**HOW TO DETERMINE AN APPROPRIATE SAMPLE SIZE**

**by Simon Moss**

|  |
| --- |
| **Introduction** |

Suppose a researcher wants to ascertain whether Echinacea can improve mood. Some participants consume Echinacea for a month. Other participants consume a placebo for a month. Then, after a month, participants are asked to indicate the extent to which they feel satisfied with life on a 10 point scale. How many participants should be allocated to these two conditions?

In research proposals, ethics applications, and other documents, research candidates need to estimate the number of participants, animals, organizations, fields, or samples they need in their study—called the sample size. This document offers some guidelines that research candidates can utilize to answer this question. This document, however, will discuss only quantitative designs.

|  |
| --- |
| **The balance between strong results and resource availability** |

If time, money, and resources were unlimited, the number of participants, animals, organizations, fields, or specimens should be as large as possible. That is, when the sample size is large, you are more likely to generate informative results. For example,

* If you plan to conduct statistical tests that generate p values, a large sample size is more likely to produce significant results.
* If you plan to conduct data mining or factor analysis, a large sample size is more likely to generate reliable results—that is, results that are likely to be replicated in the future.

But, in practice, as you have probably discovered, time, money and resources are limited. Your sample size cannot be as large as you would like. But, what is an appropriate sample size? What criteria or procedures should you apply to answer this question?

|  |
| --- |
| **Examine comparable studies** |

The simplest approach is to emulate comparable studies—studies that utilize similar procedures, measures, or designs. For example

* If these studies recruit about 100 participants, you should plan to recruit about 100 participants.
* If these studies collect data from 50 fields, you should also collect data from approximately 50 fields.

This approach is not greatly admired. Most researchers advocate a more systematic approach. But, in practice, this approach is almost as informative as more sophisticated procedures. At least, you could apply this approach to generate an initial approximation of the appropriate sample size. For example, if you apply this approach, you will discover that

* When researchers conduct factor analysis and structural equation modeling, the sample size tends to exceed 300
* When researchers conduct multiple regression, the sample size tends to exceed 150—or about 15 to 20 times the number of predictors
* When researchers conduct t-tests, ANOVAs, ANCOVAs, or MANOVAs, the sample size tends to exceed 50 participants or units in each group.
* These estimates vary considerably across disciplines and fields, however.

|  |
| --- |
| **Sample sizes that generate enough power** |

If you plan to conduct tests that generate p values, such as ANOVAs or multiple regression, you could utilize another approach to estimate the appropriate sample size. The approach that researchers tend to recommend, but perhaps overestimate, is to compute the sample size that is needed to generate enough power. So, what is power?

* Power is the probability you will generate a significant result if the conditions really differ from each other—if the measures really are related to each other.
* In this example, power is the probability you will generate a significant difference between the participants who consumed Echinacea and the participants who received the placebo on satisfaction with life if Echinacea really does improve mood

To illustrate

* If power is 0.4, the probability you will generate a significant result if Echinacea really does improve mood is 0.4 or 40%.
* If power is 0.8, the probability you will generate a significant result if the conditions really differ from each other is 0.8 or 80%. The probably you will generate a non-significant result and thus overlook the benefits of Echinacea is 0.2 or 20%

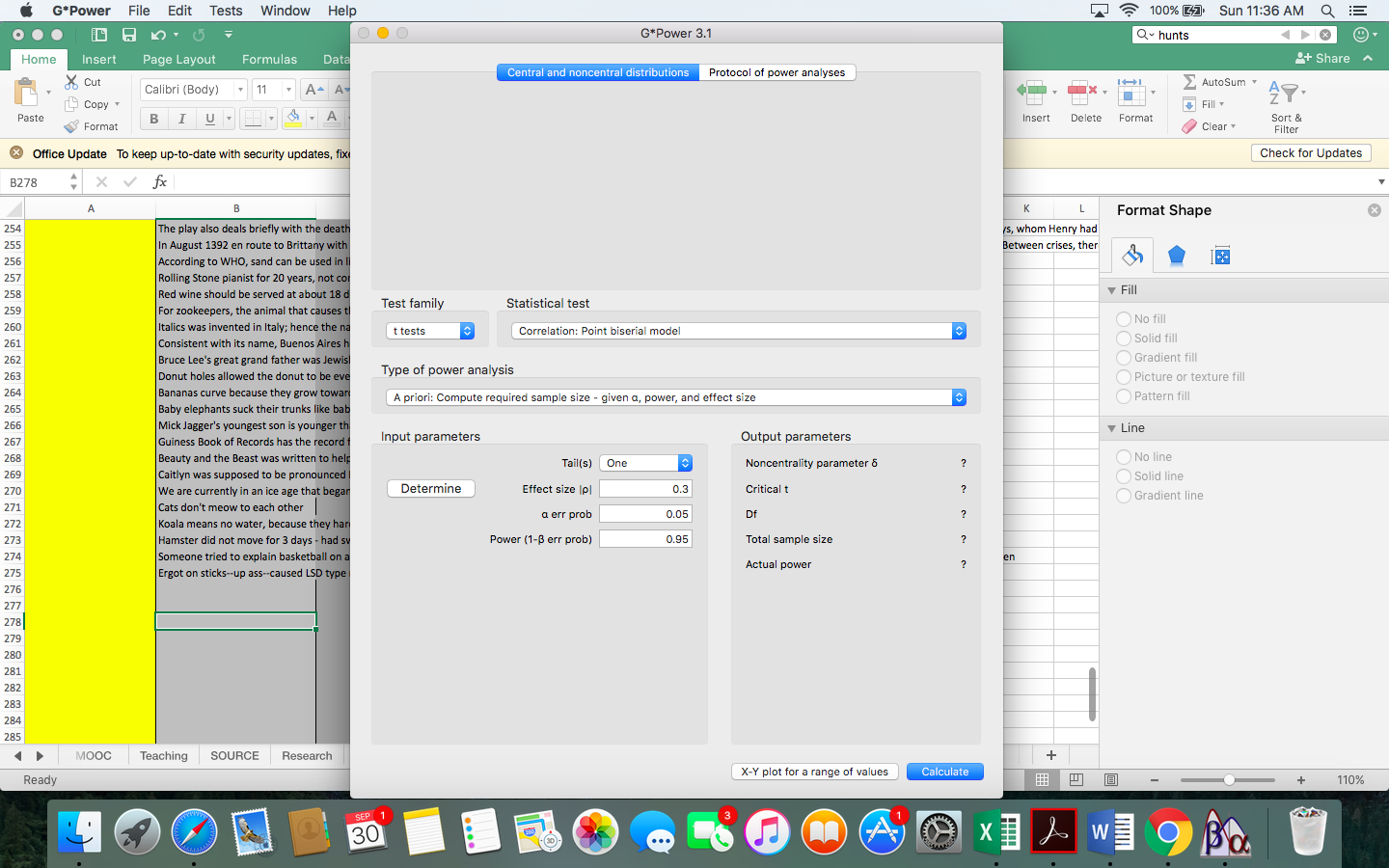
Researchers tend to agree that a suitable level of power is about 0.8 or 80%. This figure might surprise you. Surely, power should be as high as possible, such as close to 100%. Surely, researchers want to generate significant results whenever the conditions actually differ from each other—or whenever the measures, like Echinacea and mood, are related to each other. The problem is

* To generate very high levels of power, such as .99 or 99%, the sample size might need to be mammoth—such as 10 000 individuals
* To generate fairly high levels of power, such as .80 or 80%, the sample size might need to be modest only—such as 100 individuals
* So, over time, researchers have decided that, in general, a power of .80 or 80% is appropriate.

|  |
| --- |
| **How to estimate the appropriate sample size: Illustration** |

The key question, then, is how should you estimate the sample size you need to generate a power of 0.8. To answer this question, consider the example in which you want to compare participants who consumed Echinacea to participants who consumed a placebo on satisfaction with life. To estimate the appropriate sample size in this circumstance, you would first need to download some software called GPower. To achieve this goal

* Proceed to [www.gpower.hhu.de/en.html](http://www.gpower.hhu.de/en.html)
* Click either “Download G\*Power 3.1.9.2 for Windows” or “Download G\*Power 3.1.9.3 for Mac OS X” depending on your computer—and then download the file
* When activated, the software should resemble the following



To estimate the appropriate sample size, you need to specify various options. The following table specifies the options that are applicable to this example.

|  |  |
| --- | --- |
| Information you need to enter | Relevant choice for this example |
| The statistical test you would like to utilize | Press the arrows alongside “Correlation: Point biserial model” and choose “Means: Difference between two independent means (two groups)” |
| Type of power analysis | Use the default: A priori—compute required sample size |
| Tails | Choose two-tailed rather than one-tailed. Two-tailed is usually more appropriate |
| Effect size d | Enter the effect size you expect—such as 0.5 |
|  error probability | Enter 0.5 for alpha |
| Power (1 -  error probability) | Enter the power you would like—usually 0.8 |
| Allocation ratio N2/N1 | Use the default unless you know you would like more participants or specimens in one condition than in the other condition |

After you have entered this information, press Calculate. In this example, if you chose the values suggested in the table, Gpower would indicate you need to allocate 51 participants in each condition. In particular, this number indicates that

|  |
| --- |
| If the effect size really is 0.5 or medium, the probably you will generate a significant difference with a total sample size of 102 is 0.8 or 80% |

Read the previous sentence several times. As this sentence implies, you cannot determine the appropriate sample size unless you estimate the likely effect size. That is, power is the likelihood of generating a significant result, assuming a particular effect size. The probablity of generating a signficant result if the effect size is large or 0.8 is larger than is the probability of generating a significant result if the effect size is small or 0.2.

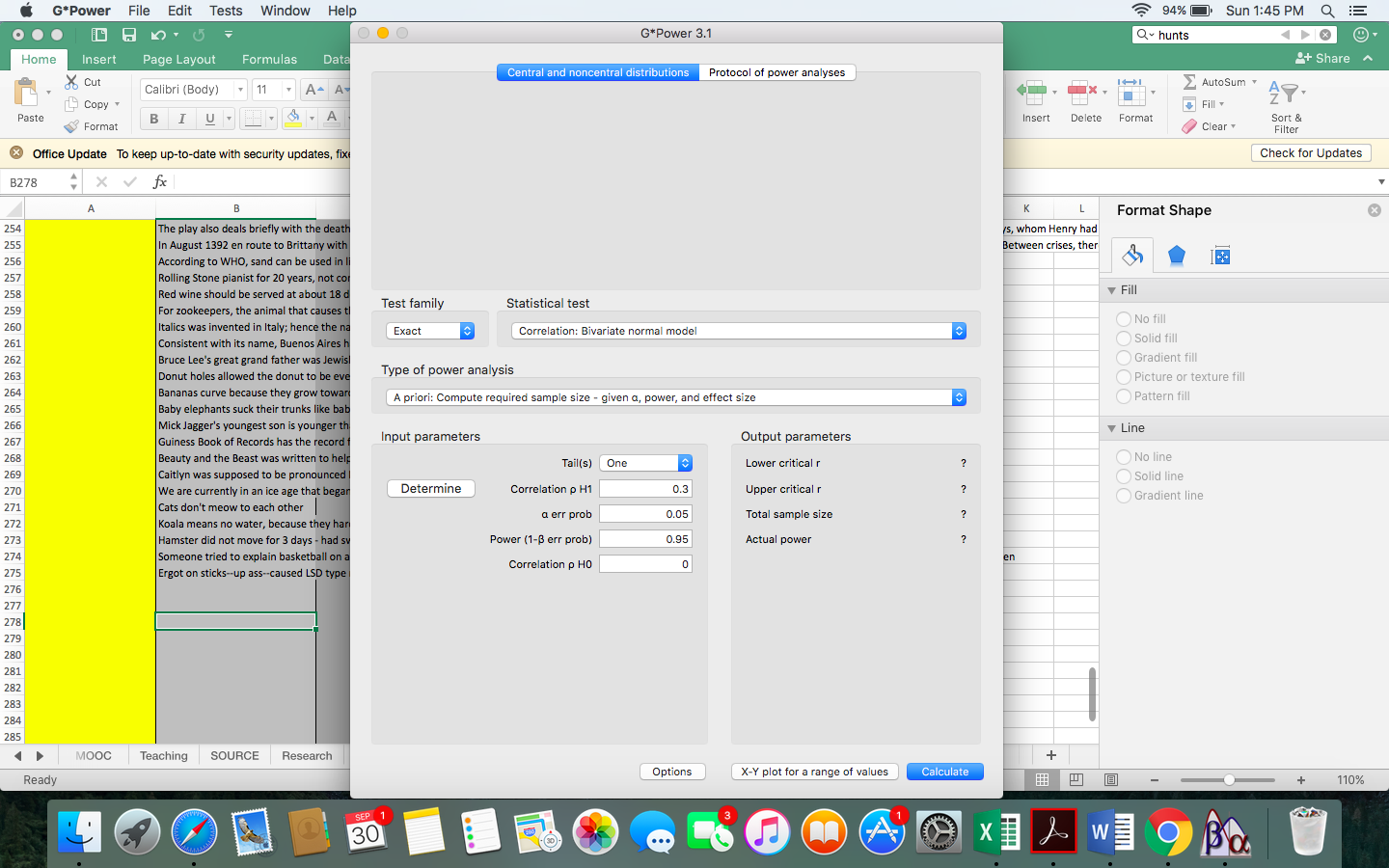
|  |
| --- |
| **How to estimate the effect size** |

This procedure is simple, except the like effect size is hard to estimate. So, how can you estimate the likely effect size? If you want to assess whether Echinacea affects satisfaction with life, is the effect size likely to be 0.2, 0.3, 0.4, 0.8, or even 1.5? The following table outlines several strategies that researchers utilize—as well as the drawbacks of each strategy. For grants, some researchers might apply two or more of these strategies.

|  |  |
| --- | --- |
| Strategy to estimate effect size | Drawback |
| **Conduct a pilot study**. You might allocate 10 participants to each group. You can then calculate the effect size from the means and standard deviations you generate | Estimate of effect size are not especially reliable. If you repeated this pilot study, the estimated effect size might be very different |
| **Utilize the effect size of similar studies**. Extract studies that are the most similar to your research. You can then calculate the effect size from the means and standard deviations of these studies | Your study must be different to previous studies—otherwise, your research would be futile. Consequently, the effect sizes of past studies would differ from the effect sizes of your study |
| **Strength of the arguments**. If you can identify many reasons the conditions should differ—such as many mechanisms for which Echinacea should improve satisfaction with life—the effect size is likely to be large. Otherwise, the effect size might be smaller. | This strategy is quite subjective |
| **Past experience**. If you have acquired experience in this field, consider whether the effects of, for example Echinacea, seem obvious to anyone, obvious to you but not the average person, or subtle—in which case the effect size is likely to be 0.8, 0.5, or 0.2 respectively | This strategy is also subjective and hard to justify |

|  |
| --- |
| **Example with correlations** |

The previous example is relevant only to circumstances in which you want to conduct an independent t-test to compare two conditions. If you want to conduct other tests, you need to adjust the procedure slightly. To illustrate, suppose you want to examine the correlation between the amount of Echinacea that people consume and the satisfaction in life. In this instance, in the box called “test families”, choose “exact”. In the list of options under “Statistical tests”, choose “Correlation: Bivariate normal model”.



To estimate an appropriate sample size,

* Specify tails as two
* Correlation  H1 is the expected correlation, such as 0.3 if medium
* The is usually 0.05 and the power is usually about 0.8
* Do not change the last box from 0
* Press calculate
* If the effect size is 0.3 or medium, the sample size will need to be 84. At this sample size, the probabiliy you will generate a significant effect is 0.8

|  |
| --- |
| **Other statistics** |

Thus far, this document has demonstrated how you can estimate the appropriate sample size if you conduct either an independent t-test or correlation. GPower can also be used to estimate the appropriate size if you need to conduct other tests as well, such as ANOVAs, ANCOVAs, and indeed hundreds of statistical procedures. The first column presents various statistical tests. The second column specifies which options you should choose when conducting this test. The third column specifies some complications for each procedure. Alternatively, you can consult the manual that is available on the GPower website.

|  |  |  |
| --- | --- | --- |
| Test | Options to chose | Some complications or suggestions |
| **Test family: t-tests** |  |  |
| Point bi-serial correlation—to explore the association between a numerical and dichotomous measure | Correlation: Point biserial model | For estimating effect sizes, use the same procedure as correlations |
| Dependent or paired-sample t-test | Means: Difference between two dependent means | To estimate the effect size, you can either   * Predict the means and standard deviations of each group as well as the correlation between the two conditions * Estimate, on average, the difference between each pair of measures—and the standard deviation of these differences   Alternatively, to estimate the effect size, follow the procedure you apply for independent t-tests but multiple the d value by 1.4 to generate a rough estimate. |
| Wilcoxon-Mann-Whitney test—nonparametric variant of the independent t-test | Means: Wilcoxon-Mann-Whitney test (two groups) |  |
| **F-tests** |  |  |
| One-way ANOVA | ANOVA: Fixed-effects, omnibus one-way | * The effect size, f, tends to be .02, .15, .35 for small, medium, and large effects respectively |

These procedures are sometimes complex and ambiguous. An alternative approach is to simplify the tests for the purpose of estimating the appropriate sample size. Here are some illustrations

**More than two conditions: ANOVAs**

Suppose you need to compare three conditions: participants who consume Echinacea, participants who consume garlic, and participants who consume a placebo. In this circumstance, you could

* determine which two conditions that are most similar
* derive the effect size d from these two conditions; for example, if these conditions differ to a small extent, d would be 0.2
* follow the procedure you applied to independent t-tests to ascertain the appropraite sample size in each condition

But later, when you conduct your analyses, you would utilize an ANOVA.

**Covariates: ANCOVAs**

Similar reasoning can be applied when your design includes covariates. For example, suppose you want to repeat the study that explores whether Echinacea improves satisfaction with life, but after controlling income. You might then

* Estimate the percentage of participants who consumed Echinacea who are more satisfied in life than the average participant who consumed the placebo
* However, assume all these participants were average in income
* You might, for example, conclude that 66% of the participants who consumed Echinacea who are more satisfied in life than the average participant who consumed the placebo
* As the following table shows, this percentage equates to a d value of 0.4.
* Using this estimate, follow the procedure you applied to independent t-tests to ascertain the appropraite sample size in each condition

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| % | 54 | 58 | 62 | 66 | 69 | 73 | 76 | 79 | 82 | 84 |
| d | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |

**Multiple measures: MANOVAS**

If researchers want to assess several measures, such as depression, anxiety, and stress in addition to satisfaction in life, they might conduct a MANOVAs. To estimate the appropriate sample size, perhaps restrict your attention to only one of these measures—perhaps the measure that might generate the smallest effect size or difference across conditions. This approach, although simple, does slightly overestimate the sample size needed. But later, when you conduct your analyses, you would utilize a MANOVA.

**Multiple regression**

When conduting multiple regression, apply the same approach that you utilize for correlations—with one exception: You need to estimate the effect size that you would have generated if participants were average on the other predictors.

**Structural equation modeling**

To ascertain whether your sample size is sufficient to generate adequate power, you may need to utilise specialised tools. For example, you could use Soper’s on-line structural equation modelling power calculator. See [www.danielsoper.com/statcalc/calculator.aspx?id=89](http://www.danielsoper.com/statcalc/calculator.aspx?id=89)

|  |
| --- |
| **Limitations of this approach** |

Many researchers utilize GPower to estimate the appropriate sample size. But, you should be aware of a paradox that limits the utility of this approach. Specifically

* This approach is restricted to tests that generate p values, such as independent t-tests or ANOVAs
* When these p values are less than 0.5, the researcher concludes the difference or relationships is significant--and can thus be quite certain the effect size differs from 0
* When these p values are greater than 0.5, the researcher concludes the difference or relationships is non-significant—and cannot be certain the effect size differs from 0.
* In other words, when researchers conduct tests that generate p values they are, in essence, conceding they are not sure the effect size differs from 0
* Yet, to estimate the appropriate sample size, they need to predict the likely effect size—a challenging task given they are not certain the effect size differs from 0

This line of reasoning implies that researchers cannot predict the effect size precisely. Hence, the estimates of sample size are also hazy.

|  |
| --- |
| **References and sources of information** |

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175-191. doi:10.3758/bf03193146

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149-1160. doi:10.3758/brm.41.4.1149