**Estimating the Sample Size Necessary to Have Enough Power**

How much data do you need -- that is, how many subjects should you include in your research. If you do not consider the expenses of gathering and analyzing the data (including any expenses incurred by the subjects), the answer to this question is very simple -- the more data the better. The more data you have, the more likely you are to reach a correct decision and the less error there will be in your estimates of parameters of interest. The ideal would be to have data on the entire population of interest. In that case you would be able to make your conclusions with absolute confidence (barring any errors in the computation of the descriptive statistics) and you would not need any inferential statistics.

Although you may sometimes have data on the entire population of interest, more commonly you will consider the data on hand as a random sample of the population of interest. In this case, you will need to employ inferential statistics, and accordingly power becomes an issue. As you already know, the more data you have, the more power you have, ceteris paribus. So, how many subjects do you need?

Before you can answer the question “how many subjects do I need,” you will have to answer several other questions, such as:

* How much power do I want?
* What is the likely size (in the population) of the effect I am trying to detect, or, what is smallest effect size that I would consider of importance?
* What criterion of statistical significance will I employ?
* What test statistic will I employ?
* What is the standard deviation (in the population) of the criterion variable?
* For correlated samples designs, what is the correlation (in the population) between groups?

In my opinion, if one considers Type I and Type II errors equally serious, then one should have enough power to make *α* = *β*. If employing the traditional .05 criterion of statistical significance, that would mean you should have 95% power. However, getting 95% power usually involves expenses too great for behavioral researchers -- that is, it requires getting data on many subjects.

A common convention is to try to get at least enough data to have 80% power. So, how do you figure out how many subjects you need to have the desired amount of power. There are several methods, including:

* You could buy an expensive, professional-quality software package to do the power analysis.
* You could buy an expensive, professional-quality book on power analysis and learn to do the calculations yourself and/or to use power tables and figures to estimate power.
* You could try to find an interactive web page on the Internet that will do the power analysis for you. I do not have a great deal of trust in this method.
* You could download and use the [G\*Power](http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/) program, which is free, not too difficult to use, and generally reliable (this is not to say that it is error free). For an undetermined reason, this program will not run on my laptop, but it runs fine on all my other computers.
* You could use the simple guidelines provided in Jacob Cohen’s “A Power Primer” (*Psychological Bulletin*, 1992, *112*, 155-159).

Here are minimum sample sizes for detecting small (but not trivial), medium, and large sized effects for a few simple designs. I have assumed that you will employ the traditional .05 criterion of statistical significance, and I have used Cohen’s guidelines for what constitutes a small, medium, or large effect.

**Chi-Square, One- and Two-Way**

Effect size is computed as . *k* is the number of cells, *P0i* is the population proportion in cell *i* under the null hypothesis, and *P1i* is the population proportion in cell *i* under the alternative hypothesis. For example, suppose that you plan to analyze a 2 x 2 contingency table. You decide that the smallest effect that you would consider to be nontrivial is one that would be expected to produce a contingency table like this, where the experimental variable is whether the subject received a particular type of psychotherapy or just a placebo treatment and the outcome is whether the subject reported having benefited from the treatment or not:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Experimental Group | |  |
| Outcome | Treatment | Control | . |
| Positive | 55 | 45 |
| Negative | 45 | 55 |

For each cell in the table you compute the expected frequency under the null hypothesis (P0)by multiplying the number of scores in the row in which that cell falls by the number of scores in the column in which that cell falls and then dividing by the total number of scores in the table. Then you divide by total *N* again to convert the expected frequency to an expected proportion. For the table above the expected frequency will be the same for every cell, . For each cell you also compute the expected proportion under the alternative hypothesis (P1) by dividing the expected number of scores in that cell by total *N*. For the table above that will give you the same proportion for every cell, 55 ÷ 200 = .275 or 45 ÷ 200 = .225. The squared difference between P1 and P0, divided by P0, is the same in each cell, .0025. Sum that across four cells and you get .01. The square root of .01 is .10. Please note that this is also the value of phi.

In the treatment group, 55% of the patients reported a positive outcome. In the control group only 45% reported a positive outcome. In the treatment group the odds of reporting a positive outcome are 55 to 45, that is, 1.2222. In the control group the odds are 45 to 55, that is, .8181. That yields an odds ratio of 1.2222 ÷ .8181 = 1.49. That is, the odds of reporting a positive outcome are, for the treatment group, about one and a half times higher than they are for the control group.

What if the effect is larger, like this:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Experimental Group | |  |
| Outcome | Treatment | Control | . |
| Positive | 65 | 35 |
| Negative | 35 | 65 |

Now the odds ratio is 3.45 and the phi is .3.

Or even larger, like this:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Experimental Group | |  |
| Outcome | Treatment | Control | . |
| Positive | 75 | 25 |
| Negative | 25 | 75 |

Now the odds ratio is 9 and the phi is .5.

Cohen considered a *w* of **.10** to constitute a small effect, .**3** a medium effect, and **.5** a large effect. Note that these are the same values indicated below for a Pearson *r*. The required total sample size depends on the degrees of freedom, as shown in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Effect Size | | |
| *df* | Small | Medium | Large |
| 1 | 785 | 87 | 26 |
| 2 | 964 | 107 | 39 |
| 3 | 1,090 | 121 | 44 |
| 4 | 1,194 | 133 | 48 |
| 5 | 1,293 | 143 | 51 |
| 6 | 1,362 | 151 | 54 |

* [The Correspondence **b**etween Phi and Odds Ratios](http://core.ecu.edu/psyc/wuenschk/StatHelp/Phi-OddsRatio.doc) – it depends the distribution of the marginals.
* [More on *w* = φ](file:///C:\Users\My%20PC\Downloads\w=phi.doc).

**Pearson *r***

Cohen considered a ρ of **.1** to be small, **.3** medium, and .**5** large. You need **783** pairs of scores for a small effect, **85** for a medium effect, and **28** for a large effect. In terms of percentage of variance explained, small is 1%, medium is 9%, and large is 25%.

**One-Sample T Test**

Effect size is computed as . A ***d*** of **.2** is considered small, **.5** medium, and **.8** large. For 80% power you need **196** scores for small effect, **33** for medium, and **14** for large.

Cohen’s *d* is not affected by the ratio of *n1* to *n2*, but some alternative measures of magnitude of effect (*rpb* and η2) are. See [this document](http://core.ecu.edu/psyc/wuenschk/StatHelp/d-r.htm).

**Independent Samples T, Pooled Variances.**

Effect size is computed as . A ***d*** of **.2** is considered small, **.5** medium, and **.8** large. Suppose that you have population with means of 10 and 12 and a within group standard deviation of 10. , a small effect. The population variance of the means here is 1, so the percentage of variance explained is 1%. Now suppose the means are 10 and 15, so *d* = .5, a medium effect. The population variance of the means is now 6.25, so the percentage of variance explained is 6%. If the means were 10 and 18, *d* would be .8, a large effect. The population variance of the means would be 16 and the percentage of variance explained 16%.

For 80% power you need, in each of the two groups, **393** scores for small effect, **64** for medium, and **26** for large.

**Correlated Samples T**

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). One could, then, define effect size and required sample size as shown above for the one sample *t*. This would, however, usually not be a good idea.

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 . If we assume equal variances, this simplifies to .

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*: . The denominator of this ratio is the standard deviation of the difference scores rather than the standard deviation of the original scores. We can then compute the required sample size as . If the sample size is large enough that there will be little difference between the *t* distribution and the standard normal curve, then we can obtain the value of δ (the noncentrality parameter) from a table found in [David Howell’s](http://www.uvm.edu/~dhowell/StatPages/StatHomePage.html) statistics texts. With the usual nondirectional hypotheses and a .05 criterion of significance, δ is 2.8 for power of 80%. You can use the [G\*Power](http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/) program to fine tune the solution you get using Howell’s table.

I constructed the table below using Howell’s table and G\*Power, assuming nondirectional hypotheses and a .05 criterion of significance.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Small effect | | | |  | Medium effect | | | |  | Large effect | | | |
| *d* | *ρ* | *dDiff* | *n* |  | *d* | *ρ* | *dDiff* | *n* |  | *d* | *ρ* | *dDiff* | *n* |
| .2 | .00 | .141 | 393 |  | .5 | .00 | 0.354 | 65 |  | .8 | .00 | 0.566 | 26 |
| .2 | .50 | .200 | 196 |  | .5 | .50 | 0.500 | 33 |  | .8 | .50 | 0.800 | 14 |
| .2 | .75 | .283 | 100 |  | .5 | .75 | 0.707 | 18 |  | .8 | .75 | 1.131 | 08 |
| .2 | .90 | .447 | 041 |  | .5 | .90 | 1.118 | 08 |  | .8 | .90 | 1.789 | 04 |

IMHO, one should not include the effect of the correlation in one’s calculation of *d* with correlated samples. Consider a hypothetical case. We have a physiological measure of arousal for which the mean and standard deviation, in our population of interest, are 50 (*M*) and 10 (*SD*). We wish to evaluate the effect of an experimental treatment on arousal. We decide that the smallest nontrivial effect would be one of 2 points, which corresponds to a standardized effect size of *d* = .20.

Now suppose that the correlation is .75. The *SD* of the difference scores would be 7.071, and the *dDiff* would be .28. If our sample means differed by exactly 2 points, what would be our effect size estimate? Despite *dDiff* being .28, the difference is still just 2 points, which corresponds to a *d* of .20 using the original group standard deviations, so, IMHO, we should estimate *d* as being .20.

Now suppose that the correlation is .9. The *SD* of the difference scores would be 4.472, and the *DDiff*  would be .45 – but the difference is still just two points, so we should not claim a larger effect just because the high correlation reduced the standard deviation of the difference scores. We should still estimate *d* as being .20.

Note that the correlated samples *t* will generally have more power than an independent samples *t* , holding the number of scores constant, as long as the *ρ12* is not very small or negative. With a small *ρ12*  it is possible to get less power with the correlated *t* than with the independent samples *t* – [see this illustration](http://core.ecu.edu/psyc/wuenschk/docs30/Power-CorrT-IndT.doc). The correlated samples *t* has only half the *df* of the independent samples *t*, making the critical value of *t* larger. In most cases the reduction in the standard error will more than offset this loss of *df.* Do keep in mind that if you want to have as many scores in a between-subjects design as you have in a within-subjects design you will need twice as many cases.

**One-Way Independent Samples ANOVA**

Cohen’s *f* (effect size) is computed as , where *μj* is the population mean for a single group, *μ* is the grand mean, *k* is the number of groups, and error variance is the mean within group variance. This can also be computed as , where the numerator is the standard deviation of the population means and the denominator is the within-group standard deviation.

We assume equal sample sizes and homogeneity of variance.

Suppose that the effect size we wish to use is one where the three populations means are 480, 500, and 520, with the within-group standard deviation being 100. Using the first formula above, . Using the second formula, the population standard deviation of the means (with *k*, not *k*-1, in the denominator) is 16.33, so *f* = 16.33 ÷ 100 = .163. By the way, David Howell uses the symbol *φ'* instead of *f*.

You should be familiar with *η2* as the treatment variance expressed as a proportion of the total variance. If *η2* is the treatment variance, then 1-*η2* is the error variance. With this in mind, we can define . Accordingly, if you wish to define your effect size in terms of proportion of variance explained, you can use this formula to convert *η2* into *f*.

Cohen considered an *f* of **.10** to be a small effect, **.25** a medium effect, and **.40** a large effect. Rearranging terms in the previous formula, . Using this to translate Cohen’s guidelines into proportions of variance, a small effect is one which accounts for about 1% of the variance, a medium effect 6%, and a large effect 14%.

The required sample size per group varies with treatment degrees of freedom, as show below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Effect Size | | |
| *df* | Small | Medium | Large |
| 2 | 393 | 64 | 26 |
| 3 | 322 | 52 | 21 |
| 4 | 274 | 45 | 18 |
| 5 | 240 | 39 | 16 |
| 6 | 215 | 35 | 14 |
| 7 | 195 | 32 | 13 |

**Correlated Samples ANOVA**

See [Power Analysis for One-Way Repeated Measures ANOVA](http://core.ecu.edu/psyc/wuenschk/MV/RM-ANOVA/Power-RM-ANOVA.doc)

**Analysis of Covariance**

See the document [Power Analysis for an ANCOV](http://core.ecu.edu/psyc/wuenschk/MV/LSANOVA/Power-ANCOV.doc).

**Multiple Correlation**

For testing the squared multiple correlation coefficient, Cohen computed effect size as . For a squared partial correlation, the same definition is employed, but the squared partial correlation coefficient is substituted for *R2*. For a squared semipartial (part) correlation coefficient, , where the numerator is the squared semipartial correlation coefficient for the predictor of interest and the denominator is 1 less the squared multiple correlation coefficient for the full model.

Cohen considered an *f2* of **.02** to be a small effect, **.15** a medium effect, and **.35** a large effect. We can translate these values of *f2* into proportions of variance by dividing *f2* by (1 + *f2* ): A small effect accounts for 2% of the variance in the criterion variable, a medium effect accounts for 13%, and a large effect 26%.

The number of subjects required varies with the number of predictor variables, as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Effect Size | | |
| # predictors | Small | Medium | Large |
| 2 | 481 | 67 | 30 |
| 3 | 547 | 76 | 34 |
| 4 | 599 | 84 | 38 |
| 5 | 645 | 91 | 42 |
| 6 | 686 | 97 | 45 |
| 7 | 726 | 102 | 48 |
| 8 | 757 | 107 | 50 |

**Where Can I Find More on Power Analysis?**

The classic source is Cohen, J. (1988). *Statistical power analysis for the behavior sciences*. (2nd ed.). Hillsdale, NJ: Erlbaum – Call number JZ1313 .D36 2002 in Joyner Library. I have parts of an earlier (1977) edition.

**Karl Wuensch, East Carolina University. Revised November, 2009.**

[Return to Karl’s Statistics Lessons Page](http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm)