GPOWER Tutorial

Before we begin this tutorial, we would like to give you a general advice for performing power analyses.

A very frequent error in performing power analyses with G*Power is to specify incorrect degrees of freedom. As a general rule, therefore, we recommend that you routinely compare the degrees of freedom as specified in G*Power with the degrees of freedom that your statistical analysis program gives you for an appropriate set of data. If you do not yet have your data set (e.g., in the case of an a priori power analysis), then you could simply create an appropriate artificial data set and check the degrees of freedom for this set.

Let us now start with the simplest possible case, a t-test for independent samples.

In a now-classic study, Warrington and Weiskrantz (1970) compared the memory performance of amnesics to normal controls. Amnesics are persons who have very serious long-term memory problems. It very often takes them weeks to learn where the bathroom is in a new environment, and some of them never seem to learn such things. Perhaps the most intriguing result of the Warrington and Weiskrantz study was that amnesics and normals differed with respect to direct, but not indirect measures of memory.

An example of a direct memory measure would be *recognition performance*. This measure is called direct because the remembering person receives explicit instructions to recollect a prior study episode ("please recognize which of these words you have seen before").

In contrast, word stem completion would be an *indirect measure of memory*. In such a task, a person is given a word stem such as "tri....." and is asked to complete it with the first word that comes to mind. If the probability of completing such stems with studied words is above base-line, then we observe an effect of prior experience.

It should be clear by now why the finding of no statistically significant difference between amnesiacs and normal in indirect tests was so exciting: All of a sudden there was evidence for memory where it was not expected, but only when the instructions did not stress the fact that the task was a memory task.

However, it may appear a bit puzzling that amnesiacs and normal were not totally equivalent with respect to the indirect word stem completion task. Rather, normal were a bit better than amnesiacs with an average of 16 versus 14.5 stems completed with studied words, respectively. Of course, in the recognition task, normal were much better than amnesiacs with correct recognition scores of 13 versus 8, respectively.

At this point, one may wonder about the power of the relevant statistical test to detect a difference if there truly was one. Therefore, let's perform a post-hoc power analysis on these Warrington and Weiskrantz (1970) data.

Post-hoc Power Analysis

For the sake of this example, let us assume that the mean word-stem completion performance for amnesics (14.5) and normals (16) as observed by Warrington and Weiskrantz (1970) reflects the population means, and let the population standard deviation of both group means be sigma = 3. We can now compute the effect size index d (Cohen, 1977) which is defined as

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

We obtain

$$d = \frac{|14.5 - 16|}{3} = 0.5$$

The resulting d = 0.5 can be interpreted as a "medium" effect according to Cohen's (1977) popular effect size conventions.

A total of

$n_1 = 4$ amnesics and

$n_2 = 8$ normal control subjects

participated in the Warrington and Weiskrantz (1970) study. These sample sizes are used by G*Power to compute the relevant noncentrality parameter of the noncentral t-distribution. The noncentral distribution of a test statistic results, for a certain sample size, if H_1 (the alternative hypothesis) is true. The noncentrality parameter delta (δ) is defined as

$$\delta = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Now we are almost set to perform our post-hoc power analysis. One more piece is missing, however. We need to decide which level of alpha is acceptable. Without much thinking, we choose alpha = .05. Given these premises, what was the power in the Warrington and Weiskrantz (1970) study to detect a "medium" size difference between amnesics and controls in the word stem completion task?

Start G*Power and

select:	Type of Power Analysis:	Post-hoc
	Type of Test:	t-Test (means), two-tailed
		Accuracy mode calculation

Next, G*Power needs the following

input:	Alpha:	.05
	Effect size "d":	0.5
	n1:	4
	n2:	8

You can now press the Calculate button and observe the following

result:	**Power (1-beta):**	**0.1148**
	Critical t:	t(10) = 2.2281
	Delta:	0.8165

This result is devastating: The relevant statistical test had virtually no power to detect a "medium" size difference between amnesics and controls in the word stem completion task.

If we were to repeat the Warrington and Weiskrantz (1970) study with more statistical power, how many participants would we need? This question is answered by an

A Priori Power Analysis

In an a priori power analysis, we know which alpha and beta levels we can accept, and ideally we also have a good idea of the size of the effect which we want to detect. We decide to be maximally idealistic and choose alpha = beta = .05. (It means a power level of 1-β = 0.95). In addition, we know that the size of the effect we want to detect is d = 0.5. We are now ready to perform our a priori power analysis.

Select:	Type of Power Analysis:	A priori
	Type of Test:	t-Test (means), two-tailed
		Accuracy mode calculation
Input:	Alpha:	.05
	Power (1-beta):	.95
	Effect size "d":	0.5
Result:	**Total sample size:**	**210**
	Actual power:	0.9500
	Critical t:	t(208) = 1.9714
	Delta:	3.6228

We are shocked. Obviously, there is no way we can recruit N = 210 subjects for our study, simply because it would be impossible to find $n_1 = 105$ amnesic patients (fortunately, very few people suffer from severe amnesia!).

Assume that we work in a hospital in which $n_1 = 20$ amnesics are treated at the moment. It seems reasonable to expect that we can recruit an equal number of control patients to participate in our study. Thus, $n_1 + n_2 = 20 + 20 = 40$ is our largest possible sample size.

What are we going to do? Well, we simply perform a

Compromise Power Analysis

Erdfelder (1984) has developed the concept of a compromise power analysis specifically for cases like the present one in which pragmatic constraints prohibit that our investigations follow the recommendations derived from an a priori power analysis. The basic idea here is that two things are fixed, the maximum possible sample size and the effect we want to detect, but that we may still opt to choose alpha and beta error probabilities in accordance with the other two parameters. All we need to specify is the relative seriousness of the alpha and beta error probabilities. Sometimes, protecting against alpha errors will be more important, and sometimes beta errors are associated with a higher cost. Which error type is more serious depends on our research question. For instance, if we invented a new, cheaper treatment of a mental disorder, then we would want to make sure that it is not worse than the older, more expensive treatment. In this case, committing a beta error (i.e., accepting both treatments as equivalent although the cheaper treatment is worse) may be considered more serious than committing an alpha error.

In basic research, both types of errors are normally considered equally serious. Thus, in our present basic-research example we choose

$$q = \alpha/\beta = 1$$

We're all set now to perform our compromise power analysis.

Select:	Type of Power Analysis:	Compromise
	Type of Test:	t-Test (means), two-tailed
		Accuracy mode calculation
Input:	n1:	20
	n2:	20
	Effect size "d":	0.5
	Beta/alpha ratio:	1
Result:	alpha:	0.2957
	Power (1-beta):	0.7043
	Critical t:	t(38) = 1.0603
	Delta:	1.5811

This is still not fantastic, but perhaps it is more reasonable than the alternatives we have. In the end, *you* will have to decide whether it is worth the trouble given these premises.

We have now arrived at the end of our tutorial. If you want to learn more about statistical power analyses, we recommend that you read Cohen's (1988) excellent book.

Referenced pages

Post-hoc power analyses

Post-hoc power analyses are done after you or someone else conducted an experiment.

You have:

- alpha,
- N (the total sample size),
- and the effect size.

You want to know

- the power of a test to detect this effect.

For instance, you tried to replicate a finding that involves a difference between two treatments administered to two different groups of subjects, but failed to find the effect with your sample of 36 subjects (14 in Group 1, and 22 in Group 2). Choose Post-hoc as type of power analysis, and t-Test on means as type of test. Suppose you expect a "medium" effect according to Cohen's effect size conventions between the two groups (delta = .50), and you want to have alpha =.05 for a two-tailed test, you punch in these values (and 14 for n 1, plus 22 for n 2) and click the "Calculate" button to find out that your test's power to detect the specified effect is ridiculously low: 1-beta = .2954.

However, you might want to draw a graph using the Draw graph option to see how the power changes as a function of the effect size you expect, or as a function of the alpha-level you want to risk.

Note that there is a list of tests for fast access to test-specific information.

Compromise Power Analysis

Compromise power analyses represent a novel concept, and only G*Power provides convenient ways to compute them. Thus, if you ever asked yourself "Why G*Power?", this is one possible answer (accuracy of the algorithms and second-to-none flexibility being other candidates for an answer to this question).

You may want to use compromise power analyses primarily in the following two situations:

1. For reasons that are beyond your control (e.g., you are working with clinical populations), your N is too small to satisfy conventional levels of alpha and beta (1-power) given your effect size.
2. Given conventional levels of significance, your N is too large (e.g., you are fitting a model to data aggregated over subjects and items) such that even negligible effects would force you to reject H0.

In compromise power analyses, users specify H_0, H_1 (i.e., the size of the effect to be detected), the test statistic to be used, the maximum possible total sample size, and the ratio

q := beta/alpha which specifies the relative seriousness of both errors (cf. Cohen, 1965, 1988, p. 5). The problem is to calculate an optimum critical value for the test statistic which satisfies beta/alpha = q. This optimum critical value can be regarded as a rational compromise between the demands for a low alpha-risk and a large power level, given a fixed sample size.

Given appropriate subroutines for computing the noncentral distributions of the relevant test statistics (i.e., the exact distributions of the test statistics if H1 is true, cf. Johnson & Kotz, 1970, chap. 28, 30, and 31), it is relatively easy to implement compromise power analyses using an efficient iterative interval dissection algorithm (cf. Press, Flannery, Teukolsky, & Vetterling, 1988, chap. 9).

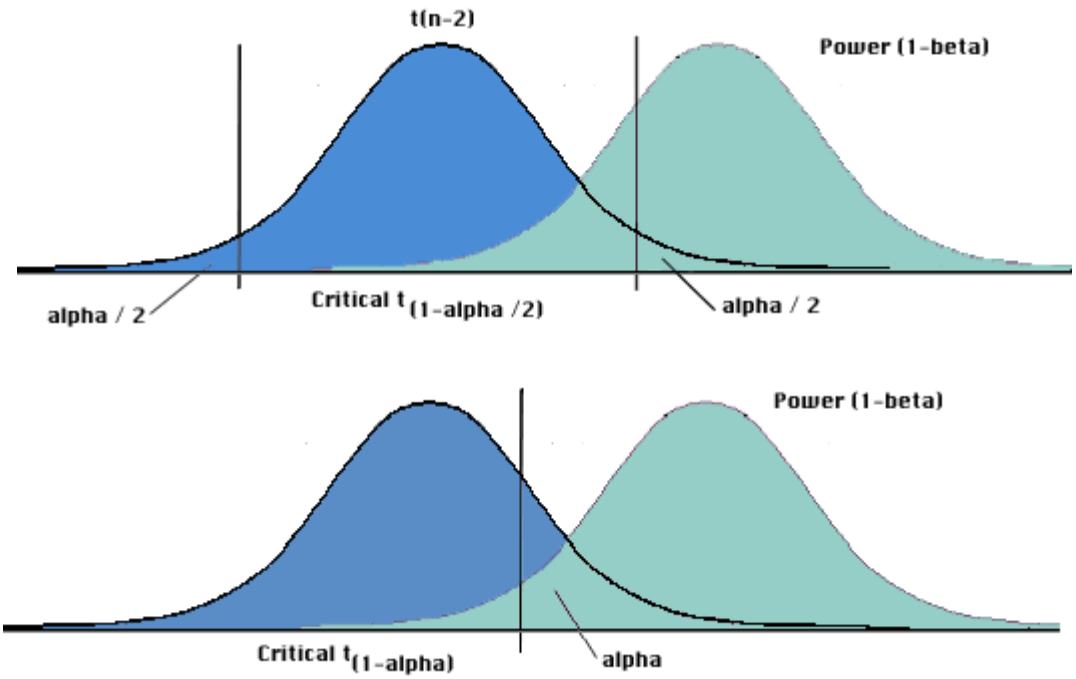
The question is, therefore, why compromise analyses are missing in the currently available power analysis software. The only reason we can think of is that non-standard results may occur, that is, results that are inconsistent with established conventions of statistical inference. Given some fixed sample size, a compromise power analysis could suggest to choose a critical value which corresponds to, say, alpha = beta = .168.

These error probabilities are indeed non-standard, but they may nevertheless be reasonable given the constraints of the research. To illustrate, consider the special case of some substantive hypothesis which implies H_0, for instance, the hypothesis of no interaction. Does it make more sense to choose alpha = beta = .168 rather than to insist on the standard level alpha = .05 associated with beta = .623? Obviously, the standard .05 alpha-level makes no sense in this situation, because it implies a risk of almost two-thirds to accept falsely the hypothesis of interest. Therefore, not only a priori and post-hoc analyses, but also compromise power analyses should be offered routinely by software which is designed to serve as a researcher's tool.

Note that there is a list of tests for fast access to test-specific information

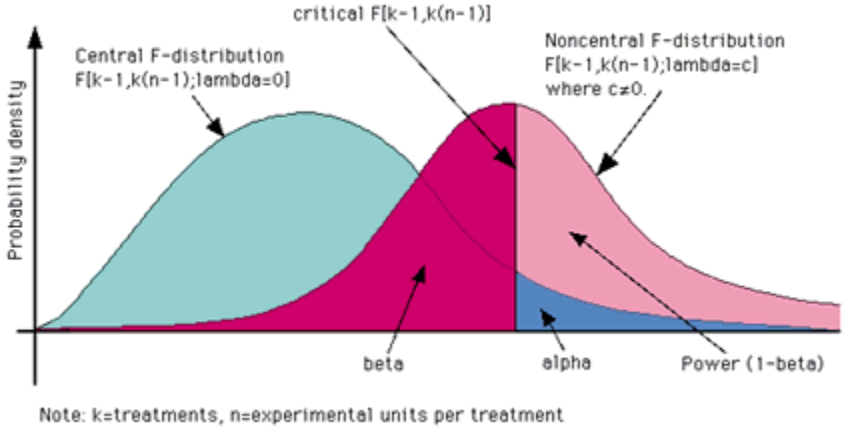
One-Tailed versus Two-Tailed Tests

If you are interested in testing two directional parameter hypotheses against each other (e.g., H0: mu1 <= mu2; H1: mu1 > mu2), a one-tailed test is more appropriate than a two-tailed test. Limiting the region of rejection to one tail of the sampling distributions of H1 provides greater power with respect to an alternative hypothesis in the direction of that tail. The figure below tries to illustrate this.



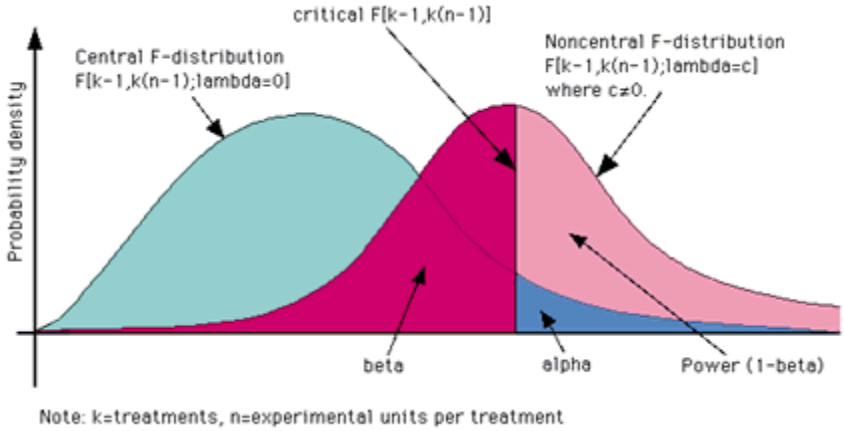
Alpha Error Probability

Alpha is the probability of falsely accepting H1 when in fact H0 is true. The figure below illustrates alpha for an F-test with respect to an alternative hypothesis that corresponds to a so-called "noncentral" F sampling distribution defined by the noncentrality parameter lambda.



Power and the Beta Error Probability

The power of a test is defined as 1-beta, and beta is the probability of falsely accepting H0 when in fact H1 is true. The figure below illustrates beta and the power of an F-test with respect to an alternative hypothesis that corresponds to a so-called "noncentral" F sampling distribution defined by the noncentrality parameter lambda.



Note: k=treatments, n=experimental units per treatment

Effect Size

Effect size can be conceived of as measures of the "distance" between H0 and H1.
Hence, effect size refers to the underlying population rather than a specific sample. In specifying an effect size, researchers define the degree of deviation from H0 that they consider important enough to warrant attention. In other words, effects that are smaller than the specified effect size are considered negligible. The effect size parameter should be specified prior to collecting (or analyzing) the data.

Which choice is considered appropriate depends on

1. the theoretical context of the research,
2. related research results published previously, and
3. cost-benefit considerations in applied research.

Cohen's (1969, 1977, 1988, 1992) effect size measures are well known and his conventions of "small," "medium," and "large" effects proved to be useful. For these reasons, we decided to render G*Power completely compatible with Cohen's measures and to display the effect size conventions appropriate for the type of test selected. These effect size indices and some of the computational procedures to arrive at effect size estimates are described in the context of the tests for which they have been defined. These are:

Cohen (1977, 1988) justifies these levels of effect sizes.

	Index	small	medium	large
t-Test on Means	d	0.20	0.50	0.80
t-Test on Correlations	r	0.10	0.30	0.50
F-Test (ANOVA)	f	0.10	0.25	0.40
F-Test (MCR)	f^2	0.02	0.15	0.35
Chi-Square Test	w	0.10	0.30	0.50

In G*Power, effect size values can either be entered directly or they can be calculated from basic parameters characterizing H1 (e.g., means, variances, and probabilities). To use the

latter option, users must click on the "Calc 'x' " button (x representing the effect size parameter of the test currently selected).

In order to prepare the appropriate G*Power input, it may sometimes be necessary to know the relation between the sample size and the effect size measure on the one hand and the noncentrality parameter of the noncentral distributions on the other hand. We have provided the relation between the sample size, the effect size measures, and the noncentrality parameters on a separate page.

Total Sample Size

In G*Power the total sample size is the number of subjects summed over all groups of the design.

In a t-test on means, the sample size may vary between groups A and B. Note, however, that in this case we want sigma to be approximately equal in both groups. Otherwise, both the t-test and the corresponding G*Power calculations may be misleading because the distributions of the test statistic under H0 and H1 will differ substantially from (central and noncentral) t-distributions.

Another problem could be unequal standard deviations in the populations underlying the two samples. In this case, Cohen (1977) recommended to adjust sigma to sigma' according to

$$\sigma' = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

According to Cohen (1977) the number of participants in both groups A and B must be equal for this correction to be acceptable. If the group sizes vary, then this adjustment is not appropriate.

Please note that you will only arrive at an approximation of the true power of the t-test if the assumption of equal variances is violated. However, Cohen (1977) argues that the approximation will be "adequate" from most purposes.

As a general warning, you should keep in mind that G*Power results are valid if the statistical assumptions underlying the tests are met (e.g., normal distributions and homogeneous variances within cells). Some work has been done on the robustness of these tests, that is, the deviation of actual and nominal alpha error probabilities when the distribution assumptions are not met. However, little is known on a test's power given a misspecified distribution model. Thus, G*Power results may or may not be useful approximations to the true power values in such cases.

In F-Test (ANOVA), we assume that there are an equal number of subjects in each group. If, in a post-hoc or compromise power analysis, the total sample size is not a multiple of the

group size, then the power analysis will be based on the average group size (a noninteger value). G*Power will inform you if this is the case.

Note also that in a priori power analyses, the sample size is usually rounded to the next multiple of the number of groups or cells in your design. This implies that the actual power of your test usually is slightly larger than the power you entered as a parameter.

The Ratio q:= beta/alpha

In a compromise power analysis, the ratio q := beta/alpha specifies the relative seriousness of both types of errors (cf. Cohen, 1965, 1988, p. 5).

For instance, if alpha errors appear twice as serious as beta errors, then you can risk a beta error which is twice as large as alpha, thus q = beta/alpha = 2/1 = 2. This value is what you would then insert as the "beta/alpha ratio" in a compromise power analysis.

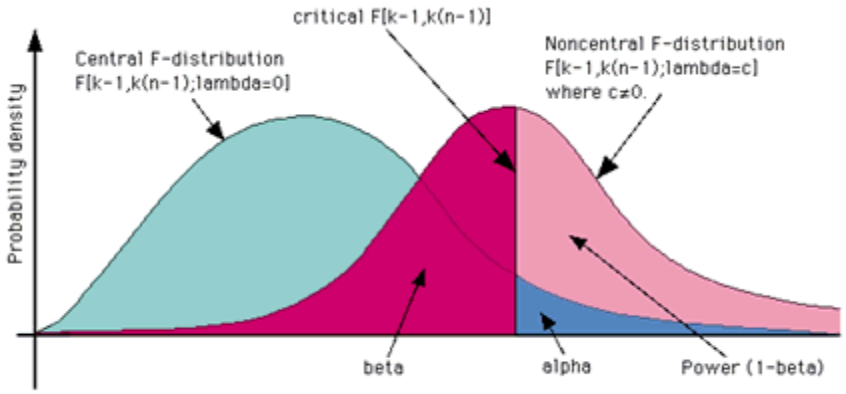
Alternatively, if you think you'd rather not risk committing a beta error (e.g., a beta error is considered three times as important as an alpha error), then you would specify q = beta/alpha = 1/3 = 0.3333.

These choices depend on the different valences you associate with either outcome of the test. However, we suspect that in basic psychological research at least, q = beta/alpha 1/1 = 1 is the rational choice most often.

Given your decision as to the relative seriousness of both types of errors, the problem is to calculate an optimum critical value for the test statistic which satisfies beta/alpha = q. This optimum critical value can be regarded as a rational compromise (hence the term "Compromise power analysis") between the demands for a low alpha-risk and a large power level, given a fixed sample size.

The Noncentrality Parameter

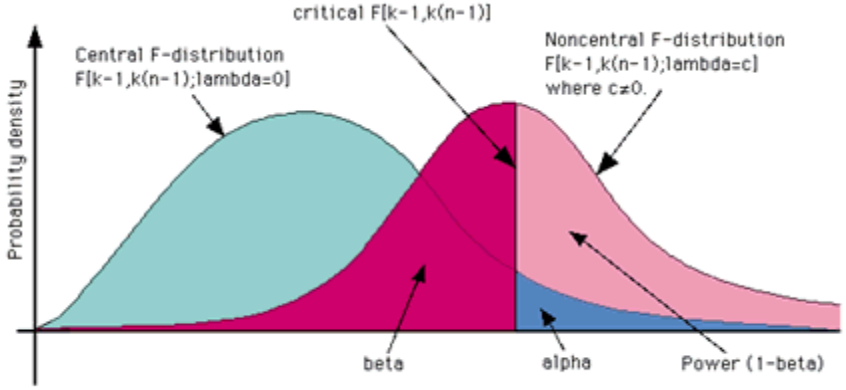
The noncentrality parameter of the t distribution is called *delta*, and that of the F and Chi^2 distributions is called *lambda*. Both measures increase as a function of N and the effect size postulated by H_1. More detailed information about the relation among sample size, effect size, and the noncentrality parameter is also available.



Note: k=treatments, n=experimental units per treatment

The Critical Value

The critical value of the test statistic (z, t, F, and Chi^2 in the cases we look at here) defines the boundary of the rejection region of H_0. Publications of power values and final decisions concerning total sample sizes or critical values should always be based on accuracy mode calculations.



Note: k=treatments, n=experimental units per treatment

The Relation Among Sample Size, Effect Size, and Noncentrality Parameter

It may sometimes be necessary to know the relation between the total sample size and the effect size measure on the one hand and the noncentrality parameter of the noncentral distributions on the other hand. Therefore, we present these relations here for all test procedures offered by G*Power.

t-Test on Means

In t-test on means, the noncentrality parameter delta is

$$\delta = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Where:

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

is Cohen's (1977, 1988, p. 40) effect size parameter for t tests for means, and n1 and n2 are the sample sizes in groups 1 and 2, respectively.

t-Test on Correlations

In t-test on correlations, the noncentrality parameter delta is

$$\delta = \sqrt{\frac{\rho^2}{1-\rho^2} N}$$

Where N is the total sample size (i.e., the number of pairs of values) and rho is the population correlation coefficient according to H_1 (i.e., Cohen's rho, see Cohen, 1977, 1988, p. 77-81).

Other t-Tests

In the Other t-Tests option we used f as an effect size measure (cf. Cohen, 1977, 1988, Chap. 8.2). The relation between delta and f is

$$\delta = f\sqrt{N}$$

F-Test (ANOVA), F-Test (MCR), and Other F-Tests

The standardized effect size measures f or f^2 are also used in power analyses for F-tests (F-Test (ANOVA), F-Test (MCR), and Other F-Tests). Their relation to the noncentrality parameter lambda of the noncentral F distribution is given by Lambda:

$$\lambda = f^2 N, \text{ where } f^2 = \frac{\rho^2}{1-\rho^2}$$

and ρ^2 denotes the coefficient of determination in the population according to H_1 (e.g., Koele, 1982, p. 514). For global ANOVA F-tests, ρ^2 is just eta^2. (**ε^2**)

For special F-tests of main effects or interactions in complex ANOVA-designs, ρ^2 equals the partial eta^2.

Analogously, ρ^2 coincides with the (partial) squared multiple correlation in multiple regression/correlation F-tests (cf. Cohen, 1988, Chap. 9.2.1).

Chi-Square Tests

For Chi-Square tests based on m-cell contingency tables (m in N), Cohen (1977, 1988, Chap. 7) uses

$$w = \sqrt{\sum_{i=1}^{m} \frac{p_{0(i)} - p_{1(i)}}{p_{0(i)}}}$$

as an effect size measure, where $p_{0(i)}$ and $p_{1(i)}$ denote the cell probabilities for the i-th cell according to H_0 and H_1, respectively. Then

$$\lambda = w^2 N$$

is the noncentrality parameter of the noncentral chi-square distribution (Cohen, 1988, p. 549).

Actual Power

When you use G*Power to perform an a priori power analysis, the program calculates the 'exact' sample size for you. Assume that this exact sample size for a t-test is 60.70. Of course, you cannot recruit 60.70 subjects. Therefore, G*Power rounds to the next reasonable integer for your t-test, which would be 62 (two groups of 31 subjects each).

However, 62 is larger than 60.70, and one way to express what this means is to say that, with 62 subjects and all other parameters being equal, your t-test has more power to detect an effect than it would have given the 'exact' number of 60.70 subjects. This 'inflated' power value is displayed as *Actual power*. Note that in this way G*Power guarantees that with the sample size computed for an a priori power analysis, the power of your test is always *at least* the power you specified.

1. **List of Tests**

    - t-Test on Means

        Two group t-test, equal group sizes, equal sigma
        Two group t-test, unequal group sizes, equal sigma
        Two group t-test, equal group sizes, unequal sigma

    - t-Test on Correlations
    - Other t-Tests

        Matched-Pairs t-Test
        One-Sample t-Test
        z-Test
        Wilcoxon-Mann-Whitney U test (plus hints for other nonparametric tests)

- F-Test (ANOVA)
    - ANOVA, Fixed Effects: Single-Factor Designs
    - ANOVA, Fixed Effects: Multi-Factor Designs
    - ANOVA, Planned Comparisons
    - Analysis of Covariance (ANCOVA)
- F-Test (MCR)
    - MCR, One Predictor Set
    - MCR, Two Predictor Sets
- Other F-Tests
    - MANOVA
    - Repeated Measures Designs, Univariate Approach
    - Repeated Measures Designs, Multivariate Approach
- Chi-Square Tests
    - Chi-Square, Goodness-of-Fit-Tests
    - Chi-Square, Contingency Tests

t-Test on Means

In this section, we refer to t-tests which are used to compare independent sample means. H_0 implies that the two means in the population are equal:

$H_0: \mu_1 - \mu_2 = 0$

For matched-pairs t-tests, use the "Other t-Tests" option.

Chose "one-tailed" or "two-tailed," depending on your hypothesis.

We have four examples on this page:

- Two group t-test, equal group sizes, equal sigma
- Unequal group sizes
- Unequal sigma

Two Group t-Test, Equal Group Sizes, Equal Sigma

You have 2 populations A and B which you want to compare with respect to x. Assume that the random variable x is normally distributed with a standard deviation of s (sigma) in both populations. Assume further that the population means of x are mu_A and mu_B in population A and B, respectively. Thus,

H_0: $\mu_A - \mu_B = 0$

H_1: $\mu_A - \mu_B = c, c \neq 0.$

Which total sample size do you need such that the probability of obtaining a t statistic equal to or larger than a critical value is alpha = 0.05 under H_0 and 1-beta = .9 under H_1?

Assume that the difference in means between the groups postulated by your H_1 is equal to one half of the standard deviation, thus d = 0.5 (e.g., $\mu_A = 10$, $\mu_B = 12$, $\sigma = 4$).

Select:	Type of Power Analysis:	A priori	
	Type of Test:	t-Test (means), two-tailed	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Power (1-beta):	.9	
	Effect size "d":	0.5	(To calculate the effect size from μ_A, μ_B, and σ, simply click "Calc d", insert the means and the standard deviation, and click "Calc & Copy".)
Result:	Total sample size:	172	
	Actual power:	0.9032	
	Critical t:	t(170) = 1.9740	
	Delta:	3.2787	

Assume further that you do not have enough money to pay 172 subjects. However, 140 would seam feasible. Which critical t would still result in a "fair" test of your H_1? We use a compromise power analysis to compute an optimum critical value for the test statistic which satisfies the ratio q := beta/alpha. This optimum critical value can be regarded as a rational compromise between the demands for a low a-risk and a large power level, given a fixed sample size.

Select:	Type of Power Analysis:	Compromise	
	Type of Test:	t-Test (means), two-tailed	
		Accuracy mode calculation	
Input:	n1:	70	
	n2:	70	
	Effect size "d":	0.5	
	Beta/alpha ratio:	2	(That is, we are willing to commit a beta error twice as large as our alpha error.)

Result:	alpha:	0.0670	
	Power (1-beta):	0.8661	
	Critical t:	t(138) = 1.8465	
	Delta:	2.9580	

Two Group t-Test, Unequal Group Sizes, Equal Sigma

We have done a study in which, for some reasons, the group sizes are not equal. In Group A we have 24 subjects; in Group B we have 33. What is the power of the t-Test comparing the means of both groups, and how much power have we lost due to the unequal group sizes?

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	t-Test (means), one-tailed (This time assume that we know the direction of the difference between the groups.)	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Effect size "d":	0.8	(We expect "large" effects according to the effect size conventions of Cohen, 1977.)
	n1:	24	
	n2:	33	
Result:	Power (1-beta):	0.9032	
	Critical t:	t(55) = 1.6730	
	Delta:	2.9821	

Two Group t-Test, Equal Group Sizes, Unequal Sigma

What do you do if $\sigma_A \neq \sigma_B$?

This is not normally a problem because the t-test is known to be quite robust, at least as long as the groups sizes are equal. Cohen (1977, p. 44) suggests to adjust sigma to sigma':

$$\sigma' = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

Simply use sigma' instead of sigma to calculate the effect size using the "Calc 'd'" option, then proceed as in the examples given above.

A word of caution is in order, however: The power values computed by G*Power will only be approximations in this case. Computer simulation results on the appropriateness of this approximation are not yet available

t-Test on Correlations

In this section, we refer to t-tests which are used to evaluate the H_0 that a product moment correlation in the population is zero

$H_0: r = 0$, and $H_1: r = c$, $c \neq 0$.

Chose "one-tailed" or "two-tailed" depending on your hypothesis. The effect size index is r, the correlation in the population itself.

Example

We want to know how many subjects it takes to detect r = .30 in the population, given alpha = beta = .05. Thus,

	H_0:	r = 0	
	H_1:	r = .30	
Select:	Type of Power Analysis:	A priori	
	Type of Test:	t-Test (correlations), one-tailed	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Power (1-beta):	.95	
	Effect size "r":	0.3	(You can calculate the effect size from r^2;simply click "Calc 'r'", insert the value for r^2, and click "Calc & Copy"; but this is obviously relatively trivial)
Result:	Total sample size:	111	
	Actual power:	0.9503	
	Critical t:	t(109) = 1.6590	
	Delta:	2.3408	

Other t-Tests

With this option, we can perform power analyses for *any* test that depends on the t-distribution. All parameters of the noncentral t-distribution can be manipulated independently. Note that with "Other t-Tests" you cannot do a priori power analyses, the

reason being that there is no definite association between N and df (the degrees of freedom). You need to tell G*Power the values of both N and df explicitly.

We consider 3 examples here:

- Matched-pairs t-tests
- One-sample t-tests
- z-Tests

In addition, we give hints on how to do power analyses for nonparametric tests such as the

- Wilcoxon-Mann-Whitney U test

Matched-Pairs t-Tests

In t-tests for matched pairs, we have differences of the values from N matched pairs,

$y_1 = x_{A1} - x_{B1}$

: : :

$y_N = x_{A1} - x_{B1}$

The H_0 we test is that the pairs do not differ, that is, the population mean μ_Y of the differences is zero. More formally,

$$H_0: \quad \mu_Y = 0$$
$$H_1: \quad \mu_Y = c,\ c \neq 0.$$

When computing the standard deviation σ_Y of the distribution of differences, we need to take into account the correlation r between A and B in the population:

$$\sigma_Y = \sqrt{\sigma_A^2 + \sigma_B^2 - 2r\sigma_A\sigma_B}$$

where σ_A and σ_B are the standard deviations of x in the populations A and B, respectively, and r is the population correlation between A and B as paired.

In matched-pairs t-tests, N is the total sample size (i.e., total number of pairs), df = N-1, and the effect size is:

$$f = \frac{\mu_Y}{\sigma_Y}$$

Where μ_Y is the difference between the means as specified by H_1.

Example

Assume that we are faced with a repeated measures design in which the same subject is observed under each of two treatments. We have data from 40 subjects. Previous research has shown that the standard deviation of the differences is approximately 20. We consider mean differences of 8 or larger as important. Thus, the effect size we need to enter is f = 8/20 = 0.4. We fix alpha at 0.05.

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	Other t-Tests, two-tailed.	
		Accuracy mode calculation	
Input:	Alpha:	0.05	
	Effect size "f":	0.4	
	N:	40	
	df:	39	(Df = N-1 in matched pairs t-tests.)
Result:	Power (1-beta):	0.6940	
	Critical t:	t(39) = 2.0227	
	Delta:	2.5298	

As we said before, you cannot perform a priori power analyses directly, but you can, of course, perform repeated post-hoc power analyses, adjusting N and df until you arrive at the power value you desire. For instance, if you want, in the above example, the power to be .95, you simply increase N and df (= N-1) until the power is as close as possible to .95 (which will be the case with N = 84 and df = 83 for the present example).

One-Sample t-Tests

We want to compare the mean of a population from which we sample to a constant c. The effect size index d is computed according to

$$f = \frac{|\mu - c|}{\sigma}$$

where μ and σ are the mean and the standard deviation in the population, respectively. As Cohen (1977, p. 46) writes, the interpretations of f (Cohen's d3') as well as the effect size conventions are identical to those for d.

N is the total sample size, and df = N-1. Thus, we're all set to do this power analysis analogously to the one for matched pairs t-tests (above).

z-Tests

We can easily do power analyses for z-tests with G*Power because, as df approaches infinity, the t-distribution asymptotically converges with the normal distribution with mean

$$d = f\sqrt{N}$$

and standard deviation 1. In other words, the critical t(32000) is virtually identical to the critical z value. As the effect size index, we use:

$$f = \frac{|\mu_1 - \mu_2|}{\sigma}$$

N is again the total sample size. μ_1 and μ_2 are the the means in populations 1 and 2, respectively.

However, for df we specify df = 32000. That's it.

In this way, you can perform power analyses for all sorts of z-tests (e.g., approximate z-test for hypotheses about binomial probabilities, comparisons of correlations between two different samples etc.). Note, however, that N and f have to be specified such that they make sense for the test you want to consider.

Nonparametric Tests

Some variants of power analyses for nonparametric tests can be conducted by adjusting the result obtained for the corresponding parametric test (cf. Bredenkamp, 1980; Singer, Lovie & Lovie, 1986).

For example, an a priori power analysis for the **Wilcoxon-Mann-Whitney U test** can be conducted by first performing an a priori power analysis for the t-test for means. If the t-test model is valid, and N_t designates the sample size necessary for the t-test to achieve some given power (1-beta), then the sample size $N_u = N_t/A.R.E.$ yields approximately the same power for the U test.

A.R.E. denotes the asymptotic relative efficiency (or Pitman efficiency) of the U test relative to the t-test which is 3/pi = .955 (see Lehmann, 1975).

The same procedure may often be used to approximate the power of randomization tests (Onghena, 1994, pp. 144-176). In this case, the A.R.E. of the randomization test relative to the corresponding parametric test is 1. For power analyses in randomization tests which do not have a corresponding parametric test, special computer software is in preparation (Onghena, 1994; Onghena & Van Damme, 1994).

F-Test for Analyses of Variance (ANOVA)

We can easily do power analyses for single-factor and multi-factor experiments. In G*Power, you select

F-Test (ANOVA), Global for

- ANOVA, fixed effects: Single-factor designs, or

F-Test (ANOVA), Special for

- ANOVA, fixed effects: Multi-factor designs,
- ANOVA, fixed effects: Planned comparisons, and
- Analyses of covariance (ANCOVA).

Note: For F-Test (ANOVA) (as well as for F-Test (MCR)), you can choose whether you want to perform power analyses for global (i.e., omnibus) tests or for special tests. Global test is the default option. This test refers to the H_0 that all means in the design are equal (ANOVA) or that all regression coefficients (next to the additive constant) are zero (MCR).

Random effects ANOVAs and mixed effects ANOVAs are not considered. We may add them at a later time, however. A discussion of how to do power analyses for repeated measures ANOVAs and MANOVAs can be found in the Other F-Tests section.

For the ANOVA designs, we will use the effect size index f (Cohen, 1977). The relation of f to the noncentrality parameter lambda is given by lambda = f^2 * N.

ANOVA, fixed Effects: Single-Factor Design

H_0 is that the population means in k conditions are identical. More formally

$$H_0: \quad \sum_{i=1}^{k}(m_i - m)^2 = 0$$

$$H_1: \quad \sum_{i=1}^{k}(m_i - m)^2 = c; c > 0$$

where k is the number of conditions, m_i is the mean in condition i, and m is the grand mean.

You can compute the effect size index from group means and s which is assumed to be constant across groups by clicking on "Calc 'f'". We will spare you the formula behind that. (Just that much: You can save a lot of time when you use the "Calc 'f'" option.)

Example

We compare 10 groups, and we have reason to expect a "medium" effect size (f = .25).

How many subjects do we need?

Select:	Type of Power Analysis:	A priori	
	Type of Test:	F-Test (ANOVA), Global	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Power (1-beta):	.95	
	Effect size "f":	.25	(Note that in G*Power you can compute f directly from the population means and the population standard deviation sigma; simply click "Calc f" after selecting "F-Test (ANOVA), Global".)
	Groups:	10	(This is a new item that pops up only when you do power analyses for ANOVAs.)
Result:	Total sample size:	390	
	Actual power:	0.9524	
	Critical F:	F(9,380) = 1.9045	
	Lambda:	24.1237	

Thus, we need 39 subjects in each of the 10 groups. What if we had only 200 subjects available? Assuming that both alpha and beta are equally serious (i.e., the ratio q := beta/alpha = 1) which probably is the default in basic research, we can compute the following compromise power analysis:

Input:	Total sample size:	200
	Effect size "f":	0.25
	Beta/alpha ratio:	1
Result:	Alpha:	0.1592
	Power (1-beta):	0.8408
	Critical F:	1.4762
	Lambda:	12.50000

ANOVA, Fixed Effects: Multi-Factor Designs

In multi-factor designs, we want to determine separately the power for the main effects and for the interactions involved.

For main effects, the H0, the interpretation of the effect size index f, and the procedure are basically the same as for single-factor designs. The major difference is that the numerator df (df = degrees of freedom) are reduced relative to a single-factor design because other factors have to be taken into account.

Thus, the only new part is that you need to specify, as the "Groups", all cells of your multi-factor design, and as numerator df (the new item for this type of power analysis) you enter i - 1, where i represents the levels of the specific factor to be tested.

Note that there may be considerable differences between the power analysis values as determined by G*Power and those determined according to the "approximations" suggested by Cohen (1977, p. 365). G*Power is correct, while Cohen's approximations systematically underestimate the power.

This problem with Cohen's approximation method is described in more detail in the context of our description of the accuracy of the algorithms used in G*Power.

Example 1

Let us assume we have a 3 x 5 design in which Factor A has 3 levels and Factor B has 5 levels. We first want to compute a power analysis for main effect A:

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	F-Test (ANOVA), Special	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Effect size "f":	.25	
	Total sample size:	270	
	Groups:	15	(That is, all cells in your 3x5 design, thus 15 because there are 3 * 5 levels for Factor A and Factor B, respectively.)
	Numerator DF:	2	(Factor A has 3 levels, thus the test of the main effect of Factor A has 3-1=2 df.)
Result:	Power (1-beta):	0.9637 (Yeah!)	
	Critical F:	F(2,255) = 3.0312	
	Lambda:	16.8750	

Example 2

Assume that your H_0 states that there is no interaction between A and B. How do you perform a power analysis for this case?

The number of Groups is again 3 * 5 = 15. The numerator df is (3-1) * (5-1) = 8. If you enter these values and leave the rest as it was for the main effect, then this is the result:

Result:	Power (1-beta):	0.8396
	Critical F:	F(8,255) = 1.9748
	Lambda:	16.8750

To extend this further, assume that you have a 3 x 4 x 6 design with factors A, B, and C. You test the main and interaction effects of this design using the following values (assuming alpha = .05, effect size f = .25, and a total sample size of 288):

Effects	Groups:	Numerator df:	Power (1-beta):
A	72	2	0.9727
B	72	3	0.9557
C	72	5	0.9197
A x B	72	6 (2 * 3)	0.9013
A x C	72	10 (2 * 5)	0.8290
B x C	72	15 (3 * 5)	0.7469
A x B x C	72	30 (2 * 3 * 5)	0.5630

Example 3

So far, we have limited the discussion to post-hoc power analyses. However, in planning a multi-factor design, we want to know how many participants we need to recruit for our experiment. How do we proceed in that case?

Essentially, our decisions involve the following steps:

Step 1:

We compute a priori power analyses for the statistical tests of the effects of all factors and interactions that are interesting from a theoretical point of view. We ignore all other factors and interactions.

Step 2:

Case 1:

One factor or interaction (henceforth our critical factor or interaction) is more important for our research question than all other factors or interactions. Two alternatives are possible:

1. Our critical factor or interaction is the one associated with the largest sample size as determined in Step 1. We use that sample size. As a rule, we will be on the safe side with all other relevant factors and interactions.

2. Our critical factor or interaction is the *not* the one associated with the largest sample size as determined in Step 1. We need to do some more work:

   We take the sample size as suggested for the critical factor or interaction and perform post-hoc power analyses for *all other* factors or interactions that are theoretically relevant. If we can live with the error probabilities associated with the statistical tests of the effects of these factors, then we are done.

   If we are not happy with the error probabilities, we try to increase the sample size up to the level at which we find both the error probabilities and the resource demands acceptable.

Case 2:

All factors and interactions are equally important. We use the largest sample size as determined in Step 1. As a rule, we will be on the safe side with all other relevant factors and interactions. (Note that Case 2 and Case 1.1 lead to the same result.)

Let us return to our 3 x 5 design in which Factor A has 3 levels and Factor B has 5 levels. For simplicity, we assume that we want to detect effects of size f = .40 for the two main effects and the interaction given alpha = beta = .05. The relevant a priori power analyses suggest the following sample sizes:

Effects	Groups:	Numerator df:	Total sample size:
A	15	2	105*
B	15	4	135*
A x B	15	8	165*

* Note that the total sample size values produced by G*Power are somewhat smaller. However, we use the next largest number that can be divided by 15 because our design has 15 cells and we wish to assure that the n's in all cells are equal

In a Case 2 situation, we would need a total sample size of 165. Given that our assumptions about alpha and the effect size remained unchanged, a total sample size of 165 would imply power values > .99 for tests of the effects of Factors A and B. This result is a dream come true!

Alternatively, let us assume a Case 1 situation in which Factor B is the most important factor from a theoretical point of view. What would the implications be of accepting 135 as the total sample size?

Given that our assumptions about alpha and the effect size remained unchanged, a total sample size of 135 would imply power values of . 9890 and .9195 for tests of the effects of Factors A and the A x B interaction, respectively. This result is certainly acceptable and we may decide to use 135 as the total sample size.

ANOVA, Planned Comparisons

With planned comparisons, the H0 is that the contrasts among the means do not explain, in the dependent variable, any variance which has not already been accounted for by other sources of the effect. The effect size f is defined as:

$$f = \sqrt{\frac{R_p^2}{1-R_p^2}}$$

where R^2_p is the partial multiple correlation between the dependent variable and the variable(s) coding the contrast among the means. In G*Power, click "Calc F" after selecting "F-Test (ANOVA), Special" to calculate f from the partial multiple correlation (referred to as partial eta-square in G*Power).

For a power analysis, it does not matter whether the contrasts are orthogonal or not. However, note that f does not only depend on the population means but also on the correlations among the contrast variables.

Example

Assume you have a Factor A with 4 levels. We want to determine whether the effect of A on our dependent variable Y is linear, but not quadratic or cubic. You can code A into 4-1=3 orthogonal contrast variables as follows.

x1, linear	-3	-1	1	3
x2, quadratic	1	-1	-1	1
x3, cubic	-1	3	-3	1

Assume that your H_1 specifies that $R^2_p = .20$ for the linear contrast (x1). Thus, f = .1667.

If you have 60 subjects in your experiment (i.e., 15 in each of the four groups), what are the alpha and beta error probabilities if both types of errors are equally important (i.e., the ratio q := beta/alpha = 1)?

Select:	Type of Power Analysis:	Compromise
	Type of Test:	F-Test (ANOVA), Special
		Accuracy mode calculation
Input:	Total sample size:	60
	Effect size "f":	0.2857
	Beta/alpha ratio:	1
Result:	Alpha:	0.1888
	Power (1-beta):	0.8112
	Critical F:	F(1,56) = 1.7700
	Lambda:	4.8975

Analyses of Covariance (ANCOVA)

In an analysis of covariance, we replace a dependent variable Y by a corrected dependent variable Y' which we arrive at by partialling out the linear relation between Y and a set X = (X_a, ... , X_q) of q covariates X_i (Cohen, 1977, p. 379).

$Y' = Y - b_i (X_i - m_{(Xi)})$, i = 1 ... q,

where

b_i is the regression weight of Y on X_i (b_i is constant across all populations),

X_i is the covariate i (i.e., X_i may be different in each of the populations), and

$m_{(Xi)}$ is the grand population mean of the concomitant variable or covariate i.

In other words, covariate X_i differs in each of the populations we look at, but its relation to Y and, hence, its regression weight b_i is the same in all of those populations.

The analysis of covariance is essentially an analysis of variance of the Y' measures. However, we need to adjust the denominator df, which is why we need to select the "F-Test (ANOVA), Special" option.

If k is the number of cells of your design, choose

groups = k + q (q is the number of covariates in your design).

In this way, the denominator df are reduced appropriately because G*Power assumes that

denominator df = N - groups.

If the correlation between Y and the covariates is substantial, then the power of your statistical test is increased. This is so because the within-population standard deviation σ_Y' in the denominator of the F ratio is smaller than σ_Y.

Specifically, where r is the (multiple) population correlation between Y' and Y, we find that

$$\sigma' = \sigma_Y \sqrt{1 - r^2}$$

The numerator does not decrease correspondingly. It may even increase.

Example

Assume a 2 x 3 design. A covariate X has been partialled out of a dependent variable Y'. We want to detect 'large' effects (f = .40) according to Cohen's effect size conventions for Factor B which has 3 levels. We had 60 subjects, and we decide that alpha = .05. What is the power of the F-test in this situation?

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	F-Test (ANOVA), Special Accuracy mode calculation	
Input:	Alpha:	.05	
	Effect size "f":	.40	
	Total sample size:	60	
	Groups:	7	(That is, 2 * 3 + 1 = 7.)
	Numerator DF:	2	(Factor B has 3 levels, thus the test of the main effect of Factor B has 3-1=2 df.)
Result:	Power (1-beta):	0.7740	
	Critical F:	F(2,53) = 3.1716	
	Lambda:	9.6000	

Other F-Tests

The "Other F-Tests" option is very powerful, but you have to know what you are doing in order to use it.

It is provided to enable you to do power analyses for *any* test based on the F-distribution which is not covered by the F-Test (ANOVA) item and the F-Test (MCR) item. Of course, you can do power analyses for standard ANOVAs and MCRs using the "Other F-Tests" option, but it is usually much more convenient (and less error-prone) to use the options we provided for these standard cases directly.

"Other F-Tests" is similar to the "Other t-Tests" item in that you can (in fact: must) specify the sample size and the degrees of freedom (both numerator and denominator) independently. Although this is important for a number of F-based tests, we think the two most important classes are

- MANOVAs,
- repeated measures analyses according to the so-called univariate approach, and
- multivariate repeated measures analyses.

In this section, we briefly sketch how you can use G*Power to perform power analyses for these types of tests.

Before we begin, please note that, as with "Other t-Tests," you cannot do a priori power analyses directly, but you can of course do repeated post-hoc power analyses, adjusting N and (simultaneously!) the df's until you arrive at the power value you desire.

MANOVAs

For reasons given in Bredenkamp and Erdfelder (1985), Olson (1976) and Stevens (1979), we prefer the Pillai-Bartlett V criterion as a multivariate test statistic. It is well known that under H_0 the transformed V statistic

$$F = \frac{\frac{V_h / s_h}{df_1}}{\frac{1 - V_h / s_h}{df_2}}$$

is approximately F(df1, df2) distributed and

V_h is the Pillai-Bartlett V for the effect to be tested,

$s_h = \min(p, n_h)$,

p = the number of dependent variables,

n_h = the number of predictors for the effect to be tested,

$df_1 = p * n_h$ (numerator degrees of freedom),

$df_2 = s_h * (N-k-p+s_h)$, and

N is the total number of subjects summed across all k groups of the design (see Pillai & Mijares, 1959; Olson, 1976).

V_h/s_h varies between 0 and 1 and can be regarded as a multivariate R^2 or eta^2.

A convenient measure for the multivariate effect size in the underlying population is

$$f^2_{mult} = \frac{\frac{V_h}{s_h}}{\frac{1-V_h}{s_h}} = \frac{V_h}{s_h - V_h}$$

where V(h) denotes the Pillai-Bartlett V in the underlying population, not in a particular sample.

Pillai and Jayachandran (1967) published exact power tables for small values of f^2_{mult} and small values of p. Stevens (1980) reported computer simulation results for a larger range of effect sizes and p values. We have compared these power tables to the power values computed by means of G*Power's "Other F-Tests" option assuming that, under H_1, the F transformation of the V statistic is approximately noncentral F(df1, df2, lambda) distributed with

numerator $df_1 = p * n_h$, and denominator $df_2 = s_h * (N-k-p+s_h)$, and the noncentrality parameter $= s_h * N * f^2_{mult}$.

In general, we found a quite good agreement, with perhaps a slight tendency to overestimate the power using the proposed approximation. Nevertheless, the approximation may often be sufficiently precise.

Note that the relation between sample size, effect size, and the noncentrality parameter lambda for MANOVAs is different from that for ANOVAs where lambda = $f^2 * N$.

For a global MANOVA test we find that $n_h = k-1$, and for special MANOVA tests we find that n_h = the number of predictors of the effect to be tested. For instance, in a MANOVA based on an AxB design, A having a levels and B having b levels, we find

$n_h = a-1$ for the main effect of A,
$n_h = b-1$ for the main effect of B, and
$n_h = (a-1)(b-1)$ for the multivariate interaction.

Example

Assume that we have a k=3 group MANOVA design with a total sample size of 3 * 20 = 60 subjects, p = 2 dependent variables, and our effect size is $f^2_{mult} = .15$.

This is how we calculate the power for this test:

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	Other F-Tests	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Effect size "f2":	0.1500	
	N:	120	Note that we enter N = (2 * total sample size) and not simply the plain total sample size because lambda $= s_h * N * f^2$ $= 2 * 60 * 0.15$ $= 18.$
	Numerator DF:	4	$p * n_h = 2 * 2 = 4$
	Denominator DF:	114	$s_h * (N-k-p+s_h)$ $= 2 * (60-3-2+2)$ $= 114$
Result:	Power (1-beta):	0.9330	
	Critical F:	F(4,114) = 2.4513	
	Lambda:	18.0000	

Repeated Measures Designs, So-Called Univariate Approach

To illustrate power analyses for the so-called univariate approach to repeated measures designs, we use an A x B design in which A is a between-subjects factor and B is a within-subject factor. Factors A and B have a and b levels, respectively.

Example

Assume that we have

a = 2 levels of Factor A,
b = 4 levels of Factor B, and
N = 2 * 10 = 20.

Between-Subjects Effect

The test for the between-subjects main effect of Factor A has

Numerator df = a - 1 = 2 - 1 = 1, and
denominator df = N - a = 20 - 2 = 18.

The power of the between-subjects effect depends on the number of repeated measures in our design, and on the correlation between the levels of the repeated measures. This can be seen when looking at the noncentrality parameter lambda for this case:

$$\eta = N \frac{m}{(1+(m-1)\rho)f^2}$$

Where:

N: is the total number of subjects,
m: is the number of levels of the repeated measures factor,
ρ: is the population correlation between the individual levels of the repeated measures factor, and
f^2 : is just the effect size for between-subject designs as used by Cohen (1977, 1988), that is, the ratio of effect variance to the error variance within cells.

Obviously, if there is no repeated measures factor (i.e., m = 1), then the above equation reduces to:

$$\eta = Nf^2,$$

which is just the noncentrality parameter G*Power uses in F-Tests (ANOVA).

Let us assume that we want to detect a "medium" effect according to Cohen's effect size conventions for ANOVA F-tests. Thus,

$f = .25$ and therefore $f^2 = 0.0625$.

Next we assume that the correlation between the levels of the repeated measures Factor B is .75. As a consequence of the so-called sphericity assumption, we must assume that the correlation between all possible pairs of repeated measurements is identical. If sphericity is not given in our data, then we have a problem. We will deal with the sphericity problem below.

Given the above assumptions, the noncentrality parameter for our design is

$$\eta = N \frac{m}{(1+(m-1)\rho)f^2} = 20 \frac{4}{(1+(4-1)0.75)0.0625} = 1.538$$

A technical point to be aware of is that G*Power computes the noncentrality parameter lambda as

$$\eta = Nf^2$$

Where f^2 is the label of the effect size slot when you select "Other F-Tests". Therefore, we need to enter

$$\frac{m}{(1+(m-1)\rho)f^2} = 0.0769$$

As the effect size term to be used in our computations:

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	Other F-Tests	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Effect size "f2":	0.0769	
	N:	20	2 * 10 = 20
	Numerator DF:	1	a - 1 = 2 - 1 = 1
	Denominator DF:	18	N - a = 20 - 2 = 18
Result:	Power (1-beta):	0.2170	
	Critical F:	F(1,18) = 4.4139	
	Lambda:	1.5380	

Within-Subject Effect

The test for the with-subjects main effect of Factor B has

Numerator df = b - 1 = 4 - 1 = 3, and
denominator df = (N - a) * (b - 1) = 18 * 3 = 54.

The power of the within-subject effect depends on the correlation between the levels of the repeated measures. This can be seen when looking at the noncentrality parameter lambda for this case:

$$\eta = \frac{Nmf^2}{1-\rho}$$

Where:

N: is the total number of subjects,
m: is the number of levels of the repeated measures factor,
ρ: is the population correlation between the individual levels of the repeated measures effect, and

f^2 : is just the effect size for between-subject designs as used by Cohen (1977, 1988), that is, the ratio of effect variance to the error variance within cells.

Let us assume that we want to detect an effect of the same size as before (i.e., $f^2 = 0.0625$). The correlation between the levels of the repeated measures Factor B is .75 (see above). As a consequence of the so-called sphericity assumption, we must again assume that the correlation between all possible pairs of repeated measurements is identical. If it is not, then we have a problem. We will deal with this problem further on.

Given the above assumptions, the noncentrality parameter for our design is

$$\eta = \frac{Nmf^2}{1-\rho} = 20 * 1 = 20.$$

As before, the technical point to be aware of is that G*Power computes the noncentrality parameter lambda as

$$\eta = Nf^2$$

Where f^2 is the label of the effect size slot when you select "Other F-Tests". Therefore, we need to enter

$$\frac{mf^2}{1-\rho} = 1$$

As the effect size term to be used in our computations.

We can now proceed as before.

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	Other F-Tests	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Effect size "f2":	1	
	N:	20	2 * 10 = 20
	Numerator DF:	3	b - 1 = 4 - 1 = 3
	Denominator DF:	54	(N - a) * (b - 1) = 18 * 3 = 54
Result:	Power (1-beta):	0.9646	
	Critical F:	F(3,54) = 2.7758	
	Lambda:	20.0000	

Interaction of Between-Subjects and Within-Subject Effect

The procedure for the within-between interaction test is basically identical to the procedure for within-subject effects. The formulae for the degrees of freedom in this case are

Numerator df = (a - 1) * (b - 1) = (2 - 1) * (4 - 1) = 3, and
denominator df = (N - a) * (b - 1) = 18 * 3 = 54.

The power of the interaction effect also depends on the correlation between the levels of the repeated measures. The noncentrality parameter lambda for this case is:

$$\eta = \frac{Nmf^2}{1-\rho}$$

Where:

N: is the total number of subjects,
m: is the number of levels of the repeated measures factor,
ρ: is the population correlation between the individual levels of the repeated measures effect, and
f^2 : is just the effect size for between-subject designs as used by Cohen (1977, 1988), that is, the ratio of effect variance to the error variance within cells.

As before, the technical point to be aware of is that G*Power computes the noncentrality parameter lambda as:

$$\eta = Nf^2$$

Where f^2 is the label of the effect size slot when you select "Other F-Tests". Therefore, we need to enter

$$\frac{mf^2}{1-\rho}$$

As the effect size term to be used in our computations. We can now proceed as before.

Problems Resulting from the Sphericity Assumption

In the so-called univariate approach, we must assume that all repeated measures have equal variances and are correlated equally with each other. This is often referred to as the sphericity assumption.

If sphericity is met, then analytic results for the power calculations of univariate repeated measures tests such as those illustrated above are available.

Unfortunately, sphericity is a very strong assumption which is very likely violated in many situations (see O'Brien & Kaiser, 1985). For instance, if five levels of a repeated measures factor represent successive points in time, then it is almost certain that the correlation of the measures taken at the first and the second level is larger than the correlation between the first and the fifth level.

If sphericity is not met, then the tests of main effects and interactions involving the within-subject factors occur at an artificially increased Type I error rate because the resulting F values are artificially inflated.

One way to react to this problem is to apply the corrected univariate tests in which the Geisser-Greenhouse or the Huynh-Feldt estimate of epsilon are used to provide improved Type I error rates.

Epsilon is 1 if sphericity is met, whereas without sphericity, we find that $\frac{1}{n} \leq \varepsilon \leq 1$

(where n represents the size of the associated residual covariance matrix, e.g., n = k-1 for a within-subject main effect with k levels).

In order to take violations of sphericity into account, both the numerator and the denominator degrees of freedom of the F test must be multiplied by epsilon, and the significance of the F ratio must be evaluated with the new degrees of freedom. The Geisser-Greenhouse epsilon tends to be relatively conservative, which is a property the Huynh-Feldt epsilon tries to correct.

How can we assess the power of corrected univariate tests?

Muller and Barton (1989) have proposed an approximation to the power of the Geisser-Greenhouse or Huynh-Feldt-corrected test. Following their approach, we compute

numerator $df_{(c)}$ = (numerator df)*(estimate of epsilon),

denominator $df_{(c)}$ = (denominator df)*(estimate of epsilon), and

lambda(c) = lambda*(estimate of epsilon).

Assume that sphericity is violated and we find that the estimate of epsilon = .6. What is the effect of this violation on the power of our within-subject test? Using the within-subject example above, we compute

numerator df(c) = 3 * .6 = 1.8

denominator df(c) = 54 * .6 = 32.4, and

lambda(c) = 20 * 0.6 = 12.

We can now reevaluate the power of this test. Note that G*Power expects df values to be integers, which is why we need to enter numerator df = 2 and denominator df = 33. We also need to make adjustments to what we enter as the effect size index in order to ensure proper calculation of lambda. More precisely, we enter 1 * 0.6 = 0.6 as the effect size index for our within-subject effect. In that way, we arrive at lambda = 20 * 0.6 = 12.

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	Other F-Tests	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Effect size "f2":	0.6	
	N:	20	2 * 10
	Numerator DF:	2	round(3 * .6) = 2
	Denominator DF:	33	round(54 * .6) = 33
Result:	Power (1-beta):	0.8506	
	Critical F:	F(2,33) = 3.2849	
	Lambda:	12.0000	

Thus, the power of the corrected test is clearly less than the power of the uncorrected test in which the sphericity problem is simply ignored.

In essence, if we insist in using the so-called univariate approach to repeated measure analyses, then we face a choice between two unattractive alternatives: Either we ignore the (non)sphericity problem (and accept that we commit an error by testing at an artificially increased Type I error rate), or accept a reduction of the power of our statistical tests.

Repeated Measures Designs, Multivariate Approach

Repeated measures designs may also be analyzed using a multivariate approach. One advantage of this approach is that MANOVAs do not require the sphericity assumption to be met (which appears to be violated quite often, see O'Brien & Kaiser, 1985).

Using the MANOVA approach, we treat the levels of the within-subject factor as different dependent variables. The univariate A x B design discussed above thus is regarded as a multivariate design with between-subjects factor A and p = b dependent variables. Let us consider the same design as above, but from a multivariate perspective.

Example

Between-Subjects Effect

First, the result for the between-subjects is identical to the result we received for the univariate approach. We can therefore proceed quickly to the

Within-Subjects Effect

The F-test for the within-subject Factor B has

Numerator df = b - 1 = 4 - 1 = 3, and

Denominator df = $s_{(h)}$ * (N-k-p+$s_{(h)}$) = 1 * (20-2-3+1) = 16.

Where

N is the number of participants,

k is the number of groups in the design (Factor A has 2 levels),

p is the number of dependent variables (The 4 levels of the within-subject factor B are recoded into 4 - 1 = 3 dependent variables using appropriate contrast variables. The recoded variables may then represent, for instance, linear, quadratic, and cubic trends in the repeated measurement. See O'Brien and & Kaiser,1985 , for details.).

The noncentrality parameter lambda for this case is identical to the one used in the univerate approach to repeated measures analyses (see Davidson, 1972):

$$\eta = \frac{Nmf^2}{1-\rho}$$

Where

N is the total number of subjects,

m is the number of levels of the repeated measures factor,

ρ is the population correlation between the individual levels of the repeated measures effect, and

f^2 is just the effect size for between-subject designs as used by Cohen (1977, 1988), that is, the ratio of effect variance to the error variance within cells.

We consider again an effect size of $f^2 = 0.0625$ in the following example. Given ρ = .75 (as before), we need to enter

$$\frac{mf^2}{1-\rho} = 1$$

as the effect size term to be used in our computations.

Select:	Type of Power Analysis:	Post-hoc	
	Type of Test:	Other F-Tests	
	Accuracy		
Input:	Alpha:	.05	
	Effect size "f2":	1	
	N:	20	2 * 10 = 20
	Numerator DF:	3	b - 1 = 4 - 1 = 3
	Denominator DF:	16	(s(h) * (N-k-p+s(h)) = 1 * (20-2-3+1)= 16
Result:	Power (1-beta):	0.9270	
	Critical F:	F(3,16) = 3.2389	
	Lambda:	20.0000	

Thus, the power for the multivariate approach (0.9270) is slightly smaller than that for the univariate approach (0.9646). However, this small advantage of the univariate approach is present if and only if the sphericity assumption is met. If not, the multivariate approach usually has more power (see O'Brien & Kaiser, 1985). In our example, the power of the corrected univariate test was 0.8506.

As with the so-called univariate approach, interactions of within-subject and between-subjects factors are treated just like within-subject effects.

Chi-Square Tests

There are two major categories of Chi^2 tests:

- Goodness-of-fit-tests and
- Contingency tests.

In both cases we have 2 distributions over m categories which are to be compared, one posited by H_0 and one by H_1. We use the effect size index w (Cohen, 1977).

The noncentrality parameter lambda of the noncentral Chi^2 distribution is given by

$$\eta = w^2 N$$

Goodness-of-Fit-Tests

H_0 postulates a multinomial distribution across the m disjoint categories with probabilities $p_{0(1)}, p_{0(2)}, \ldots, p_{0(m)}$, with

$$\sum_{i=1}^{m} p_{0_i} = 1$$

H_1 posits a different multinomial distribution with probabilities $p_{1(1)}$, $p_{1(2)}$, ... $p_{1(m)}$, with

$$\sum_{i=1}^{m} p_{1_i} = 1$$

The effect size index w is given by

$$w = \sqrt{\sum_{i=1}^{m} \frac{\left(p_{0_i} - p_{1_i}\right)^2}{p_{0_i}}}$$

You can easily calculate the effect size index w from these probabilities using the "Calc 'w'" option after you select "Type of Test: Chi-square Test." You will save time if you decide to use this option.

The number of df (degrees of freedom) is m-1 if all probabilities have fixed values according to H_0 and no parameter needs to be estimated. More df's can be lost in the process of parameter estimation. For instance, if the fit of empirical data to the normal distribution is tested, then 2 more df's are lost because the mean and the standard deviation have to be determined. Thus df = m - 1 - 2 in this case.

Example 1

We test how well an empirical distribution fits the normal distribution. First, we determine the theoretical probabilities for 10 intervals. We want to detect "small" deviations from the theoretical distribution according to Cohen's effect size conventions, thus w = 0.1. How many subjects do we need, given alpha = beta = .05?

Select:	Type of Power Analysis:	A Priori	
	Type of Test:	Chi-square test	
		Accuracy mode calculation	
Input:	Alpha:	.05	
	Power (1 - beta):	0.95	
	Effect size "w":	0.100	To calculate conveniently the effect size from the probabilities defining H0 and H1, click "Calc 'w'", insert the probabilities and click "calc and copy")
	DF for Chi:	7	(m - 1 - 2 = 10 - 3 = 7)

Result:	Total sample size:	2184
	Actual power:	0.9500
	Critical Chi^2:	Chi^2(7) = 14.0671
	Lambda:	21.8400

Example 2

Compromise power analyses can be of particular value when performing goodness-of-fit tests. For instance, it may be that we have very many data points such that, given alpha and beta = .05, even tiny and negligible deviations of the H0 and H1 probability distributions would result in rejections of the model. For instance, we could have 3500 data points and a 1 df model test, in which case the question would be which level of alpha = beta guarantees that only effects of at least w = 0.1 are detected. Let us suppose the relative seriousness of alpha and beta is given by the ratio q := beta/alpha = 1.

Select:	Type of Power Analysis:	Compromise
	Type of Test:	Chi-square test
		Accuracy mode calculation
Input:	Total sample size:	3500
	Effect size "w":	0.1000
	beta / alpha ratio:	1
	DF for Chi:	1
Result:	Alpha:	0.0022
	Power (1 - beta):	0.9978
	Critical Chi^2:	Chi^2(1) = 9.3934
	Lambda:	35.0000

Contingency Tests

Suppose we have a two-dimensional I x J contingency table with i * j = m cells. H_0 postulates that the random variables J and I are stochastically independent. In other words, the cell probabilities are determined by the associated column and row probabilities. H_1, in contrast, posits that the distribution of the probabilities across the m cells is not determined by the column and row probabilities. Again, w is computed from the probability distributions according to H0 and H1 (see above). The degrees of freedom are given by

```
df = (i-1) * (j-1).
```

Example

Let us test the independence assumption for a 2 x 3 table, that is, df = 1 * 2 = 2. Given a total sample size of 180, alpha = .05, and w = 0.327: What is the power of this test?

Select:	Type of Power Analysis:	Post hoc
	Type of Test:	Chi-square test
		Accuracy mode calculation
Input:	Alpha:	.05
	Effect size "w":	0.3270
	Total sample size:	180
	DF for Chi:	2
Result:	Power (1-beta):	0.9818
	Critical Chi^2:	Chi^2(2) = 5.9915
	Lambda:	19.2472