An Overview of Power Analysis

Power is the conditional probability that one will reject the null hypothesis given that the null hypothesis is really false by a specified amount and given certain other specifications, such as sample size and criterion of statistical significance (alpha). I shall introduce power analysis in the context of a one sample test of the mean. After that I shall move on to statistics more commonly employed.

There are several different sorts of power analyses – see Faul, Erdfelder, Lang, & Buchner (*Behavior Research Methods*, 2007, *39*, 175-191) for descriptions of five types that can be computed using G*Power 3. I shall focus on "a priori" and "a posteriori" power analysis.

A Priori Power Analysis. This is an important part of planning research. You determine how many cases you will need to have a good chance of detecting an effect of a specified size with the desired amount of power. See my document Estimating the Sample Size Necessary to Have Enough Power for required number of cases to have 80% for common designs.

A Posteriori Power Analysis. Also know as "post hoc" power analysis. Here you find how much power you would have if you had a specified number of cases. Is it "a posteriori" only in the sense that you provide the number of number of cases, as if you had already conducted the research. Like "a priori" power analysis, it is best used in the planning of research – for example, I am planning on obtaining data on 100 cases, and I want to know whether or not would give me adequate power.

Retrospective Power Analysis. Also known as "observed power." There are several types, but basically this involves answering the following question: "If I were to repeat this research, using the same methods and the same number of cases, and if the size of the effect in the population was exactly the same as it was in the present sample, what would be the probability that I would obtain significant results?" Many have demonstrated that this question is foolish, that the answer tells us nothing of value, and that it has led to much mischief. See this discussion from Edstat-L. I also recommend that you read Hoenig and Heisey (*The American Statistician*, 2001, *55*, 19-24). A few key points:

- Some stat packs (SPSS) give you "observed power" even though it is useless.
- "Observed power" is perfectly correlated with the value of *p* – that is, it provides absolutely no new information that you did not already have.
- It is useless to conduct a power analysis AFTER the research has been completed. What you should be doing is calculating confidence intervals for effect sizes.

One Sample Test of Mean

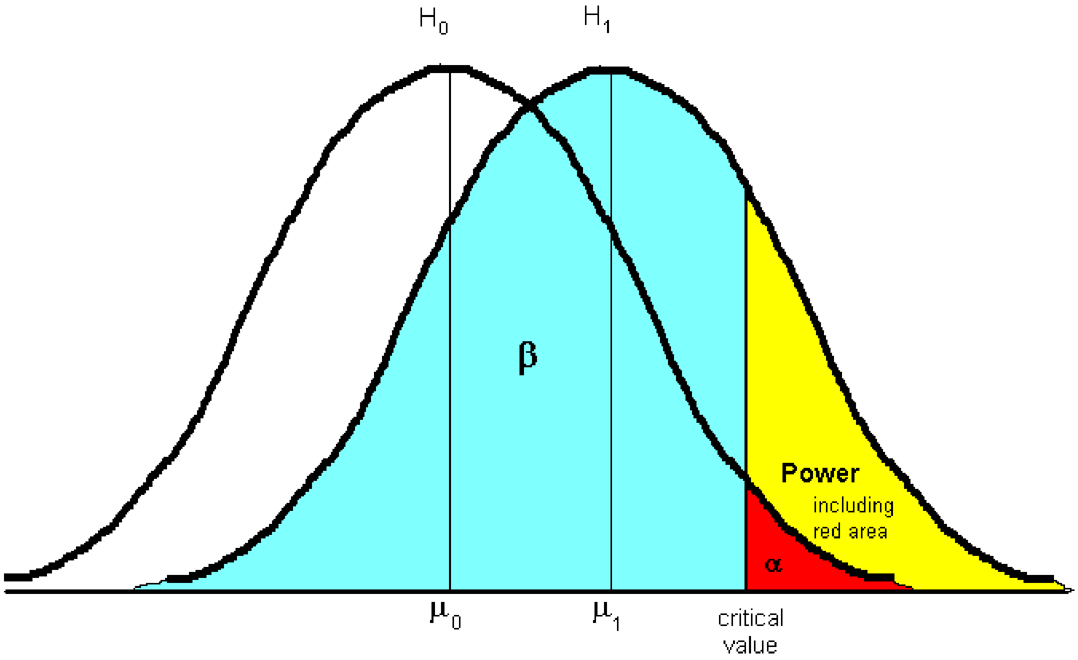
Imagine that we are evaluating the effect of a putative memory enhancing drug. We have randomly sampled 25 people from a population known to be normally distributed with a μ of 100 and a σ of 15. We administer the drug, wait a reasonable time for it to take effect, and then test our subjects' IQ. Assume that we were so confident in our belief that the drug would either increase IQ or have no effect that we entertained directional hypotheses. Our null hypothesis is that after administering the drug $\mu \leq 100$; our alternative hypothesis is $\mu > 100$.

These hypotheses must first be converted to exact hypotheses. Converting the null is easy: it becomes $\mu = 100$. The alternative is more troublesome. If we knew that the effect of the drug was to increase IQ by 15 points, our exact alternative hypothesis would be $\mu = 115$, and we could compute power, the probability of correctly rejecting the false null hypothesis given that μ is really equal to 115 after drug treatment, not 100 (normal IQ). But if we already knew how large the effect of the drug was, we would not need to do inferential statistics.

One solution is to decide on a **minimum nontrivial effect size**. What is the smallest effect that you would consider to be nontrivial? Suppose that you decide that if the drug increases μ_{iq} by 2 or more points, then that is a nontrivial effect, but if the mean increase is less than 2 then the effect is trivial.

Now we can test the null of μ = 100 versus the alternative of μ = 102. Let the left curve represent the distribution of sample means if the null hypothesis were true, μ = 100. This sampling distribution has a μ = 100 and a $\sigma_{\bar{x}} = \frac{15}{\sqrt{25}} = 3$. Let the right curve represent the sampling distribution if the exact alternative hypothesis is true, μ = 102. Its μ is 102 and, assuming the drug has no effect on the variance in IQ scores, $\sigma_{\bar{x}} = \frac{15}{\sqrt{25}} = 3$.

The red area in the upper tail of the null distribution is α. Assume we are using a one-tailed α of .05. How large would a sample mean need be for us to reject the null? Since the upper 5% of a normal distribution extends from 1.645 σ above the μ up to positive infinity, the sample mean IQ would need be 100 + 1.645(3) = 104.935 or more to reject the null. What are the chances of getting a sample mean of 104.935 or more if the alternative hypothesis is correct, if the drug increases IQ by 2 points? The area under the alternative curve from 104.935 up to positive infinity represents that probability, which is power. Assuming the alternative hypothesis is true, that μ = 102, the probability of rejecting the null hypothesis is the probability of getting a sample mean of 104.935 or more in a normal distribution with μ = 102, σ = 3. Z = (104.935 – 102)/3 = 0.98, and $P(Z > 0.98)$ = .1635. That is, power is about 16%. If the drug really does increase IQ by an average of 2 points, we have a 16% chance of rejecting the null. If its effect is even larger, we have a greater than 16% chance.



Suppose we consider 5 the minimum nontrivial effect size. This will separate the null and alternative distributions more, decreasing their overlap and increasing power. Now, Z = (104.935 – 105)/3 = –0.02, $P(Z > -0.02)$ = .5080 or about 51%. **It is easier to detect large effects than small effects.**

Suppose we conduct a 2-tailed test, since the drug could actually decrease IQ; α is now split into both tails of the null distribution, .025 in each tail. We shall reject the null if the sample mean is 1.96 or more standard errors away from the μ of the null distribution. That is, if the mean is 100 + 1.96(3) = 105.88 or more (or if it is 100 – 1.96(3) = 94.12 or less) we reject the null. The probability of that happening if the alternative is correct (μ = 105) is: Z = (105.88 – 105)/3 = .29, $P(Z > .29)$ = .3859, power = about 39%. We can ignore $P(Z < (94.12 - 105)/3) = P(Z < -3.63)$ = very, very small. Note that our power is less than it was with a one-tailed test. **If you can correctly predict the direction of effect, a one-tailed test is more powerful than a two-tailed test.**

Consider what would happen if you increased sample size to 100. Now the $\sigma_{\bar{x}} = \frac{15}{\sqrt{100}} = 1.5$. With the null and alternative distributions less plump, they should overlap less, increasing power. With $\sigma_{\bar{x}} = 1.5$, the sample mean will need be 100 + (1.96)(1.5) = 102.94 or more to reject the null. If the drug increases IQ by 5 points, power is : Z = (102.94 – 105)/1.5 = –1.37, $P(Z > -1.37)$ = .9147, or between 91 and 92%. **Anything that decreases the standard error will increase power. This may be achieved by increasing the sample size or by reducing the σ of the dependent variable.** The σ of the criterion variable may be reduced by reducing the influence of extraneous variables upon the criterion variable (eliminating “noise” in the criterion variable makes it easier to detect the signal, the grouping variable’s effect on the criterion variable).

Now consider what happens if you change α. Let us reduce α to .01. Now the sample mean must be 2.58 or more standard errors from the null μ before we reject the null. That is, 100 + 2.58(1.5) = 103.87. Under the alternative, Z = (103.87 – 105)/1.5 = –0.75, $P(Z > -0.75)$ = 0.7734 or about 77%, less than it was with α at .05, ceteris paribus. **Reducing α reduces power.**

Please note that all of the above analyses have assumed that we have used a normally distributed test statistic, as $Z = \frac{\bar{X} - \mu_{\phi}}{\sigma_{\bar{x}}}$ will be if the criterion variable is normally distributed in the population or if sample size is large enough to invoke the CLT. Remember that using Z also requires that you know the population σ rather than estimating it from the sample data. We more often estimate the population σ, using Student’s t as the test statistic. If N is fairly large, Student’s t is nearly normal, so this is no problem. For example, with a two-tailed α of .05 and N = 25, we went out ± 1.96 standard errors to mark off the rejection region. With Student’s t on N –1 = 24 df we should have gone out ± 2.064 standard errors. But 1.96 versus 2.06 is a relatively trivial difference, so we should feel comfortable with the normal approximation. If, however, we had N = 5, df = 4, critical t = ± 2.776, and the normal approximation would not do. A more complex analysis would be needed.

One Sample Power the Easy Way

Hopefully the analysis presented above will help you understand what power analysis is all about, but who wants to have to do so much thinking when doing a power analysis? Yes, there are easier ways. These days the easiest way it to use computer software that can do power analysis, and there is some pretty good software out there that is free. I like free!

I shall illustrate power analysis using the G∗Power program. I am planning on conducting the memory drug study described above with 25 participants. I have decided that the minimum nontrivial effect size is 5 IQ points, and I shall employ nondirectional hypothesis with a .05 criterion of statistical significance.

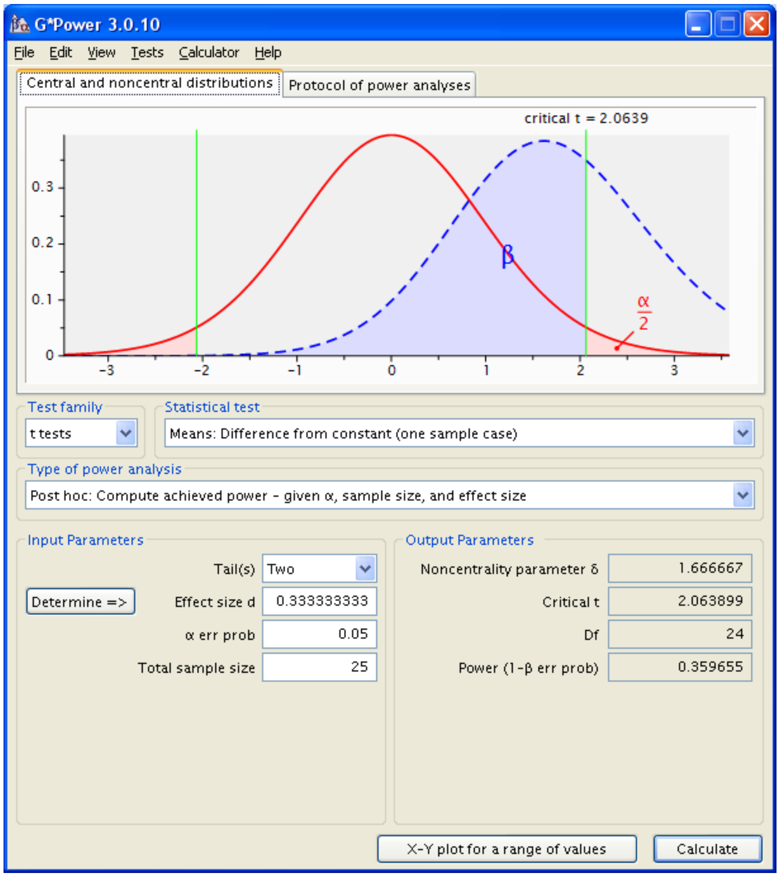
I boot up G*Power and select the following options:

- Test family: t tests
- Statistical test: Means: Difference from constant (one sample case)
- Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size
- Tails: Two
- Effect size d: 0.333333 (you could click "Determine" and have G*Power compute d for you)
- α error prob: 0.05
- Total sample size: 25

Click "Calculate" and you find that power = 0.360.

At the top of the window you get a graphic showing the distribution of *t* under the null and under the alternative, with critical *t*, β, and power indicated.

If you click the "Protocol" tab you get a terse summary of the analyses you have done, which can be printed, saved, or cleared out.



Test family: t tests

Statistical test: Means: Difference from constant (one sample case)

Type of power analysis: Post hoc: Compute achieved power - given α, sample size, and effect size

Input Parameters

Tail(s): Two

Effect size d: 0.333333333

α err prob: 0.05

Total sample size: 25

Output Parameters

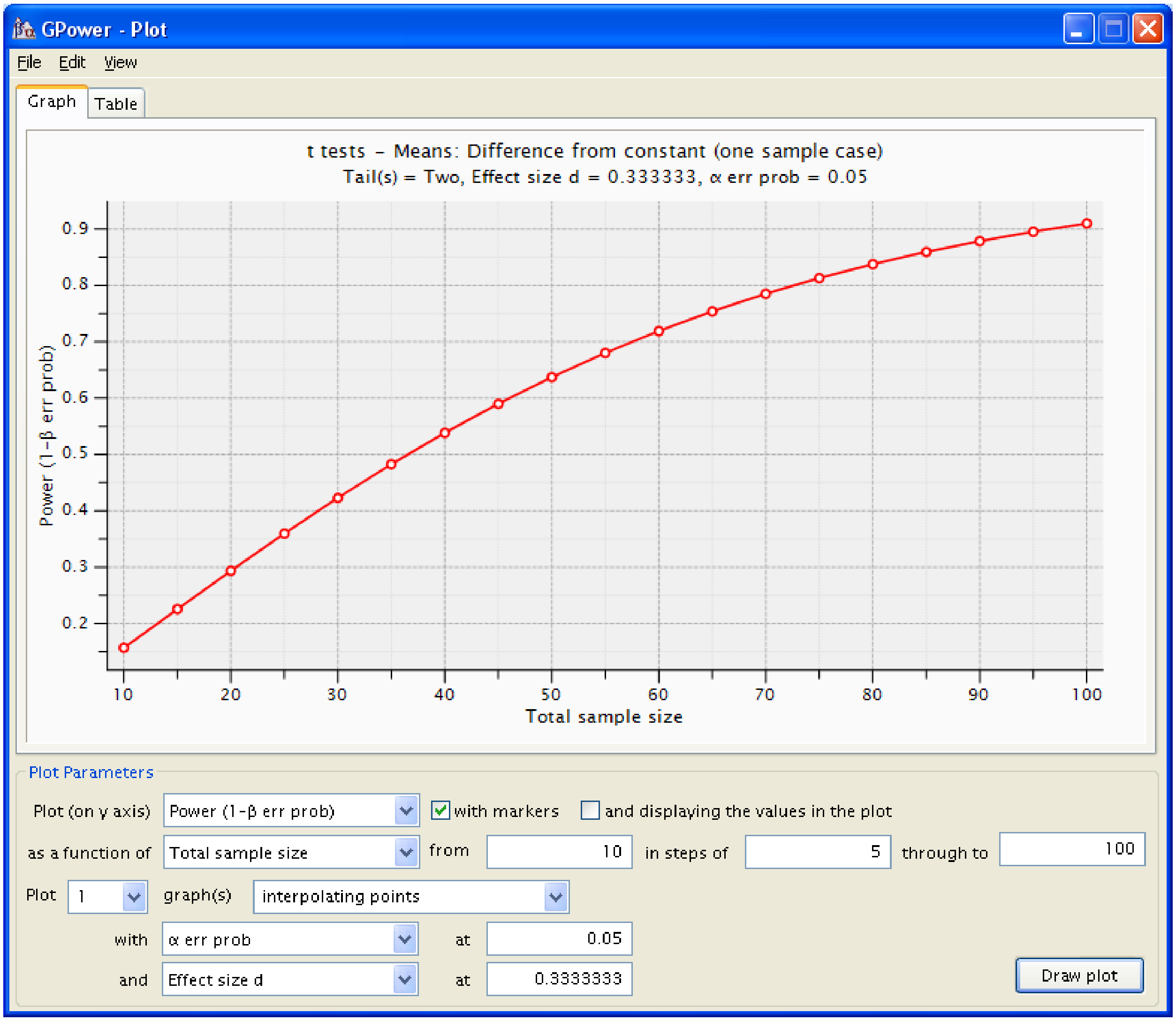
Noncentrality parameter δ: 1.666667

Critical t: 2.063899

Df: 24

Power (1-β err prob): 0.359655

At the bottom of the window you can click "X-Y plot for a range of values." Select what you want plotted on Y and X and set constants and then click "Draw plot." Here is the plot showing the relationship between sample size and power. Clicking the “Table” tab gives you same information in a table.



t tests – Means: Difference from constant (one sample case)
Tail(s) = Two, Effect size d = 0.333333, α err prob = 0.05

Plot Parameters

Plot (on y axis) Power (1-β err prob) ☑ with markers □ and displaying the values in the plot

as a function of Total sample size from 10 in steps of 5 through to 100

Plot 1 graph(s) interpolating points

with α err prob at 0.05

and Effect size d at 0.3333333

Having 36% power is not very encouraging – if the drug really does have a five point effect, there is a 64% chance that you will not detect the effect and will make a Type II error. If you cannot afford to get data from more than 25 participants, you may go ahead with your research plans and hope that the real effect of the drug is more than five IQ points.

If you were to find a significant effect of the drug with only 25 participants, that would speak to the large effect of the drug. In this case you should not be hesitant to seek publication of your research, but you should be somewhat worried about having it reviewed by “ignorant experts.” Such bozos (and they are to be found everywhere) will argue that your significant results cannot be trusted because your analysis had little power. It is useless to argue with them, as they are totally lacking in understanding of the logic of hypothesis testing. If the editor cannot be convinced that the reviewer is a moron, just resubmit to a different journal and hope to avoid ignorant expert reviewers there. I should add that it really would have better if you had more data, as your estimation of the size of the effect would be more precise, but these ignorant expert reviewers would not understand that either.

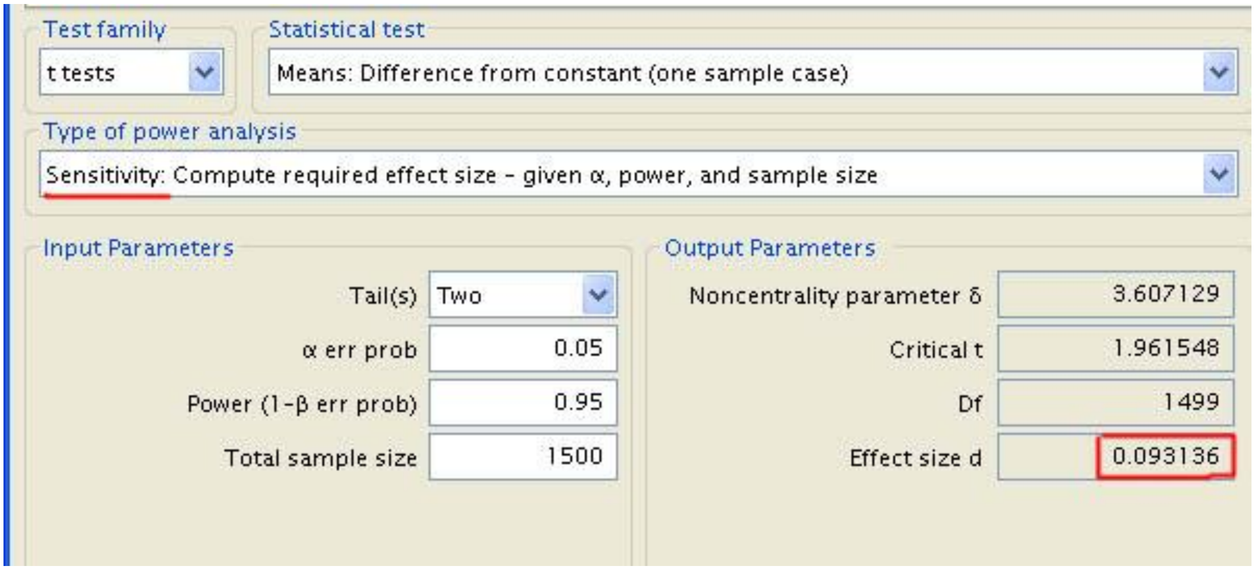
If you were not able to reject the null hypothesis in your research on the putative IQ drug, and your power analysis indicated about 36% power, you would be in an awkward position. Although you could not reject the null, you also could not accept it, given that you only had a relatively small (36%) chance of rejecting it even if it were false. You might decide to repeat the experiment using an *n* large enough to allow you to **accept the null if you cannot reject it**. In my opinion, if 5% is a reasonable risk for a Type I error (α), then 5% is also a reasonable risk for a Type II error (β), [unless the serious of one of these types of errors exceeds that of the other], so let us use power = 1 – β = 95%.

How many subjects would you need to have 95% power? In G*Power, under Type of power analysis, select "A priori: Compute required sample size given α, power, and effect size." Enter ".95" for "Power (1- β err prob)." Click "Calculate." G*Power tells you that you need 119 subjects to get the desired power. Now write that grant proposal that will convince the grant reviewers that your research deserves funding that will allow you get enough data to be able to make a strong statement about whether or not the putative memory enhancing drug is effective. If it is effective, be sure that ignorant reviewers are the first to receive government subsidized supplies of the drug for personal use.

If we were to repeat the experiment with 119 subjects and still could not reject the null, we can "accept" the null and conclude that the drug has no nontrivial (≥ 5 IQ points) effect upon IQ. The null hypothesis we are accepting here is a "**loose null hypothesis**" [95 < μ < 105] rather than a "**sharp null hypothesis**" [μ = exactly 100]. Sharp null hypotheses are probably very rarely ever true.

Others could argue with your choice of the minimum nontrivial effect size. Cohen has defined a small effect as *d* = .20, a medium effect as *d* = .50, and a large effect as *d* = .80. If you defined minimum ***d*** at .20, you would need even more subjects for 95% power.

A third approach, called a **sensitivity analysis** in G∗Power, is to find the smallest effect that one could have detected with high probability given *n*. If that *d* is small, and the null hypothesis is not rejected, then it is accepted. For example, I used 1500 subjects in the IQ enhancer study. Consider the null hypothesis to be -0.1 ≤ d ≤ +0.1. That is, if *d* does not differ from zero by at least .1, then I consider it to be ≅ 0.



Test family: t tests

Statistical test: Means: Difference from constant (one sample case)

Type of power analysis: Sensitivity: Compute required effect size - given α, power, and sample size

Input Parameters

Tail(s): Two

α err prob: 0.05

Power (1-β err prob): 0.95

Total sample size: 1500

Output Parameters

Noncentrality parameter δ: 3.607129

Critical t: 1.961548

Df: 1499

Effect size d: 0.093136

For power = 95%, *d* = .093. If I can't reject the null, I accept it, concluding that if the drug has any effect, it is a trivial effect, since I had a 95% chance of detecting an effect as small as . 093 σ. I would prefer simply to report a confidence interval here, showing that *d* is very close to zero.

Install G∗Power on Your Personal Computer

If you would like to install G∗Power on your Windows computer, you can download it from Universität Duesseldorf.

Two Independent Samples Test of Means

If $n_1 \neq n_2$, the effective sample size is the harmonic mean sample size, $\tilde{n} = \dfrac{2}{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$.

For a fixed total *N*, the harmonic mean (and thus power) is higher the more nearly equal n_1 and n_2 are. This is one good reason to use equal *n* designs. Other good reasons are computational simplicity with equal *n*'s and greater robustness to violation of assumptions. The effective (harmonic) sample size for 100 subjects evenly split into two groups of 50 each is 50; for a 60:40 split it is 48; for a 90:10 split it is 18.

Consider the following a priori power analysis. We wish to compare the Advanced Psychology GRE scores of students in general psychology masters programs with that of those in clinical psychology masters programs. We decide that we will be satisfied if we have enough data to have an 80% chance of detecting an effect of 1/3 of a standard deviation, employing a .05 criterion of significance. How many scores do we need in each group, if we have the same number of scores in each group?

Select the following options:

- Test family: ttests
- Statistical test: Means: Difference between two independent means (two groups)
- Type of power analysis: A priori: Compute required sample size given α, power, and effect size
- Tails: Two
- Effect size d: 0.333333 (you could click "Determine" and have G*Power compute *d* for you)
- α error prob: 0.05
- Power (1- β err prob): .8
- Allocation ratio N2/N1: 1

Click "Calculate" and you see that you need 143 cases in each group, that is, a total sample size of 286.

Change the allocation ratio to 9 (nine times as many cases in the one group than in the other) and click "Calculate" again. You will see that you would need 788 subjects to get the desired power with such a lopsided allocation ratio.

Consider the following a posteriori power analysis. We have available only 36 scores from students in clinical programs and 48 scores from students in general programs. What are our chances of detecting a difference of 40 points (which is that actually observed at ECU in 1981) if we use a .05 criterion of significance and the standard deviation is 98?

Change the type of power analysis to Post hoc. Enter d = 40/98 = .408, n_1 = 36, n_2 = 48. Click "Calculate." You will see that you have 45% power.

```
Output:   Noncentrality parameter δ     =  1.850514
          Critical t                    =  1.989319
          Df                            =  82
          Power (1−β err prob)          =  0.447910
```

Two Related Samples, Test of Means

The correlated samples *t* test is mathematically equivalent to a one-sample *t* test conducted on the difference scores (for each subject, score under one condition less score under the other condition). The greater *ρ12*, the correlation between the scores in the one condition and those in the second condition, the smaller the standard deviation of the difference scores and the greater the power, ceteris paribus. By the variance sum law, the standard deviation of the difference scores is $\sigma_{Diff} = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2}$. If we assume equal variances, this simplifies to $\sigma_{Diff} = \sigma\sqrt{2(1-\rho)}$.

When conducting a power analysis for the correlated samples design, we can take into account the effect of *ρ12* by computing *dDiff*, an adjusted value of *d*: $d_{Diff} = \frac{\mu_1 - \mu_2}{\sigma_{Diff}} = \frac{d}{\sqrt{2(1-\rho_{12})}}$, where *d* is the effect size as computed above, with independent samples.

Please note that using the standard error of the difference scores, rather than the standard deviation of the criterion variable, as the denominator of *dDiff***,** is simply a means of incorporating into the analysis the effect of the correlation produced by matching. If we were computing estimated *d* (Hedges' *g*) as an estimate of the standardized effect size given the obtained results, we would use the standard deviation of the criterion variable in the denominator, not the standard deviation of the difference scores. I should admit that on rare occasions I have argued that, in a particular research context, it made more sense to use the standard deviation of the difference scores in the denominator of *g*.

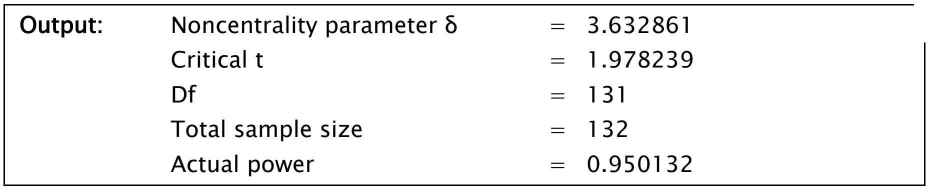
Consider the following a priori power analysis. I am testing the effect of a new drug on performance on a task that involves solving anagrams. I want to have enough power to be able to detect an effect as small as 1/5 of a standard deviation (*d* = .2) with 95% power – I consider Type I and Type II errors equally serious and am employing a .05 criterion of statistical significance, so I want beta to be not more than .05. I shall use a correlated samples design (within subjects) and two conditions (tested under the influence of the drug and not under the influence of the drug). In previous research I have found the correlation between conditions to be approximately .8.

$$d_{Diff} = \frac{d}{\sqrt{2(1-\rho_{12})}} = \frac{.2}{\sqrt{2(1-.8)}} = .3162 .$$

Use the following settings:

- Statistical test: Means: Difference between two dependent means (matched pairs)
- Type of power analysis: A priori: Compute required sample size given α, power, and effect size
- Tail(s): Two
- Effect size dz: .3162
- α error prob: 0.05
- Power (1- β err prob): .95

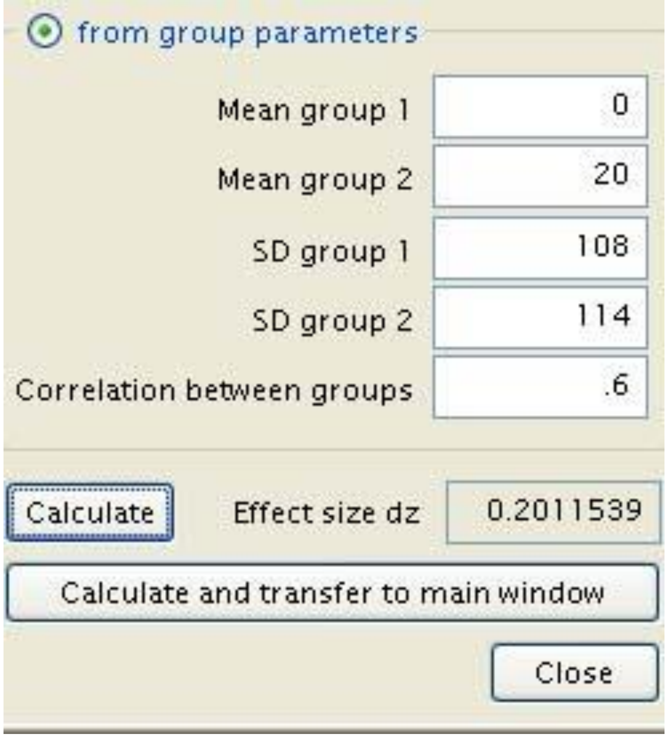
Click "Calculate." You will find that you need 132 pairs of scores.



```
Output:    Noncentrality parameter δ    =  3.632861
           Critical t                   =  1.978239
           Df                           =  131
           Total sample size            =  132
           Actual power                 =  0.950132
```

Consider the following a posteriori power analysis. We assume that GRE Verbal and GRE Quantitative scores are measured on the same metric, and we wish to determine whether persons intending to major in experimental or developmental psychology are equally skilled in things verbal and things quantitative. If we employ a .05 criterion of significance, and if the true size of the effect is 20 GRE points (that was the actual population difference the last time I checked it, with quantitative > verbal), what are our chances of obtaining significant results if we have data on 400 persons? We shall assume that the correlation between verbal and quantitative GRE is .60 (that is what it was for social science majors the last time I checked). We need to know what the standard deviation is for the "dependent variable," GRE score. The last time I checked, it was 108 for verbal, 114 for quantitative.

Change type of power analysis to "Post hoc." Set the total sample size to 400. Click on "Determine." Select "from group parameters." Set the group means to 0 and 20 (or any other two means that differ by 20), the standard deviations to 108 and 114, and the correlation between groups to .6. Click "Calculate" in this window to obtain the effect size dz, .2100539.



● from group parameters

Mean group 1: 0

Mean group 2: 20

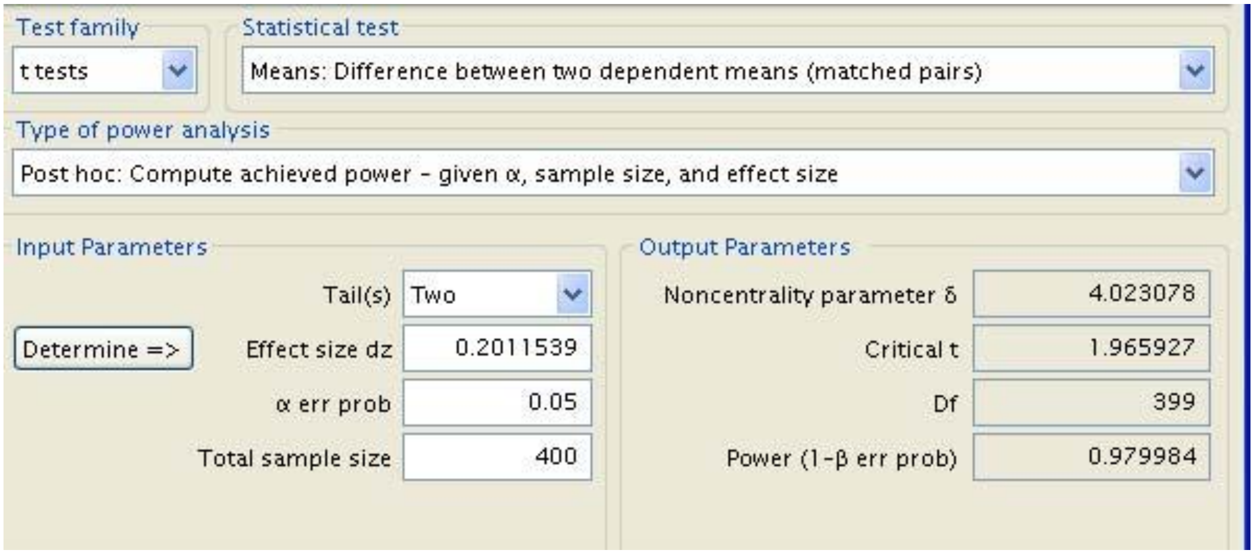
SD group 1: 108

SD group 2: 114

Correlation between groups: .6

Effect size dz: 0.2011539

Click "Calculate and transfer to main window" to move the effect size dz to the main window. Click "Calculate" in the main window to compute the power. You will see that you have 98% power.



Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input Parameters

Tail(s): Two

Effect size dz: 0.2011539

α err prob: 0.05

Total sample size: 400

Output Parameters

Noncentrality parameter δ: 4.023078

Critical t: 1.965927

Df: 399

Power (1-β err prob): 0.979984

Type III Errors and Three-Choice Tests

Leventhal and Huynh (*Psychological Methods*, 1996, *1*, 278-292) note that it is common practice, following rejection of a nondirectional null, to conclude that the direction of difference in the population is the same as what it is in the sample. This procedure is what they call a "directional two-tailed test." They also refer to it as a "three-choice test" (I prefer that language), in that the three hypotheses entertained are: parameter = null value, parameter < null value, and parameter > null value. This makes possible a Type III error: correctly rejecting the null hypothesis, but incorrectly inferring the direction of the effect - for example, when the population value of the tested parameter is actually more than the null value, getting a sample value that is so much below the null value that you reject the null and conclude that the population value is also below the null value. The authors show how to conduct a power analysis that corrects for the possibility of making a Type III error. See my summary at: **http://core.ecu.edu/psyc/wuenschk/StatHelp/Type_III.htm**

Bivariate Correlation/Regression Analysis

Consider the following a priori power analysis. We wish to determine whether or not there is a correlation between misanthropy and support for animal rights. We shall measure these attributes with instruments that produce scores for which it is reasonable to treat the variables as continuous. How many respondents would we need to have a 95% probability of obtaining significant results if we employed a .05 criterion of significance and if the true value of the correlation (in the population) was 0.2?

Select the following options:

- Test family: ttests
- Statistical test: Correlation: Point biserial model (that is, a regression analysis)
- Type of power analysis: A priori: Compute required sample size given α, power, and effect size
- Tails: Two
- Effect size |r|: .2
- α error prob: 0.05
- Power (1- β err prob): .95

Click “Calculate” and you will see that you need 314 cases.

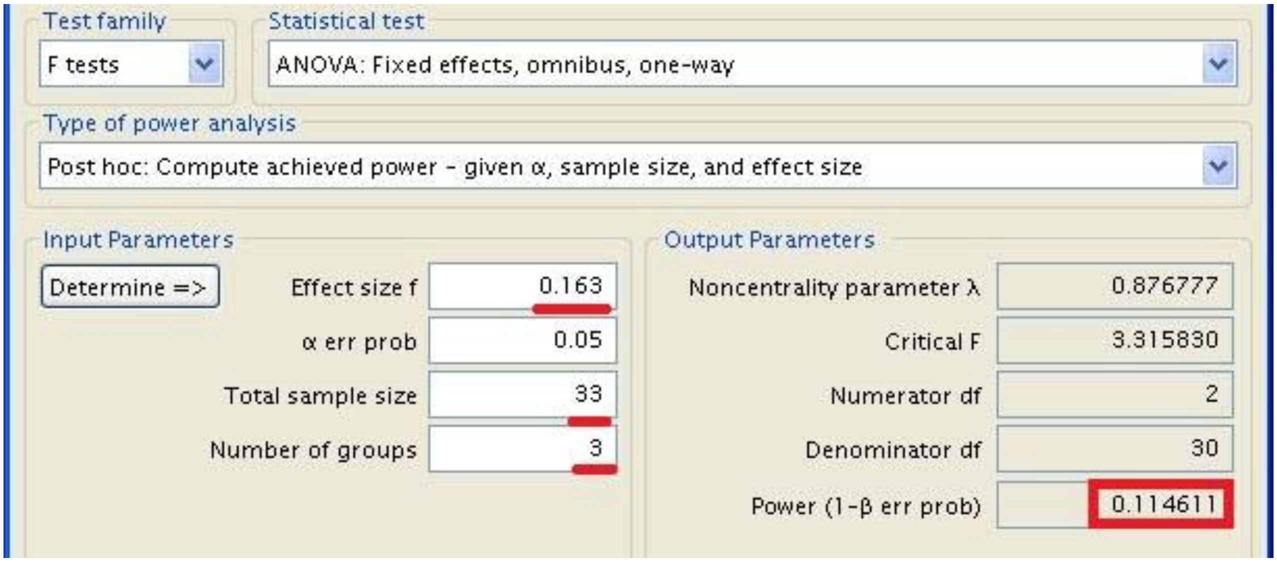
```
t tests – Correlation: Point biserial model
Analysis:   A priori: Compute required sample size
Input:      Tail(s)                          =  Two
            Effect size |r|                  =  .2
            α err prob                       =  0.05
            Power (1–β err prob)             =  0.95
Output:     Noncentrality parameter δ        =  3.617089
            Critical t                       =  1.967596
            Df                               =  312
            Total sample size                =  314
            Actual power                     =  0.950115
```

Check out Steiger and Fouladi’s R2 program, which will do power analysis (and more) for correlation models, including multiple correlation.

One-Way ANOVA, Independent Samples

The effect size may be specified in terms of ***f***: $f = \sqrt{\frac{\sum_{j=1}^{k} (\mu_j - \mu)^2}{k\sigma^2_{error}}}$. Cohen considered an *f* of .10 to represent a small effect, .25 a medium effect, and .40 a large effect. In terms of percentage of variance explained η^2, small is 1%, medium is 6%, and large is 14%.

Suppose that I wish to test the null hypothesis that for GRE-Q, the population means for undergraduates intending to major in social psychology, clinical psychology, and experimental psychology are all equal. I decide that the minimum nontrivial effect size is if each mean differs from the next by 20 points (about 1/5 σ). For example, means of 480, 500, and 520. The sum of the squared deviations between group means and grand mean is then $20^2 + 0^2 + 20^2 = 800$. Next we compute *f*. Assuming that the σ is about 100, $f = \sqrt{800/3/10000} = 0.163$. Suppose we have 11 cases in each group.



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input Parameters

Effect size f: 0.163

α err prob: 0.05

Total sample size: 33

Number of groups: 3

Output Parameters

Noncentrality parameter λ: 0.876777

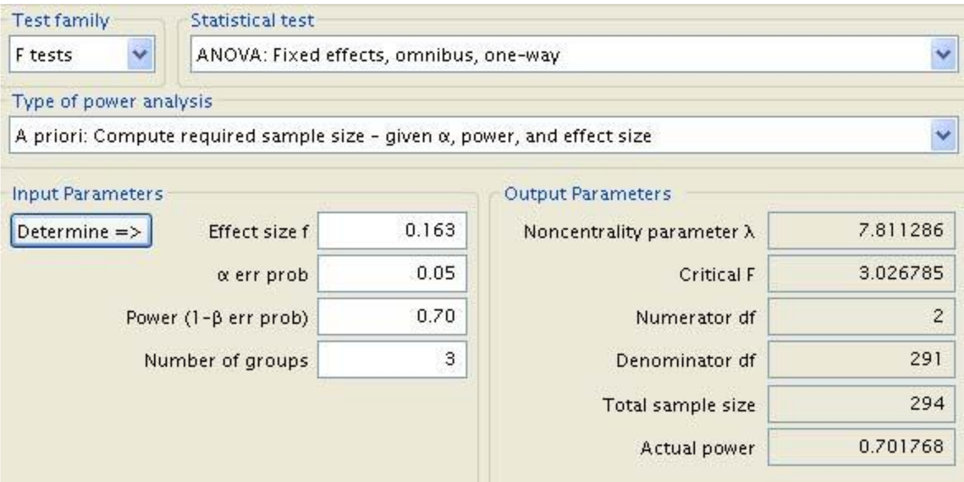
Critical F: 3.315830

Numerator df: 2

Denominator df: 30

Power (1-β err prob): 0.114611

OK, how many subjects would you need to raise power to 70%? Under Analysis, select A Priori, under Power enter .70, and click Calculate.



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters

Effect size f: 0.163

α err prob: 0.05

Power (1-β err prob): 0.70

Number of groups: 3

Output Parameters

Noncentrality parameter λ: 7.811286

Critical F: 3.026785

Numerator df: 2

Denominator df: 291

Total sample size: 294

Actual power: 0.701768

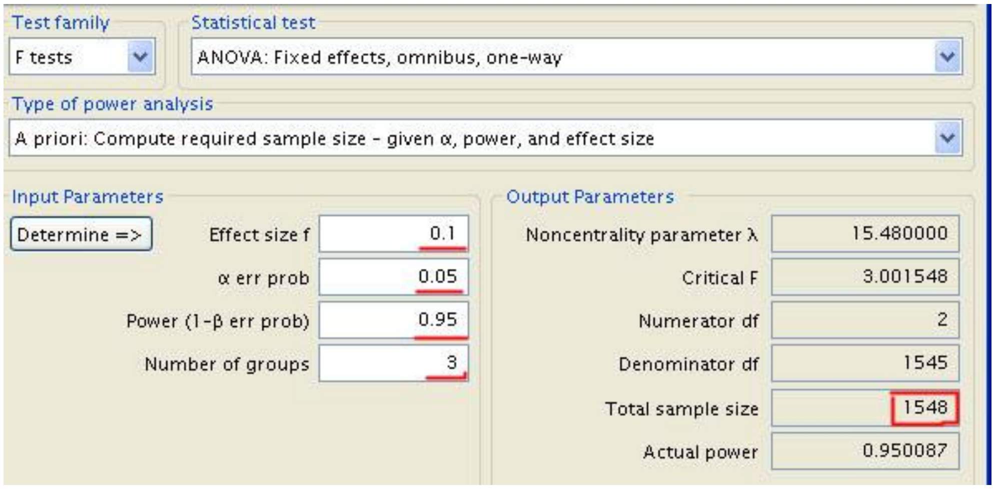
G*Power advises that you need 294 cases, evenly split into three groups, that is, 98 cases per group.

Analysis of Covariance

If you add one or more covariates to your ANOVA model, and they are well correlated with the outcome variable, then the error term will be reduced and power will be increased. The effect of the addition of covariates can be incorporated into the power analysis in this way:

- Adjusting the effect size statistic, *f*, such that the adjusted *f*, $f' = \frac{f}{\sqrt{1-r^2}}$, where *r* is the correlation between the covariate (or set of covariates) and the outcome variable.
- Reducing the error *df* by one for each covariate added to the model.

Consider this example. I am using an ANOVA design to compare three experimental groups. I want to know how many cases I need to detect a small effect (*f* = .1). G∗Power tells me I need 1,548 cases. Ouch, that is a lot of data.



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters

Effect size f: 0.1

α err prob: 0.05

Power (1-β err prob): 0.95

Number of groups: 3

Output Parameters

Noncentrality parameter λ: 15.480000

Critical F: 3.001548

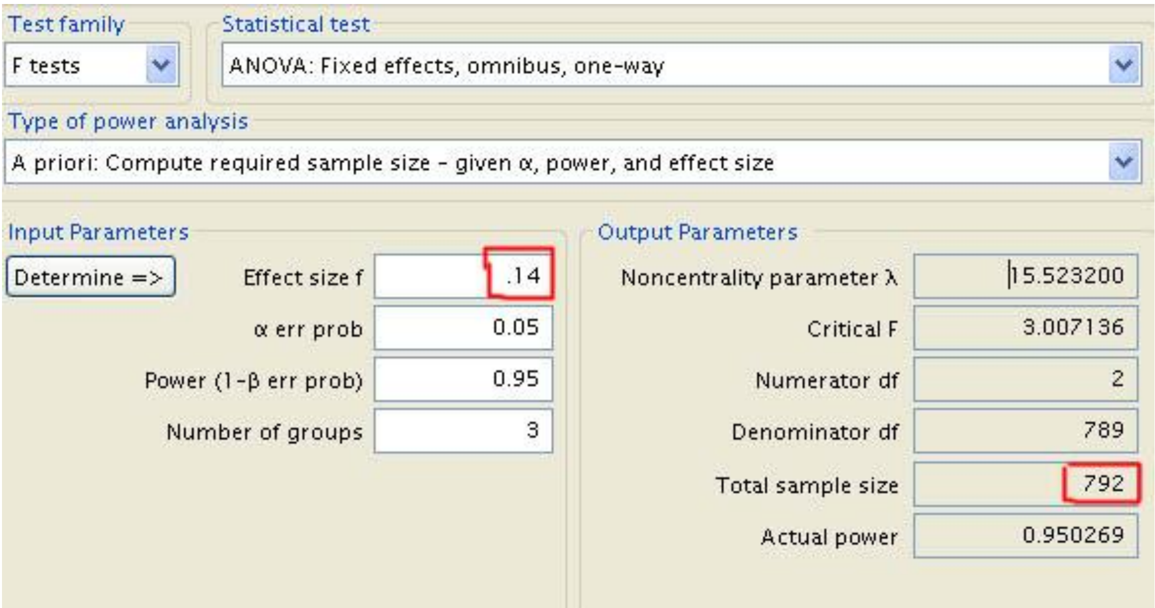
Numerator df: 2

Denominator df: 1545

Total sample size: 1548

Actual power: 0.950087

Suppose I find a covariate that I can measure prior to manipulating the experimental variable and which is known to be correlated .7 with the dependent variable. The adjusted *f* for a small effect increases to $f' = \frac{.1}{\sqrt{1-.49}} = .14$.



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters

Effect size f: .14

α err prob: 0.05

Power (1-β err prob): 0.95

Number of groups: 3

Output Parameters

Noncentrality parameter λ: 15.523200

Critical F: 3.007136

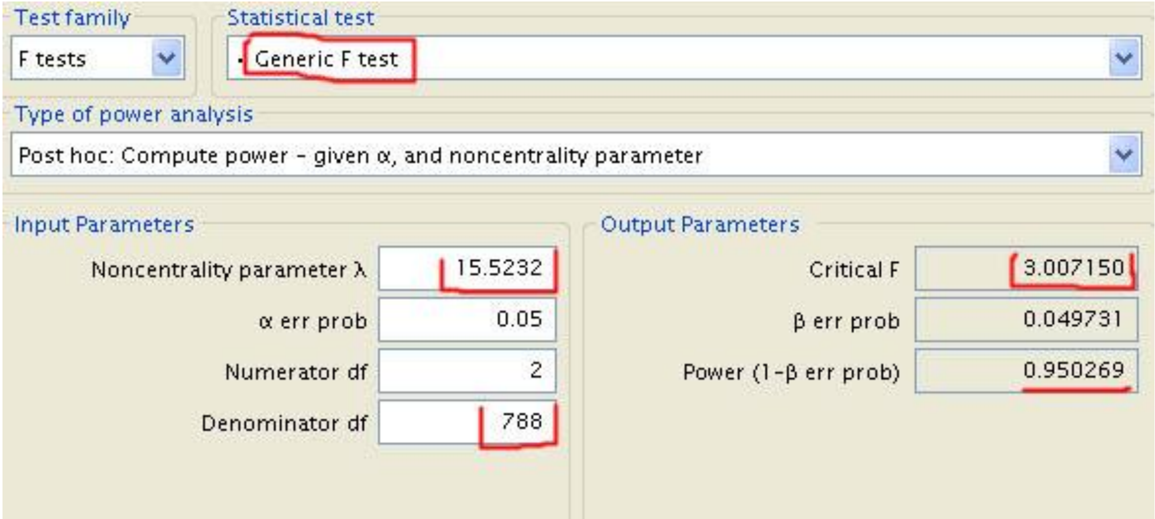
Numerator df: 2

Denominator df: 789

Total sample size: 792

Actual power: 0.950269

Now I only need 792 cases. Do note that the error *df* here should be 788, not 789, but that one *df* is not going to make much difference, as shown below.



Test family: F tests

Statistical test: Generic F test

Type of power analysis: Post hoc: Compute power - given α, and noncentrality parameter

Input Parameters

Noncentrality parameter λ: 15.5232

α err prob: 0.05

Numerator df: 2

Denominator df: 788

Output Parameters

Critical F: 3.007150

β err prob: 0.049731

Power (1-β err prob): 0.950269

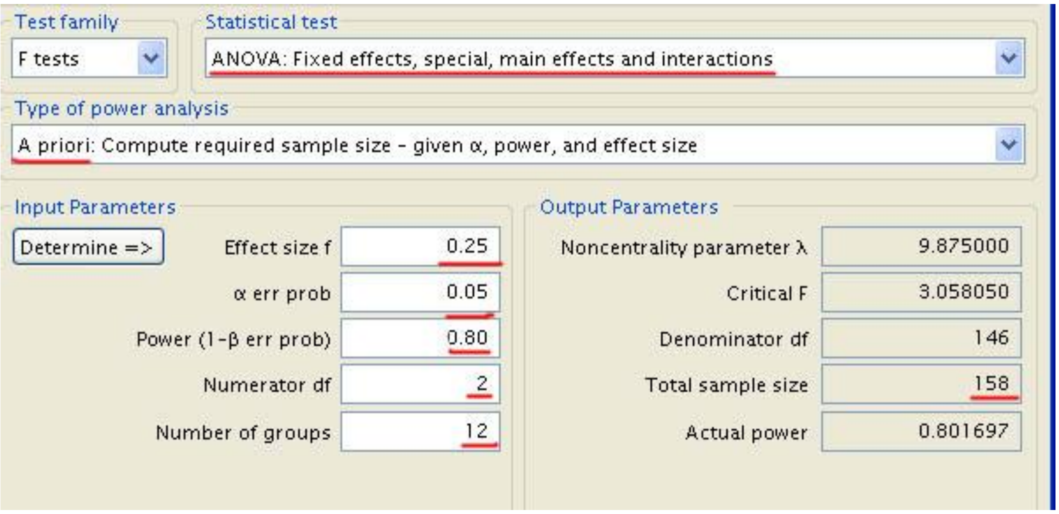
I used the Generic *F* Test routine with the noncentrality parameter from the earlier run, and I dropped the denominator *df* to 788. The value of the critical *F* increased ever so slightly, but the power did not change at all to six decimal places.

Factorial ANOVA, Independent Samples

The analysis is done pretty much the same as it is with a one-way ANOVA. Suppose we are planning research for which an A x B, 3 x 4 ANOVA would be appropriate. We want to have enough data to have 80% power for a medium sized effect. The omnibus analysis will include three *F* tests –

one with 2 *df* in the numerator, one with 3, and one with 6 (the interaction). We plan on having sample size constant across cells.

Boot up G*Power and enter the options shown below:



Test family: F tests
Statistical test: ANOVA: Fixed effects, special, main effects and interactions
Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

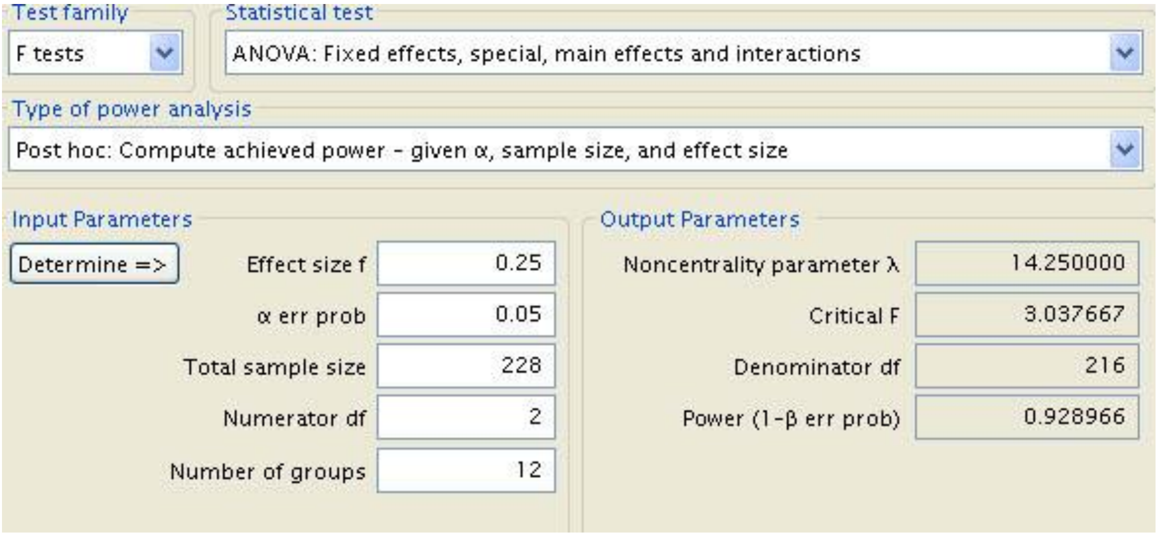
Input Parameters
Effect size f: 0.25
α err prob: 0.05
Power (1-β err prob): 0.80
Numerator df: 2
Number of groups: 12

Output Parameters
Noncentrality parameter λ: 9.875000
Critical F: 3.058050
Denominator df: 146
Total sample size: 158
Actual power: 0.801697

Remember that Cohen suggested .25 as the value of *f* for a medium-sized effect. The numerator *df* for the main effect of **Factor A** is (3-1)=2. The number of groups here is the number of cells in the factorial design, 3 x 4 = 12. When you click “Calculate” you see that you need a total *N* of 158. That works out to 13.2 cases per cell, so bump the *N* up to 14(12) = 168.

What about Factor B and the interaction? There are (4-1)=3 *df* for the main effect of Factor A, and when you change the numerator *df* to 3 and click “Calculate” again you see that you need an *N* of 179 to get 80% power for that effect. The interaction has 2(3)=6 *df*, and when you change the numerator *df* to 6 and click “Calculate” you see that you need an *N* of 225 to have 80% power to detect a medium-sized interaction. With equal sample sizes, that means you need 19 cases per cell, 228 total *N*.

Clearly you are not going to have the same amount of power for each of the three effects. If your primary interest was in the main effects, you might go with a total *N* that would give you the desired power for main effect but somewhat less than that for the interaction. If, however, you have reason to expect an interaction, you will go for the total *N* of 228. How much power would that give you for the main effects?



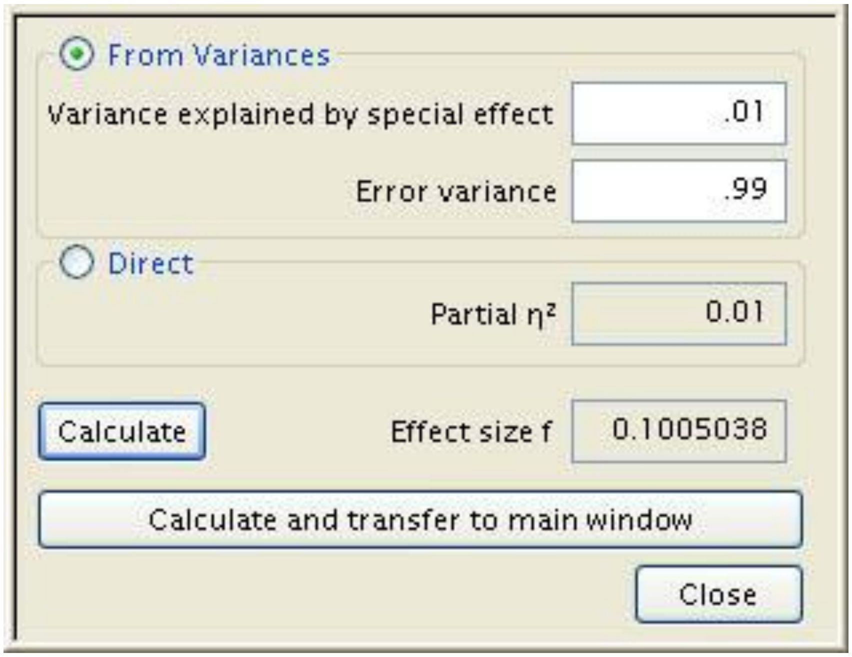
Test family: F tests
Statistical test: ANOVA: Fixed effects, special, main effects and interactions
Type of power analysis: Post hoc: Compute achieved power - given α, sample size, and effect size

Input Parameters
Effect size f: 0.25
α err prob: 0.05
Total sample size: 228
Numerator df: 2
Number of groups: 12

Output Parameters
Noncentrality parameter λ: 14.250000
Critical F: 3.037667
Denominator df: 216
Power (1-β err prob): 0.928966

As you can see, you would have almost 93% power for A. If you change the numerator *df* to 3 you will see that you would have 89.6% power for B.

If you click the “Determine” button you get a second window which allows you select the value of *f* by specifying a value of η^2 or partial η^2. Suppose you want to know what *f* is for an effect that explains only 1% of the total variance. You tell G*Power to that the “Variance explained by special effect” is .01 and “Error variance” is .99. Click “Calculate” and you get an *f* of .10. Recall that earlier I told you that an *f* of .10 is equivalent to an η^2 of .01.



● From Variances

Variance explained by special effect .01

Error variance .99

○ Direct

Partial η² 0.01

Effect size f 0.1005038

If you wanted to find *f* for an effect that accounted for 6% of the variance, you would enter .06 (effect) and .94 (error) and get an *f* of .25 (a medium-sized effect).

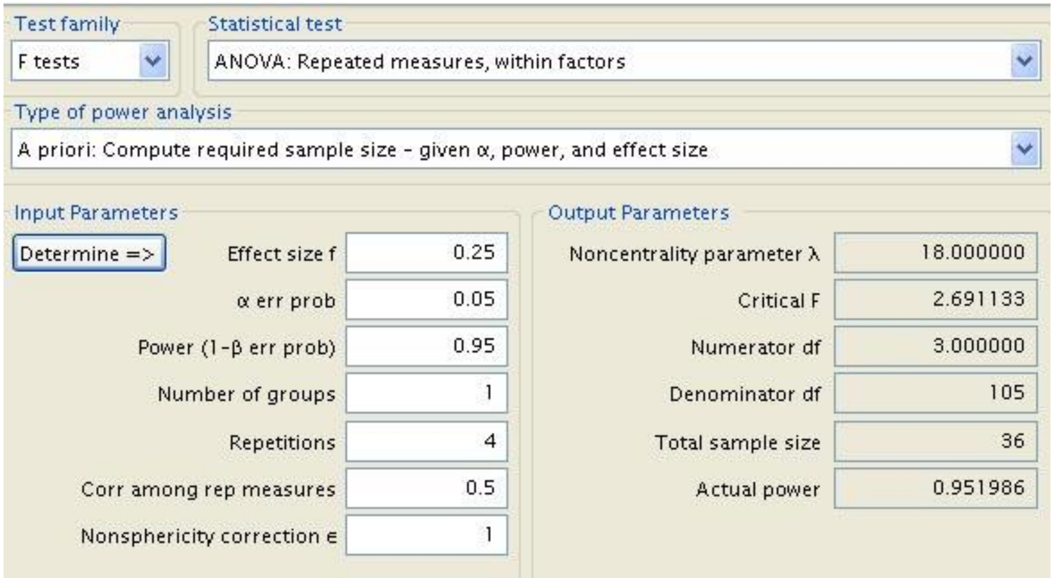
Wait a minute. I have ignored the fact that the error variance in the factorial ANOVA will be reduced by an amount equal to variance explained by the other factors in the model, and that will increase power. Suppose that I have entered Factor B into the model primarily as a categorical covariate. From past research, I have reason to believe that Factor B will account for about 14% of the total variance (a large effect, equivalent to an *f* of .40). I have no idea whether or not the interaction will explain much variance, so I play it safe and assume it will explain no variance. When I calculate *f* I should enter .06 (effect) and .80 (error – 1 less .06 for A and another .14 for B). G*Power gives an *f* of .27, which I would then use in the power analysis for Factor A.

ANOVA With Related Factors

The analysis here can be done with G*Power in pretty much the same way described earlier for independent samples. There are two new parameters that you will need to provide:

- the value of the correlation between scores at any one level of the related factor and any other level of the repeated factor. Assuming that this correlation is constant across pairs of levels is the sphericity assumption.
- ε -- this is a correction (applied to the degrees of freedom) to adjust for violation of the sphericity assumption. The *df* are literally multiplied by ε, which has a upper boundary of 1. There are two common ways to estimate ε, one developed by Greenhouse and Geisser, the other by Huynh and Feldt.

Here is the setup for a one-way repeated measures or randomized blocks ANOVA with four levels of the factor:



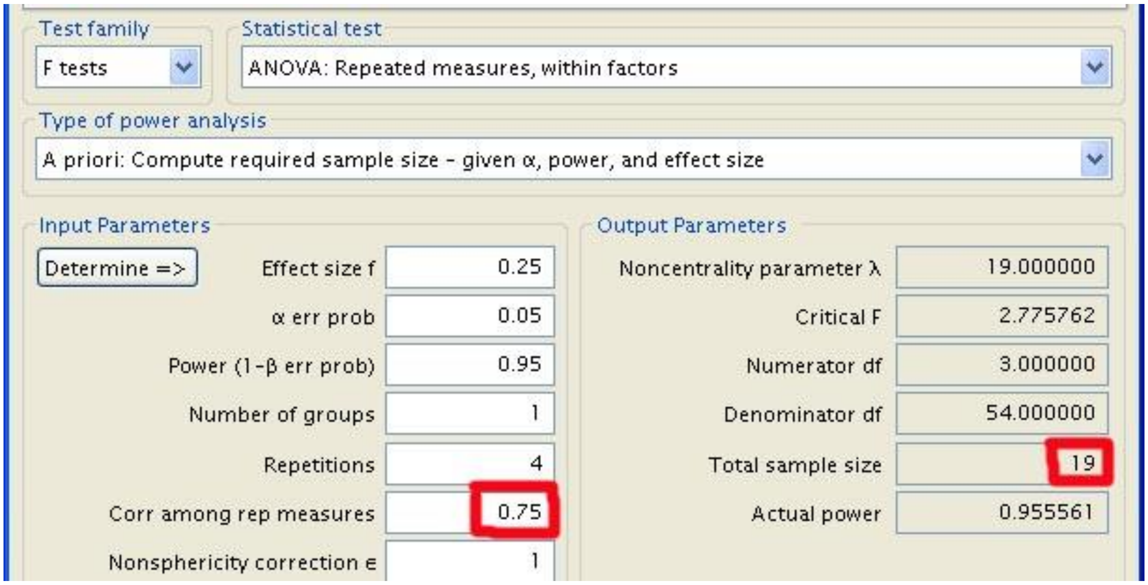
Test family: F tests

Statistical test: ANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters		Output Parameters	
Effect size f	0.25	Noncentrality parameter λ	18.000000
α err prob	0.05	Critical F	2.691133
Power (1-β err prob)	0.95	Numerator df	3.000000
Number of groups	1	Denominator df	105
Repetitions	4	Total sample size	36
Corr among rep measures	0.5	Actual power	0.951986
Nonsphericity correction ε	1		

We need 36 cases to have 95% power to detect a medium sized effect assuming no problem with sphericity and a .5 correlation between repeated measures. Let us see what happens if we have a stronger correlation between repeated measures:



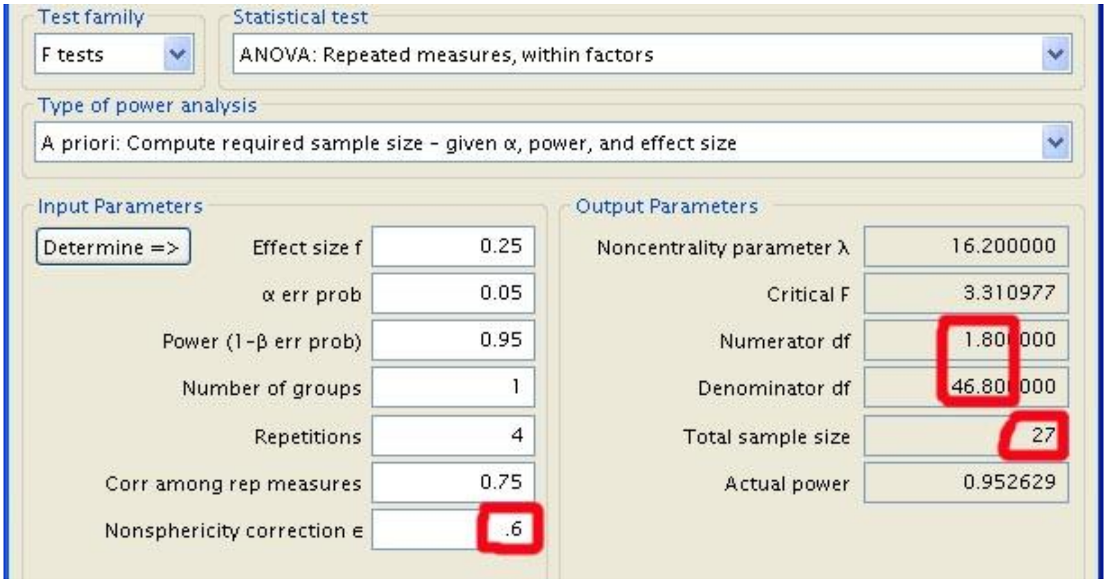
Test family: F tests

Statistical test: ANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters		Output Parameters	
Effect size f	0.25	Noncentrality parameter λ	19.000000
α err prob	0.05	Critical F	2.775762
Power (1-β err prob)	0.95	Numerator df	3.000000
Number of groups	1	Denominator df	54.000000
Repetitions	4	Total sample size	19
Corr among rep measures	0.75	Actual power	0.955561
Nonsphericity correction ε	1		

Very nice. I guess your stats prof wasn't kidding we she pointed out the power benefit of having strong correlations across conditions – but what if you have a problem with the sphericity assumption. Let us assume that you suspect (from previous research) that epsilon might be as low as .6.



Test family: F tests

Statistical test: ANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

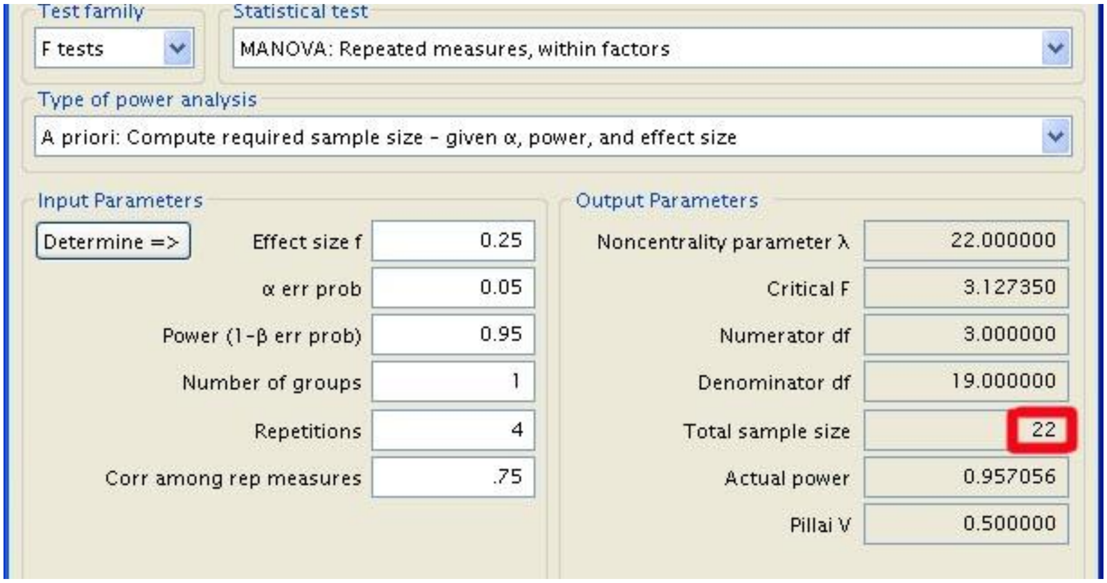
- Effect size f: 0.25
- α err prob: 0.05
- Power (1-β err prob): 0.95
- Number of groups: 1
- Repetitions: 4
- Corr among rep measures: 0.75
- Nonsphericity correction ε: .6

Output Parameters

- Noncentrality parameter λ: 16.200000
- Critical F: 3.310977
- Numerator df: 1.800000
- Denominator df: 46.800000
- Total sample size: 27
- Actual power: 0.952629

Notice the reduction and the degrees of freedom and the associated increase in number of cases needed.

Instead of the traditional “univariate approach” ANOVA, one can analyze data from designs with related factors with the newer “multivariate approach,” which does not have a sphericity assumption. G*Power will do power analysis for this approach too. Let us see how many cases we would need with that approach using the same input parameters as the previous example.



Test family: F tests

Statistical test: MANOVA: Repeated measures, within factors

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

- Effect size f: 0.25
- α err prob: 0.05
- Power (1-β err prob): 0.95
- Number of groups: 1
- Repetitions: 4
- Corr among rep measures: .75

Output Parameters

- Noncentrality parameter λ: 22.000000
- Critical F: 3.127350
- Numerator df: 3.000000
- Denominator df: 19.000000
- Total sample size: 22
- Actual power: 0.957056
- Pillai V: 0.500000

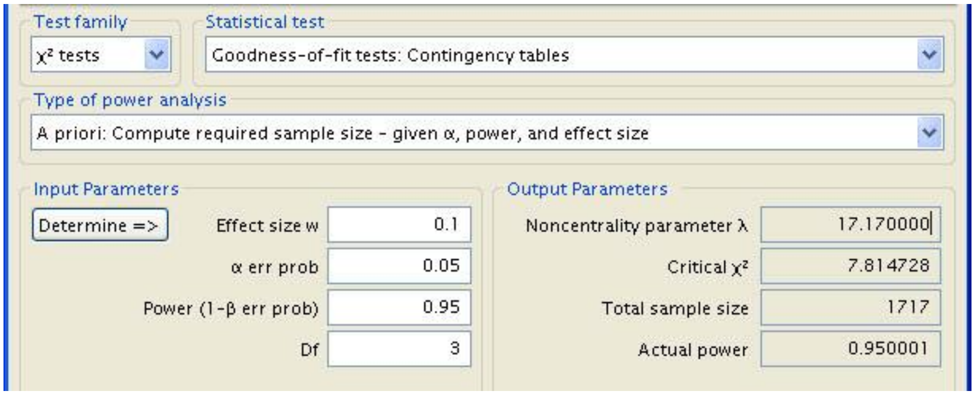
Contingency Table Analysis (Two-Way)

Effect size is computed as $\boxed{w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}}$. *k* is the number of cells, P_{0i} is the population proportion in cell *i* under the null hypothesis, and P_{1i} is the population proportion in cell *i* under the alternative hypothesis. Cohen’s benchmarks are

- .1 is small but not trivial
- .3 is medium
- .5 is large

When the table is 2 x 2, *w* is identical to ϕ.

Suppose we are going to employ a 2 x 4 analysis. We shall use the traditional 5% criterion of statistical significance, and we think Type I and Type II errors equally serious, and, accordingly, we seek to have 95% power for finding an effect that is small but not trivial. As you see below, you need a lot of data to have a lot of power when doing contingency table analysis.



Test family: χ² tests

Statistical test: Goodness-of-fit tests: Contingency tables

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters		Output Parameters	
Effect size w	0.1	Noncentrality parameter λ	17.170000
α err prob	0.05	Critical χ²	7.814728
Power (1-β err prob)	0.95	Total sample size	1717
Df	3	Actual power	0.950001

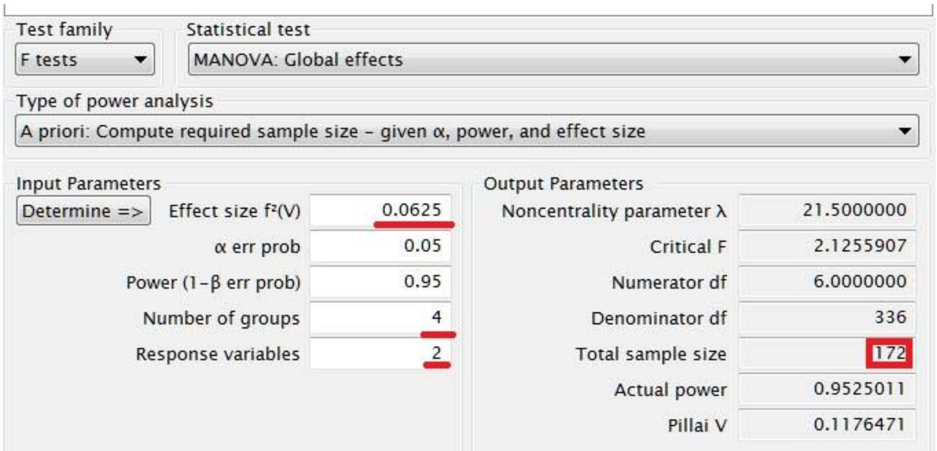
MANOVA and DFA

Each effect will have as many roots (discriminant functions, canonical variates, weighted linear combinations of the outcome variables) as it has treatment degrees of freedom, or it will have as many roots as there are outcome variables, whichever is fewer. The weights maximize the ratio $\frac{SS_{among\_groups}}{SS_{within\_groups}}$. If you were to compute, for each case, a canonical variate score and then conduct an ANOVA comparing the groups on that canonical variate, you would get the sums of squares in the ratio above. This ratio is called the eigenvalue (λ).

Theta is defined as $\theta = \frac{\lambda}{\lambda + 1}$. Commonly employed test statistics in MANOVA are the Hotelling's trace, Wilks' lambda, Pillai's trace, and Roy's greatest characteristic root. Hotelling's trace is simply the sum of the eigenvalues. To get Wilks lambda you subtract each theta from 1 and then calculate the product of the differences. To get Pillai's trace you simply sum the thetas. Roy's gcr is simply the eigenvalue for the first root.

G*Power uses f^2 as the effect size parameter ($.1^2 = .01$ is small, $.25^2 = .0625$ is medium, and $.40^2 = .16$ is large), and allows you convert a value of Pillai's trace (or other trace) to *f* if you wish.

Suppose you are going to conduct a one-way MANOVA comparing four groups on two outcome variables. You want to have 95% power for detecting a medium-sized effect.



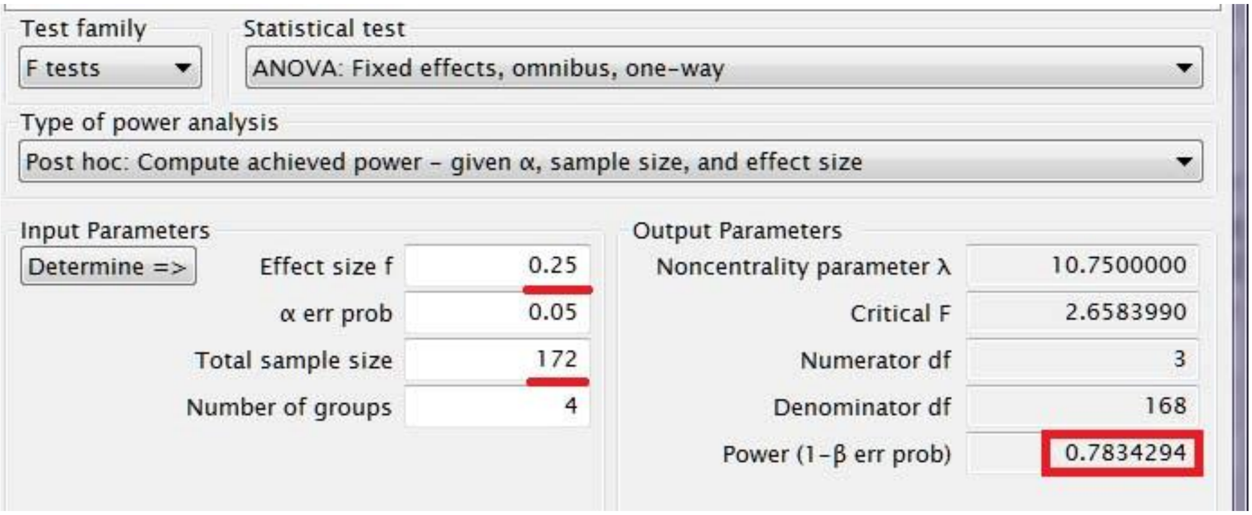
Test family: F tests

Statistical test: MANOVA: Global effects

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters		Output Parameters	
Effect size f²(V)	0.0625	Noncentrality parameter λ	21.5000000
α err prob	0.05	Critical F	2.1255907
Power (1-β err prob)	0.95	Numerator df	6.0000000
Number of groups	4	Denominator df	336
Response variables	2	Total sample size	172
		Actual power	0.9525011
		Pillai V	0.1176471

As you can see, you need 172 cases (43 per group). You might actually need more than 172 cases, depending on what you intend to do after the MANOVA. Many researchers want to follow a significant MANOVA with univariate ANOVAs, one on each outcome variable. How much power would you have for a such an ANOVA if you had 172 cases?



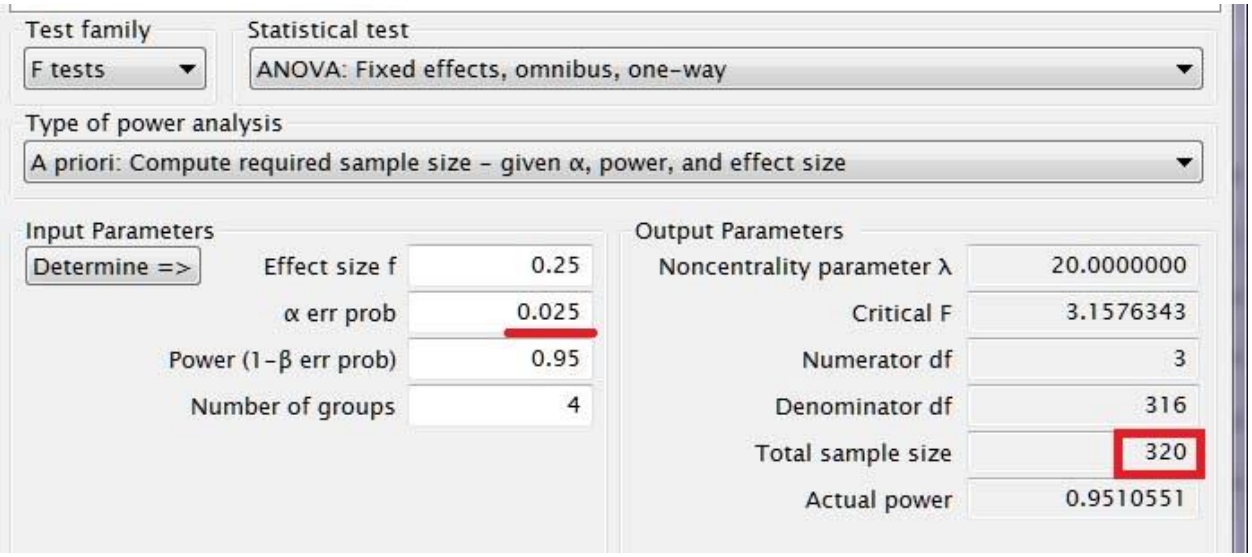
Test family: F tests
Statistical test: ANOVA: Fixed effects, omnibus, one-way
Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input Parameters
Effect size f: 0.25
α err prob: 0.05
Total sample size: 172
Number of groups: 4

Output Parameters
Noncentrality parameter λ: 10.7500000
Critical F: 2.6583990
Numerator df: 3
Denominator df: 168
Power (1–β err prob): 0.7834294

Oh my, only 78% power. When planning research you really should consider how much power you will have for the follow-up analyses you will employ after the initial (usually more powerful) analysis.

So, how many cases would we need to have 95% power for those ANOVAs? Here is the G*Power solution. Note that I have dropped the α to .025, assuming that the researcher has applied a Bonferroni correction to cap the familywise error at .05 across these two ANOVAs (I do not think doing so is a very good idea, but that is another story).



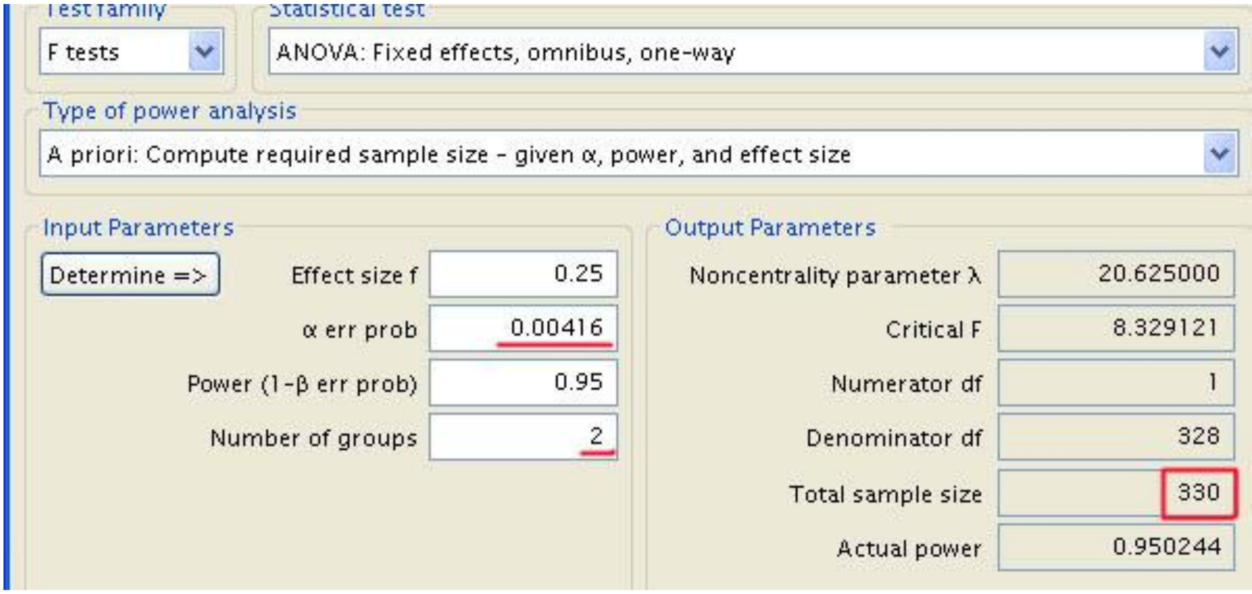
Test family: F tests
Statistical test: ANOVA: Fixed effects, omnibus, one-way
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters
Effect size f: 0.25
α err prob: 0.025
Power (1–β err prob): 0.95
Number of groups: 4

Output Parameters
Noncentrality parameter λ: 20.0000000
Critical F: 3.1576343
Numerator df: 3
Denominator df: 316
Total sample size: 320
Actual power: 0.9510551

Now, what are you going to do after a significant univariate ANOVA? Likely you are going make some pairwise comparisons among the group means. Suppose you are going to compare groups 1 with 2, 1 with 3, 1 with 4, 2 with 3, 2 with 4, and 3 with 4. That is six contrasts for each Y variable. Since you have two Y variables, you could make as many as 12 contrasts. To cap familywise error at .05 via Bonferroni, you must now use a per-comparison alpha of .05/12 = .00416.

Furthermore, suppose you will not pool variances and degrees of freedom across groups. How many cases will you need?



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input Parameters

Effect size f	0.25
α err prob	0.00416
Power (1-β err prob)	0.95
Number of groups	2

Output Parameters

Noncentrality parameter λ	20.625000
Critical F	8.329121
Numerator df	1
Denominator df	328
Total sample size	330
Actual power	0.950244

You need 330/2 = 165 cases per group. Since you have four groups, you need a total of 660 cases.

Since DFA is equivalent to MANOVA, you can use G*Power for power analysis for DFA as well.

Links

- Assorted Stats Links
- G*Power 3 – download site
  - User Guide – sorted by type of analysis
  - User Guide – sort by test distribution
- Internet Resources for Power Analysis
- List of the analyses available in G*Power
- Power Conventions: What effect size should I assume and how much power should I want?

Karl L. Wuensch
Dept. of Psychology
East Carolina University
Greenville, NC 27858-4353 USA

September, 2020.