

I n d i a n a U n i v e r s i t y
University Information Technology Services

Hypothesis Testing and Statistical Power of a Test*

Hun Myoung Park, Ph.D.
statmath@indiana.edu

© 2004-2010
Last modified on June 2010

University Information Technology Services
Center for Statistical and Mathematical Computing
Indiana University
410 North Park Avenue Bloomington, IN 47408
(812) 855-4724 (317) 278-4740
<http://www.indiana.edu/~statmath>

* The citation of this document should read: "Park, Hun Myoung. 2008. *Hypothesis Testing and Statistical Power of a Test*. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University." <http://www.indiana.edu/~statmath/stat/all/power/index.html>

*This document provides fundamental ideas of statistical power of a test and illustrates how to conduct statistical power analysis using SAS/STAT and G*Power 3.0.*

1. Hypothesis Testing
2. Type II Error and Statistical Power of a Test
3. Components of Statistical Power Analysis
4. An Example of Statistical Power
5. Software Issues
6. Applications: Comparing Means and Proportions
7. Applications: ANOVA and Linear Regression
8. Conclusion

How powerful is my study (test)? How many observations do I need to have for what I want to get from the study? You may want to know statistical power of a test to detect a meaningful effect, given sample size, test size (significance level), and standardized effect size. You may also want to determine the minimum 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 research resources. This document summarizes basics of hypothesis testing and statistic power analysis, and then illustrates how to do using SAS 9, Stata 10, G*Power 3.

1. Hypothesis Testing

Let us begin with discussion of hypothesis and hypothesis testing.

1.1 What Is a Hypothesis?

A hypothesis is a specific conjecture (statement) about a property of a population of interest. A hypothesis should be interesting to audience and deserve testing. A frivolous or dull example is “the water I purchased at Kroger is made up of hydrogen and oxygen.” A null hypothesis is a specific baseline statement to be tested and it usually takes such forms as “no effect” or “no difference.”¹ An alternative (research) hypothesis is denial of the null hypothesis. Researchers often, but not always, expect that evidence supports the alternative hypothesis. 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 A two-tailed test considers both extremes (left and right) of a probability distribution.

A good 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

¹ Because it is easy to calculate test statistics (standardized effect sizes) and interpret the test results (Murphy 1998).

² μ (Mu) represents a population mean, while \bar{x} denotes a sample mean.

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

1.2 Hypothesis Testing

Hypothesis testing is a scientific process to examine if a hypothesis is plausible or not. In general, hypothesis testing follows next five steps.

- 1) State a null and alternative *hypothesis* clearly (one-tailed or two-tailed test)
- 2) Determine a *test size (significance level)*. Pay attention to whether a test is one-tailed or two-tailed to get the right critical value and rejection region.
- 3) Compute a *test statistic* and *p-value* or construct the confidence interval, depending on testing approach.
- 4) Decision-making: reject or do not reject the null hypothesis by comparing the subjective criterion in 2) and the objective test statistic or p-value calculated in 3)
- 5) Draw a conclusion and interpret substantively.

There are three approaches of hypothesis testing (Table 1). Each approach requires different subjective criteria and objective statistics but ends up with the same conclusion.

Table 1. Three Approaches of Hypothesis Testing

Step	Test Statistic Approach	P-Value Approach	Confidence Interval Approach
1	State H_0 and H_a	State H_0 and H_a	State H_0 and H_a
2	Determine test size α and find the critical value	Determine test size α	Determine test size α or $1 - \alpha$, and a hypothesized value
3	Compute a test statistic	Compute a test statistic and its p-value	Construct the $(1 - \alpha)100\%$ confidence interval
4	Reject H_0 if $TS > CV$	Reject H_0 if $p\text{-value} < \alpha$	Reject H_0 if a hypothesized value does not exist in CI
5	Substantive interpretation	Substantive interpretation	Substantive interpretation

* TS (test statistic), CV (critical value), and CI (confidence interval)

The classical *test statistic approach* computes a test statistic from empirical data and then compares it with a critical value. If the test statistic is larger than the critical value or if the test statistic falls into the rejection region, the null hypothesis is rejected. In the *p-value approach*, researchers compute the p-value on the basis of a test statistic and then compare it with the significance level (test size). If the p-value is smaller than the significance level, researchers reject the null hypothesis. A *p-value* is considered as amount of risk that researchers have to take when rejecting the null hypothesis. Finally, the *confidence interval approach* constructs the confidence interval and examines if a hypothesized value falls into the interval. The null hypothesis is rejected if the hypothesized value does not exist within the confidence interval.

1.3 Type I Error and Test Size (Significance Level)

The *size of a test*, often called *significance level*, is the probability of committing a Type I error. A Type I error occurs when a null hypothesis is rejected when it is true (Table 2). This test size

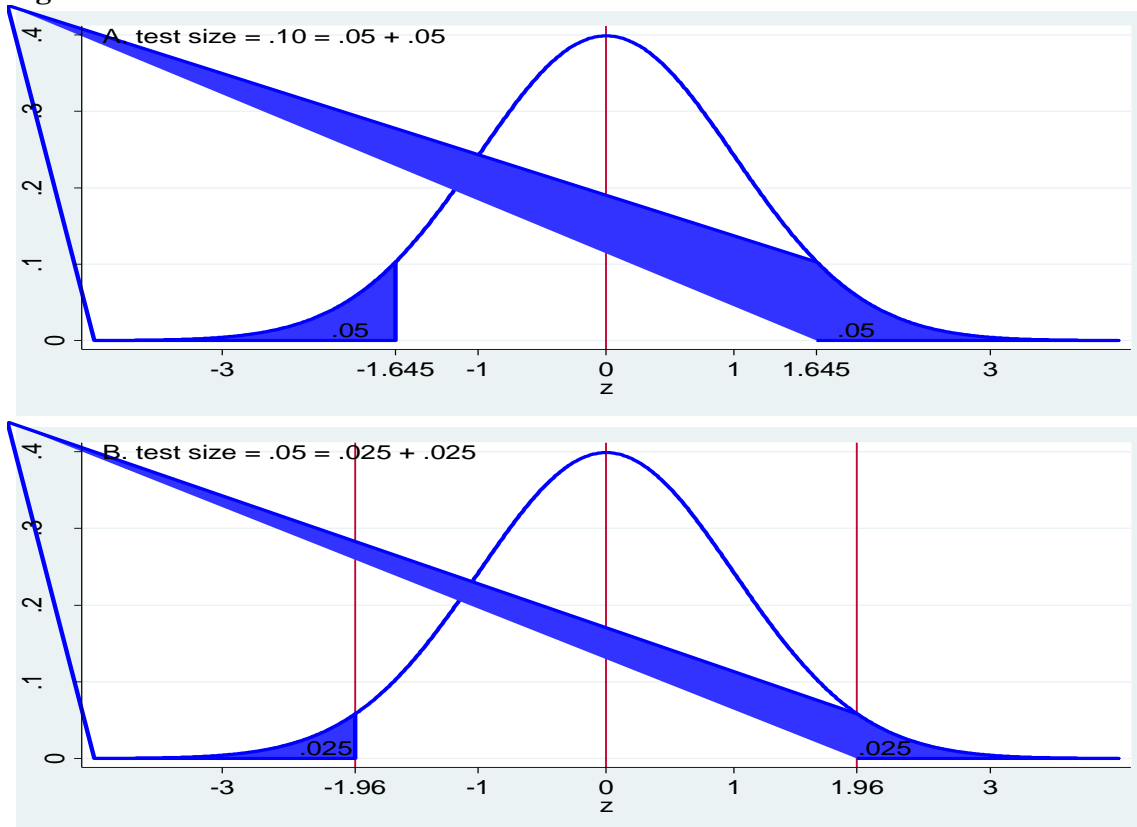
³ This behavior, often called “data fishing,” just hunts a model that best fits the sample, not the population.

is denoted by α (*alpha*). The $1 - \alpha$ is called the *confidence level*, which is used in the form of the $(1 - \alpha) \cdot 100$ percent confidence interval of a parameter.

Table 2. Size and Statistical Power of a Test

	<i>Do not reject H_0</i>	<i>Reject H_0</i>
H_0 is true	Correct Decision $1 - \alpha$: Confidence level	Type I Error α : Size of a test (Significance level)
H_0 is false	Type II Error β	Correct Decision $1 - \beta$: Power of a test

Figure 1. Test Size and Critical Value in the Standard Normal Distribution



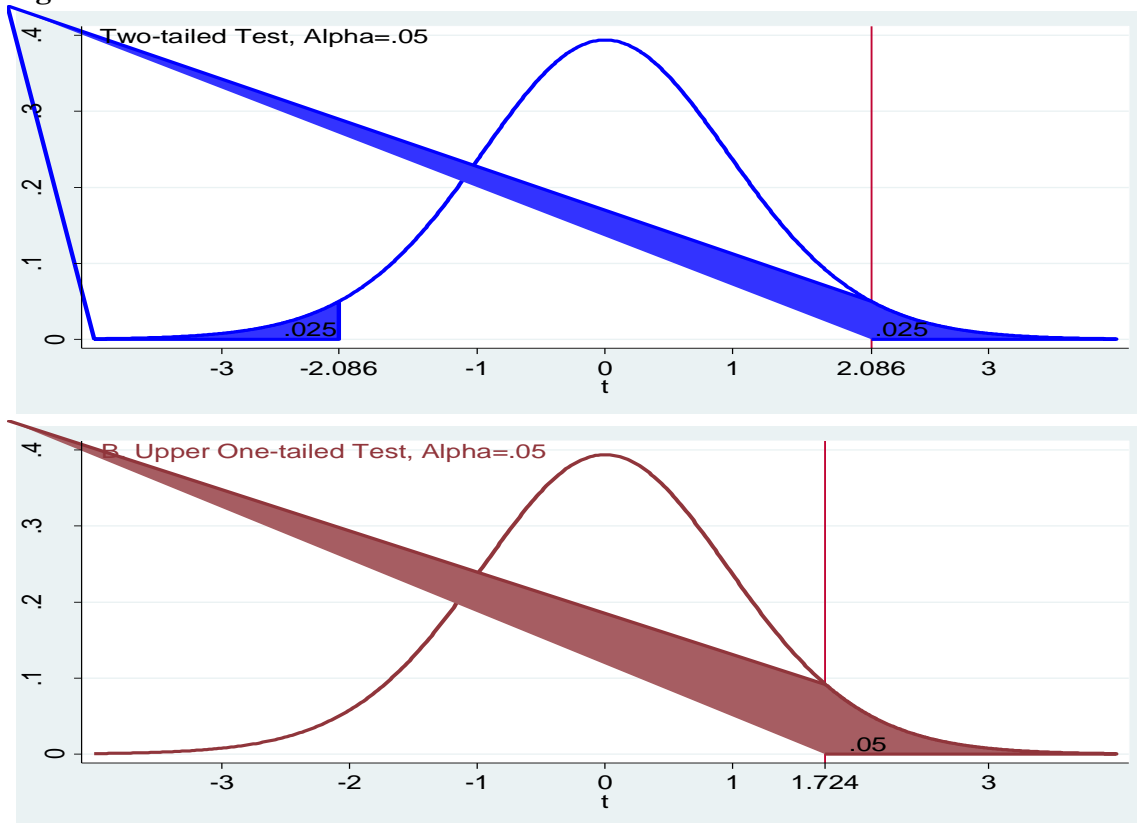
In a two-tailed test, the test size (significance level) is the sum of two symmetric areas of both tails of a probability distribution. See the shaded areas of two standard normal distributions in Figure 1. These areas, surrounded by the probability distribution curve, x-axis, and a particular value (critical value), are called *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 and .10, .05, and .01 levels are conventionally used.

Think about a test size of .10 (see A in Figure 1). .10 is the sum of two rejection regions. The area underneath the probability distribution from a particular value, say 1.645, to the positive infinity is .05. So is the area from -1.645 to the negative infinity. Such a particular value is called the *critical value* of the significance level. In B of Figure 1, ± 1.96 are the critical values that form two rejection areas of .05 in the standard normal probability distribution. The critical

values for the .01 level of the two-tailed test are ± 2.58 . As test size (significance level) decreases, critical values shift to the extremes; the rejection areas become smaller; and thus, it is less likely to reject the null hypothesis.

Critical values depend on probability distributions (with degrees of freedom) and test types (one-tailed versus two-tailed) as well. For example, the critical values for a two-tailed t-test with 20 degrees of freedom are ± 2.086 at the .05 significance level (Figure 2). When considering the positive side only (one-tailed test), the critical value for the .05 significance level is 1.724. Note that the area underneath the t-distribution from 1.724 to 2.086 is .025; accordingly, the area from 1.724 to the positive infinity is $.05 = .025 + .025$.

Figure 2. Test Size and Critical Value in the T Distribution



How do we substantively understand the test size or significance level? Test size 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 .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 an effect size (deviation from the baseline) is large; instead, the conclusion is more convincing (less risky).

For example, a test statistic of 1.90 in A of Figure 2 is not statistically discernable from zero at the .05 significance level since 7.19 percent (p-value), areas from ± 1.90 to both extremes, is riskier than our significance level criterion (5%) for the two-tailed test; thus, we do not reject the null hypothesis. However, the test statistic 1.90 is considered exceptional at the .10 significance level, because we are willing to take a risk up to 10 percent but the 1.90 says only

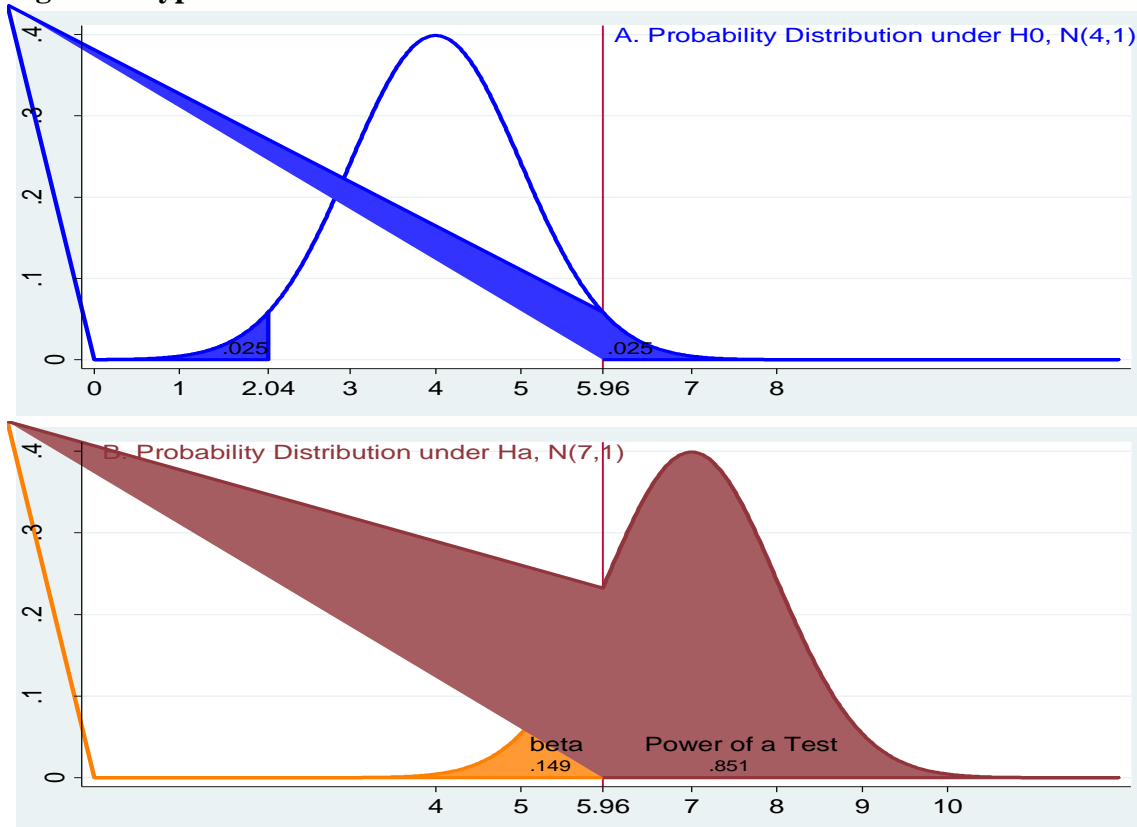
about 7 percent of being wrong. Empirical data indicate that it is less risky (about 7%) than we thought (10%) to reject the null hypothesis.

When performing a one-tailed test at the .05 level (B of Figure 2), 1.90 is larger than the critical value of 1.724 or 1.90 falls within the rejection region from 1.724 to the positive infinity. The p-value, the area from 1.90 to the positive infinity, is .036, which is smaller (not riskier) than the test size (significance level) .05. Therefore, we can reject the null hypothesis at the .05 level.

2. Type II Error and Statistical Power of a Test

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). A prospective analysis is often used to determine a required sample size to achieve target statistical power, while a retrospective analysis computes the statistical power of a test given sample size and effect size.⁴

Figure 3. Type II Error and Statistical Power of a Test



The probability of Type I errors is represented by two blue areas in Figure 3 (A). Note that 5.96 is the critical value at the .05 significance level. Suppose the true probability distribution is the normal probability distribution with mean 7 and variance 1, $N(7,1)$, but we mistakenly believe that $N(4,1)$ is the correct distribution. Look at the left-hand side portion of the probability distribution B from the negative infinity to 5.96. What does this portion mean? When we get a test statistic smaller than 5.96, we mistakenly do not reject the null hypothesis because we believe probability distribution A is correct.⁵ Therefore, the portion in orange is the probability

⁴ http://en.wikipedia.org/wiki/Statistical_power

⁵ For example, if we get a test statistic of 5, we will not reject the null hypothesis; 5 is not sufficiently far away from the mean 4. However, this is not a correct decision because the true probability distribution is $N(7,1)$ and the

of failing to reject the null hypothesis, $H_0 : \mu = 4$, when it is false. This probability of committing *Type II error* is denoted by β (*beta*)

The right-hand side portion of the probability distribution B is the area underneath the probability distribution B from 5.96 (a critical value in probability distribution A but a particular value in B) to the positive infinity. This is $1 - \beta$, which is called the statistical power of a test. As mentioned before, this portion means the probability of rejecting the false null hypothesis $H_0 : \mu = 4$ or detecting an effect that actually exists there. This interpretation is indeed appealing to many researchers, but it is neither always conceptually clear nor easy to put it into practice. Conventionally a test with a power greater than .8 (or $\beta \leq .2$) is considered statistically powerful (Mazen, Hemmasi, Lewis 1985).

null hypothesis should have been rejected; the test statistic of 5 is sufficiently far away from the mean 7 at the 05 level.

3. Components of Statistical Power Analysis

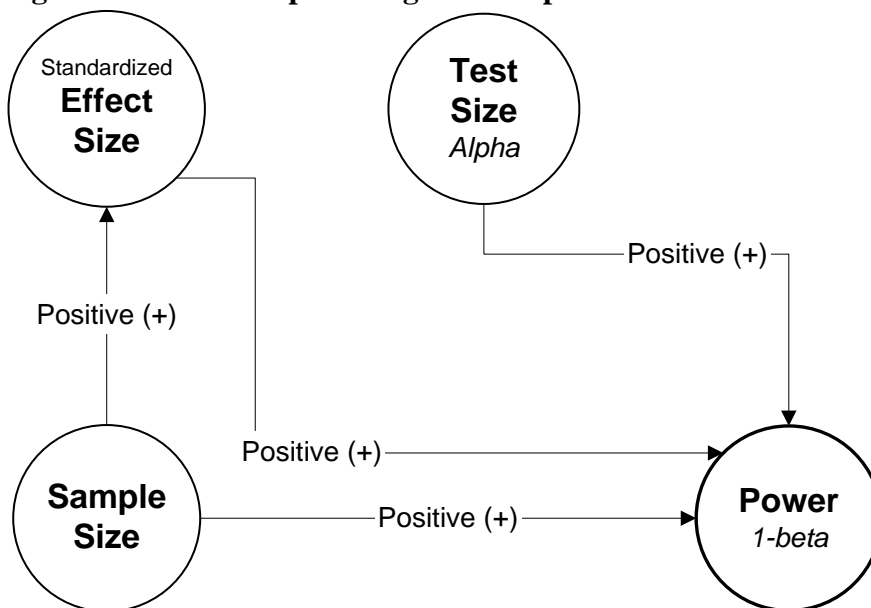
Statistical power analysis explores relationships among following four components (Mazen, Hemmasi, Lewis 1985). Keep in mind that a specific model (test) determines a probability distribution for data generation process and formulas to compute test statistics. For example, a t-test uses the T distribution, while ANOVA computes F scores.

1. Standardized effect size: (1) effect size and (2) variation (variability)
2. Sample size (N)
3. Test size (significance level α)
4. Power of the test ($1 - \beta$)

A *standardized effect size*, a test statistic (e.g., T and F scores) is computed by combining (unstandardized) effect and variation.⁶ In a t-test, for example, the standardized effect is effect (deviation of a sample mean from a hypothesized mean) divided by standard error, $t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$.

An effect size in actual units of responses is the “degree to which the phenomenon exists” (Cohen 1988). Variation (variability) is the standard deviation or standard error of population. Cohen (1988) calls it the reliability of sample results. This variation often comes from previous research or pilot studies; otherwise, it needs to be estimated.

Figure 4. Relationships among the Components



Sample size (N) is the number of observations (cases) in a sample. When each group has the same number of observations, we call them balanced data; otherwise, they are unbalanced data. As discussed in the previous section, the *test size* or *significance level* (α) is the probability of

⁶ In statistics, the terminology is effect size like Cohen’s *d* without taking sample size into account. I intentionally added “standardized” to illustrate clearly how N affects standardized effect size.

rejecting the null hypothesis that is true. The *power of the test* ($1 - \beta$) is the probability of correctly rejecting a false null hypothesis.

How are these components related each other? Figure 4 summarizes their relationships. First, as a standardized effect size increases, statistical power increases (positive relationship), holding other components constant. A large standardized effect means that a sample statistic is substantially different from a hypothesized value; therefore, it becomes more likely to detect the effect. Conversely, if the standardized effect size is small, it is difficult to detect effects even when they exist; this test is less powerful.

How does test size (significance level) affect statistical power? There is a trade-off between Type I error (α) and Type II error (β). Shift the vertical line in Figure 5 to the left and then imagine what will happen in statistical power. Moving the line to the left increases the Type I error and simultaneously reduces the Type II error. If a researcher has a lenient significance level like .10, the statistical power of a test becomes larger (positive relationship). The critical values in A and A' are shifted to the left, increasing rejection regions; β decreases; and finally statistical power increases. Conversely, if we have a stringent significance level like .01, the Type II error will increase and the power of a test will decrease. We are moving the line to the right!

How does sample size N affect other components? When sample size is larger, variation (standard error) becomes smaller and thus makes standardized effect size larger. In t-test, for instance, standard error is sample standard deviation divided the square root of N, $s_{\bar{x}} = \frac{s_x}{\sqrt{N}}$. A larger standardized effect size increases statistical power.

In general, the most important component affecting statistical power is sample size in the sense that the most frequently asked question in practice is how many observations need to be collected. In fact, there is a little room to change a test size (significance level) since conventional .05 or .01 levels are widely used. It is difficult to control effect sizes in many cases. It is costly and time-consuming to get more observations, of course.

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, a hypothesis test will be weak and less convincing.⁷ Accordingly, there may be little chance to detect a meaningful effect even when it exists there.

Statistical power analysis often answers these questions. What is the statistical power of a test, given N, effect size, and test size? At least how many observations are needed for a test, given effect size, test size, and statistical power? How do we answer these questions?

⁷ However, 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.

4. An Example of Statistical Power

Imagine a random variable that is normally distributed with mean 4 and variance 1 (see A in Figure 5). The positive critical value of a two-tailed test at the .05 significance level is 5.96 since $1.96 = (5.96-4)/1$ is the corresponding critical value in the standard normal distribution (see A' in Figure 5). The null hypothesis of z test is the population mean is a hypothesized value 4, $H_0 : \mu = 4$. The rejection regions of this test are the areas from the negative infinity to 2.04 and from 5.96 to the positive infinity in A of Figure 5). The standardized z score is computed as follows.

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{5.96 - 4}{1} = 1.96$$

Now, suppose the true probability distribution is not $N(4,1)$, but $N(7,1)$, which is illustrated in B of Figure 4. In other word, we mistakenly believe that the random variable is normally distributed with a mean 4; accordingly, statistical inferences based on this wrong probability distribution are misleading.

Look at the alternative probability distribution and think about the area underneath the distribution curve from the negative infinity to a particular value 5.96 in B, which is the positive critical value on the wrong probability distribution in A). The area in fact means the probability that we do not mistakenly reject the null hypothesis, $H_0 : \mu = 4$ when the alternative hypothesis, $H_a : \mu = 7$, is true, since we did not realize the true probability distribution, $N(7,1)$. This is a Type II error.

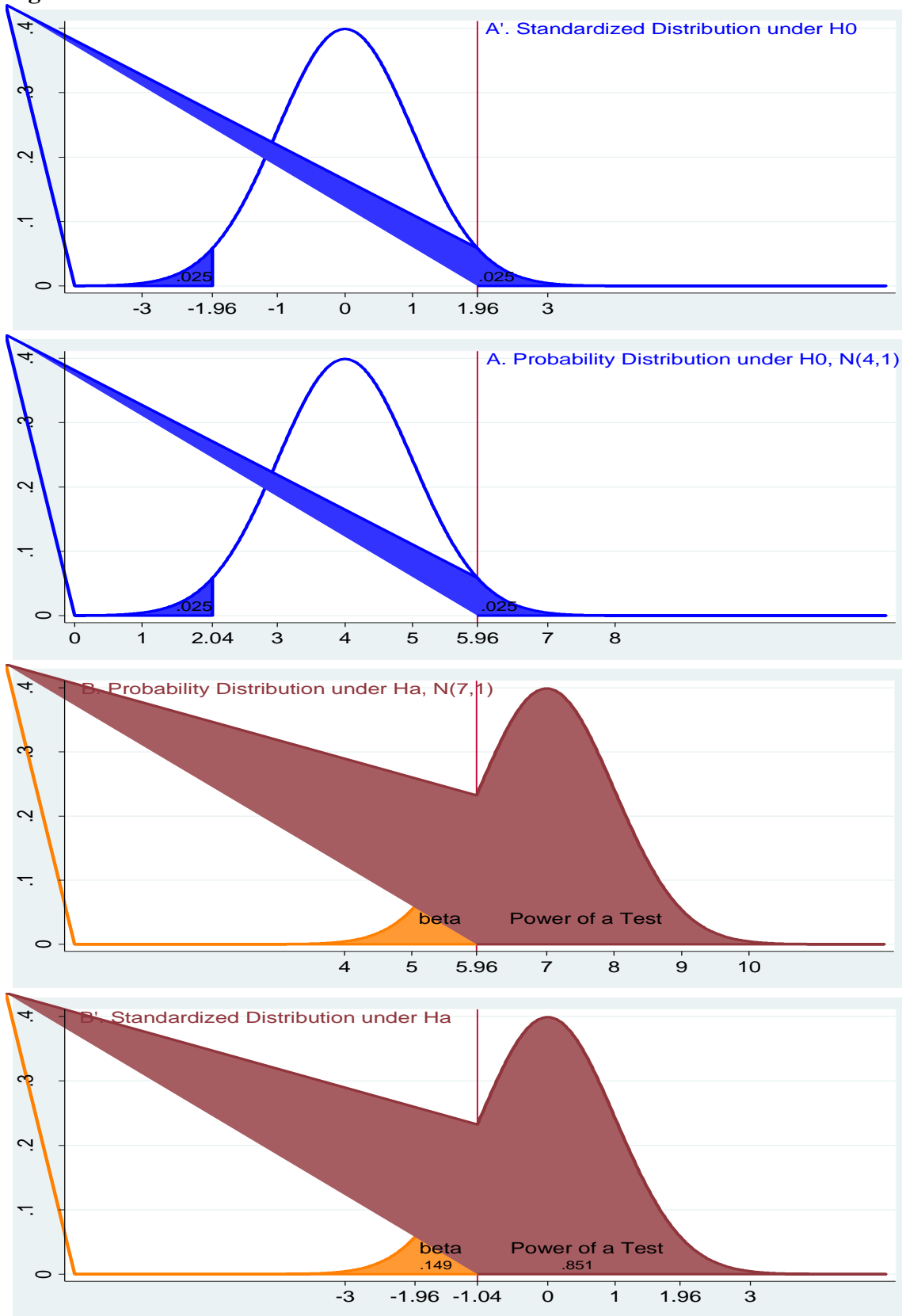
Suppose we get a sample mean 4.5. The corresponding z score is $.5 = (4.5-4)/1$, which is smaller than the critical value 1.96. Therefore, we do not reject the null hypothesis at the .05 significance level. If the alternative probability distribution is $N(7,1)$, the absolute value of the z score is $2.5 = |(4.5-7)/1|$, which is much larger than 1.96. That is, 4.5 is substantially far away from the mean 7 so that we undoubtedly reject the hypothesis $H_0 : \mu = 7$ at the .05 level.

The question here is to what extent we will commit a Type II error when the true probability distribution, although not easy to know in reality, is $N(7,1)$. We need to convert 5.96 into a z score under the true probability distribution with mean 7 and unit variance.

$$z = \frac{\bar{x} - \mu}{\sigma} = \frac{5.96 - 7}{1} = -1.04$$

Then, compute the area from the negative infinity to -1.04 in the standard normal distribution. From the standard normal distribution table, we can get .149 as shown in B' of Figure 5. This probability of not to reject the false null hypothesis is β . If the true probability distribution is $N(7,1)$, we are running about 15 percent risk of committing a Type II error.

Figure 5. Statistical Power of a Z-test



Researchers may want to know the extent that a test can detect an effect, if any. So, the question here is, “How powerful is this z test?” As explained in the previous section, the statistical power of a test is $1-\beta$, 1 minus the probability of committing a Type II error. Here we get $.851=1-.149$. Put it differently, we can detect an effect, if any, on average 85 times per 100 trials. Since this statistical power is larger than .8, this z test is considered powerful.

Let me summarize steps to compute a statistical power of a test.

- 1) Find critical values in a probability distribution, give a significance level. In the example, 2.04 and 5.96 are the critical values at the .05 significance level (see A and A' in Figure 5).⁸
- 2) Imagine the true (alternative) probability distribution and standardize the critical value in step 1 on the alternative probability distribution (see B and B' in Figure 5). We got $-1.04 = (5.96 - 7)/1$.
- 3) Read the probability from the standard normal distribution table. This is β or the probability of committing Type II error (see the left-hand side of B'). $\beta = .149$.
- 4) Finally, compute the statistical power of the test, $1 - \beta$. The power is $.851 (=1-.149)$. If power is larger than .80, conclude that a test is highly likely to detect an effect if it actually exists.

⁸ $1.96 = \left| \frac{?-4}{1} \right|$. Therefore, we get 2.04 and 5.96 from 4 ± 1.96 .

5. Software Issues

There are various software packages to perform statistical power analyses. Some examples include POWER and GLMPOWER of SAS/STAT, Stata, R, SPSS SamplePower 2.0, and G*Power.⁹ These software packages vary in scope, accuracy, flexibility, and interface (Thomas and Krebs 1997). This document focuses on SAS, Stata, and G*Power.

SAS POWER and GLMPOWER procedures perform prospective power and sample size analysis in a flexible manner. They support t-test, tests of binomial proportions, ANOVA, bivariate and partial correlation, and linear regression model. GLMPOWER can handle more complex linear models but unlike POWER this procedure requires an actual SAS data set. SAS Power and Sample Size (PSS) is a Web-based application that requires Web server software and Microsoft Internet Explorer. POWER and GLMPOWER can conduct sensitivity analysis and present results in plots.

G*Power is a free software package and is downloadable from <http://www.pscho.uni-duesseldorf.de/aap/projects/gpower/>. The latest version 3 runs under Microsoft Windows XP/Vista and MacOS 10 (Tiger and Leopard).¹⁰ This software support various tests including t-test, z-test, ANOVA, and chi-square test. Its point-and-click method makes analysis easy.

Stata `.sampsi` command supports t-test, tests of equality of proportions, and repeated measurements. This command can calculate minimum sample size and statistical power using a single command.

⁹ SamplePower is an add-on package for SPSS. See 6.1 for an example.

¹⁰ G*Power 2 runs only under Microsoft DOS and MacOS version 7-9.

6. Applications: Comparing Means and Proportions

This section illustrates how to perform sample size and statistical power analysis for comparing means and proportions using SAS POWER and Stata `.sampsi`.

6.1 Statistical Power of a One-sample T-test

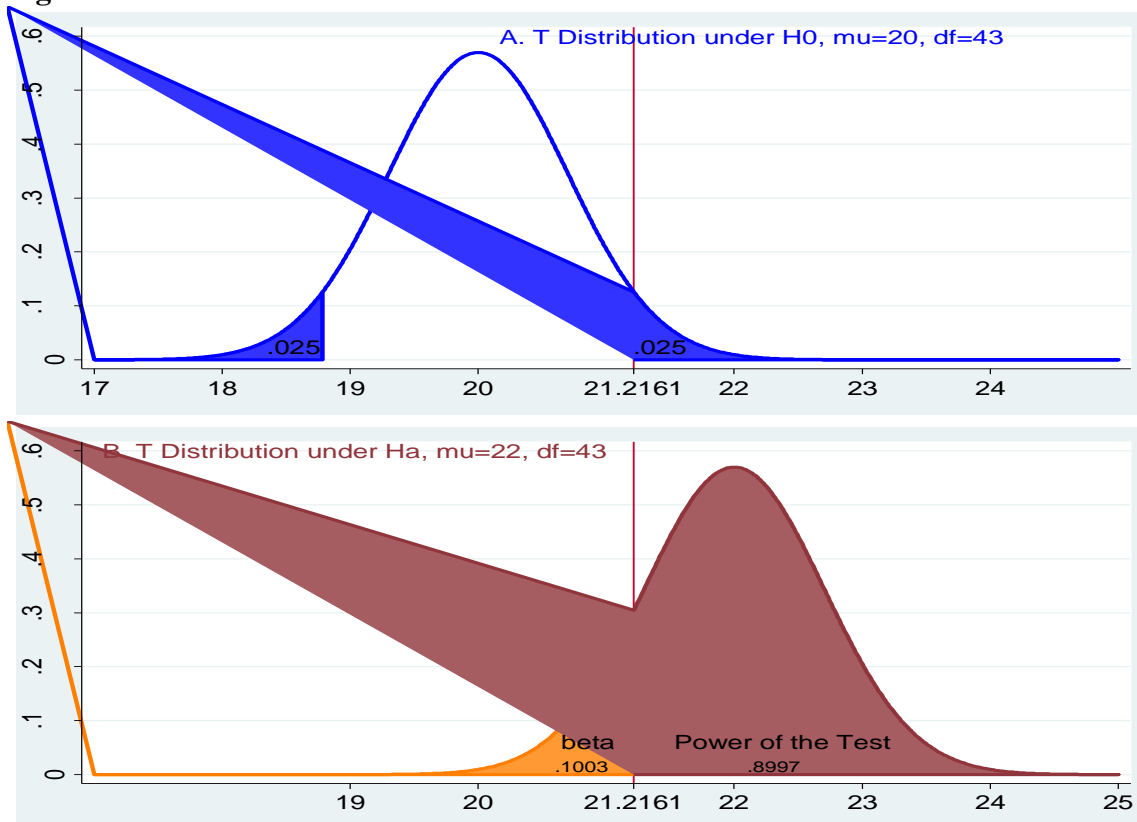
Suppose we conducted a survey for a variable and got mean 22 and standard deviation 4 from 44 observations. We hypothesize that the population mean is 20 and want to test at the .05 significance level. This is a one-sample t-test with $H_0 : \mu = 20$. The question here is “what is the statistical power of this test?” Let us summarize components of statistical power analysis first.

Table 3. Summary Information for a Statistical Power Analysis

Test	Effect Size	Variation (SD)	Sample Size	Test Size	Power
T-test	2 (=22-20)	4	44 (d.f.=43)	.05	?

From the t distribution table, we found that 2.0167 is the critical value at the .05 level when degrees of freedom is 43 (=44-1). The corresponding value is $21.2161 = 2.0167 * 4 / \sqrt{44} + 20$ since $2.0167 = (21.2161 - 20) / (4 / \sqrt{44})$. By definition, the area underneath the hypothesized probability distribution with mean 20 from 21.2161 to the positive infinity is .025, which is the right rejection region (see A in Figure 6).

Figure 6. Statistical Power in the T Distribution



On the claimed probability distribution with mean 22, the area from 21.2160 to the positive infinity is the statistical power of this test (see B in Figure 6). The area from the negative infinity to 21.2161 is defined as β or the probability of a Type II error. The standardized t score for 21.2161 on the claimed distribution is $-1.2999 = (21.2161 - 22) / (4 / \sqrt{44})$. β , $P(t < -1.2999)$ when $df=43$, is .1003 and statistical power is $.8997 = 1 - .1003$.

This computation is often tedious and nettlesome. The SAS POWER procedure makes it easy to compute the statistical power of this test. Consider the following syntax.

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

The ONESAMPLEMEANS statement below indicates this test is a one-sample t-test. The ALPHA and SIDES options say a two-tailed test at the .05 significance level (test size). You may omit these default values. NULLM=20 means the hypothesized mean in H_0 while MEAN=22 means the claimed mean of 22. The STDDEV and the NTOTAL options respectively indicate standard deviation (variation) and sample size (N). Finally, the POWER option with a missing value (.) requests a solution for POWER, the statistical power of this test.

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 information of the test and then returns the value of statistical power. The power of this one-sample t-test is .900, which is almost same as .8997 we manually computed above.

Stata `. sampsi` makes the computation much simpler. In this command, hypothesized and claimed (alternative) means are followed by a series of options for other information. The `onesample` option indicates that this test is a one-sample test.

```
. sampsi 20 22, onesample alpha(.05) sd(4) n(44)
```

```
Estimated power for one-sample comparison of mean
to hypothesized value
```

```
Test Ho: m = 20, where m is the mean in the population
```

```
Assumptions:
```

```
alpha = 0.0500 (two-sided)
alternative m = 22
sd = 4
sample size n = 44
```

```
Estimated power:
```

```
power = 0.9126
```

Stata gives us .9126, which is slightly larger than what SAS reported. You may look up manuals and check the methods and formulas Stata uses.

Figure 7. Power Analysis of a One-sample T-test in SamplePower 2.0

The screenshot shows the SamplePower 2.0 software interface. The main window is titled "SamplePower - [One-sample t-test]". It features a menu bar (File, View, Options, Tools, Scenarios, Help) and a toolbar with various icons. The main area contains a table with the following data:

Group	Population Mean	Standard Deviation	N of Cases	Standard Error	95% Lower	95% Upper
Expected mean	20.0	4.0	44	0.60	18.79	21.21
Test against the constant	22.0					

Below the table, it shows "Alpha= 0.050, Tails= 2" and "Power" is 90%.

Two pop-up windows are visible:

- Welcome:** This interactive guide will lead you through the... and precision. To move this box use the title bar above. To close or reactivate this panel, select Help from the menu.
- Summary - Power:** For the given effect size (population mean= 20.0, tested against a constant of 22.0), sample size (44), and alpha (0.050, 2-tailed), power is 0.900. This means that 90% of studies would be expected to yield a significant effect, rejecting the null hypothesis that the population mean is 22.0.

Figure 7 illustrates an example of SPSS SamplePower 2.0, which reports the same statistical power of .90 at the .05 significance level.

6.2 Minimum Sample Size for a One-sample T-test

Suppose we realize that some observations have measurement errors, and that the population standard deviation is 3. We want to perform an upper one-tailed test at the .01 significance level and wish to get a statistical power of .8. Now, how many observations do we have to collect? What is the minimum sample size that can satisfy these conditions?

Table 4. Summary Information for a Sample Size Analysis

Test	Effect Size	Variation (SD)	Sample Size	Test Size	Power
T-test	2 (=22-20)	3	?	.01	.80

The SIDES=U option below indicates an upper one-tailed test with an alternative value greater than the null value (L means a lower one-sided test). The test size (significance level) is .01 and standard deviation is 3. POWER specifies a target statistical power. Since a missing value is placed at NTOTAL, SAS computes a required sample size.

```
PROC POWER;
  ONESAMPLEMEANS
  ALPHA=.01 SIDES=U
  NULLM=20 MEAN=22
  STDDEV=3
  NTOTAL=.
  POWER=.8;
```

```
RUN;
```

SAS tells us at least 26 observations are required to meet the conditions.

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

Stata `. sampsi` command here has the `power` option without `n` option. Stata returns 23 as a required sample size for this test. However, 23 observations produce a statistical power of .7510 in this setting. Note that `onesided` indicates this test is one-sided.

```
. sampsi 20 22, onesample onesided alpha(.01) sd(3) power(.8)
```

```
Estimated sample size for one-sample comparison of mean
to hypothesized value
```

```
Test Ho: m = 20, where m is the mean in the population
```

```
Assumptions:
```

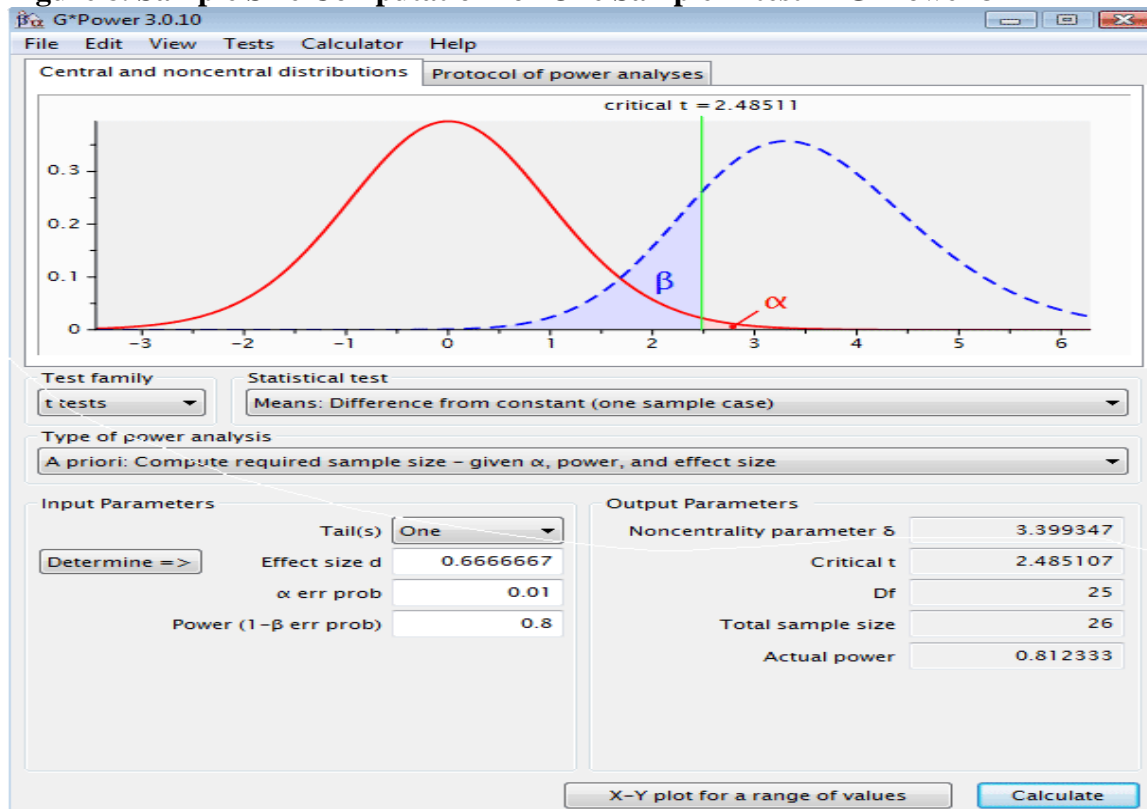
```
alpha = 0.0100 (one-sided)
power = 0.8000
alternative m = 22
sd = 3
```

```
Estimated required sample size:
```

```
n = 23
```

In G*Power 3, click Tests → Means → One group: Difference from constant to get the pop-up. Alternatively, choose `t test` under Test family, Means: Difference from constant (one sample case) under Statistical test, and `a priori: Compute required sample size - given α , power, and effect size` under Type of power analysis as shown in Figure 8.

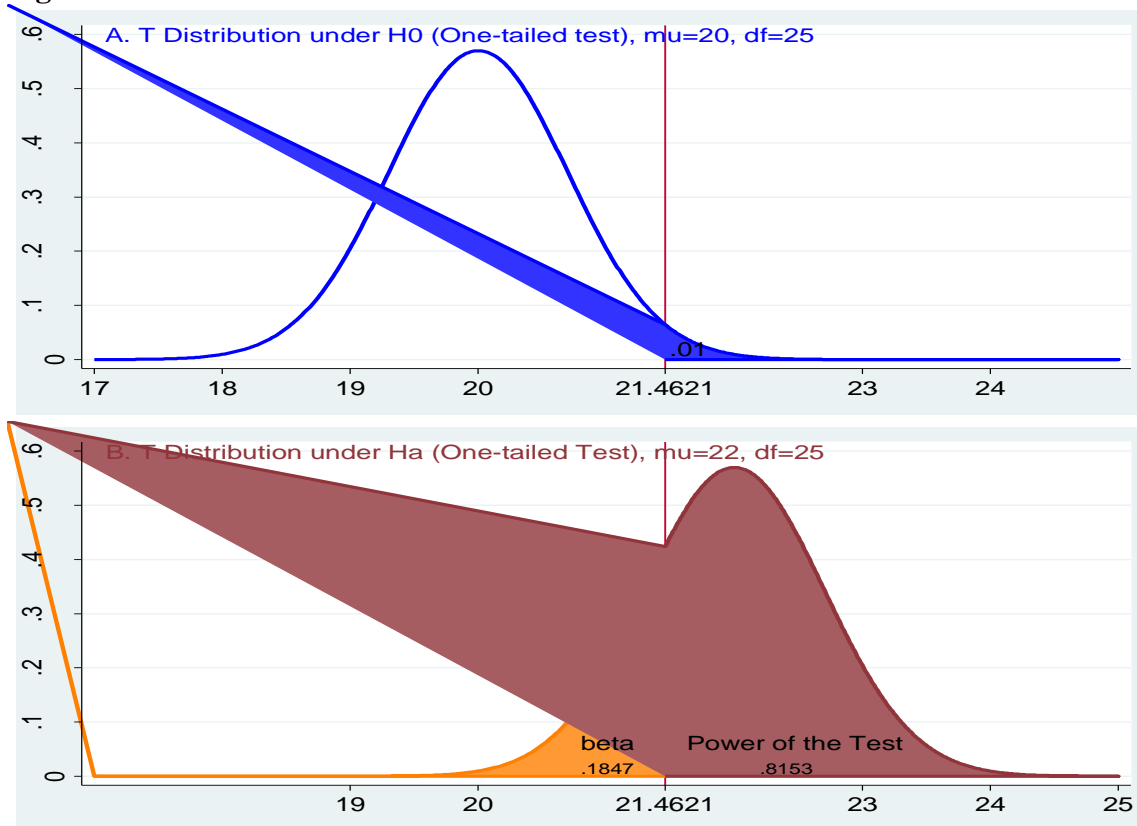
Figure 8. Sample Size Computation for One Sample T-test in G*Power 3



Choose one-tailed test and then provide “effect size”, significance level, and target statistical power. Note that “effect size” in G*Power is Cohen’s d , which is effect divided by standard deviation, NOT by standard error; accordingly, .6667 is $(22-20)/3$. Don’t be confused with standardized effect size I call here. G*Power suggests 26 observations as does SAS.

G*Power reports the critical value of the upper one-tailed test at the .01 level under the null hypothesis $H_0 : \mu = 20$. The value corresponds to 21.462109 on the original t distribution since $2.4851 = (21.462109 - 20) / 3 * \sqrt{26}$. The number 3.399347 at Noncentrality parameter δ is the t statistic under the alternative hypothesis $H_a : \mu = 22$, $3.3993 = (21.462109 - 22) / 3 * \sqrt{26}$. Figure 9 illustrates this computation. The power of this test is .8123 which is almost the same as SAS and G*Power computed.

Figure 9. Statistical Power in the T Distribution



6.3 Statistical Power and Sample Size of a Paired-sample T-test

Consider a paired sample t-test with mean difference of 3 and standard deviation of 4. Now, we want to compute statistical power when sample size increases from 20 to 30 and 40 (Table 5).

The PAIREDMEANS statement with TEST=DIFF performs a paired-sample t-test. The CORR option is required to specify the correlation coefficient of two paired variables. Two variables here are correlated with their moderate coefficient of .5. In NPAIRS, 20, 30, and 40 are listed to see how

statistical power of this test varies as the number of pairs changes. Alternatively, you may list numbers as “NPAIRS = 20 To 40 BY 10”

Table 5. Summary Information for a Statistical Power Analysis

Test	Effect Size	Variation (SD)	Sample Size	Test Size	Power
T-test	3	4	20, 30, 40	.01	?

```
PROC POWER;
  PAIREDMEANS TEST=DIFF
  ALPHA = .01 SIDES = 2
  MEANDIFF = 3
  STDDEV = 4
  CORR = .5
  NPAIRS = 20 30 40
  POWER = .;
```

```
RUN;
```

The statistical power of this two-tailed test is .686 when N=20, .902 when N=30, and .975 when N=40.

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	4
Correlation	0.5
Null Difference	0

Computed Power

Index	N Pairs	Power
1	20	0.686
2	30	0.902
3	40	0.975

Stata produces a bit higher statistical power of .782 when N=20, .937 when N=30, and .985 when N=40.

```
. sampsi 0 3, onesample alpha(.01) sd(4) n(20)
```

Estimated power for one-sample comparison of mean
to hypothesized value

Test Ho: m = 0, where m is the mean in the population

Assumptions:

```

alpha = 0.0100 (two-sided)
alternative m = 3
sd = 4
sample size n = 20

```

Estimated power:

```
power = 0.7818
```

Now, compute required sample sizes to get statistical powers of .70, .80, and .90 (Table 6).

Table 6. Summary Information for a Sample Size Analysis

Test	Effect Size	Variation (SD)	Sample Size	Test Size	Power
T-test	3	4	?	.01	.70, .80, .90

SAS suggests 21, 25, and 30, respectively. The POWER option below can be also written as

“POWER = .70 .80 .90;”

PROC POWER;

```

PAIREDMEANS TEST=DIFF
ALPHA = .01 SIDES = 2
MEANDIFF = 3
STDDEV = 4
CORR= .5
NPAIRS = .
POWER =.70 TO .90 BY .10;

```

RUN;

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	4
Correlation	0.5
Null Difference	0

Computed N Pairs

Index	Nominal Power	Actual Power	N Pairs
1	0.7	0.717	21
2	0.8	0.819	25
3	0.9	0.902	30

Stata suggests that 21 and 27 observations are required for statistical powers of .80 and .90, respectively.

```
. sampsi 0 3, onesample alpha(.01) sd(4) power(.80)
```

Estimated sample size for one-sample comparison of mean

<http://www.indiana.edu/~statmath>

to hypothesized value

Test Ho: $m = 0$, where m is the mean in the population

Assumptions:

```
alpha = 0.0100 (two-sided)
power = 0.8000
alternative m = 3
sd = 4
```

Estimated required sample size:

```
n = 21
```

6.4 Statistical Power and Sample Size of an Independent Sample T-test

The following example conducts a statistical power analysis for an independent sample t-test. The `TWOSAMPLEMEANS` statement indicates an independent sample t-test. The `GROUPMEANS` option below specifies means of two groups in a matched format in parentheses. We want to know how statistical power will change when the pooled standard deviation increases from 1.5 to 2.5 (see `STDDEV`). Note the default option `SIDES=2` is omitted.

```
PROC POWER;
  TWOSAMPLEMEANS
  ALPHA = .01
  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;
```

The `PLOT` statement with `X=N` option draws a plot of statistical power as N in the X -axis changes.

The `POWER` Procedure
Two-sample t Test for Mean Difference

Fixed Scenario Elements

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

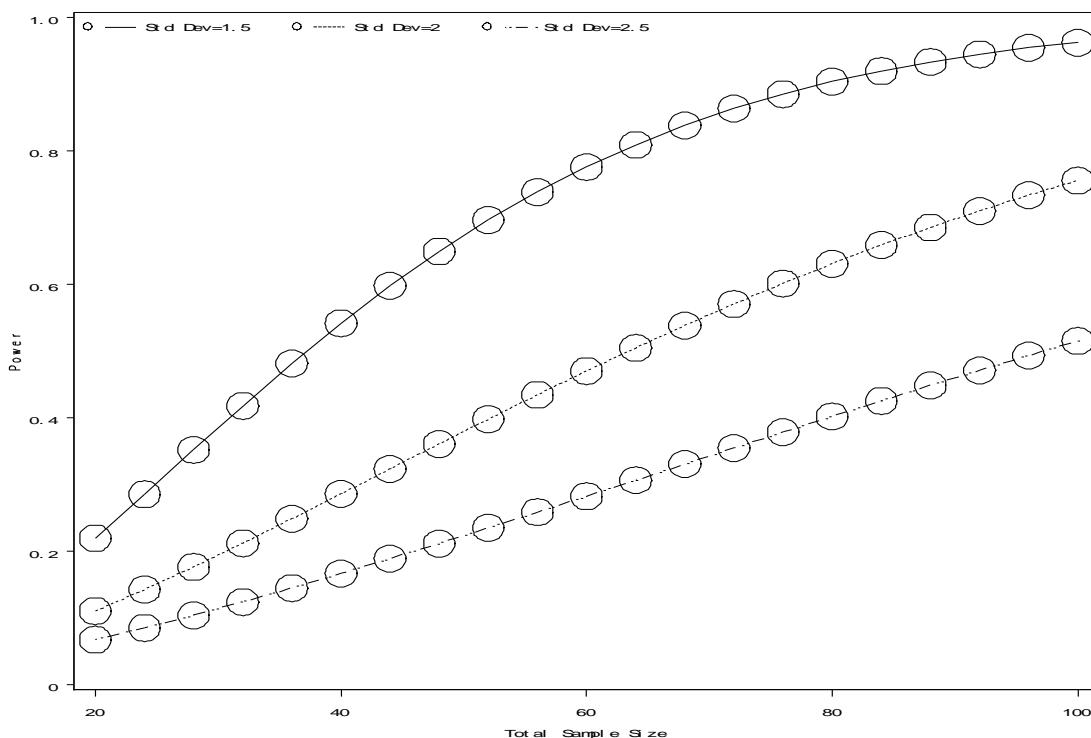
Computed Power

Std

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

Figure 10 visualizes the relationship between statistical power and sample size when sample size changes from 20 to 100. The statistical power for standard deviation 1.5 is more sensitive to sample size than those for 2.0 and 2.5 when N is less than 60 (compare slopes of three curves).

Figure 10. A Plot of Statistical Power and Sample Size



In Stata, you need to provide standard deviations and numbers of observations of two groups. Stata reports a bit larger statistical power when standard deviation is 1.5.

```
. sampsi 2.79 4.12, alpha(.01) sd1(1.5) sd2(1.5) n1(22) n2(22)
```

Estimated power for two-sample comparison of means

Test Ho: $m_1 = m_2$, where m_1 is the mean in population 1
and m_2 is the mean in population 2

Assumptions:

```
alpha = 0.0100 (two-sided)
m1 = 2.79
m2 = 4.12
sd1 = 1.5
sd2 = 1.5
sample size n1 = 22
n2 = 22
```



```
n2/n1 = 1.00
```

Estimated power:

```
power = 0.6424
```

Now, let us compute the minimum number of observations to get statistical powers of .80 and .90. Individual standard deviations (as opposed to pooled standard deviations) of two groups are listed in the GROUPSTDDEVS option. Since ALPHA and SIDES were omitted, this test is a two-tailed test at the .05 significance level.

```
PROC POWER;
  TWOSAMPLEMEANS TEST=DIFF_SATT
  GROUPMEANS = (2.79 4.12)
  GROUPSTDDEVS = (1.5 3.5)
  NTOTAL=.
  POWER=.80 .90;
RUN;
```

SAS suggests that 132 and 176 observations are needed for statistical powers of .80 and .90, respectively.

The POWER Procedure
Two-sample t Test for Mean Difference with Unequal Variances

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Group 1 Mean	2.79
Group 2 Mean	4.12
Group 1 Standard Deviation	1.5
Group 2 Standard Deviation	3.5
Number of Sides	2
Null Difference	0
Nominal Alpha	0.05
Group 1 Weight	1
Group 2 Weight	1

Computed N Total

Index	Nominal Power	Actual Alpha	Actual Power	N Total
1	0.8	0.05	0.801	132
2	0.9	0.05	0.901	176

Stata by default reports the minimum number of observations for statistical power .90 at the .05 level. The required observation of 174 (87+87) is almost similar to 176 that SAS suggests.

```
. sampsi 2.79 4.12, sd1(1.5) sd2(3.5)
```

Estimated sample size for two-sample comparison of means

Test Ho: $m_1 = m_2$, where m_1 is the mean in population 1
and m_2 is the mean in population 2

Assumptions:

<http://www.indiana.edu/~statmath>

```

alpha = 0.0500 (two-sided)
power = 0.9000
m1 = 2.79
m2 = 4.12
sd1 = 1.5
sd2 = 3.5
n2/n1 = 1.00

```

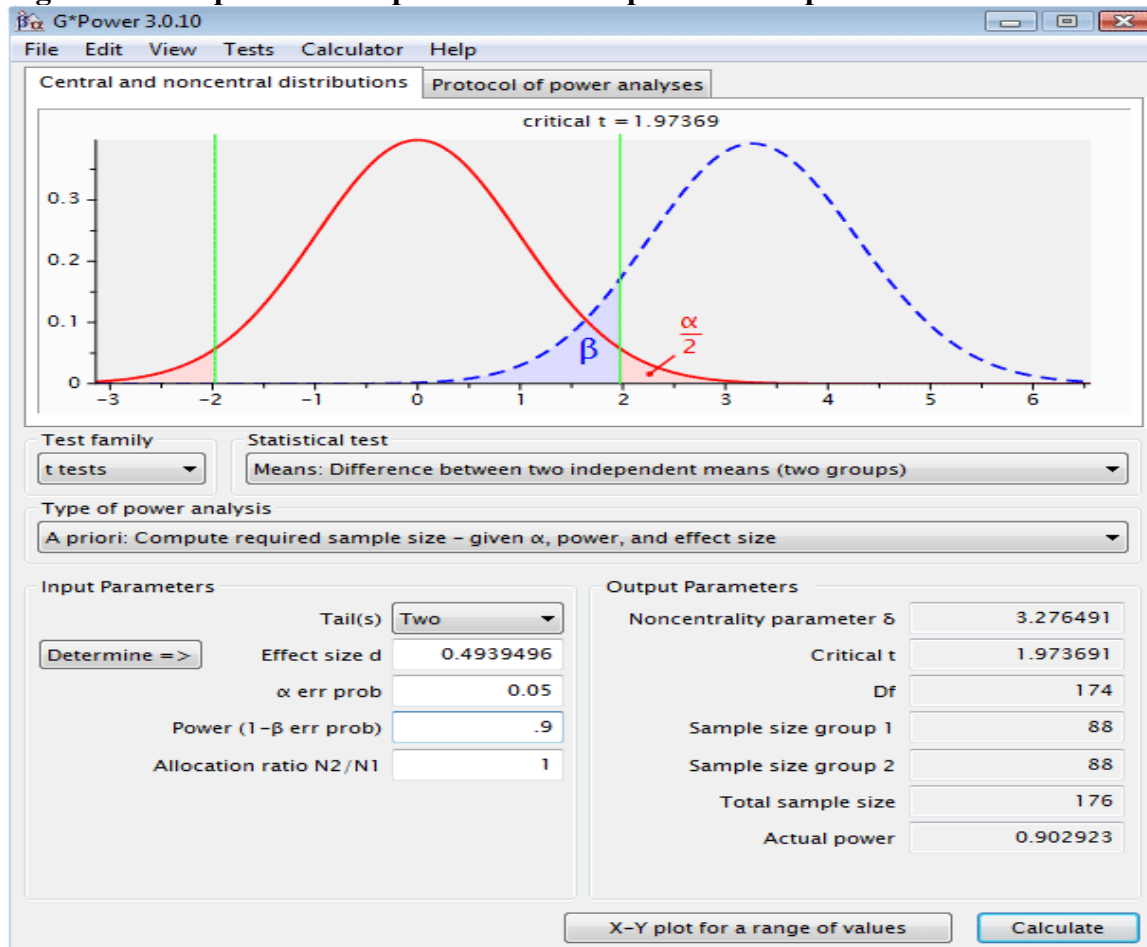
Estimated required sample sizes:

```

n1 = 87
n2 = 87

```

Figure 11. Sample Size Computation for Independent Sample T-test in G*Power 3



In G*Power 3, click Tests→Means→Two independent groups. In order to get “Effect size d”, click Determine => to get a pop-up. Then, provide sample means and standard deviations. Click Calculate to compute Effect size f and then click Calculate and transfer to main window to copy the “effect size” to the main window in Figure 11. Finally, click Calculate to see the minimum number of observations computed. G*Power reports 176 observations as does SAS. Note that degrees of freedom are $174 = 176 - 2$.

6.5 Advanced Features of SAS POWER and GLMPOWER

SAS provides highly flexible ways of specifying values of interest. You may list means and standard deviations of more than one pair using parentheses or a vertical bar (|) of the cross grouped format. “GROUPMEANS = 1 2 | 3 4” is equivalent to “GROUPMEANS = (1 3) (1 4) (2 3) (2 4)” Conversely, “GROUPSTDDEV = (.7 .8) (.7 .9)” can be written as “GROUPSTDDEV = .7 | .8 .9”

```
PROC POWER;
  TWOSAMPLEMEANS TEST=DIFF_SATT
  GROUPMEANS = 1 2 | 3 4
  GROUPSTDDEV = (.7 .8) (.7 .9)
  NTOTAL = .
  POWER = .80;
```

RUN;

SAS report minimum sample sizes for eight scenarios consisting of four pairs of means and two pairs of standard deviations. For example,

The POWER Procedure
Two-sample t Test for Mean Difference with Unequal Variances

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Nominal Alpha	0.05
Group 1 Weight	1
Group 2 Weight	1

Computed N Total

Index	Mean1	Mean2	Std Dev1	Std Dev2	Actual Alpha	Actual Power	N Total
1	1	3	0.7	0.8	0.0414	0.840	8
2	1	3	0.7	0.9	0.0450	0.910	10
3	1	4	0.7	0.8	0.0354	0.881	6
4	1	4	0.7	0.9	0.0363	0.833	6
5	2	3	0.7	0.8	0.0489	0.840	22
6	2	3	0.7	0.9	0.0492	0.823	24
7	2	4	0.7	0.8	0.0414	0.840	8
8	2	4	0.7	0.9	0.0450	0.910	10

The SAS GROUPNS option allows us to compute a required sample size of a group given a sample size of the other group. Note that NTOTAL assumes balanced data where two groups have the same number of observations. Suppose we want to know the minimum sample size of group 1 when sample sizes of group 2 are given 5 and 10. Pay attention to the missing values specified in GROUPNS.

PROC POWER;

```
TWOSAMPLEMEANS TEST=DIFF
GROUPMEANS = ( 2 3)
STDDEV = .7
GROUPNS = (. 5) (. 10)
POWER = .80;
```

RUN;

SAS report 25 observations of group 1 as a required sample size when group 2 has 5 observations, and 8 observations when group 2 have 10. Total numbers of observations in two scenarios are 30 (=25+5) and 18 (=8+10), respectively.

The POWER Procedure
Two-sample t Test for Mean Difference

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Group 1 Mean	2
Group 2 Mean	3
Standard Deviation	0.7
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Alpha	0.05

Computed N1

Index	N2	Actual Power	N1
1	5	0.804	25
2	10	0.807	8

6.6 Statistical Power and Sample Size of Comparing a Proportion

Imagine a binary variable of a Yes/No or True/False type question. Suppose we have 50 observations and 30 percent answered Yes. We want to know if the true proportion is .50, $H_0 : \pi = .5$. Let us conduct normal approximated z-test (TEST=Z) in stead of the exact binomial test (TEST=EXACT). The ONESAMPLEFREQ statement compares a proportion with a hypothesized proportion, which is specified in NULLPROPORTION.

```
PROC POWER;
ONESAMPLEFREQ TEST=Z
NULLPROPORTION = .5
PROPORTION = .3
NTOTAL = 50
POWER = .;
```

RUN;

SAS reports .859 as the statistical power of comparing the proportion of a binary variable.

The POWER Procedure
Z Test for Binomial Proportion

Fixed Scenario Elements

Method	Exact
Null Proportion	0.5
Binomial Proportion	0.3
Total Sample Size	50
Number of Sides	2
Nominal Alpha	0.05

Computed Power

Lower Crit Val	Upper Crit Val	Actual Alpha	Power
18	32	0.0649	0.859

Stata returns a power of .828, which is similar to .859 above.

```
. sampsi .5 .3, onesample n(50)
```

Estimated power for one-sample comparison of proportion
to hypothesized value

Test Ho: $p = 0.5000$, where p is the proportion in the population

Assumptions:

```
alpha = 0.0500 (two-sided)
alternative p = 0.3000
sample size n = 50
```

Estimated power:

```
power = 0.8283
```

Let us compute the minimum number of observations to get statistical powers of .80 and .90. SAS suggests that 71 and 88 observations are respectively need.

```
PROC POWER;
  ONESAMPLEFREQ TEST=Z METHOD=NORMAL
  ALPHA=.01 SIDES=2
  NULLPROPORTION = .5
  PROPORTION = .3
  NTOTAL = .
  POWER = .80 .90;
RUN;
```

The POWER Procedure
Z Test for Binomial Proportion

Fixed Scenario Elements

```

Method                Normal approximation
Number of Sides              2
Null Proportion            0.5
Alpha                     0.01
Binomial Proportion        0.3

```

```

                Computed N Total
Index      Nominal Power  Actual Power  N Total
1          0.8          0.807         71
2          0.9          0.900         88

```

The following `.sampsi` command returns the same required sample sizes 71 and 88 for the two statistical powers.

```
. sampsi .5 .3, onesample power(.80) alpha(.01)
```

```
Estimated sample size for one-sample comparison of proportion
to hypothesized value
```

```
Test Ho: p = 0.5000, where p is the proportion in the population
```

```
Assumptions:
```

```

alpha = 0.0100 (two-sided)
power = 0.8000
alternative p = 0.3000

```

```
Estimated required sample size:
```

```
n = 71
```

6.7 Statistical Power and Sample Size of Comparing Two Proportions

Now, consider two binary variables. Suppose their proportions of Yes or Success are .25 and .45, respectively. We want to know there is no difference in population proportions, $H_0: \pi_1 = \pi_2$. The `TEST=FISHER` option tells SAS to conduct the Fisher's exact conditional test for two proportions. Note "`NPERGROUP = 80 TO 120 BY 10`" is equivalent to "`NPERGROUP = 80 90 100 110 120`"

```

PROC POWER;
  TWOSAMPLEFREQ TEST=FISHER
  GROUPPROPORTIONS = (.25 .45)
  NPERGROUP = 80 TO 120 BY 10
  POWER = .;
RUN;

```

The above SAS procedure computes statistical power when N of one variable changes from 80 to 120, assuming balanced data. SAS reports statistical power ranging from .71 to .88.

```

The POWER Procedure
Fisher's Exact Conditional Test for Two Proportions

```

Fixed Scenario Elements

Distribution Method	Exact conditional Walters normal approximation
Group 1 Proportion	0.25
Group 2 Proportion	0.45
Number of Sides	2
Alpha	0.05

Computed Power

Index	N Per Group	Power
1	80	0.708
2	90	0.764
3	100	0.811
4	110	0.850
5	120	0.881

Alternatively, you may use `GROUPNS` to specify numbers of observations of two groups. This option is very useful when data are not balanced. The following procedure gives us the same answered as we got above.

```
PROC POWER;
  TWOSAMPLEFREQ TEST=FISHER
  GROUPPROPORTIONS = (.25 .45)
  GROUPNS = (80 80) (90 90) (100 100) (110 110) (120 120)
  POWER = .;
RUN;
```

In Stata, you just need to list two proportions followed by numbers of observations. Stata returns .705 when $N=80$, which is almost same as .708 above.

```
. sampsi .25 .45, n1(80) n2(80)

Estimated power for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
              and p2 is the proportion in population 2
Assumptions:
      alpha = 0.0500 (two-sided)
      p1 = 0.2500
      p2 = 0.4500
sample size n1 = 80
              n2 = 80
              n2/n1 = 1.00

Estimated power:
      power = 0.7048
```

Conversely, we may get required numbers of observations to get statistical powers of .80 and .90 by omitting the value of `NPERGROUP`. SAS suggests that each group should have at least 89 and 128 observations to meet the .80 and .90 level.

```

PROC POWER;
  TWOSAMPLEFREQ TEST=FISHER
  GROUPPROPORTIONS = (.25 .45)
  NPERGROUP = .
  POWER = .80 .90;
RUN;

```

The POWER Procedure
Fisher's Exact Conditional Test for Two Proportions

Fixed Scenario Elements

Distribution Method	Exact conditional Walters normal approximation
Group 1 Proportion	0.25
Group 2 Proportion	0.45
Number of Sides	2
Alpha	0.05

Computed N Per Group

Index	Nominal Power	Actual Power	N Per Group
1	0.8	0.802	98
2	0.9	0.902	128

In Stata, provide a target statistical power. Both SAS and Stata suggest that at least 256 (= 128 × 2) observations are need for the statistical power of .90.

```
. sampsi .25 .45, power(.90)
```

Estimated sample size for two-sample comparison of proportions

Test Ho: $p_1 = p_2$, where p_1 is the proportion in population 1
and p_2 is the proportion in population 2

Assumptions:

```

alpha = 0.0500 (two-sided)
power = 0.9000
p1 = 0.2500
p2 = 0.4500
n2/n1 = 1.00

```

Estimated required sample sizes:

```

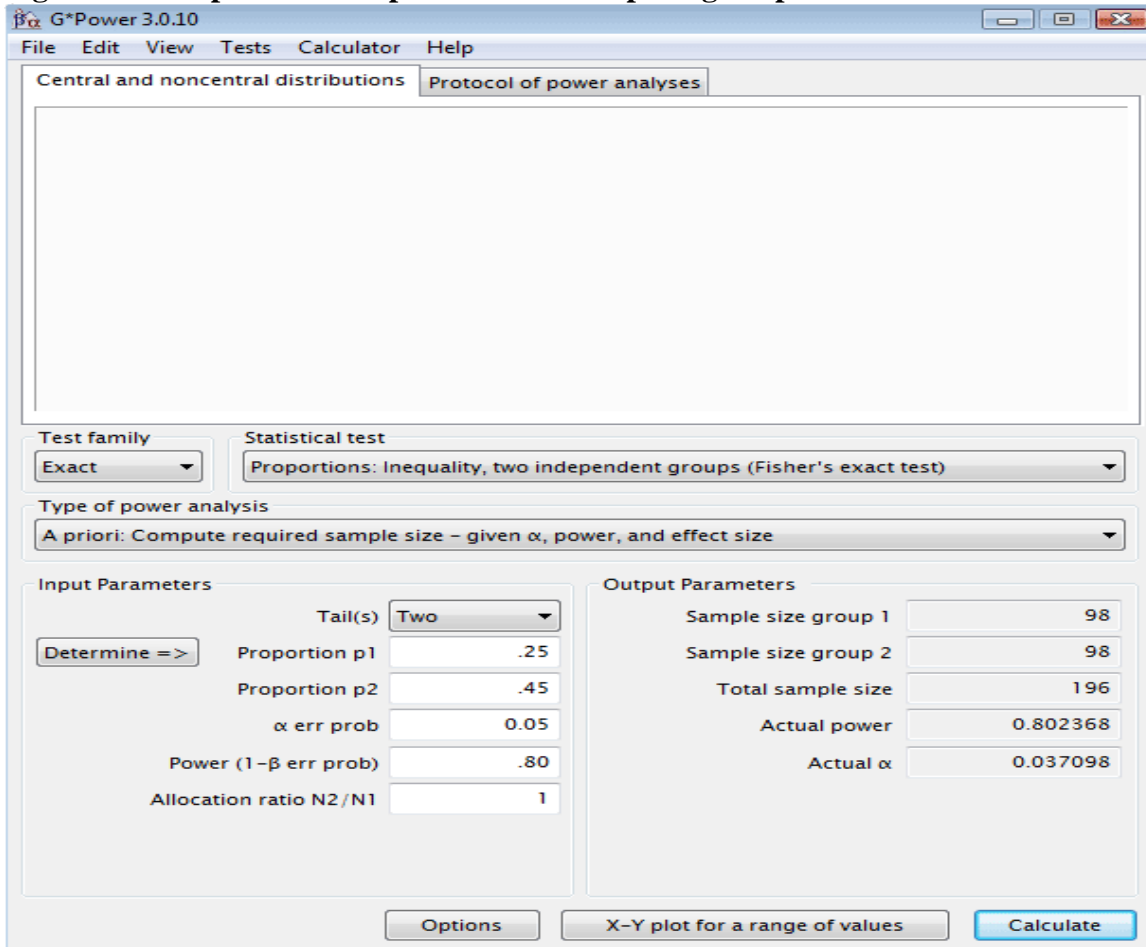
n1 = 128
n2 = 128

```

In G*Power, click Tests → Proportions → Two independent groups: Inequality, Fisher's exact test to get a pop-up shown in Figure 12. Note that other methods may return different results. Choose the two-tailed test and then provide two proportions and a target statistical power. Finally, click Calculate to compute the required sample size and actual statistical power.

Like SAS, G*Power 3 reports 196 observations for a statistical power of .80 but suggest 254 observations for a power of .90.

Figure 12. Sample Size Computation for Comparing Proportions in G*Power 3



7. Applications: ANOVA and Linear Regression

SAS POWER and GLMPOWER procedures perform statistical power and sample size analysis for analyses of variance (ANOVA) and linear regression, while Stata `.sampsi` does not.

7.1 Statistical Power and Sample Size of a One-way ANOVA

The POWER procedure has the ONEWAYANOVA statement for a one-way ANOVA. Note that NPERGROUP specifies the number of observations of a group, assuming balanced data.

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

RUN;

The above POWER procedure lists statistical powers for different standard deviations and numbers of observations. For example, if a standard deviation is 4 and each group has 10 observations, statistical power is .972 at the .05 level.

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

Index	Std Dev	N Per 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 performs power analysis of 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 for ANOVA in the GLMPOWER procedure. In other word, the POWER statement is encompassed in this procedure.

```
PROC GLMPOWER DATA=power.car;
  CLASS mode;
  MODEL income=mode;
```

```

POWER STDDEV = 15
ALPHA = .01
NTOTAL = 100 TO 200 BY 50
POWER = .;

```

RUN;

The power of this one-way ANOVA is .924 when N is 200.

The GLMPOWER Procedure

Fixed Scenario Elements

Dependent Variable	income
Source	mode
Alpha	0.01
Error Standard Deviation	15
Test Degrees of Freedom	3

Computed Power

Index	Nominal N Total	Actual		Error DF	Power
		N Total			
1	100	100		96	0.542
2	150	100		96	0.542
3	200	200		196	0.924

Conversely, we can compute required sample size by specifying target statistical powers. Since ALPHA and SIDES are omitted, this is a two-tailed test at the .05 level.

```

PROC GLMPOWER DATA=power.car;
CLASS male;
MODEL income=male;
POWER STDDEV=15
NTOTAL= .
POWER= .70 .80 .90;

```

RUN;

SAS suggests from 1,900 to 3,100 depending on target statistical power.

The GLMPOWER Procedure

Fixed Scenario Elements

Dependent Variable	income
Source	male
Alpha	0.05
Error Standard Deviation	15
Test Degrees of Freedom	1

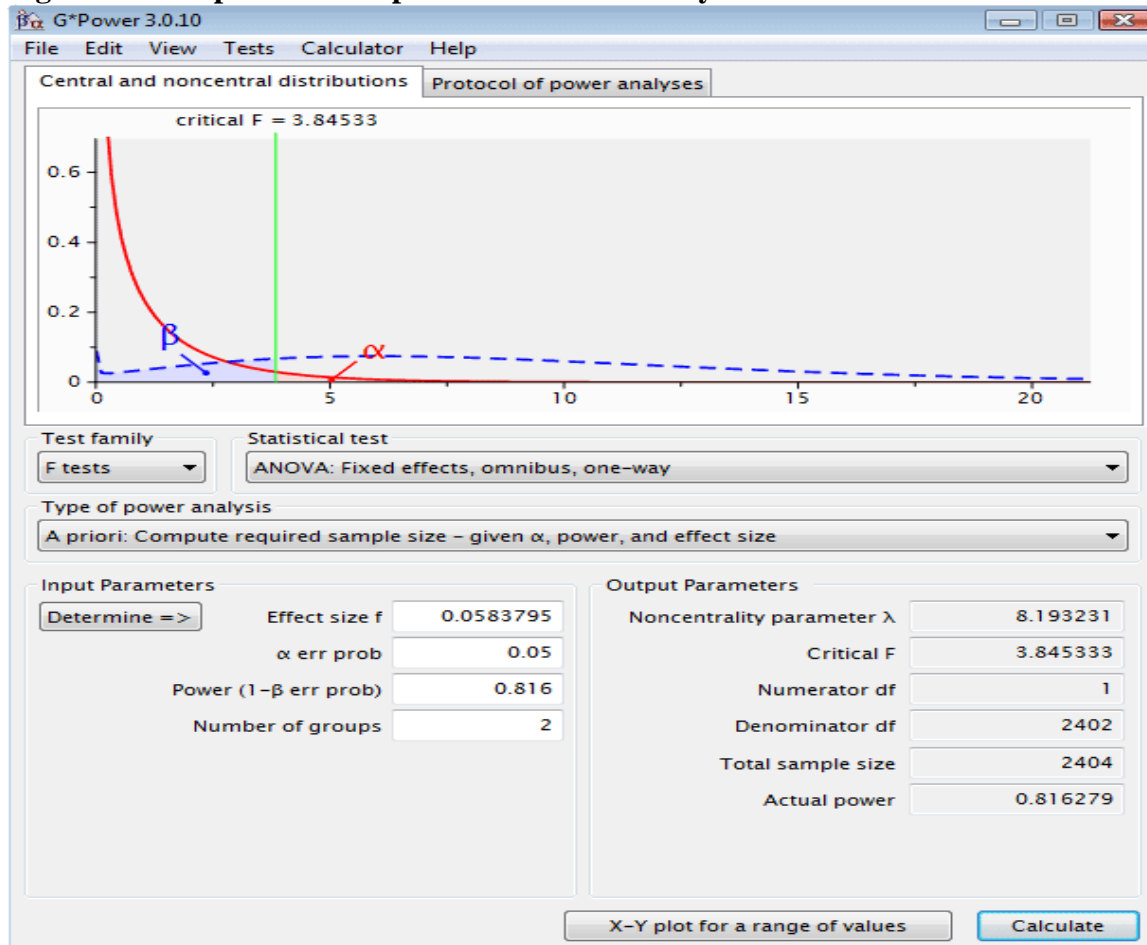
Computed N Total

Nominal	Error	Actual	N
---------	-------	--------	---

Index	Power	DF	Power	Total
1	0.7	1898	0.720	1900
2	0.8	2398	0.816	2400
3	0.9	3098	0.901	3100

In G*Power, click Tests→Variances→One Group to get a pop-up for a one-way ANOVA. Do not forget to choose F-tests under Test family (Figure 13). G*Power 3 reports 2,404 observations for a statistical power of .816, which is almost the same as 2,400 that SAS suggests. Interestingly, G*Power illustrates how significance level and β are related in this setting.

Figure 13. Sample Size Computation for a Two-way ANOVA in G*Power 3



In order to compute “Effect size f ”, click Determine=> to open a pop-up in Figure 14. Choose Effect size from means under Select procedure. Provide the number of groups and a hypothesized standard deviation (not sample standard deviation) of each group. This number determines the size of a workable spreadsheet below.

In the spreadsheet, provide sample means and number of observations of groups. Click Calculate to compute Effect size f and then click Calculate and transfer to main window to copy the “effect size” to the main window in Figure 13. Finally, click Calculate in

Figure 13 to see the required number of observations computed and actual statistical power given the sample size.

Figure 14. Providing Group Information and Standard Deviation

Group	Mean	size
1	61.0375	40
2	59.25	60

7.2 Statistical Power and Sample Size of a Two-way ANOVA

Since `POWER` does not support two-way ANOVA, we have to use `GLMPower`. As mentioned before, this procedure can handle more complicated models. Variables `owncar` and `male` here are used as fixed factors with their interaction allowed. The `PLOT` statement draws a plot of statistical power and sample size (Figure 15).

```
PROC GLMPower DATA=power.car;
  CLASS owncar male;
  MODEL income= owncar | male;
  POWER STDDEV=10
    ALPHA=.01
    NTOTAL= 200
    POWER= .;
  PLOT X=N MIN=150 MAX=250;
RUN;
```

The `GLMPower` procedure lists statistical powers when standard deviation is 10 and $N=200$ and then draw a plot in order to illustrate how statistical power changes when N changes from 100 to 200.

The `GLMPower` Procedure

Fixed Scenario Elements

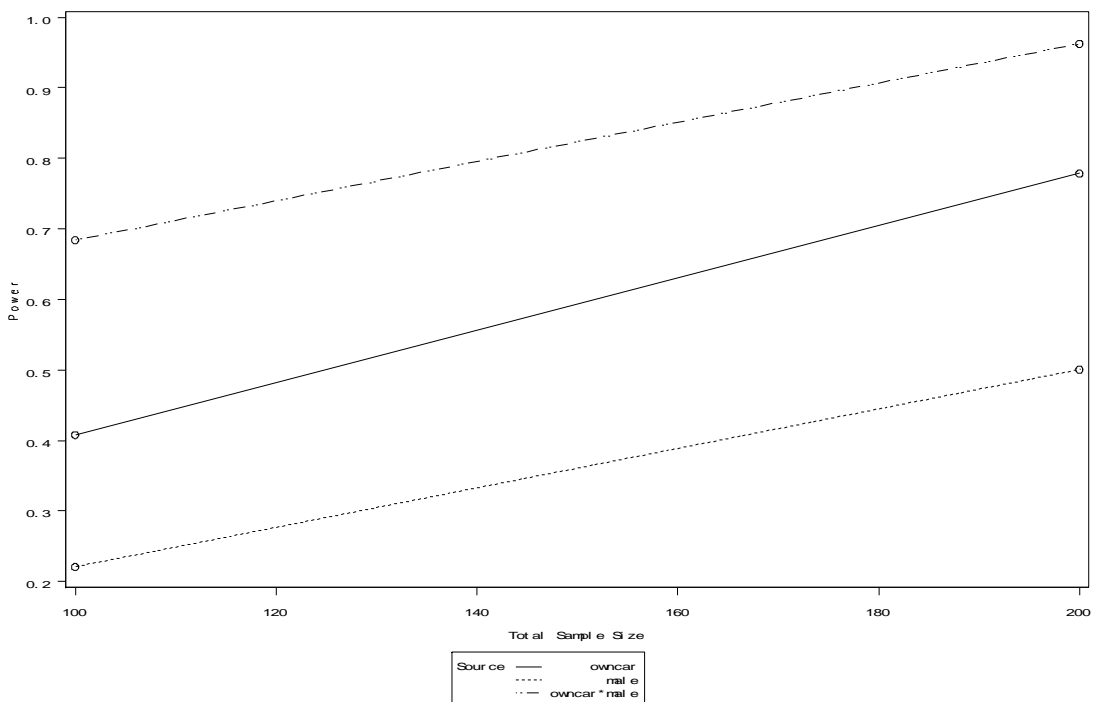
Dependent Variable	income
Alpha	0.01

Error Standard Deviation 10
 Total Sample Size 200
 Error Degrees of Freedom 196

Computed Power

Index	Source	Test DF	Power
1	owncar	1	0.779
2	male	1	0.501
3	owncar*male	1	0.963

Figure 15. A Plot of Statistical Power and Sample Size



Finally, let us get the minimum number of observations by omitting the value of NTOTAL.

```
PROC GLMPOWER DATA=power.car;
  CLASS owncar male;
  MODEL income= owncar male;
  POWER STDDEV=10
    ALPHA=.05
    NTOTAL= .
    POWER=.80;
RUN;
```

The GLMPOWER Procedure

Fixed Scenario Elements

Dependent Variable income

Alpha	0.05
Error Standard Deviation	10
Nominal Power	0.8

Computed N Total					
Index	Source	Test DF	Error DF	Actual Power	N Total
1	owncar	1	197	0.977	200
2	male	1	397	0.821	400

If you want to perform other statistical power analysis for more complicated model and linear regression model, see the SAS manuals.

8. Conclusion

A hypothesis is a specific conjecture about an aspect of a population of interest. Hypothesis testing is a scientific process to investigate if a null hypothesis is true or not, although we never know if the hypothesis is actually true or not. A Type I error is a wrong decision of rejecting a true null hypothesis, while Type II error occurs when not rejecting false null hypothesis. The probability of committing Type I error is the size of a test or significance level α , whereas the probability of committing Type II error is β . $(1 - \alpha)$ is called confidence level, while $(1 - \beta)$ is statistical power of a test. The statistical power is the ability of a test to detect an effect, if the effect actually exists. But it is notable that this concept is not always clear. See Section 1 (Figure 2, in particular) and 2 for review.

Despite the presence of many statistical packages for statistical power analysis, I would recommend SAS and G*Power 3. SAS POWER and GLMPOWER procedures support various models with many options. Its (online) documentation is a good resource for statistical power analysis. G*Power 3 is a free software package that supports various models. Its point-and-click method makes it very easy to use. SAS and G*Power returns almost the same results. The Stata `.sampsi` command is handy to perform simple power analysis but does not support ANOVA and complicated linear models. Stata's results are slightly different from those of SAS and G*Power in some circumstances. Finally, you can also try with R and Minitab.

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. Upper Saddle River, NJ: Prentice Hall.
- Greene, William H. 2003. *Econometric Analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall.
- High, Robin. 2000. Important Factors in Designing Statistical Power Analysis Studies.
<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.
- Mazen, A. Magid M., Masoud Hemmasi, and Mary Frances Lewis. 1985. "In Search of Power: A Statistical Power Analysis of Contemporary Research in Strategic Management." *Academy of Management Proceedings*, 30-34.
- 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.
- Stata Press. 2007. *Stata Base Reference Manual 3*. College Station, TX: StataCorp LP.
- Thomas, Len, and Charles J. Krebs. 1997. "A Review of Statistical Power Analysis Software," *Bulletin of the Ecological Society of America*, 78 (2): 126-139.

<http://www.psych.uni-duesseldorf.de/aap/projects/gpower/>

Revision History

- 2004. Memo
- 2008. First draft
- 2010. Revised by Dani Marinova.