothesis that the population mean is 22 ($H_a : \mu = 22$). How can we test our conjecture (alternative hypothesis)? Suppose we took a random sample with 44 observations from the population.

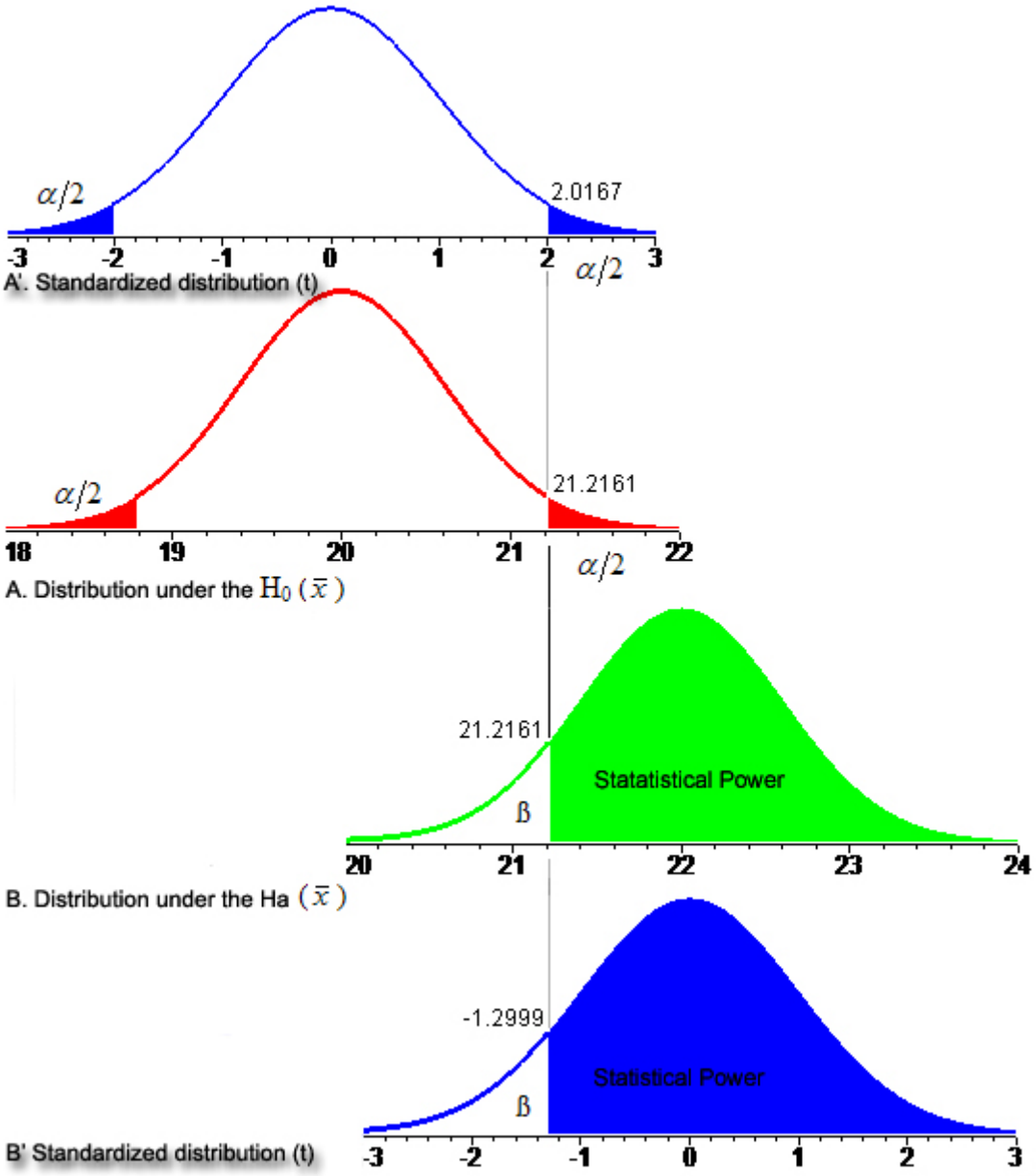
Think about the ordinary T-test first. We need to know how far the 22 is deviated from the baseline 20. Of course, the distance (effect size) is 2 (=22-20). But we do not know how big the effect size 2 is. Put differently, we do not know exactly how likely such a sample mean 22 can be observed if the true mean is 20. This is why people try to take advantage of using standardized probability distributions (e.g., T, F, and Chi-squared). By looking through these distribution tables, we are able to know the likelihood of observing a sample statistic in fairly easy manner. The A' in Figure 2 is a standardized probability distribution of A. The 20 in A corresponds to 0 in A' and 21.2161 in A is equivalent to 2.0167 in A'.

The t statistic here is 3.3166248: $t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{22 - 20}{4}\sqrt{44}$. This value is located all the way to the right in A', indicating the p-value is extremely small.[5] If the null hypothesis of population mean 20 is true, it is quite unlikely to observe the sample mean 22 (p<=.01). Obviously, the conjecture of population mean 20 is not likely. Thus, the null hypothesis is undoubtedly rejected in favor of the alternative hypothesis.

[5] The p-value of a test statistic is the sum of probabilities that the statistic and more exceptional (closer to both extremes) sample statistics are observed when the null hypothesis is true.

However, this test does not tell anything about the power of the test. We may want to know the extent that a test can detect the effect, if any. So, the question here is, "How powerful is this T-test?" There are four steps to compute the statistical power.

Figure 2. Statistical Power of a T-test



First, find the critical value in the original probability distribution with mean 20 and standard deviation 4. Let us use the standardized probability distribution A' to know the corresponding critical value in the original distribution A. By looking at the T distribution table (df=43 at the .05 level), we know that the probability of being greater than or equal to 2.0167 is .025 (see A' in Figure 2). Note that another .025 area is located at the

opposite direction because it is a two-tailed test. The 2.0167 is equivalent to $21.2161 = 2.0167 * 4/\sqrt{44} + 20$ in the original probability distribution A.[6]

Next, imagine an alternative probability distribution with mean 22 and standard deviation 4 (see B in Figure 2). It is the original probability distribution shifted to the right by 2 units. What is the probability of being less than the 21.2161 in the alternative probability distribution B? The probability is ß. In order to know ß, again let us convert the original value into the standardized probability distribution of the alternative probability distribution. The question is, "What is the T value for the 21.21611 in the alternative probability distribution?" (B and B' in Figure 2) The computations are made by one of the two approaches. We got -1.2999329.

$$t_a = 2.0166919 - \frac{|20-22|}{4}\sqrt{44} = -1.2999329$$
$$t_a = \frac{21.216111-22}{4}\sqrt{44} = -1.2999328$$

Third, find ß from the T distribution table with 43 (=N-1) degree of freedom. We can read .10027466. This ß is the probability of falling into the region less than or equal to the -1.2999328 in the standardized alternative probability distribution (see B' in Figure 2).

Finally, compute the statistical power, $1-\beta = P(t \geq t_a)$. The power is .89972532 (=1-.10027466). See the shaded areas of B and B' in Figure 2. This high statistical power indicates that this T-test is highly likely to detect the effect or reject the null hypothesis that the population mean is 20.

5. Relationship among the Components

Now, let us talk about how the components mentions in section 3 are related to each other. First, a model (or a test) dictates the formula for standardized effect sizes (test statistics). For example, T-test computes a standardized T score as $t = \frac{\bar{x}-\mu}{s}\sqrt{n}$, where $\frac{s}{\sqrt{n}}$ is an estimated population standard deviation (variation).

How do standardized effect sizes affect the statistical power? As a standardized effect size increases, the power increases (positive relationship). Imagine an alternative probability distribution with mean 23. In order word, shift the alternative probability distribution B to the right by one unit, holding original probability distribution constant. Effect size becomes larger (3=23-20). It is obvious that ß becomes smaller and the statistical power, in turn, increases. By the same token, the power becomes smaller if we

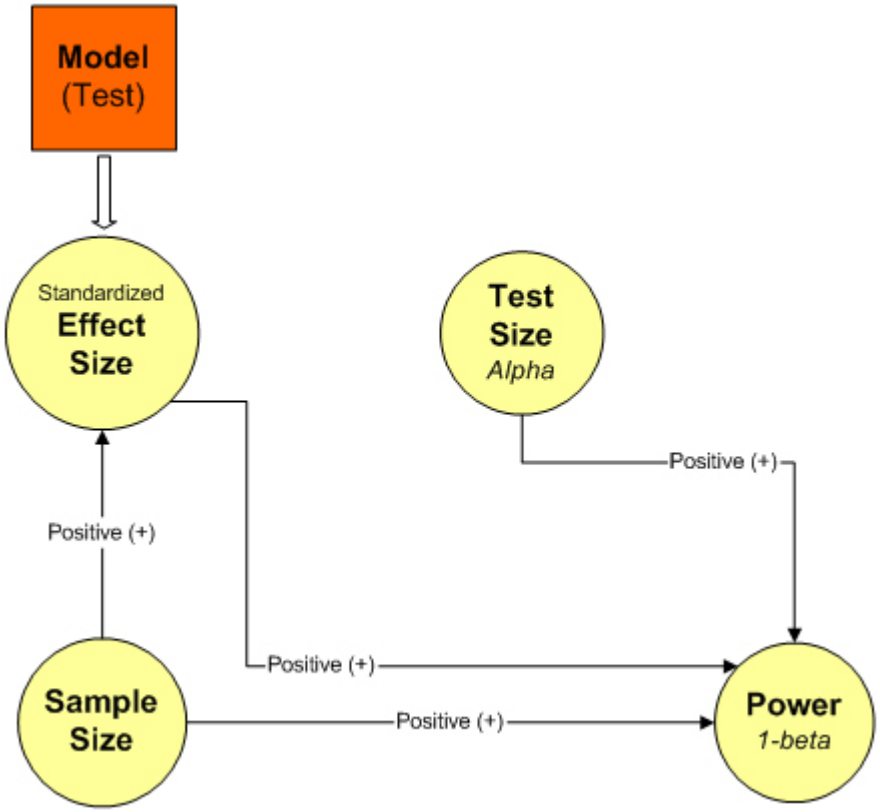
[6] $2.0167 = \frac{?-20}{4}\sqrt{44}$

shift the alternative distribution to the left by one to have mean 21. If the standardized effect size is small, it is difficult to detect effects even when they exist; it is less powerful.

Now, how does the size of a test (significance level) affect the statistical power? If a researcher has a lenient significance level like the .10 rather than the .05 or .01, the statistical power of the test becomes larger (positive relationship).

Imagine a line connecting 2.0167 in A', 21.2161 in A and B, and -1.2999 in B'. And then shift the line to the left, leaving all probability distributions untouched. What will happen? The test size (alpha=significance level) increases; critical values in A and A' are shifted to the left, increasing rejection areas; ß decreases; and finally statistical power increases.[7] Conversely, if we have a stringent significance level, the power of the test decreases; we are moving the line to the right!

Figure 3. Relationships among the Components



How about the sample size? A larger sample size generally leads to a parameter estimate with smaller variances, a larger standardized effect size, eventually, a greater ability to detect a significant difference (positive relationship). Look at the T statistic formula.

In general, the most important component affecting the statistical power is the sample size. In fact, there is a little room to change a test size (significance level). It is also

[7] There is a trade-off between the Type I error (alpha) and Type II error (beta). Moving the line to the left increases the Type I error, reducing the Type II error.

difficult to control effect sizes in many cases. It is costly and time-consuming to get more observations, of course. But the frequently asked question in practice is how many observations need to be collected.

However, if too many observations are used (or if a test is too powerful with a large sample size), even a trivial effect will be mistakenly detected as a significant one. Thus, virtually anything can be proved regardless of actual effects. (High 2000). By contrast, if too few observations are used, the hypothesis test will result in low statistical power.[8] There may be little chance to detect a meaningful effect even when it exists there. How do we know if the number of our observations is reasonable? Sample size analysis can answer.

6. Applications

We have conceptually discussed what the statistical power is. Now, let us analyze statistical power and sample size in several models using the SAS POWER procedure.

a. Power analysis of a one-sample T-test

Go back to the example mentioned in section four. Let me summarize the components of the statistical power analysis first. The goal is to compute statistical power of information about other components.

Table 2. Summary Information for a Statistical Power Analysis

Test	Sample Size	Test Size	Power	Effect Size	Variation
T-test	44	.05	?	2 (=22-20)	4

Let us use the SAS POWER procedure to conduct the same power analysis. You have to specify the type of a test first. The `ONESAMPLEMEANS` statement in the following example indicates the one-sample T-test.

```
PROC POWER;
        ONESAMPLEMEANS
        ALPHA=.05 SIDES=2
        NULLM=20 MEAN=22
        STDDEV=4
        NTOTAL=44
        POWER=.;
RUN;
```

The `ALPHA` and `SIDES` options say a two-tailed test at the .05 significance level (test size). The `NULLM` option specifies the mean value of the null hypothesis (H_0), while `MEAN` option specifies hypothesized mean value (H_a). The `STDDEV` and the `NTOTAL` options respectively

[8] Note that there is no clear cut-point of "too many" and "too few", since it depends on models and specifications. For instance, if a model has many parameters to be estimated, or if a model uses the maximum likelihood estimation method, the model needs more observations than otherwise.

indicate the standard deviation (variation) and sample size (N). Finally, the POWER option ending with a period (.) asks SAS to compute the statistical power of this test. Take a look at the SAS output.

The POWER Procedure
One-sample t Test for Mean

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Number of Sides	2
Null Mean	20
Alpha	0.05
Mean	22
Standard Deviation	4
Total Sample Size	44

Computed Power

Power
0.900

SAS summarizes the information about core components, and then returns the value of statistical power of the test.

b. Sample size analysis of a one-sample T-test

Let us change the scenario for a sample size analysis. Suppose we realize that some observations have unreliable information due to measurement errors, and that the population standard deviation is 3, not 4. Boss's guideline requires the .01 significance level in the upper one-tailed test and a lower power level .8. Now, we want to know the minimum sample size that can satisfy these conditions.

Table 3. Summary Information for a Sample Size Analysis

Test	Sample Size	Test Size	Power	Effect Size	Variation
T-test	?	.01	.800	2 (=22-20)	3

Look at the following SAS POWER procedure and its output. Note that the U of the SIDES option represents the upper one-tailed test with alternative value greater than the null value (L means a lower one-sided test). The test size (significance level) and standard deviation are corrected as requested. The NTOTAL option has a period (.), while the POWER option specifies the target level of statistical power.

```
PROC POWER;
        ONESAMPLEMEANS
        ALPHA=.01 SIDES=U
        NULLM=20 MEAN=22
        STDDEV=3
        NTOTAL=.
```

```
        POWER=.8;
RUN;
```

The following output says that only 26 observations are needed to reach the targeting power.

```
                    The POWER Procedure
                 One-sample t Test for Mean

                  Fixed Scenario Elements

Distribution                        Normal
Method                               Exact
Number of Sides                          U
Null Mean                               20
Alpha                                 0.01
Mean                                    22
Standard Deviation                       3
Nominal Power                          0.8

                  Computed N Total

                  Actual        N
                   Power    Total

                   0.812       26
```

c. Power analysis of a paired-sample T-test

Let us move on to the paired sample T-test. The PAIREDMEANS statement and the TEST=DIFF option are needed. The CORR option is also required to specify the correlation coefficient of the two paired variables, while the NPAIRS option specifies the number of pairs. You may list hypothesized numbers of pairs to see how statistical powers are sensitive to the number of pairs.

```
PROC POWER;
        PAIREDMEANS TEST=DIFF
        ALPHA=.01 SIDES=2
        MEANDIFF=3
        STDDEV=3.5
        CORR= .2
        NPAIRS=20 30 40
        POWER=.;
RUN;
```

Look at the following SAS output that produces three statistical powers according to the hypothesized sample sizes.

```
                    The POWER Procedure
              Paired t Test for Mean Difference

                  Fixed Scenario Elements

Distribution                        Normal
```

Method	Exact
Number of Sides	2
Alpha	0.01
Mean Difference	3
Standard Deviation	3.5
Correlation	0.2
Null Difference	0

Computed Power

Index	N Pairs	Power
1	20	0.575
2	30	0.821
3	40	0.936

d. Power analysis of a two independent samples T-test

The following example conducts a statistical power analysis for the two independent samples T-test. The `TWOSAMPLEMEANS` statement indicates the two independent samples T-test. The `GROUPMEANS` option species the means of two groups. The `PLOT` statement with `X=N` option draws a plot of statistical powers as N in the X-axis changes.

```
PROC POWER;
        TWOSAMPLEMEANS
        ALPHA=.01 SIDES=2
        GROUPMEANS= (2.79 4.12)
        STDDEV= 1.5 2.0 2.5
        NTOTAL=44
        POWER=.;
        PLOT X=N MIN=20 MAX=100 KEY=BYCURVE(NUMBERS=OFF POS=INSET);
RUN;
```

You may list more than one hypothesized standard deviation in the `STDDEV` option in order to know how statistical power of the test is sensitive to the standard deviations. The output is as follows.

The POWER Procedure
Two-sample t Test for Mean Difference

Fixed Scenario Elements

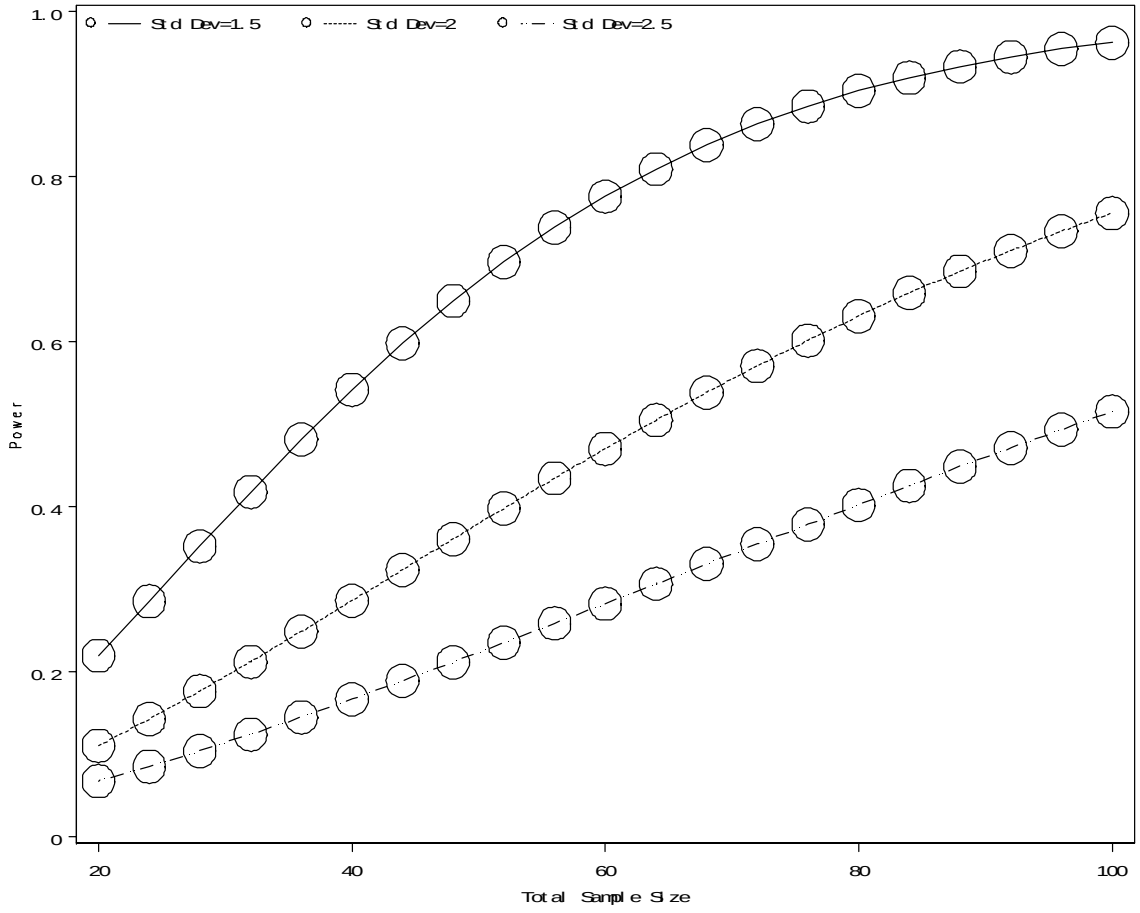
Distribution	Normal
Method	Exact
Number of Sides	2
Alpha	0.01
Group 1 Mean	2.79
Group 2 Mean	4.12
Total Sample Size	44
Null Difference	0
Group 1 Weight	1
Group 2 Weight	1

Computed Power

Index	Std Dev	Power
1	1.5	0.598
2	2.0	0.324
3	2.5	0.189

A plot of statistical powers and sample sizes visualizes the sensitivity of a test. The following plot shows that the power for standard deviation 1.5 is more sensitive to sample size than those for larger standard deviations when N is less than 60.

Figure 4. A Plot of Statistical Power and Sample Size



e. Power analysis of a one-way ANOVA

Now, consider a one-way ANOVA. The POWER procedure has the `ONEWAYANOVA` statement for its analysis. Note that the `NPERGROUP` option specifies the number of observations of a group, assuming balanced data.[9]

[9] When each group has the same number of observations, we call them balanced data; otherwise, unbalanced data.

```
PROC POWER;
        ONEWAYANOVA TEST=OVERALL
        ALPHA=.05
        GROUPMEANS= (5 7 3 11)
        STDDEV= 4 5 6
        NPERGROUP= 10 15
        POWER=.;
RUN;
```

The above POWER procedure analyzes ANOVA with four different groups. Listing several standard deviations and numbers of observations conducts sensitivity analysis of the statistical power, producing the six combinations (=3 X 2).

```
                    The POWER Procedure
              Overall F Test for One-Way ANOVA

                 Fixed Scenario Elements

              Method                 Exact
              Alpha                   0.05
              Group Means         5 7 3 11

                    Computed Power

                      Std    N Per
              Index    Dev    Group    Power

                  1      4       10    0.972
                  2      4       15    0.999
                  3      5       10    0.858
                  4      5       15    0.972
                  5      6       10    0.696
                  6      6       15    0.886
```

The SAS GLMPOWER procedure also conducts a power analysis for one-way ANOVA. Unlike the POWER procedure, this procedure requires an existing SAS data set. Like the ANOVA procedure, the CLASS and the MODEL statements are required in this GLMPOWER procedure. Note that the POWER is not an option, but a statement in this procedure.

```
PROC GLMPOWER DATA=power.car;
        CLASS mode;
        MODEL credits=mode;
        POWER STDDEV=10
              ALPHA=.01
              NTOTAL=1000
              POWER=.;
RUN;
```

The output of the GLMPOWER procedure is similar to that of the POWER procedure.

```
                    The GLMPOWER Procedure

                 Fixed Scenario Elements
```

```
                    Dependent Variable              credits
                    Source                             mode
                    Alpha                              0.01
                    Error Standard Deviation             10
                    Total Sample Size                  1000
                    Test Degrees of Freedom               3
                    Error Degrees of Freedom            996

                              Computed Power

                                  Power

                                  0.856
```

f. Power analysis of a two-way ANOVA

In order to conduct power analysis of the two-way ANOVA in SAS, we need to use the `GLMPOWER` procedure because the `POWER` procedure does not support this model. Again, note that the `CLASS` and `MODEL` statements are requited.

```
PROC GLMPOWER DATA=power.car;
        CLASS owncar sex;
        MODEL credits= owncar sex;
        POWER STDDEV=10
          ALPHA=.01
          NTOTAL= .
          POWER=.8;
RUN;
```

Here is the output of the `GLMPOWER` procedure.

```
                          The GLMPOWER Procedure

                         Fixed Scenario Elements

                    Dependent Variable              credits
                    Alpha                              0.01
                    Error Standard Deviation             10
                    Nominal Power                       0.8

                              Computed N Total

                              Test                Actual
          Index    Source       DF    Error DF     Power     N Total

              1    owncar        1     1854997     0.800     1855000
              2    sex           1       31997     0.804       32000
```

7. Software Issues

There are various software packages for statistical power and sample size analyses. Among them are the POWER and GLMPOWER of SAS/STAT, SPSS SamplePower 2.0,

and G*Power.[10] See the following for the list and review of statistical power analysis software.

http://www.zoology.ubc.ca/~krebs/power.html

These software packages vary in scope, accuracy, flexibility, and interface (Thomas and Krebs 1997). Some packages may support a test, while others may not. They may be general purpose statistical software with such functions embedded (e.g., SAS) or professional software with specialty in statistical power and/or sample size analysis (e.g., G*Power). Some software package like SAS/STAT Power and Sample Size (PSS) is not stand-alone, but Web-based. Following Web site has a statistical power calculator than enables you to compute power online.

http://calculators.stat.ucla.edu/powercalc/

Some software like G*Power runs under DOS mode. The majority of existing power analysis software works in Microsoft Windows. Some are written for UNIX machine. Most software uses the point-and-click interface, while others depend on command line. They may adopt different algorithms so that their results may be different.

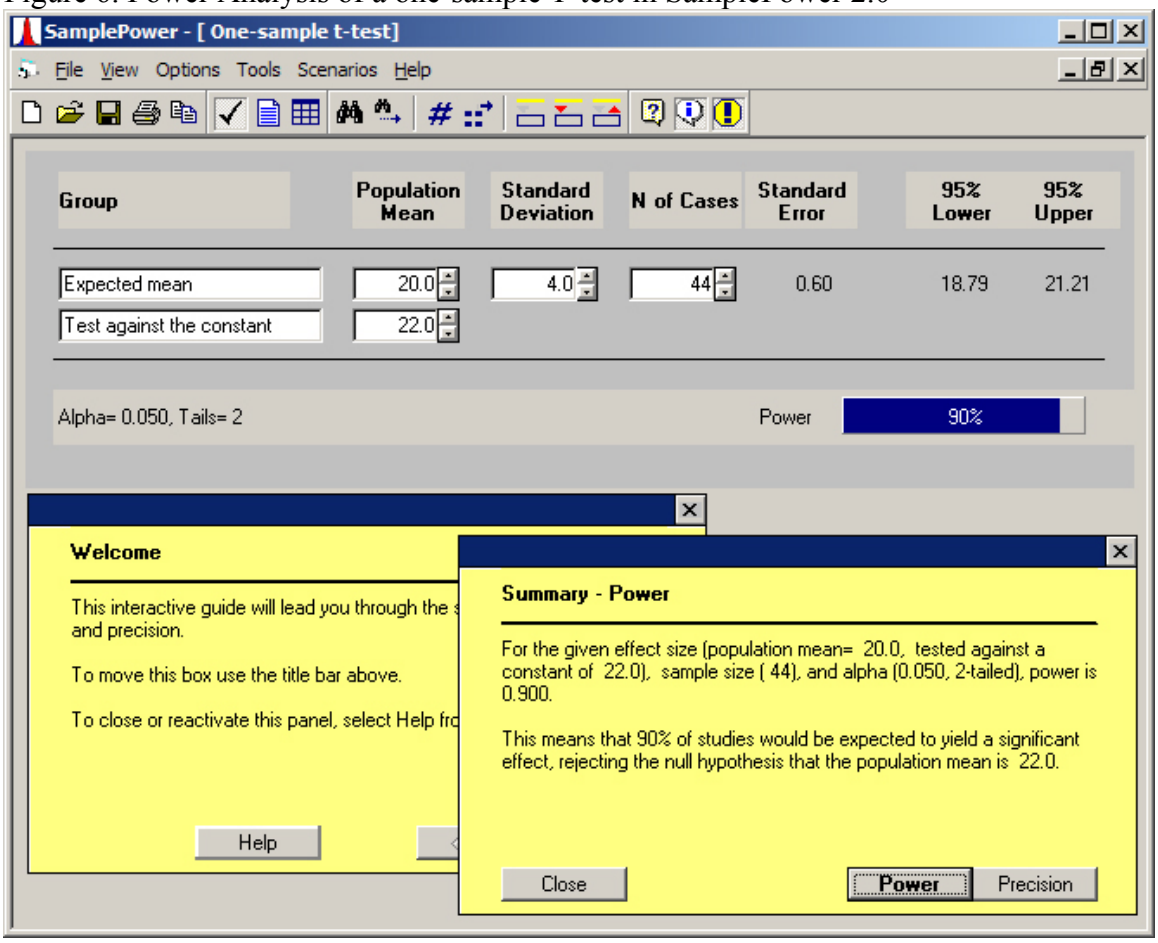
Figure 5 illustrates how to conduct the power analysis of a one-way ANOVA (Section 6.e) using G*Power. The result .9723 corresponds to the first case that has standard deviation 4 and 10 observations in each group.

Figure 5. Power Analysis of a one-way ANOVA in G*Power

Figure 6 is a screenshot of SamplePower 2.0, replicating the power analysis of a one-sample T-test (Section 6.a). The summary pop-up window wraps up the power analysis.

[10] G*Power is available on http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

Figure 6. Power Analysis of a one-sample T-test in SamplePower 2.0



References

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. Hillsdale, NJ: L. Erlbaum Associates.

Greene, William H. 2000. *Econometric Analysis,* 4th ed. Prentice Hall.

High, Robin. 2000. http://cc.uoregon.edu/cnews/summer2000/statpower.html

Kirk, Roger E. 1995. *Experimental Design: Procedures for the Behavioral Science,* 3rd ed. Pacific Grove, CA: Brooks/Cole Publishing.

Murphy, Kevin R. 1998. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*. Mahwah, NJ: L. Erlbaum Associates.

SAS Institute. 2004. *Getting Started with the SAS Power and Sample Size Application*. Cary, NC: SAS Institute.

SAS Institute. 2004. *SAS/STAT User's Guide, Version 9.1*. Cary, NC: SAS Institute.

Thomas, Len, and Charlies J. Krebs. 1997. "A Review of Statistical Power Analysis Software," *Bulletin of the Ecological Society of America*, 78 (2): 126-139. Available at http://www.zoology.ubc.ca/~krebs/power.html